

# 1

## Biological Background and Introduction

### 1.1 Lymphoma

The aetiology of lymphoma and lymphoid leukaemia currently remains enigmatic. Lymphoma diagnoses can be housed under the umbrella term of ‘haematological neoplasms’ - a broad definition that represents a complex hierarchy, containing an array of classes with an increasing degree of homogeneity as the hierarchy is traversed downwards. Classification systems have evolved from the Lukes-Butler classification (Lukes and Butler, 1966) to the 1982 Working Formulation (WHO, 1982) for non-Hodgkin lymphoma, the Revised European-American Lymphoma (REAL) Classification, and most recently, the World Health Organisation’s (WHO) classification system (WHO, 2008). Representing approximately 5.3% of all cancer cases (Horner et al., 2008) in the United States, treatments range from radiotherapy and chemotherapy to bone-marrow

transplants. Given the heterogeneous spread of lymphoma subtypes and cases, however, the generality of these treatments have highlighted their shortcomings. In recent years haematologists and medical professionals have concluded that a suite of highly discerning diagnostic methods is vital to delivering targeted treatments. Molecular and statistical analysis is currently seen as an important aid of traditional pathological methods in this diagnostic inventory. Techniques such as classification allow for better delineation of lymphoma subtypes, which in turn allow treatments to be tailored more specifically to patients, prolong their survival, reduce side-effects and help clinicians develop new treatments. Since at least 7% (Kelley et al., 2005) of the human genome is constituted of immune system genes (and at least 20% of these have known disease association), a deep knowledge of the genetic basis for an immune system disorder such as lymphoma proves invaluable.

## 1.2 Gene Expression Profiling

The process by which a gene, embedded in the DNA (deoxyribonucleic acid) of an organism, is synthesised into a gene product, is called gene expression and constitutes one of the central tenets of molecular biology (Crick, 1970). All gene products exist as RNA (ribonucleic acid) molecules at some point in their synthesis. Messenger RNAs (mRNAs) are translated into proteins through a ribosome; others (non-coding RNAs) assist with translation, gene splicing and regulation (Riddihough, 2005). The interactions of a preinitiation complex of molecules precipitate the transcription of the DNA genetic code into RNA molecules. An enzyme called RNA polymerase recognises a particular short sequence called a promoter region and, with the help of other regulating molecules, identifies a position downstream at which transcription can begin. Another enzyme called DNA helicase breaks the hydrogen bonds between the two strands of the double helix, ‘unzipping’ the DNA template strand from its complementary coding strand. RNA polymerase assembles the new RNA strand as a complement of the template strand before terminating the transcript after a hairpin loop followed by a weakly-bonded poly-A tail on the DNA renders its interaction with the RNA

mechanically unstable. This results in a free RNA strand.

Gene expression profiling (NCBI, 2007) is a technique that simultaneously measures the expression (in RNA output) of a large number of genes from multiple samples. The goal of gene expression profiling is to build a comprehensive view of the gene expression of a particular class of biological tissue, such as one that is diseased, or has undergone a particular treatment. A major tool employed in this field is DNA microarray technology. Physically, a microarray is a substrate (usually a glass plate, although it can be made of silicon or nylon) onto which purified RNA from tissue samples are deposited. The plate contains tens of thousands of individual pits. Attached to each pit is a unique DNA probe: an oligonucleotide from a specific short region in the genome.

Most microarray protocols involve reverse-transcribing mRNAs from the sample of interest into its complement: cDNA. Microarrays exploit the ability of these cDNA molecules to bind to the DNA probe with the intention of measuring the amount of mRNA produced by a sample of diseased tissue in question, relative to an identical amount of healthy tissue. Complementary sequences bind to each other in the famous double-helix formation *in vivo*; on the microarray substrate *in vitro* the process is called hybridization. Each probe is unique, and each pit functions as an individual northern blot experiment: detecting and measuring the amount of RNA present in the sample (Kevil et al., 1997). Thus a single microarray plate can perform tens of thousands of gene expression measurements in parallel.

Traditional microarray technology is ‘two-channel’; it uses a relative measurement to determine the degree of up or downregulation in each feature. DNA from both diseased and healthy tissue samples are mixed in the same northern blot with a fluorescent label, which, in the scanning stage, is excited by a laser. With the use of a microscope and camera, a heat map is generated (Figure 1.1), where all features can be simultaneously observed.

Such synoptic visualisation can be an important initial tool for biologists to identify genes of interest for further research, in complement to the employment of the statistical methods described in this thesis. Visualisation of high-dimensional data is a field in its own right (Grinstein et al., 2001); tools such as Visumap© and GGobi contain a suite

of resources that produce an intuitive display of selected features in a high-dimensional dataset. As well as heat maps, visualisations such as 2D and 3D scatter plots (with a feature plotted along each axis) help with drawing discriminants (when the data is supervised) and identifying clusters. The colour and shape of the plotted observations on a scatter plot can be used to denote class labels or—from phenotypic analysis—other categorical descriptions.

More recent microarray technologies such as Affymetrix<sup>TM</sup> and ChIP-chip (Chromatin Immunoprecipitation) are known as ‘one-channel’ arrays, where the tissue in question is analysed individually, and not mixed with another. Hence, measurements represent absolute amount of RNA present in the tissue, rather than relative amounts.

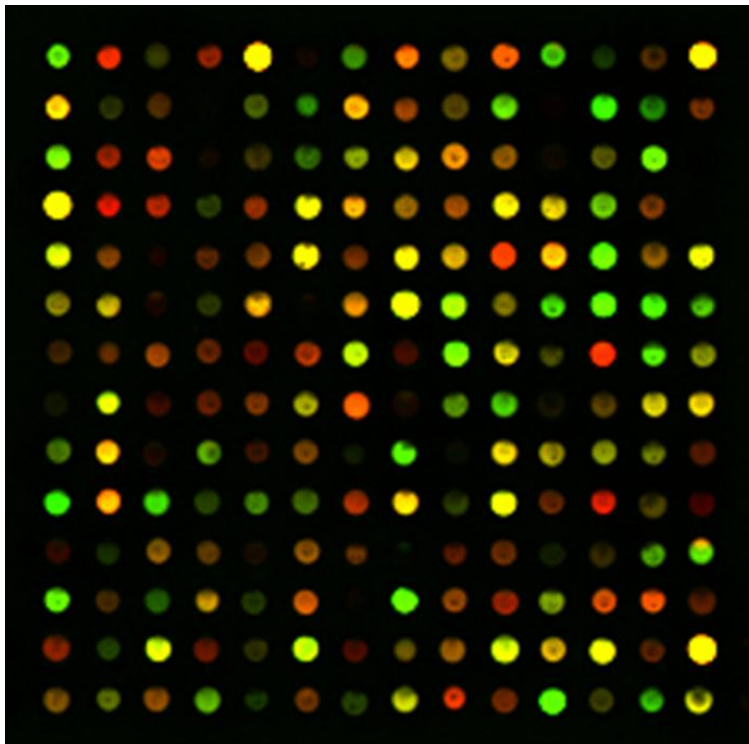


FIGURE 1.1: Subsection of a two-colour microarray heatmap. Each dot represents the expression value of a single gene, based on its intensity. Red dots represent up-regulated genes, and green dots down-regulated ones. Image obtained under Creative Commons licence from Department of Biology at James Madison University.

A research group (often, but not always, connected to a hospital where patient

tissue samples are available) will usually be able to present a statistician with a number of gene expression profiles for analysis, with the subsequent work published by a group of affiliates from different research hubs. Some examples of such studies include Alizadeh et al. (2000), Shipp et al. (2002) and Loi et al. (2011). The number of samples  $n$  usually varies from around 30 to over 100, but the important distinguishing factor of the resulting dataset is the number of variables  $P$ , which routinely stretches into the thousands, and is often over 20000<sup>1</sup>. A mandatory pre-processing step (for statistical analysis) is normalisation of the data collected from multiple arrays. Array readings—even those performed by the same machine—may vary in overall intensity between individual samples. Since statistical analysis requires standardised values, ensuring inter-array uniformity is a vital step. This step is usually performed using readily available public software (Gentleman et al., 2004), before the statistician begins dimension reduction procedures.

The knowledge base of cancer aetiology is rapidly growing. However, it is far from complete. The characterisation of lymphoma subtypes from multiple transcription signatures heralds a greater specificity in both diagnosis and treatment. For example, (Monti et al., 2005) identified three discrete subsets of Diffuse Large B-Cell Lymphoma using a combination of ranking features by their differential expression values and a biologically-informed method called Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). It is anticipated that the discovery of further subtypes via computational methods may even subvert the current medical diagnostic paradigm, which is reliant solely on phenotypic markers. Transcriptome screening may even become commonplace as a pre-treatment step for patients, though obviously not rendering the current orthodoxy obsolete.

This thesis concentrates on the statistical end of gene expression analysis, which, in contrast to molecular biology, is the realm in which a standard protocol is profoundly lacking. A sound statistical analysis provides buttresses on which scientists

---

<sup>1</sup>Recently, customized microarrays have been developed to analyse alternative splices of select genes. This technology is called RNASeq, and is discussed further in Chapter 6. The bioinformatic implication of the ability to analyse one gene with multiple probes is a potentially exponential increase in the number of features in the dataset in question, thereby further increasing the number of features in the dataset to be reduced.

and clinicians can draw confidence to put forward putative bellwether genes for cancer diagnosis. We do not intend to be dogmatic in our bid to convince statisticians to follow our proposed methods; rather we aim to provide an exploration of the challenges encountered when performing high-dimensional data analysis, and suggest workable remedies to them.

The primary question that this thesis aims to answer is ‘What kind of protocols must a bioinformatician be aware of when performing dimension reduction?’ Under the auspices of this question are a number of important secondary questions:

- Which of the existing, available, state-of-the-art dimension reduction techniques are the most likely to return pointers to biomarkers useful in medical diagnostics?
- What pitfalls, caveats and contexts need to be considered when applying these techniques?
- What are their strengths and weaknesses?
- How much consideration, by way of analysis, needs to be extended to each feature in order to be confident of its relative contribution to class demarcation?
- Of what worth is a concordance analysis (one that combines results from different dimension-reduction techniques)?

Chapter 2 is an overview of the most widely-used machine learning techniques in dimension reduction. It starts with traditional methods such as the  $t$ -statistic and Principal Component Analysis (PCA), and moves on to describe techniques used when data points are plotted explicitly in the feature space, such as Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs). We then describe the need for a broader algorithmic framework to attack the problems caused by high-dimensionality, review various approaches and discuss their merits and drawbacks. Important concepts such as regularisation and constraint relaxation are introduced and examples are given of how they can be integrated into an algorithmic framework suitable for data mining. An in-depth review of the stand-alone methods shrinkage regression and Decision Trees

is also provided. Lastly, we explore how these techniques can be used for both feature selection and classification, and suggest strategies for implementing both.

Chapters 3, 4 and 5 are presented in journal article format and all present original work. Chapter 3 details a study that demonstrates *gestalt* discriminatory power of features in consort, Chapter 4 details the need for consideration of regularisation in applying machine learning techniques and Chapter 5 investigates the ability of an ensemble learner to identify biomarkers. A more detailed preamble, separate to the article, can be found at the beginning of each of these chapters. Chapter 6 discusses some miscellaneous issues that arise in the practical application of the techniques discussed in previous chapters, analyses some results from the study in Chapter 5 with a view to the further directions the research may take, and includes a summary and conclusion.





# 2

## The Evolution of Dimension Reduction Techniques

The dimension reduction problem is colloquially referred to as the  $P \gg n$  problem, since the number of measurable features  $P$  on a microarray plate far exceeds the number of available patient samples  $n$ . Dimension-reduction aims to isolate a small set of features, size  $q$ , that produces some surface with dimension  $q - 1$  that neatly separates the data points, by their diagnostic class, in hyperspace. The methodology of quantifying differences between classes has intuitively mirrored the development of statistics. Statistical platforms for 21st century dimension reduction methods can be found in the work of eminent statisticians such as Sir Francis Galton, Karl Pearson and Ronald Fisher. This section attempts to describe the most commonly used techniques for statistical dimension reduction, from the most rudimentary methods to contemporary,

state-of-the-art algorithms.

The archetypal framework most familiar to statisticians is a linear model, with  $X$  as an input vector containing  $\log_2$ -transformed expression values for each feature, and  $Y$  as a response vector representing (in a medical context) the diagnosis of the sample. This can be represented as:

$$Y = XB + \epsilon \quad (2.1)$$

where  $X \in \mathbb{R}^P$ ,  $Y \in \{a, b\}$  (or  $Y \in \{1, -1\}$  if framed as a regression),  $B$  is the coefficient vector of the features in  $X$ , and  $\epsilon$  is the residual vector from the fit. Log-transformed values for  $X$  are preferable because their corresponding ‘raw’ values, derived from the intensity of the coloured spots from the heat map, follow a long-tailed distribution, but after log-transformation they are symmetrical about 0 (Parmigiani et al., 2003, p. 55). For the purposes of this thesis, we assume  $Y$  only contains two values. These are usually categorical (such as  $Y \in \{\text{Hodgkin’s lymphoma, Non-Hodgkin’s lymphoma}\}$ ) in a classification paradigm, although the same problem can be reframed simply as a regression where  $Y \in \{1, -1\}$ . Although it is mathematically possible for more than 2 classes to be compared simultaneously, given  $k$  classes, it is computationally faster to calculate all individual class pairs  $\frac{k(k+1)}{2}$ , and assign a weighting system to resolve the class demarcations (Pranckeviciene and Somorjai, 2006). On the issue of variable dependence, it is well known that, given the nature of gene regulatory networks (GRNs), there will indeed be dependencies that, to a certain degree, can be quantified. However, the repository of information on GRNs is incomplete, and attempting to incorporate a set of numerical models based on biological research that favour particular features over others into the dimension reduction protocol does not necessarily guarantee a clearer picture. Hence, in this thesis, we analyse gene expression data only.

## 2.1 Fold Change and $t$ -tests

The simplest method to compare the expression difference between two classes  $a$  and  $b$  for feature  $k$  is the fold change, which is simply the difference in their means (Tibshirani,

2007):

$$\mu(X_k|Y = a) - \mu(X_k|Y = b) \quad (2.2)$$

Note that this value is the same as the log-transformed ratio of the ‘raw’ expression data (Lönnstedt and Speed, 2002):

$$\log_2 \frac{\mu(X_k^{\text{raw}}|Y = a)}{\mu(X_k^{\text{raw}}|Y = b)} \quad (2.3)$$

This measurement, however, does not take into account underlying noise in  $X_k$ . A noisy feature  $X_k$  on its own tells us little about the expression pattern for a gene; however if  $(X_k|Y = a)$  and  $(X_k|Y = b)$  have low variances, they are more likely to inhabit two tightly packed domains and be easily separable, and hence their biological implication is likely to be significant.

Since a larger variance means less separable classes, dividing the difference in the means of the classes by the standard error of the samples will give us a better indication of the separability of the resulting predictor. This value is called the  $t$ -statistic:

$$t_k = \frac{\mu(X_k|Y = a) - \mu(X_k|Y = b)}{s.e.(X_k)} \quad (2.4)$$

Sometimes the fold change is small, but the standard error also small enough to make the  $t$ -statistic large enough to be considered alongside features with a large fold change. As a compromise between the fold change and  $t$ -statistic values, a constant  $S_o$  is added to the denominator to guard against this phenomenon. The resulting statistic is called the modified  $t$ -statistic (Tusher et al., 2001; Tibshirani et al., 2002):

$$t(\text{modified})_k = \frac{\mu(X_k|Y = a) - \mu(X_k|Y = b)}{s.e.(X_k) + S_o} \quad (2.5)$$

The selection of the value  $S_o$  should be such that it removes, from the top echelon of a list of ranked features (by  $t$ -statistic), those with a small fold change.

More exotic variations on the  $t$ -statistic have been proposed in recent years, such as the moderated  $t$ -statistic (Smyth, 2004), where a Bayesian posterior variance is

substituted into the denominator in place of the usual sample variance. In this case, the moderated  $t$ -statistic is able to ‘borrow’ information from the other features in the dataset, thereby contextualising the predictive strength of the selected feature set. Further examples involve computation of  $t$ -statistics informed by other features that are correlated with the feature in question (Tibshirani and Wasserman, 2006; Zuber and Strimmer, 2009).

At this juncture, it is crucial to understand that ‘no feature is an island’, and, despite the need to tease out individual features as diagnostic candidates, robust class differentiation and prediction relies on feature synergy in statistical analysis, as genes do in biology. A lengthier discussion on this topic can be found in Chapter 3.

## 2.2 Principal Components of a High-Dimensional Dataset

When features are considered as statistical agents in tandem, the earliest method of determining the most informative aspects of high-dimensional data is through an orthogonal linear transformation called Principal Component Analysis (PCA), invented in 1901 (Pearson, 1901). By ranking the magnitude of the eigenvectors of the data input variable  $X$ , PCA rotates the coordinate system, so as to diagonalise its covariance matrix  $C$ .

Eigenvectors of  $C$  are selected to form a basis of explanatory variables for the original data to be projected onto. This allows the data to be ‘seen’ from a viewpoint where only the most informative dimensions are retained. An appropriate metaphor would be a lower-dimensional ‘shadow’ of the original high-dimensional dataset.

The response variable is then introduced into the model via Principal Components Regression (Jolliffe, 1982). This is most often done with an ordinary least squares regression, wherein the factors most correlated with the response are selected.

In practice, PCA is computationally unwieldy, given that the covariance matrix calculation needs  $nP^2$  iterations. Two known methods avoid this step by estimating

the principal components through an iterative algorithm. The first is the non-linear iterative partial least squares (NIPALS) algorithm (Geladi, 1986), which, given that dimension-reduction analyses frequently only require a handful of critical genes, only calculates the first few principal components. The second method (Roweis, 1998) estimates the first principal component, also by an iterative algorithm, and then calculates the remaining principal components (Golub and Van Loan, 1996) via the Gram-Schmidt process.

Despite both of these improvements drastically reducing the computational time needed to determine the largest components in the predictor set  $X$ , PCA and PCR have fallen out of favour with many computational statisticians in recent years. This is because the rotation of the data space inherent in PCA is arbitrary to the biological aims of feature selection. Instead of a subset of predictor features being regressed against the response, the implicated components are regressed. These components may still contain information from all of the features in the original dataset, and hence solutions derived from PCA may be useful only from a purely statistical point of view.

## 2.3 Learning Explicitly in the Feature Space

One of the routines needed in carrying out feature selection is separation of differently-labelled data, which is analogous to the demarcation of blood cancer phenotypes. The explicit manifestation of this separation can be achieved in the  $n$ -dimensional space, by plotting the log-transformed values of gene expression data into the feature space. After these points are plotted, the goal is to draw a line (2 feature model), surface (3 features) or hyperplane (4 or more features), linear or nonlinear, that best separates the two labelled groups. This line is called the *decision boundary*, or sometimes the *discriminant*.

### 2.3.1 Fisher's Linear Discriminant and Linear Discriminant Analysis

The linear classifier metaheuristic has its roots in Fisher's linear discriminant, which can be measured as the ratio of the variance between labelled classes to the variance within them (Fisher, 1936). More formally, Fisher's  $F$ -statistic can be written like this, where:

$$F(X, Y) = \frac{\sigma_{between}^2}{\sigma_{within}^2} = \frac{(w \cdot (\mu(X_k|Y = a) - \mu(X_k|Y = b)))^2}{w^T \cdot (Cov(X_k|Y = a) + Cov(X_k|Y = b)) \cdot w} \quad (2.6)$$

where  $w$  is a vector of weights normal to the discriminant that maps  $X$  onto  $\mathbb{R}$ . The samples can then be projected along an axis  $Y = wX$ . The related Linear Discriminant Analysis (LDA) attempts to simplify Fisher's discriminant by introducing the assumption that the covariances of classes  $a$  and  $b$  are equal. When this condition is satisfied, the predictor function is simply whether the dot product  $w \cdot X$  is greater or less than some constant  $c$ , indicating the class (or in the context of the feature space, the side of the decision boundary) the unknown sample belongs to (McLachlan, 1992; Duda et al., 2001).

The probability density functions of the two classes can also be modelled by a Naive Bayes classifier. Instead of being modelled explicitly in the function space, parameter estimation takes place using the Maximum Likelihood method. The Naive Bayes method also includes the assumption that any feature or pattern discovered in the data is completely independent of any other. Naive Bayes and LDA are what are called *generative models*, which are characterised by their reliance on conditional density functions (Mitchell, 2010; Duda et al., 2001). Despite generative models, and Naive Bayes in particular, performing better than would be intuitively expected (Hand and Yu, 2001; Webb et al., 2005; Rennie et al., 2003), generative models are usually outperformed by more recently developed classification methods (Caruana, 2006).

A different philosophy is used with *discriminative models*. These models seek to maximise the separation power of the classifier by minimising the *risk* (usually defined

by the observed rate of wrongly classified samples) on the training set by parameterisation. Such models include logistic regression (Hosmer and Lemeshow, 2000), the perceptron (Rosenblatt, 1957) and the support vector machine (SVM) (Cortes and Vapnik, 1995). The SVM represents, in our opinion, the most explicable and versatile algorithm of the linear classifier epoch in 21st century machine learning, and will be a major element of the research (see Chapters 4 and 5) contained in this thesis.

### 2.3.2 Regularisation and the Support Vector Machine

A classifier must not only separate the two classes on the training set, it must also do so in a way that draws a compromise between the influences of the classes' respective clusters. The decision boundary is ideally one that is sensitive enough to be contoured to the 'no-man's land' between these clusters, but generalised to the point of being easily described. However, in the case of ill-posed problems (which may manifest in the feature space as overlapping clusters and unorthodox cluster shapes), there is a friction between these two important goals which has not always been recognised. During the 1960s, many researchers believed that the minimisation of the training error was the best way to construct the most accurate classifier (Vapnik, 2000, pp. 6-7). This was subsequently proven to be wrong, as will be shown later, and has been usurped by a theory more in line with the parsimony principle named Structural Risk Minimisation (Vapnik and Chervonenkis, 1974). The main idea behind SRM can be described as a trade-off between the quality of the approximation of the given data and the complexity of the approximating function (Vapnik, 2000, p. 95). In the case of ill-posed problems, a by-product of training error minimisation is an intricate approximating function. As well as being more difficult to describe mathematically, approximating functions drawn by minimising training error frequently become *less* accurate at predicting the classes of unlabelled data points past a certain threshold.

By way of a real-world analogy, consider the case of a farmer whose property has a rabbit infestation. A certain region of this property is rabbit-free, evidenced by the lack of rabbit burrows, and he wants to build a fence to keep the rabbits from spreading to this area. He also wants to salvage as much land as possible that has not been infested,

so he builds curved fences tightly around the areas that he can see contain rabbit holes. However, this fencing approach fails, since there are a few rabbit holes outside the enclosed area that he has missed, through which the rabbits can escape. These ‘missed burrows’ can be thought of as a *test set* (See Section 2.8) of unlabelled data points that have been misclassified as a result of the farmer trying to follow the contours of the rabbit infestation too closely with his fences. Taking the goal of maximising non-infested land to the logical extreme, one can see that the approach of building a circular fence, the diameter of the hole, around each and every rabbit hole that the farmer finds would result in a large amount of useless work. More formally, the farmer has drawn a function that has *overfit* the data; a phenomenon where the training algorithm follows the anomalies of the data so closely that the resulting predictor loses the ability to generalise, resulting in decreased performance. To solve this problem the farmer needs to build a fence that follows a smoother locus. Although, by doing this, the farmer must compromise by sacrificing a broader area of land to be designated as infested, he will have a better chance of isolating the infestation within this broad area than if a smaller area was ‘trained harder’ towards the observed rabbit holes.

When separating data points in a function space, the generalisation or ‘smoothing’ of the decision boundary is known as *regularisation*. In the simplest terms, regularisation can be seen as a method of reducing model complexity. Adjusting the degree of regularisation of a model can be seen as a trade-off between biasing the model towards an easily described function, and being sensitive enough to describe the variance of the data. An intermediate zone between these two extremes will likely represent a degree of model complexity that best predicts unlabelled data points. Regularisation is a key component of building a robust learning machine.

The support vector machine (SVM) is machine learning tool that attempts to perform regularisation explicitly in the feature space. Like LDA, SVMs draw a decision boundary as an attempt to separate the two classes, but two parallel and equidistant margins either side of this boundary are drawn as well. The SVM algorithm actually draws the two margins first: for example, in  $\mathbb{R}^2$  (see Figure 2.1), one margin touching two data points belonging to the same class is drawn first, a parallel margin through a



data point from the opposite class is drawn second, and the decision boundary is then drawn halfway between. For linearly separable classes, the maximisation of the distance between the two margins, represented by the normed vector  $\|V\|$ , will confer the hyperplane that minimises the risk of the problem, as defined in SRM tradition. The proof for this is derived through Lagrangian multipliers under Karush-Kuhn-Tucker complementarity conditions and can be found in Smola et al. (2000). The distances to the decision boundary from the data points on the margin are called the support vectors, with lengths  $\frac{\|V\|}{2}$ . However, when classes overlap, extra constraints must enter the equation. The risk minimisation is then, in the support vector context, given by the minimisation of:

$$\frac{1}{\|V\|} + \frac{\sum \xi_i}{\lambda} \quad (2.7)$$

where  $\xi_i$  are the  $i$  ‘slack variables’, representing the distances of data points inside their corresponding margin, to that margin (See Figure 2.2).  $\lambda$  is a parameter that controls the degree of regularisation. A large  $\lambda$  will minimise the influence of the slack variables and result in a model biased towards maximising the margin width. Shrinking the value of  $\lambda$  will result in a more sensitively drawn decision boundary.

### 2.3.3 Non-linear Decision Boundaries and the Kernel Trick

In some cases, when the data is not linearly separable, the data can be transformed into a more tractable set. For example, the data may be transformed via a polynomial, or other non-linear function into a set which confers a lower risk than the original set. This method of data manipulation is called the *kernel trick* (Aizerman et al., 1964). It is often used in lieu of addition of further information through other features, as a possible solution to overfitting and density problems associated with a large feature subset. However, given the abundance of features at hand in a high-dimensional dataset, as well as the advent of coefficient shrinkage methods (see section 2.5), the kernel trick is not frequently used in bioinformatics. The kernel trick itself can be seen as a contributor, as well as a remedy, to overfitting, since it adds a degree of complexity to the model and

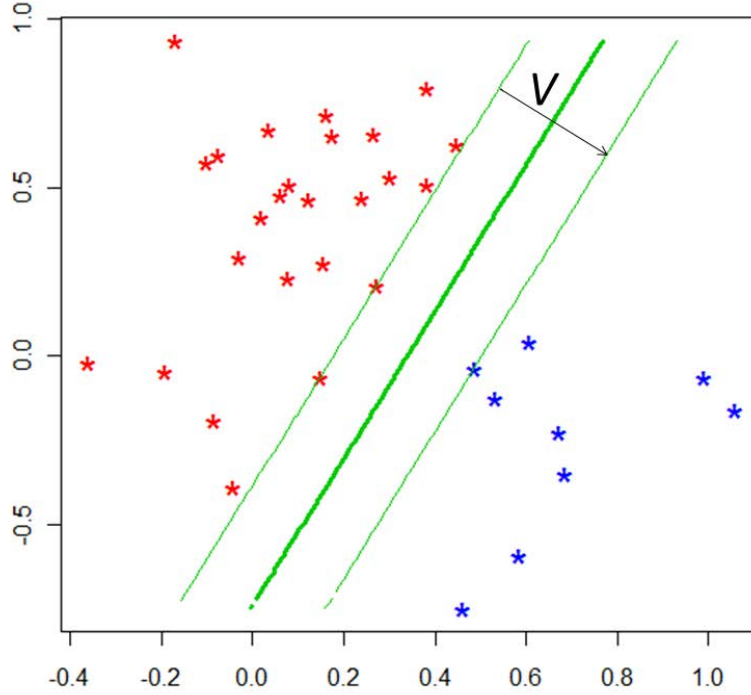


FIGURE 2.1: A support vector machine in  $\mathbb{R}^2$  separating two (red and blue) classes. Note the two red data points lying along one margin, and one blue along the opposite margin. Image created using R package *svmpath*.

increases computational time for the model's construction. In this thesis, no non-linear SVMs have been used to train data, although they provide a powerful visual aid for demonstrating regularisation. Figure 2.3 shows a two-dimensional data set with heavily overlapping classes. The kernel trick uses a nonlinear projection of the data onto the real line, and then draws classification boundaries and margins on the transformed line. For instance, each data point  $X$  can be represented by a combination of radial basis functions:

$$F(X) = \sum_{i=1}^N w_i G(X, c_i) \quad (2.8)$$

where  $G(X, c_i)$  is a fixed function radially symmetric around  $c_i$  and  $N$ ,  $w$  and  $c$  define the parameters of the transformation. These are sinuous and difficult to describe when under-regularized and, in the case of the decision boundary, non-contiguous. However,

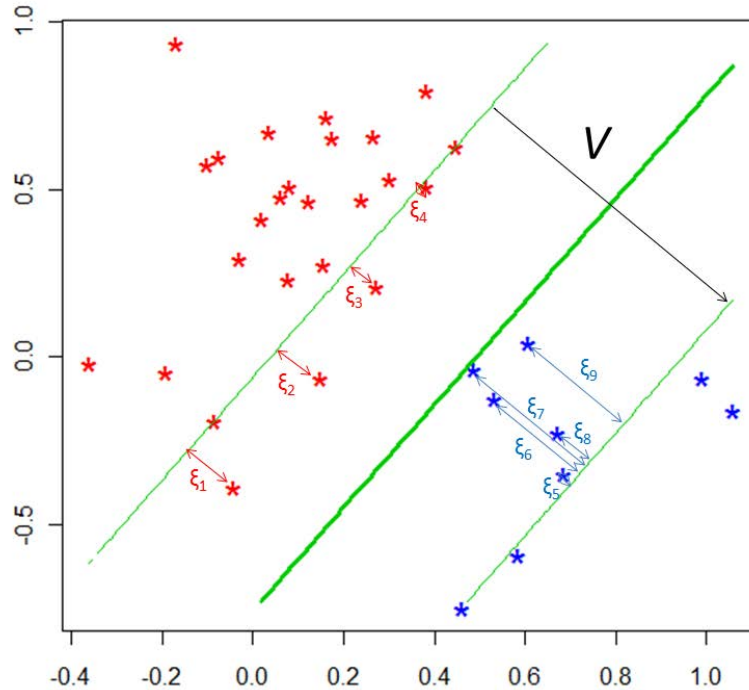


FIGURE 2.2: A more regularized support vector machine using the same dataset in Figure 2.1. The parameter  $\xi_i$  denotes the slack variables, whose sum is divided by  $\lambda$  and added to the norm inverse of  $V$ . Note that in the first case it is not clear that there will much gain in widening the margin since the classes are linearly separable. Figure 2.3, however, demonstrates the opposite and more common effect of regularisation. Image created using R package *svmpath*.

they are smoother and more easily described when  $\lambda$  is increased, and this decreases the LOO error by over 25%.

Unfortunately, there is no single value of  $\lambda$  that works best for all datasets, hence the optimal value must be estimated from data samples.

## 2.4 Architecture for Dimension Reduction Algorithms

In the context of this thesis, we define the *objective function* as a function to be optimised from an input consisting of a set of feature sets. A well-constructed objective function can inform the statistician about the properties of the input, such as how effectively labelled data points can be separated. Objective functions may be calculated

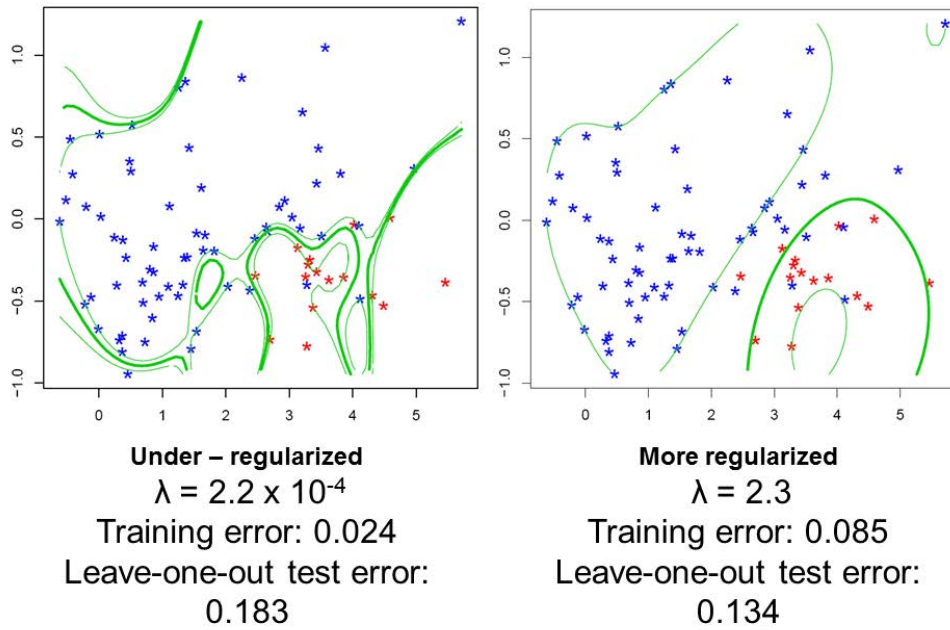


FIGURE 2.3: Comparison of two SVMs drawn on two identical dataset scatterplots. The decision boundary (thick green line) is heavily overfitted through an attempt to follow the training data too closely. The more regularized SVM on the right has a higher training error but a lower test error. Image created using R package *svmpath*.

by  $t$ -tests and the like, but they are only the bricks and mortar that constitute a statistical model; the model itself must have a contrived architecture which is constructed algorithmically. The process by which a feature selection model is built may take a number of forms.

Methods that select parameters to produce an objective function, such as SVMs, need a computationally fast method of finding the feature subset that confers an extremum of that function. Brute force selection, though guaranteed to find the global optimum, results in an unmanageable combinatorial explosion. For example, say we want to find the optimal dimensionally-reduced feature set whose size is 10, from the complete set of features whose size is 20000, subject to some objective function. This means that approximately  $\binom{20000}{10} \approx 3 \times 10^{36}$  objective function calls need to be computed. Fortunately, a number of effective alternatives to exhaustive search are

available.

We want to use our chosen objective function, built into an optimisation algorithm, to select a small group of features that best predict unlabelled classes. Building a classifier with all features present is likely to result in a horrendously overfit model, large computational costs, and no guarantee of high prediction accuracy. The simplest method of selecting a feature subset is forward selection, sometimes called forward stepwise regression (Efroymson, 1960). This is where features are added, one by one, to the model based on which feature confers the maximum (or minimum) criterion value. This can be seen as a *greedy* or *hill climbing* method. While the most rudimentary version of forward selection is a simple regression against the response variable, the incorporation of a more sophisticated objective function (such as SVM or LDA) in lieu of the regression is likely to yield better results. The *cardinality* (hereby defined as the number of features in the predictor set) of the model will vary according to the data set and hence must be estimated. This can be achieved by the construction of a wrapper, whereby cardinality is regressed against test error and the model conferring the lowest error is chosen, or by a stopping criterion such as Mallows' Cp (Mallows, 1973). Further discussion on this subject can be found in Chapter 6.

Another way of obtaining a highly predictive feature set is through backward elimination. As the name suggests, all features are present in the model at the start of the learning process and most features—ideally all uninformative and redundant ones—are eliminated from the feature set in a stepwise fashion. This has been done with SVMs (Guyon et al., 2002) with promising results.

### 2.4.1 Filters, Wrappers and Embedded Algorithms

*Filters* are models where the objective function is assumed to provide an accurate indication of the separability of classes, and as such, filter out the features with individual objective function (criterion) values that either do not pass a particular threshold, such as the top or bottom quantile of a ranked list of correlation coefficients. Filtering variables is often used as a data preprocessing step (Guyon, 2003) in order to create a smaller, and thus more workable, subset of genes for further reduction. A filter routine

will simply take the top individually-ranked features in a dataset by their criterion value, with the assumption that they constitute a superlative predictor. In Chapter 3 we show this to be an inaccurate assumption.

*Wrappers*, however, are more sophisticated methods that can evaluate features as components of a set, rather than individually. Wrappers, as considered here, assume nothing about the way a feature set is initially selected, and treat the reduced feature set as a black box (Guyon, 2003; Kohavi and John, 1997), where the models they build are assessed on their posterior predictive power on unseen data. The defining characteristic of the wrapper heuristic is that it contains a nested, or ‘wrapped’ structure, which can be, for example, a cross-validation (see Section 2.8.2) within a cross-validation. The posterior performances (given by a predictive algorithm) of the inner validation routine is itself used as the training criterion to select a feature set for assessment by the outer validation routine. Intuitively, a wrapper will produce more reliable classifiers because of this inbuilt validation mechanism. Yet, since a classifier must be both built and assessed for each feature subset (hence, the nested, or *wrapped* structure), it is unsurprising that wrapper learning is, generally speaking, computationally burdensome. As an additional drawback, wrappers are also prone to overfitting, since their method of risk estimation is unregularised (Kohavi and Sommerfield, 1995; Guyon, 2003; Loughrey and Cunningham, 2004; Reunanen, 2003).

An *embedded method* is one that produces both a feature subset and a classifier during the same process. The process is ‘active’ in the sense that the algorithm contains iterations where information on the classifier’s current performance on iteration  $i$ , using a user-defined objective function as a proxy, is fed back into the algorithm on iteration  $i + 1$ , in an attempt to improve on the performance until some stopping criterion is reached. Unlike the filter method, the objective function is calculated as a measure of the performance of the whole subset at the current point in the routine, instead of each of its individual features. The method can then actively ratchet up the separation and prediction power of the classifier, depending on the criterion value and algorithm heuristics. Such methods are the choice of most statisticians, and include modern methods such as the LASSO (Tibshirani, 1996a) and Random Forests (Breiman, 2001).

### 2.4.2 Stochastic Methods

The problem of finding the most informative subset of features is one of *combinatorial optimisation*: finding the combination of elements conferring the optimal result out of a set of possible configurations. Forward selection and backward elimination find the optimal subset via a *deterministic* technique, that is, there is only one solution for each dataset the algorithm is applied to. The solution also has only one optimisation route; each process will be identical every time the routine is run. *Stochastic*, or *Monte Carlo* methods (Ripley, 1987) introduce an element of randomness into the selection process, which means that a different feature set is possible with different runs with identical parameterisations. The main advantage of stochastic methods as they apply to feature selection is that they allow greater flexibility in exploring the space of possible feature combinations. Stochastic methods allow, to a certain degree, the algorithm to ‘backtrack’ on a particular feature-finding path if another is found to render a deeper criterion value. In contrast, forward selection and backward elimination do not allow feature replacement; once a feature has been selected (or omitted) it remains that way. Two stochastic optimisation methods will be discussed here.

#### Simulated Annealing

Simulated Annealing (SA) (Kirkpatrick et al., 1983) is a method which attempts to circumvent the tendency of greedy methods to halt at local optima in the feature space. The classic analogy is that found in metallurgy: heat allows atoms to become unbound from their current energy state and find more stable energy states through slow cooling. SA temporarily permits moderate deterioration in the current objective function value while the routine is ‘hot’, but this permissiveness is gradually rescinded during the cooling schedule. The computational equivalent involves substituting one feature for another based on the gain or loss in the objective function value as a routine within a nested iterative procedure. The number of substitutions allowed, and the size of the subset of neighbour candidates at each substitution is at the discretion of the programmer, but may be informed by the available computational power. SA subverts

the ‘push-pop’ memory stack style of adding and subtracting features from the model used by forward and backward selection algorithms, by allowing substitution at any point in the stack. Despite the total number of features being large, the number of available neighbours is limited by the small size of the target feature set, making simulated annealing a viable option for feature subset search.

### **Constraint Relaxation: The Cross-Entropy Method**

The nature of the data often presents problems for the statistician. A single feature seems insignificant in an entire dataset, but as an  $n$ -sized vector contributing to a small feature set, it can still contain redundant information and exhibit a high degree of variance (Hastie et al., 2009, p. 61). Hence the dichotomy between including a feature in a model, and not including it, is often unwieldy. Implicit in the architecture of the most recent dimension-reduction methods is a recognition that assigning equal coefficients, or ‘weights’, to each feature present in a model is an arbitrary constraint. In essence, we want to select only the useful pieces of data from  $X$  without being forced, necessarily, to select whole rows or columns in the data frame, but also without compromising the integrity of the data. The power of SVMs lies in their ability to be partial to selected *data points* in the feature space when demarcating classes, but this is done only when features have already been selected. The idea of using ‘partial’ *features* can be implemented in a number of different ways, and represents a form of mathematical relaxation (Goffin, 1980). By way of a linear programming (LP) analogy, the integrality constraint is removed, and features are allowed to possess weights between the values of 0 and 1. The statistician can then implement a method that iteratively updates the weights of the candidate features by increments less than 1. Such an approach is analogous to an interior point method (Wright, 2004), where a discrete constraint (for example, setting the weights of all features to an equal value with some upper bound on their sum) is set in the routine initialisation, and this constraint is gradually imposed on the parameters of the problem in the continuous feature space. Returning to the LP analogy, an interior point method (with relaxed parameters) provides an alternative to the coarser approach of hopping between the extremes of the feasible region, as the



simplex method (Cormen et al., 2001b, pp. 790-804) does.

The Cross-Entropy (CE) method (Rubinstein and Kroese, 2004) is a stochastic method for combinatorial optimisation developed especially to aid rare-event probability estimation. Through parameter tuning, it can be used to display emergent properties of a system through adaptive updating of an iterative procedure. The CE algorithm has applicability for a large and diverse array of problems, so we will only describe it in the context of feature selection here. Consider a vector  $V$  size  $P$ , the tally of all candidate features for our dimensionally reduced feature set  $Q$ , and an integer  $q$  (preferably a factor of  $P$ ) the size of the set we want. All elements of  $V$  are set to the value (weight)  $\frac{q}{P}$ , such that the sum of the elements is  $q$ .  $V$  can be seen as a vector of the probabilities of some feature  $k$  being included in  $Q$ , or  $V(k) = P[\text{'k is included in Q'}]$ . More formally:

$$V(k) = P[k \in Q], k = 1, \dots, P \quad (2.9)$$

This paragraph will describe one whole iteration of the CE algorithm. Random samples from the complete feature set based on  $V(k)$  are divided into  $\frac{P}{q}$  subsets, (or such that each and every feature is given a reasonable chance to be assessed) size  $q$  with replacement after the creation of each subset, and the performance of each subset is rated by a given objective function. The best performing feature sets (given by a user-defined percentile), are retained to influence the next generation: The vector of probabilities  $V_t$  is shifted towards the features' sample frequencies in the retained sets. The amount of shifting is controlled by a smoothing parameter  $\alpha$ ; if, in the current generation,  $V$  represents the vector of probabilities used and  $V^*$  represents the retained sets' sample frequencies, then the next generation's vector of probabilities is given by:

$$V_{t+1} = \alpha V_t^* + (1 - \alpha) V_t \quad (2.10)$$

Iterations are performed until some stopping criterion is reached, usually defined as when the values of all  $q$  features in  $V$  approach 1. The smoothing parameter  $\alpha$  is an important user-defined variable used to control how quickly the algorithm reaches

convergence. Small values of  $\alpha$  are seen as ‘fairer’ on all features since they allow slower convergence, and hence more opportunities exist for features to gain a foothold in the ratcheting process typical of feature behaviour in this algorithm (see Figure 2.4). Large values of  $\alpha$  run a higher risk of  $Q$  becoming stuck in a shallow local optimum, since fewer features are considered overall. The CE method can be computationally demanding, however, and hence minuscule values of  $\alpha$  are inadvisable (for a discussion on parameter selection and computational costs, see Chapter 6). The issue of complete feature coverage will be discussed more fully in Chapter 3. Conferring a similar effect on algorithm behaviour to  $\alpha$ , albeit inversely, is the size of the percentile used to select elite feature sets. Test runs are usually performed in order to estimate appropriate values for both.

## Progression of $V$ through one Cross-Entropy run with $\alpha=0.02$

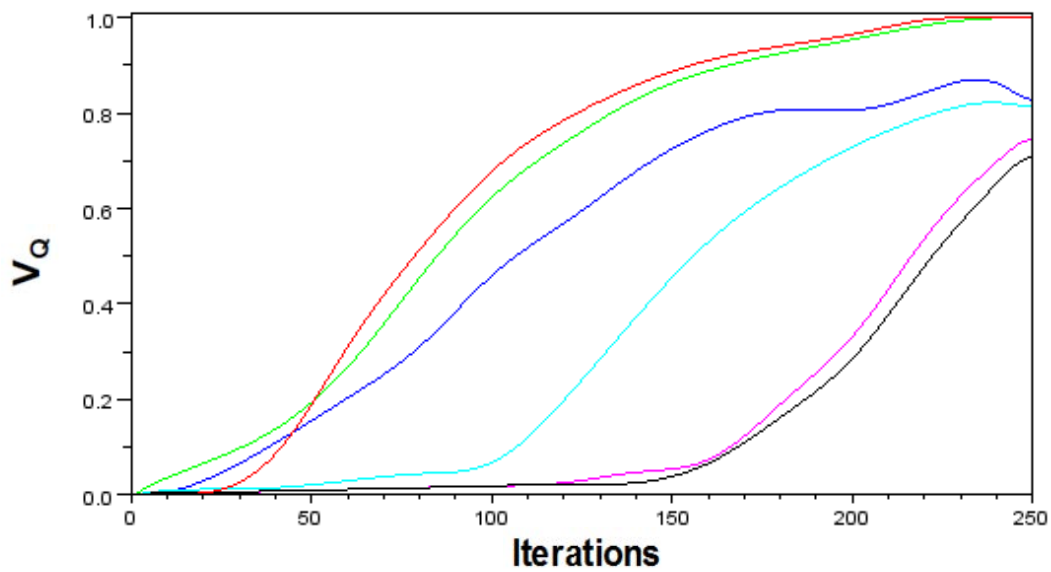


FIGURE 2.4: Paths of the top six feature weights in the probability vector  $V$  in the first 250 iterations of a CE method run where  $q = 6$ ,  $\alpha = 0.02$  and the top decile of feature sets are extracted to create  $V^*$ . This is a very small value of  $\alpha$ , given that Rubinstein and Kroese (2004) often use values upward of 0.7. Of particular importance is how the features represented by the magenta and black series do not gain a significant foothold until around iteration 150. This implies earlier iterations of  $V$  contained other features not shown here, whose weight values decreased as they were replaced. These other features may well have ended up in  $Q$  if a larger value of  $\alpha$  were used. Also of note is the sudden rise of the red feature at iteration 30 as the most favoured feature, overtaking the green, and the apprehensive behaviour of the blue feature, showing a slight decrease even near iteration 250. It is likely in this scenario that feature competition is fierce, and the value of  $q$  may need to be revised upwards.

## 2.5 Deterministic Feature Selection via Coefficient Shrinkage

The CE method assesses features for their suitability for inclusion in the final feature set as probabilities based on their performance. An alternative strategy to this is viewing these suitability scores as *coefficients* in a covariance matrix. An early implementation of this strategy for the dimension-reduction problem can be found in Forward Stagewise Regression (Weisberg, 1980). Beginning in the manner of forward selection when choosing the first feature, the algorithm then selects subsequent features most correlated with the current residual. Each feature's coefficient is updated at every step by the addition of the linear regression coefficient on the current residual until no more features are correlated with the residuals. Forward Stagewise Regression has been superseded in both computational time and accuracy by shrinkage regression methods (Hastie et al., 2009, p. 73), which we will discuss now.

Consider a simple regression of the feature matrix  $X$  against the binary response variable  $Y$ . In the case of stepwise selection using residual sum of squares (RSS) as the objective function, the vector that contains the coefficients of all features will also be binary, simply determining whether a feature is in or out of the dimensionally-reduced feature set. Shrinkage methods apply the relaxation metaheuristic such that the coefficient values can be intermediate of 0 and 1, providing a more continuous solution than stepwise regression, and varying the degree of influence of each feature by its corresponding coefficient as each feature is introduced into the model. Line graphs may be helpful in visualising the coefficient paths of each feature as the model is constructed, since their functions are piecewise continuous.

Shrinkage methods are usually presented in the form given by the least squares error estimate:

$$F(B) = \sum_{i=1}^n (Y_i - B_0 - \sum_{j=1}^P X_{ij}B_j)^2 \quad (2.11)$$

where  $B_0$  is the intercept. A constraint is then imposed on this function:

$$\sum_{j=1}^P |B_j|^\gamma \leq q \quad (2.12)$$

where  $\gamma$  is a penalty parameter, and  $q$  is an upper bound on the sum of the transformed coefficient. The parameter  $q$  acts in a similar fashion to  $q$  in the CE method—providing an upper limit on the amount of information in the model for the purposes of feature selection—except that in shrinkage methods it does not necessarily imply the number of non-zero features the model contains. The parameter  $\gamma$  is perhaps the most influential in determining the outcome of  $F(B)$  (see Figure 2.5); it allows a non-Euclidean metric for penalised residuals, and hence affects solution density and manifoldness. The constraint  $\sum_{j=1}^P |B_j|^\gamma \leq q$  bounds the  $\ell^\gamma$ -norm of the parameter vector  $B$ . When  $\gamma = 2$ , the resulting method is called ridge regression (Hoerl and Kennard, 1970), where the Euclidean penalty *is* used and feature coefficients are shrunk proportionally. When  $\gamma = 1$  it is known as LASSO (Tibshirani, 1996a), where coefficients are shrunk by constant factor and truncated at zero. LASSO is ‘least absolute shrinkage’ because it represents the smallest value of  $\gamma$  for which shrinkage can be performed while the constraints remain *convex* (see Figure 2.5). The primary advantage of a convex constraint region is its guarantee of conferring one, and only one, solution. Shrinkage using concave constraint regions (where  $0 \leq \gamma < 1$ ) runs the risk of obtaining multiple solutions, which must be independently evaluated, increasing computational time. Conversely, as  $\gamma$  increases when it is greater than 1, the solutions become denser.

Other forms of penalisation can be found in algorithms such as include RLAD (Wang et al., 2006), the Dantzig Selector (Candes and Tao, 2007) and MC+ (Zhang, 2007). Such methods are effective, but lack a flexibility that allows the penalisation form to be varied within the same framework. In terms of the shrinkage framework the parameter  $\gamma$  is clearly critical in determining the outcome of the model. Similar to the way  $\lambda$  regulates the trade-off between minimising model bias and minimising variance,  $\gamma$  can also be placed on a continuum regulating two desirable, yet

opposed model characteristics, namely sparsity and convexity.

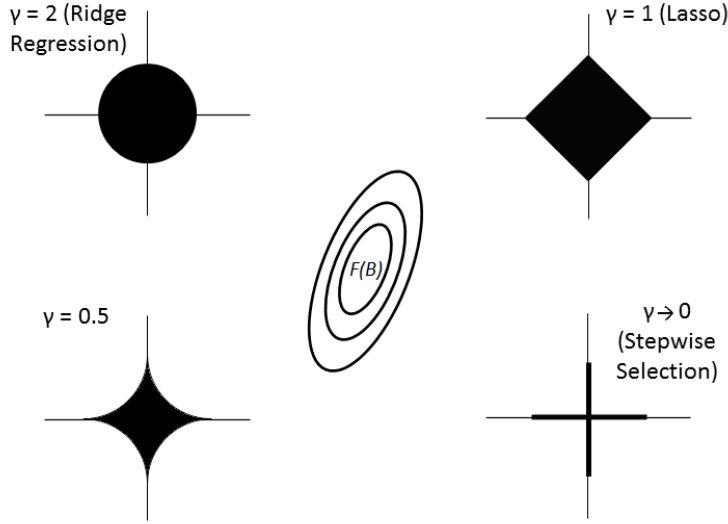


FIGURE 2.5: A selection of estimation pictures for  $|B_1|^\gamma + |B_2|^\gamma \leq q$  for  $\gamma = \{2, 1, 0.5\}$ ,  $\gamma \rightarrow 0$ . In each subfigure, the black area is the unit disc according to the norm defined by  $\gamma$ . Our aim is to minimise the ordinary least squares error function, around which the ellipses are centred, subject to the constraints shown. As the least squares error function touches the constraints, we can see that we get a unique solution for  $\gamma \geq 1$ , but an increased risk of a dense solution (not touching at an axis) as  $\gamma$  grows. A sparse solution is one that has a few non-zero coefficients. In this 2D diagram, such a solution would be represented by a point lying on one of the axes. However, if  $\gamma < 1$ , the ellipse has a greater probability of touching two points on the constraint circle simultaneously, resulting in multiple solutions,  $q$  solutions, in fact in  $\mathbb{R}^q$ , as  $\gamma \rightarrow 0$ . Figure is based on the Lasso paper by Tibshirani (1996).

The aforementioned point is of vital consequence when undertaking feature selection. Recall that our objective is to select a small subset of features, where the final feature vector only has a handful of non-zero coefficients. Such a solution can be described as *sparse*. Most scientists desire a concise, qualitative report; perhaps describing what the statistician sees as features with large coefficients as key biomarkers, and those with lower ones as ancillary biomarkers. A long list of putative biomarkers with differing coefficients is unhelpful to a scientist wishing to illuminate the governing mechanisms of disease.

There is also a statistical reason for tending towards a sparse model. When the number of features with non-zero coefficients is equal to or exceeds the sample size, the estimate of the covariance matrices becomes singular (Schäfer and Strimmer, 2005), resulting in a trivial solution, since the sample space has as many dimensions as data points themselves. Such a property provides a natural and elegant stopping criterion, as well as a solution path for shrinkage methods such as Least-Angle Regression, which we will discuss now.

### 2.5.1 Least Angle Regression

A few more recent variations on the regression shrinkage metaheuristic have been developed. Least Angle Regression (Efron et al., 2004) attempts to further alleviate the cumbersome task of fitting an entire feature to the model by continuously moving the value of its coefficient toward its least squares value, until another feature becomes more correlated to the response variable, at which it is added to the model along with the existing features. This is opposed to fitting the entire feature to the model as per LASSO (albeit with a coefficient less than 1), and thus LARS is superior in terms of both eliminating noise and ensuring ‘fairness’ to all competing candidate features.

### 2.5.2 Elastic Net and Generalised Path Seeking

There is a tension between the tendency of LASSO to make arbitrary choices among groups of highly correlated features with high class-separation power for inclusion in the active model, and the tendency of ridge regression to shrink the coefficients towards each other. These events may produce undesirable results in both cases (Hastie et al., 2009, p. 662). The dual objective of quickly isolating these features and determining their coefficients via a nuanced analysis needs a compromise, and so the Elastic Net penalty (Zou and Hastie, 2005) was developed, using  $\sum_{j=1}^P (\alpha |B_j| + (1 - \alpha) |B_j|^2)$  instead of  $\sum_{j=1}^P |B_j|^\gamma$ ,  $\gamma = \{1, 2\}$ , where  $\alpha$  acts as a tuning parameter (similar to the smoothing parameter in the CE method) depending on the degree of sparsity the problem needs.

However, there are circumstances, such as when features have a much lower degree of correlation, where an even sparser solution than the LASSO produces superior results. Generalized Path Seeking (Friedman, 2008) provides a broad framework by incorporating a penalty  $E(B)$  (such as  $\sum_{j=1}^P |B_j|^\gamma$ , but may also include the elastic net penalty and others) into the least-squares loss function  $L(B) = \sum_{i=1}^n (Y_i - B_0 - \sum_{j=1}^P X_{ij} B_j)^2$  such that  $R(B) = L(B) + \lambda E(B)$ , with  $\lambda$  functioning as a regularisation parameter. Note how this form mirrors that of that of the SVM objective function (see Equation 2.7, albeit with  $\lambda$  acting inversely here), with the loss function  $L(B)$  analogous to the margin width and the penalty function  $E(B)$  comparable to calculating the sum of the margin overlaps. Through linear programming with the use of Karush-Kuhn-Tucker conditions (the formal procedure is shown in Friedman (2008)), GPS can approximate the coefficient paths, as  $\lambda$  is varied between 0 and  $\infty$ , of all penalty functions satisfying  $\frac{\partial E(B)}{\partial |B_j|} > 0$  (which applies to both the classic penalisation method and the Elastic Net) for all samples,  $0 \leq \gamma \leq 2$ . Penalties that satisfy these conditions include Elastic Net, RLAD, MC+ and many others, for which GPS earns its ‘generalised’ epithet. It is our opinion that, at the time of publication, GPS provides the most flexible implementation of the coefficient shrinkage metaheuristic, and hence is ideally suited to bioinformatic dimension reduction tasks.

## 2.6 Decision Trees

The final feature selection metaheuristic to be discussed in this chapter is decision trees. As SVMs make the decision boundary explicit in the feature space, and shrinkage methods use a coefficient model to quantify each feature’s relative contribution to the final model, decision trees make explicit the criteria that separate the binary variable  $Y$ . It is possible to map decision boundaries of trees in the feature space, but they are frequently non-contiguous and too complicated to describe as a plotted function. They are visualised as multiple rectangular hyperplanes of varying sizes in the feature space, but are much more easily interpreted in tree form. Decision tree methods have



their genesis in a technique called Automatic Interaction Detection (AID) (Morgan and Sonquist, 1963) and its extensions Theta AID (THAID) (Morgan and Messenger, 1973) and Chi-squared AID (CHAID) (Kass, 1980), but their most well-known implementation is the Classification and Regression Tree method (Breiman et al., 1984). CART dichotomously splits the samples in a recursive fashion (Figure 2.6), determining data splits with simple relational operators ( $<$ ,  $\leq$ ,  $\geq$ ,  $>$ ), using only one feature for each node. The standard tree is similar to that shown in Figure 2.7, except the splitting criteria are mathematical and not qualitative.

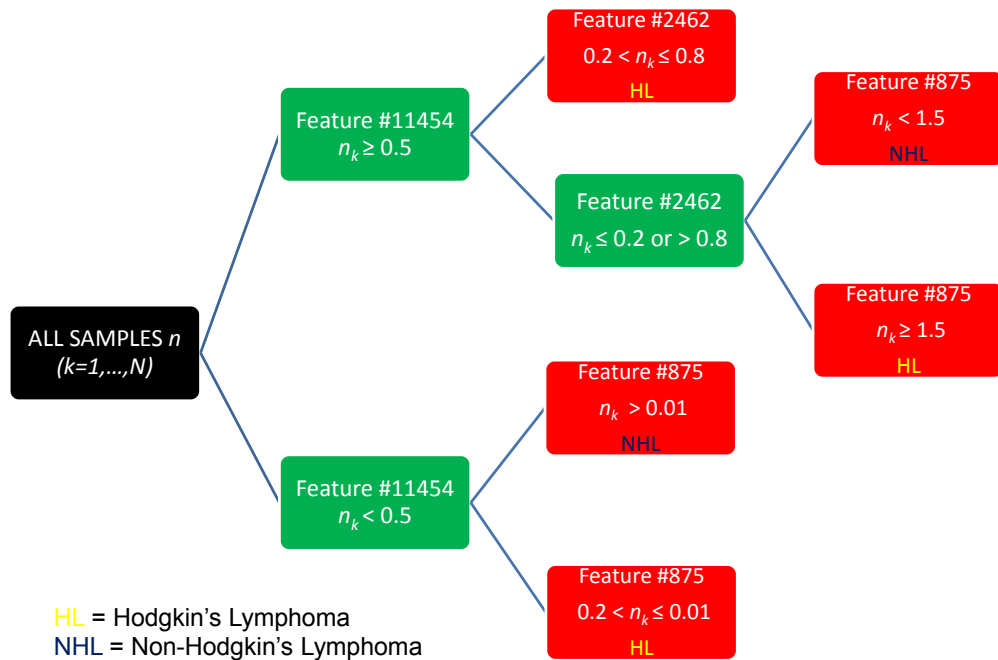


FIGURE 2.6: A simple decision tree distinguishing between Hodgkin's and Non-Hodgkin's Lymphoma. Terminal nodes are in red and decision nodes are in green.

Just like SVMs, decision trees are prone to both under and overfitting and need to be parameterised for optimal results. A tree with too few nodes may fail to capture the governing structure of the model, so a lower bound on the depth of the tree is usually set (Hastie et al., 2009, p. 308). More common, however, is construction of trees with too many nodes which contain splits that provide little or no improvement on the training accuracy (or equivalent measurement) of the tree. These trees should

be recursively pruned by removing the node that incurs the least reduction in this value (this is called weakest link pruning, Hastie et al. (2009, p. 308)), after assessing all branches and their corresponding error rates along that path. The branch with the error rate corresponding to a pre-defined threshold should then be selected. In practice, most trees are grown then pruned in a trial-and-error phase before this threshold is chosen.

Most modern implementations of CART do not use the training error per se in their evaluation of which nodes to prune. Instead, they use the Gini index as a measure of model risk, which can be calculated by simply multiplying the proportion  $P_k$  of samples from the majority class in the region defined by the node, by the minority proportion, producing the term  $P_k(1 - P_k)$ . Since the Gini index is a ratio analysis rather than an absolute, it is more sensitive to node purity: a shorter term for how well the classes are separated by the node split.

## 2.7 Random Forests, Bootstrapping and Stability Selection

Despite its elegance and intuitive design, one of the disadvantages of the CART method is its inherent instability. If a dataset is updated or changed, even a small adjustment may result in a completely different model, since a change in a criterion near the root can be propagated all the way down the tree (Hastie et al., 2009, p. 312). This problem is solved by Random Forests (Breiman, 2001), an ensemble learning method that grows multiple trees, and averages their prediction output through a process called bootstrap aggregation, or *bagging*. Each tree acts as a predictor, and the classification is achieved through a voting majority from this group of trees. As the name suggests, each tree in the Random Forest stochastically picks a subset of features, typically size  $\sqrt{P}$ , where  $P$  is the total number of features. Care is taken to ensure every variable is given a chance to be evaluated. With a random selection of features, each tree is constructed in a similar fashion to CART, but with one major difference: a proportion of samples

are left out. These are called the out-of-bag (OOB) samples. Like a wrapper algorithm with an internal cross-validation routine, the performance of each tree is evaluated by testing on these OOB samples before voting takes place. Importantly in the context of this thesis, Random Forests also calculate an importance index for each feature. This is achieved by initially passing the OOB samples down the trees and recording the prediction accuracy for each tree, followed by a random permutation of the values in the feature vector and then passing these values down the trees again. The variable importance is estimated by the degree by which the prediction accuracy drops on the second pass.

Random Forests are viewed, along with SVMs and the latest shrinkage methods, as state-of-the-art learning algorithms. This is because they are able to incorporate three of the most desirable facets of statistical learning: an internal validation mechanism, a stochastic element, and variance reduction via bootstrapping. The latter attribute is of increasing importance to statisticians who are involved in bioinformatics, since the stability of the results often has a bearing on the bona fides of the biomarkers found (Meinshausen and Bühlmann, 2010). Cross-validation is established as the best way to evaluate the performance of a learning method, but with feature selection the goals are different. Ultimately, the validation of the meaningfulness of features selected in a bioinformatics context takes place in the laboratory and in clinics, but there is a further level of quality control that the statistician is able to implement before his or her results are reported to a team of scientists. A single training estimate—especially a deterministic one—on a dataset may seem sufficient to determine the appropriate coefficients of its feature set. However, variable dependencies and noise will inevitably interfere with this estimation, and give a result that is a distortion of the true significance of the features chosen. A remedy called Stability Selection (Meinshausen and Bühlmann, 2010) has recently been proposed which suggests taking multiple bootstrap samples from the dataset of size  $\frac{n}{2}$ , applying a dimension-reduction routine separately for each one, and aggregating the results. By way of example, using a motif dataset, Stability Selection in conjunction with the LASSO has been used to successfully isolate two genes—one governing and one ancillary—implicated in transcription binding

sites, whereas a single run of the LASSO yields 26 genes whose coefficients are much less discernible (Meinshausen and Bühlmann, 2010). One of the most important messages to come from this example is that, especially in the case of the ancillary gene, the value of the aggregated LASSO coefficient of the genes is less important than the *frequency* with which the genes appear in the dimensionally-reduced sets. Stability selection highlights the need for consistency of feature performance across all, or nearly all, bootstrap samples before consideration as a truly influential biomarker.

## 2.8 Machine Learning and Classification

The development of dimension-reduction algorithms has its genesis in the intuition of their architects, but such creativity has little worth if it fails to produce real-world results. In software engineering, the ability of a method to produce meaningful results is primarily assessed by its performance in validation procedures. In the context of bioinformatic diagnosis, the appropriate procedures are supervised machine learning and classification. Machine learning encompasses a broad range of methods designed to ‘learn’ the contours of a dataset, thereby generating a model that incorporates, at the very least, its most distinguishing features. The learning algorithm is said to be *supervised* if the class of the samples is already known (in other words, it is accompanied by a response variable), and *unsupervised* if it is not. A typical microarray-generated dataset will already have its samples diagnosed by a pathologist based on their phenotype, and thus a model can be generated using a bipartite response variable  $Y \in \{a, b\}$  that finds features that best separate the two classes. Often, a class hierarchy with dichotomous splits is used to represent multiple levels of relevant sample classes (such as in Figure 2.7), and an analysis is performed on each split, each characterised by a different set of features.

Unsupervised learning algorithms tend to favour clustering methods, so that generalised class demarcations can be drawn (Duda et al., 2001). In supervised learning the goals are very different: the learning technique needs to be validated through the prediction of unlabelled data. A model is generated by a given learning technique

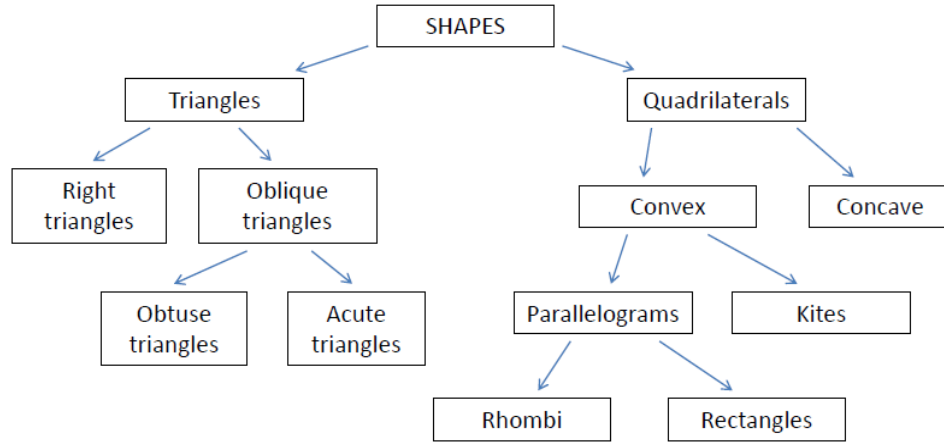


FIGURE 2.7: A non-exhaustive dichotomous taxonomy of well-known two-dimensional shapes.

using a set of samples called the *training set*, where the response variable is known. Subsequently, a separate, usually smaller *test set* of samples, whose response value is unknown to the model, is classified by the model. The outcomes of these predictions can then be checked against the actual response values to evaluate its performance, which can be expressed in a number of ways.

A sample with a response value of  $a$  that has been classified as  $a$  is labelled a True Positive (TP); if it is classified as  $b$  it is a False Positive (FP or Type I error). Similarly a sample with response value  $b$  classified as  $b$  is a True Negative (TN) and  $a$  a False Negative (FN or Type II error). Accuracy can be expressed as a single value  $\frac{TP + TN}{TP + FP + TN + FN}$ , however, a more nuanced and informative approach is usually taken. Sensitivity  $\frac{TP}{TP + FN}$  and specificity  $\frac{TN}{FP + TN}$  are used to construct a receiver-operating characteristic (ROC, see Figure 2.8), which, instead of indicating the test accuracy, evaluates how well a classifier can separate the two data classes. One ROC

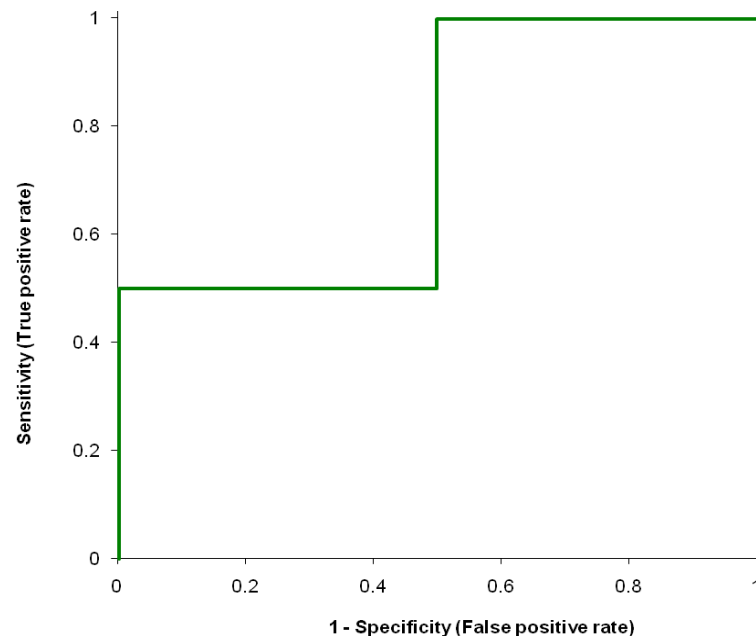
curve can represent a family of classifiers that differ only in their decision threshold; this threshold can be plotted at various points along the curve as a way of ‘tuning’ the classifier.

The construction of a ROC in the case of bipartite sample classification is usually calculated by ranking the scores, predicted by the classifier, given to each of the tested samples. The two axes represent, from 0 to 1, the TP rate and  $1 -$  the FP rate as the classifier traverses this ranked list.

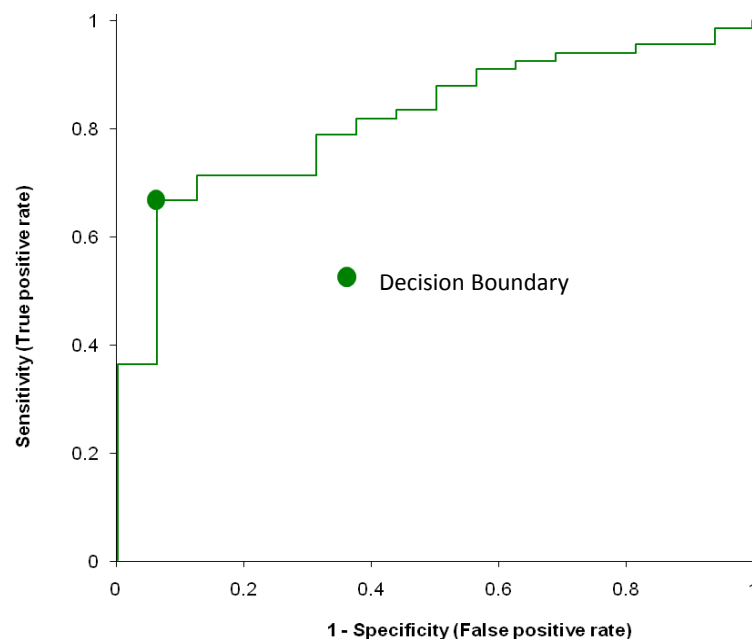
The following example is illustrated by the ROC in Figure 2.8(a). A test set of samples  $k_1, k_2, k_3$  and  $k_4$  may have response values 1,  $-1$ , 1 and  $-1$  respectively, but the classifier predicts their scores as 0.3, 0.1,  $-0.1$  and  $-0.3$  respectively. The path traced from  $(0, 0)$  to  $(1, 1)$  represents the traversal of these ranked scores from largest to smallest. When encountering a sample whose actual response is  $Y = 1$ , the ROC curve moves upwards a measure of  $\frac{1}{\text{TP} + \text{FP}}$ , and when encountering a sample whose actual response is  $Y = -1$ , the curve moves a measure of  $\frac{1}{\text{TN} + \text{FN}}$  to the right. Thus, a perfect ROC will move upwards and pass through  $(0, 1)$  before moving to the right. It will also have an Area Under Curve (AUC) of 1. The shape of a ROC and its AUC do not necessarily inform the accuracy of a classifier; they instead inform the *separability* of the two classes. Measuring the  $\ell^1$  distance from the decision threshold point (which represents a predicted score of 0 when  $Y \in \{1, -1\}$ ) to  $(0, 1)$  provides a reliable indication of the accuracy of a classifier when performing comparative tests (see Figure 2.8(b)).

### 2.8.1 Nearest-neighbour Classification

The  $k$ -Nearest neighbour (KNN) method is a common classification algorithm used in supervised learning that explicitly uses the feature space to determine the class of an unknown sample (Dasarathy, 2002). Given one of these samples, its classification is determined by a ‘majority voting’ system, where the  $k$  nearest training samples, from the sample in question, are identified in the feature space. When there are  $\frac{k}{2}$  samples from each class, the sum of the Euclidean distances from the sample in question to the



(a)



(b)

FIGURE 2.8: (a) A Receiver Operating Characteristic (ROC). (b) ROC with a dot representing the decision boundary.

training samples for each class is taken, and the tested sample is awarded to the smaller of these sums. Given that the classification only depends on the local structure of the data, the computational time can be limited by smaller values of  $k$ . However, KNN classification is sensitive to noise, and should never be used as a classifier on  $P \gg n$  datasets without performing dimension reduction first (Duda et al., 2001; Parvin et al., 2008). Care must also be taken when class numbers are unbalanced; a weighting system should be employed in this case (see Section 6.1).

Although KNN has been superseded by other methods with respect to classification, it has a practical use for statisticians as an imputation tool for missing values in a high-dimensional dataset (Troyanskaya et al., 2001). If there is enough real data for a particular feature (say, a minimum of  $\frac{n}{2}$  samples have actual gene expression values), then the remaining values can be ‘filled in’ using the KNN algorithm, relying on more complete feature vectors with similar profiles to the one in question.

## 2.8.2 Cross Validation

The performance of one learning algorithm relative to another cannot be evaluated analytically. Only through repeated exposure to real-world datasets can a statistician be confident of the training technique he or she puts into practice will yield meaningful results. As mentioned before, assessing the performance of a training algorithm involves the partitioning of the data into a training set and validation (test) set. Similar to bootstrap estimation (Efron, 1979), a more accurate estimate of the predictive power of a model is achieved with a resampling technique. In the context of machine learning, this method is called *cross validation* (Lachenbruch and Mickey, 1968; Geisser, 1993; Kohavi, 1995; Devijver and Kittler, 1982).

A finite amount of available data demands its repeated partitioning into training and validation samples, which are recycled as either training or test points for each new round of validation. There are a number of ways to do this:

- $k$ -fold cross validation splits the total number of samples  $n$  into  $k$  subsamples, where each subsample is used as the test set for each of the  $k$  validation iterations,



and the remaining  $n - \frac{n}{k}$  samples are used for training. The results are then averaged across all subsamples. The default for most routines is usually  $k = 5$  or  $k = 10$  (Fushiki, 2009). Each sample is resampled an equal number of times; however, this method is susceptible to outliers clustering in one subsample group. Care must be taken to ensure that each subsample is *stratified*, that is, they contain equal proportions of samples labelled  $a$  and  $b$ .

- Leave-one-out cross validation (LOOCV) is where a single sample is used as the validation set for each iteration, and the training set constitutes the remaining  $n - 1$  samples. LOOCV is more thorough in that it removes the problem of potential outlier clustering, but is always computationally expensive due to its  $n$  iterations.
- Random resampling cross validation (RRCV) randomly selects the samples used in the training and test sets with replacement, with a user-defined test/train ratio (Tibshirani, 1996b). Although it is able to overcome outlier clustering and is able to be evaluated in reasonable computational time, some samples are likely to be over-represented, and others under-represented. Hence, the number of resampling rounds should be suitably high.

Cross-validation has another role in classification: the testing of hypotheses suggested by the data. All data sets, when trained, usually contain some artifacts that are by-products of the training process, which may suggest a biological trend or phenomenon where none actually exists. When performing cross-validation, a statistician must also possess the healthy scepticism of a scientist, and run a large number of tests over a range of parameter values before accepting that an observed statistical phenomenon is genuine, and not apophenic. As an additional level of validation, liaison with wet-lab scientists during this process is also advisable.

### 2.8.3 Machine Learning ‘Hygiene’

One very important caveat that seems obvious, but is nonetheless commonly overlooked by machine learning novices, is to keep the training and test samples separate at all

times during the learning process. Using training samples to test the model positively biases the performance of the model (Ambroise and McLachlan, 2002), and thus any validation procedures must be tested out on data unseen by the classifier.

## 2.9 Closing Remark

Dimension reduction is an exciting statistical field that has gained a significant amount of traction in the last decade or so. This is a result of the acceleration in development of large-scale gene expression profiling technology, which has subsequently demanded the employment of high-throughput data analysis. However, this explosion of activity seems to suffer from a decentralised bailiwick. Many high-performing and robust algorithms are being proposed, but the lack of a unifying protocol or pedagogy to guide novice statisticians in their first forays into this field is apparent. This chapter has attempted to highlight some of the phenomena that high-dimensional datasets possess, the corresponding caveats that need to be taken into account when analysing them, and remedial techniques that should be focal in the mind of any statistician serious about performing dimension reduction, academically or professionally. The following chapters will attempt to show how important the application of such techniques can be, and provide guidance towards the development of protocols that ensure meaningful biological output.

# 3

## Two-Step Cross-Entropy Feature Selection for Microarrays — Power Through Complementarity

This chapter constitutes a journal article that was published in the July/August 2011 issue of *Transactions on Computational Biology and Bioinformatics*. Wording has been changed slightly for this thesis by recommendation from examiners' comments. The citation is:

Peters T., Bulger D.W., Loi T., Yang J.Y., Ma D. *IEEE/ACM Trans Comput Biol Bioinform.* 2011 Jul-Aug;8(4):1148-51. doi:10.1109/TCBB.2011.30

### 3.1 Introduction

This chapter describes a study in which we present evidence for the danger of discarding features based on a filtering process, as described in Chapter 2. Through implementation of a Cross-Entropy method, we show that many features which might otherwise be discarded through rankings based on individually-derived statistics may nevertheless be worthy enough for consideration in dimensionally-reduced feature subsets.

The motivation behind this article was a discomfort with the way some dimension reduction routines were taking place in the field of transcriptomics. Appreciative of the enormous volumes of data involved in such routines, we were nevertheless surprised at the lack of detail in many studies of the nature of the preprocessing steps that were involved. After reading many feature selection articles we found that a large number of studies discarded the vast majority of features based on a calculation performed in isolation from other features, such as the classic or modified  $t$ -statistic and  $F$ -statistic (for example, in Xu and Li (2003)). Advances such as the moderated  $t$ -statistic (Smyth, 2004) and Significance Analysis of Microarrays (Tusher et al., 2001) have attempted, with some success, to contextualise the indices of ranked features while retaining the dimension-reduction strategy of a filter (selecting the top percentile of features from a ranked list). However, we wanted to uncover a deeper structure in high-dimensional datasets than filters would allow, and suspected that any filtering process would be too coarse, and a finer, more gradual process—such as an iterative algorithm that revealed emergent interaction effects between features—would be superior. This suspicion was strengthened by the finding that features that are seemingly redundant and possess little or no separation power on their own, in fact produce remarkable separation in tandem with other features (Guyon, 2003), and hence a more sophisticated dimension reduction method is needed to reflect this. Guyon used synthetic data to prove her point, but we wanted to use real data from transcriptomes to suggest, perhaps in an iconoclastic way, that discarding any features without assessing them in their proper context has the potential to limit accurate analysis and skew results, both statistically and biologically. It should be noted that the choice of smoothing parameter used

---

in this paper is relatively arbitrary, and its chosen value may not be optimal. It is important that  $\alpha$  be small enough so as the algorithm does not overshoot a local optimum, but employment of a very small  $\alpha$  is very costly in computational terms. After an exploratory analysis with differing values of  $\alpha$ , we decided on 0.1 as a value that made an appropriate trade-off between these two conflicting issues. Therefore, we believe that the value chosen confers a fair implementation of the algorithm.



# TWO-STEP CROSS-ENTROPY FEATURE SELECTION FOR MICROARRAYS — POWER THROUGH COMPLEMENTARITY

Tim Peters<sup>1,\*</sup>, David Bulger<sup>1,\*</sup>, To-ha Loi<sup>2,\*</sup>, Jean Y.H. Yang<sup>3,\*</sup>, David Ma<sup>2,\*</sup>

<sup>1</sup> Department of Statistics, Macquarie University, NSW 2109, Australia <sup>2</sup> Department of Haematology and Bone Marrow Transplant Unit, St Vincent’s Hospital, Darlinghurst, Sydney, Australia <sup>3</sup> School of Mathematics and Statistics, University of Sydney, Australia

\* E-mail: t.peters@mq.edu.au

## Abstract

Current feature selection methods for supervised classification of tissue samples from microarray data generally fail to exploit complementary discriminatory power that can be found in *sets* of features (Guyon, 2003). Using a feature selection method with the computational architecture of the Cross-Entropy method (Rubinstein and Kroese, 2004), including an additional preliminary step ensuring a lower bound on the number of times any feature is considered, we show when testing on a human lymph node dataset that there are a significant number of genes that perform well when their complementary power is assessed, but ‘pass under the radar’ of popular feature selection methods that only assess genes individually on a given classification tool. We also show that this phenomenon becomes more apparent as diagnostic specificity of the tissue samples analysed increases.

## Background

The interdependence of attributes in a high-dimensional dataset, such as a collection of microarray samples where  $P$  (number of features)  $\gg n$  (number of samples), has the potential to yield new insights, both computationally and biologically. The sparseness of a dataset can be estimated by its sample/feature ratio; this figure is often 1:200

or smaller (Somorjai et al., 2003): For statistical insights to be gleaned from these datasets, dimensionality reduction, also known as feature selection, must be conducted. A useful application of feature selection is to identify biologically significant genes or sets of genes with a view to building diagnostic tools (Tibshirani et al., 2002; Dettling, 2004). However, this has proven difficult, due in no small part to the large amount of biological and technical noise in the datasets in question (Aris et al., 2004; Li et al., 2004). A variety of feature selection methods has been employed with the aim of circumventing this noise and creating classifiers which are robust and insensitive to outliers (Somorjai et al., 2003). The success of these feature selection algorithms has been varied—no ‘gold standard’ has currently been accepted.

The literature on this subject is large and growing. Decentralised studies are not held to a common standard; they use differing datasets and training algorithms, and are consequently optimistically biased (Berrar et al., 2006).

Comparisons between types of cancer (represented by leaves on the classification tree in Figure 3.1) are more reliable when done in a pairwise fashion (Somorjai et al., 2003; Li et al., 2004) and in fact decrease in accuracy as the number of classes to be compared increases (Li et al., 2004). Somorjai et al. (2003) argues that classifiers need to have two properties: firstly, they need to be *robust*, that is, insensitive to outlying data and generalised enough to classify unknown (that is, not used in the training set) samples correctly; secondly, they need to have some *biological significance*, where the set of features in the classifier has a degree of similarity with an established molecular phenomenon, such as a pathway or cascade. Ideally, the classifier should be derived from the smallest possible number of discriminatory features, whilst maintaining robustness. Most classifiers do not satisfy either of these two ideals; this is mostly due to noise in the dataset. Noise reduction takes place in the feature selection step, where the aim is to produce a small subset of genes that possesses both high generalisation ability and insensitivity to outliers. A comparative overview of the variety of approaches to feature selection can be found in Dash and Liu (1997). Feature selection can be described as either *forward* (where features are added one at a time on the basis of their performance until the desired number is reached), *backward* (where features are



removed from the entire group one at a time until a desired number is reached) (Sewell, 2007), or a multi-step method combining both. Seeking the optimal gene set for use by a given classification tool will involve some sort of optimisation algorithm.

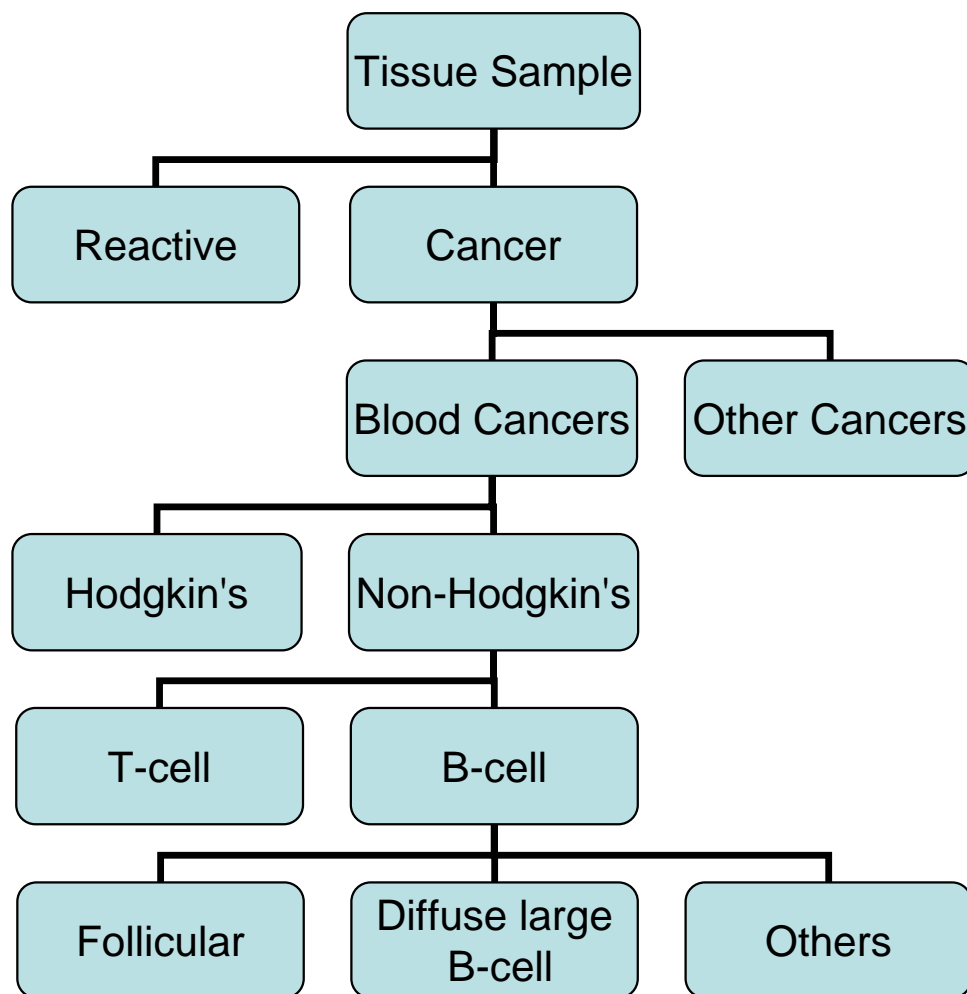


FIGURE 3.1: Diagnostic tree for lymphoma or suspected lymphoma samples. Even though one node may have more than two children, classification accuracy degrades rapidly once 3 or more classes are considered simultaneously (Somorjai et al., 2003). The two bivariate splits used in this study are Cancer vs. Reactive (Level 1) and Follicular vs. Diffuse Large B-cell Lymphoma (Level 5). Note ‘Reactive’ describes non-cancerous lymph nodes that react to some external irritant.

Exhaustive (or brute-force) optimisation, which has a Landau notation of  $O(\text{const}^n)$ , where  $n$  is the number of features (or genes), is infeasible due to the unreasonable

amount of computational time needed to evaluate the optimal set, since  $n$  is on the order of 20 000. A problem such as this is relatively computationally intractable (Garey and Johnson, 1979). Hence, a heuristic method will need to be employed to seek an optimal set.

One frequently overlooked property of microarray datasets is the interdependence of the gene expression values. For example, Xing *et al.* (2001) (Xing and Karp, 2001) and Ng (1998) (Ng, 1998) use feature selection algorithms where the discriminatory power for each gene is computed individually. Though unacceptably superficial, analysis frequently entails the ranking individual genes instead of gene *sets*. With the establishment in recent years that cancer aetiology comprises a complex interdependence of gene products, a feature selection strategy that reflects the contextual nature of the expression levels of each gene is therefore indispensable. Although published studies infrequently report the details of preprocessing in feature selection, our practical experience is that the feature set is usually dramatically reduced on the basis of individual  $F$ -statistics before any interactions are considered. Guyon (2003) proves that features that appear redundant when considered individually, can in fact be useful for noise reduction and class separation when they are considered in complement to one another. Exploitation of this property can be found in the study of unsupervised classification of multivariate time-series (Yoon et al., 2005), however the field of supervised microarray classification appears largely unexplored in this context. An exception is the Random Forest method (Breiman, 2001), however, in general, linear discriminant analyses (the classification method we have employed in this study) outperform classification trees when the sample size is small (Dudoit et al., 2002). In this paper we have implemented a feature selection algorithm based on the Cross-Entropy (CE) method (Rubinstein and Kroese, 2004). With each feature given an initial equal weighting, the algorithm simulates a competitive environment which, in this case, allows sets of genes that show a consistent ability to separate classes in the training set to be drawn out as emergent properties over a number of iterations. This ‘snowballing’ effect provides the basis upon which the contextual performance of each feature can be evaluated.

We aim to show that discarding features from a microarray dataset based on individual discriminatory power is a poor method of dimensionality reduction, due to many features' being valuable in a complementary context.

## Data and Methodology

### Patient samples, Microarray and Pre-processing

Lymph node biopsies and mobilised normal peripheral blood stem cells (reference samples) were collected following written informed consent. The use of these specimens was approved by the Human Ethics and Research Committee of St Vincent's Hospital, Sydney. The diagnosis of each biopsy was made based on standard histological, immunological and cytogenetic analysis. RNA was isolated using TRIzol reagent (Invitrogen, Victoria, Australia) followed by RNeasy micro column purification (Qiagen, Victoria, Australia).

Labelled cDNA of lymph node and reference RNA were synthesised, labelled with either Cy3 or Cy5 fluorophores and hybridised onto human 19K Compugen 70mer oligonucleotide microarrays (Adelaide Microarray Facility, Australia) using the 3DNA Array 900MPX labelling kit (Genisphere, Hatfield, PA) according to the manufacturer's protocol. Arrays were scanned using a GenePix 4000A scanner and fluorescent signals quantified using GenePix Pro 3.0 image analysis software (Molecular devices, Sunnyvale CA). Data from GenePix results files were pre-processed without background subtraction followed by within-array print-tip loess normalization using algorithms from the Bioconductor packages (Gentleman et al., 2004). Missing values were imputed using the GEPAS online preprocessing tool at <http://gepas.bioinfo.cipf.es/cgi-bin/preprocess> using  $k$ -nearest neighbour imputation with  $k=10$ . A final count  $P$  of 18 664 features were used in this study.

## Analysis

All machine learning in this study was *supervised*, meaning that the samples were already classified (diagnosed). We implemented a 210-iteration CE method over 82 training sets, and aggregated the number of iterations (out of  $82 \times 210$ ) in which each feature was used. This serves as a rough estimate of each gene’s discriminatory utility in a complementary context. The process was as follows:

1. Devise a diagnostic ‘tree’ (see Figure 3.1) that defines grouping of the tissues to be classified.
2. Choose a two group split and assign each tissue sample to the corresponding group. Even though one node may have more than two children, classification accuracy degrades rapidly once three or more classes are considered simultaneously (Somorjai et al., 2003).
3. Designate a subset of the data as a training set.
4. Apply a feature selection algorithm that chooses the discriminatory features based on the training data set.
5. Repeat Steps 3–4 for several training sets.

We ranked each gene by the number of inclusion events, and compared these ranks with the genes’  $F$ -value ranks.

This procedure was performed on two pairwise splits: the Cancer/Reactive (Level 1) (see Figure 3.1) and the Follicular Lymphoma (FL)/Diffuse Large B-cell Lymphoma (DLBCL) (Level 5), representing endpoints of specificity. The Level 5 split was chosen amongst other classes of B-cell lymphoma because the two classes had the largest available sample sizes. It was repeated 19 times within each split for 19 values of the gene set size parameter  $2 \leq q \leq 20$ . 82 tissue samples were classified at Level 1 (66 cancerous, 16 reactive) and 36 were classified at Level 5 (24 follicular and 12 diffuse large B-cell).

All computational analysis was implemented in R Version 2.5 (R Development Core Team, 2010). Our novel feature selection method is an implementation based on the

Cross-Entropy (CE) method (Rubinstein and Kroese, 2004). This is a robust iterative method ideal for feature set reduction.

### Preliminary Forward-selection Phase

The CE optimization process can be seen as a backward selection process, where candidate genes are discarded until the final set of size  $q$  is obtained. However, preliminary trial runs showed that a straightforward implementation of the CE method, with parameters chosen to give convergence within a reasonable timeframe, discarded the majority of features without ever sampling them. To solve this problem, we implemented a preliminary forward selection process to ensure each gene is given sufficient consideration for inclusion in the main CE phase. The implementation of this preliminary process can be summarized as follows:

1. Set  $v$  (individual weighting) of all genes to 0.
2. Assign **all** genes to  $P/q$  gene sets **with no repeats** randomly.
3. Quantify the discriminatory power of each gene set. In this study we have used Fisher's  $F$ -statistic (Fisher, 1936).
4. Rank all sets by this discriminatory power and identify the top 10% of these sets.
5. Increase  $v$  by the value of  $(q/P)/10$  for each representation that a particular gene in this top 10%.
6. Repeat steps 2–5 nine more times.

This forward selection procedure produces a shortlist of genes with varying  $v$ s which will be used on the next, backward selection process. This preliminary forward-selection step provides as many individual opportunities for genes to enter the shortlist as there are genes in the complete list. However, higher-performing genes take more opportunities, and the  $v$  values of all genes in the resulting shortlist are proportional to their performance in the forward-selection step.

### Main Backward-Selection CE Phase

The preliminary forward-selection step serves to initialise  $v^0$  (superscripts counting loop iterations) for the CE method, which then proceeds iteratively as follows. At each iteration  $t$ , a random sample of  $N$  gene sets is drawn according to  $v^{t-1}$  (normalised to a probability distribution by dividing by the required gene set size  $q$ ). The genes within each gene set are drawn without replacement, but the gene sets are drawn independently and so may intersect. Each gene set is evaluated, and the best decile  $G_1, \dots, G_n$  are retained.

The update rule for  $v^t$ , based on the CE (Kullback–Leibler distance) between  $v^{t-1}$  and the unknown optimal value, reduces to a simple form here: without smoothing, for each gene  $k$ ,  $v^t(k)$  is updated to equal  $\#\{G_j : k \in G_j\}/n$ , the proportion of retained gene sets containing  $k$ .

Without smoothing, a feature not appearing in the retained genesets in one iteration obtains a  $v$  value of 0, and can never be sampled again. The large number of genes makes this too strict. The update rule can be smoothed by setting

$$v^t = (1 - \alpha) \times v^{t-1} + \alpha \times \#\{G_j : k \in G_j\}/n,$$

where  $\alpha$  is a smoothing parameter. No smoothing occurs if  $\alpha = 1$ , whereas if  $\alpha = 0$ , the vector  $v$  is not updated at all.

The parameters we used are: sample size  $N = 100$  and gene set retention proportion  $\rho = 0.1$  (i.e., at each iteration we drew 100 gene sets and retained the best ten), number of iterations  $T = 200$ , and smoothing parameter  $\alpha = 0.1$  for a fairly conservative update rule.

The ideal final value of  $v$  would take the value 1 at each of the  $q$  genes in the optimal gene set, and 0 everywhere else. Our algorithm always approached this ideal very closely, with all  $q$  features holding a  $v$  in excess of 0.997 at the end of 200 iterations.

## Results and Discussion

When the relative performance of each gene in the two-step method is compared to its individual  $F$ -value, a band of genes with modest  $F$ -values can be seen to have superior performance when analysed with the contextual method. This band is clearer to the naked eye when the analysis is performed at the Level 1 split (Figure 3.2a), however the phenomenon of genes performing under a complementary comparison is more pronounced numerically in the Level 5 split (Figure 3.2b). Roughly a third (304 of 933) of the top 5% performing genes in Level 1 appear in the bottom 85% of genes ranked by  $F$ -value individually. The corresponding proportion in Level 5 is well over a half (573 of 933).

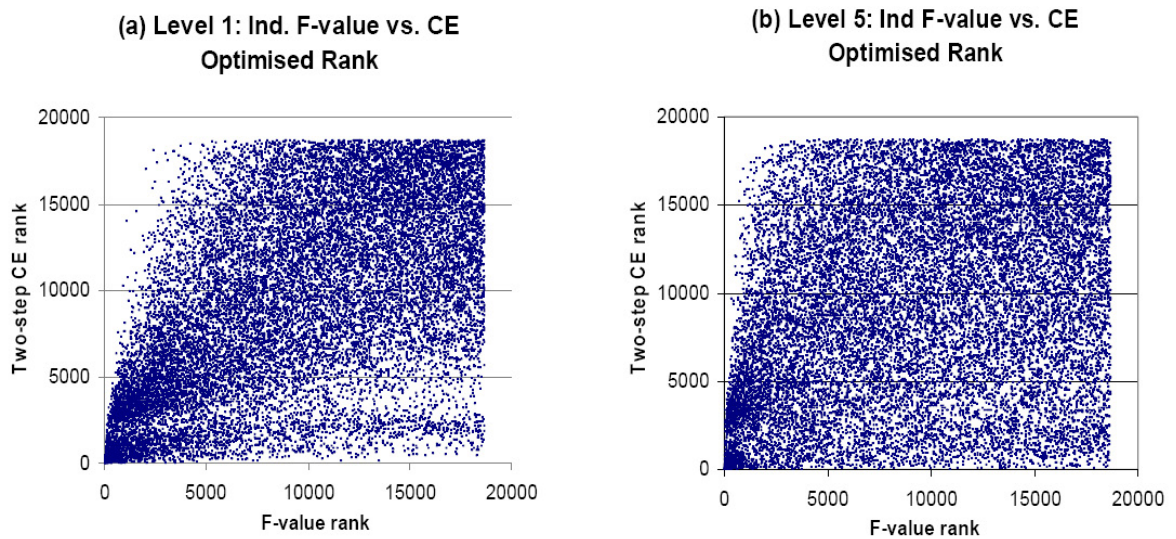


FIGURE 3.2: Scatter plots comparing CE performance rank with individual  $F$ -value rank for Level 1 (a) and Level 5 (b). Each spot represents a microarray feature. The vertical axis represents all features ranked by an aggregate of the performance of each gene in all 19 CE phases performed ( $q = 2, 3, \dots, 20$ ). Both plots show many genes ranking very poorly by  $F$ -value but performing very well in multi-gene classifiers. Also interesting is the starker bimodality of the distribution of CE rank conditioned on  $F$ -value rank in Level 1 (a) than in Level 5 (b).

Given these results, it is clear that genetic interdependence cannot be ignored when applying feature selection. Some studies (Subramanian et al., 2005; Liu et al., 2006) have approached the feature selection problem from the biological end, finding gene sets with constituents that share common biological function, chromosomal location,

or regulation, although this method is self-limiting since the premises for this method are restricted to current knowledge in this field. Further work on the complementary discriminatory power that two or more high-performing genes possess, and their biological parallels, is an obvious direction that this study points to. A disadvantage of this algorithm is that cumulation of  $v(k)$  after a certain number of iterations for a given feature  $k$  is difficult for other features to compete with, although prudent choice of the smoothing parameter  $\alpha$  can ensure the most deserving features are included. This disadvantage is evidenced in the poor performance of the two-step CE method (using  $\alpha = 0.1$ ) as a training algorithm for class prediction, in comparison to a forward selection control method in which features were added according to their independent  $F$ -values. Using leave-one-out cross-validation on the Level 1 split, the CE method yielded a maximum 66 out of 82 samples predicted correctly, using 5 features, whereas the control method correctly predicted 73 out of 82, using 9 features. Similar analysis on the Level 5 split yielded 33 out of 36 samples predicted correctly by the CE method using 2 features, whereas the control method predicted the class of all 36 samples correctly using 5 features. For this reason we do not recommend the use of the CE method for feature selection without thorough tuning of  $\alpha$ . It is likely that our arbitrary selection of  $\alpha = 0.1$  is too large, resulting in an over-regularized model and subpar predictive performance. Using an algorithm with an immediate (as opposed to an aggregated) reward for superior contextual performance, such as simulated annealing, may be more beneficial. We recommend the use of the Two-Step Cross-Entropy feature selection method in tandem with established feature selection methods such as Random Forest (Breiman, 2001) and the Lasso (Tibshirani, 1996), with subsequent meta-analysis. The most important insight from the above results is that feature selection methods that consider genes where the discriminatory power or information gain of a gene is indexed individually are of limited use in classification and subsequent biological analysis. These feature selection methods are ‘throwing the baby out with the bathwater’ in that they discard potentially valuable information on the nature of the samples that produced the dataset. Furthermore, the greater complementary power a gene set has (as seen in the more specific pairwise split of Level 5), the more dire



the consequences of superficial gene analysis. It may be suggested that the complementarity phenomenon in the dataset analysed may be coincidental rather than the result of biological interdependence, but its greater pronunciation at the more specific level (Figure 3.2b) supports the intuition that this interdependence can be more easily identified at such levels.

For both CE analysis on gene sets and individual gene evaluation, we used linear discriminant analysis, because it was a simple statistical method for assessing discriminatory power. However, other classification tools have been shown to provide higher classification accuracy than LDA, in particular support vector machines (SVMs) (Li et al., 2004; Statnikov et al., 2004). One study has also suggested that a linear discriminator feature selection method combined with a SVM to train classifiers produces superior results (Guyon, 2003). It is our aim in further research to find the most effective combination of an objective function for class discrimination, and optimization algorithm, for use in the construction of robust classifiers for diagnostic use.

## Acknowledgements

Generation of the in-house microarray data was funded in part by Australian grants from The Sydney Foundation for Medical Research, St Vincent's Hospital Haematology Research Fund and the Arrow Bone Marrow Transplant Foundation. Thanks to Michael Baxter, for IT help.

## References

- Aris, Virginie M., Michael J. Cody, Jeff Cheng, James J. Dermody, Patricia Soteropoulos, Michael Recce, and Peter P. Tolias. "Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer." *BMC Bioinformatics* 5: (2004) 185.
- Berrar, Daniel, Ian Bradbury, and Werner Dubitzky. "Avoiding model selection bias in small-sample genomic datasets." *Bioinformatics* 22, 10: (2006) 1245–1250.
- Breiman, Leo. "Random Forests." *Machine Learning* 45, 1: (2001) 5–32.

- Dash, M., and H. Liu. “Feature selection for classification.” *Intelligent Data Analysis* 1, 3: (1997) 131–156.
- Dettling, M. “Finding predictive gene groups from microarray data.” *Journal of Multivariate Analysis* 90, 1: (2004) 106–131.
- Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. “Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data.” *Journal of the American Statistical Association* 97, 457: (2002) 77–87.
- Fisher, R. A. “The use of multiple measurements in taxonomic problems.” *Annals of Eugenics* 7, 2: (1936) 179–188.
- Garey, M. R., and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, volume 44 of *Books in the Mathematical Sciences*. W. H. Freeman, 1979.
- Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. “Bioconductor: open software development for computational biology and bioinformatics.” *Genome Biol* 5, 10: (2004) R80.
- Guyon, Isabelle. “An Introduction to Variable and Feature Selection 1 Introduction.” *Journal of Machine Learning Research* 3: (2003) 1157–1182.
- Li, Tao, Chengliang Zhang, and Mitsunori Ogihara. “A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.” *Bioinformatics* 20, 15: (2004) 2429–2437.
- Liu, Chun-Chi, Wen-Shyen E Chen, Chin-Chung Lin, Hsiang-Chuan Liu, Hsuan-Yu Chen, Pan-Chyr Yang, Pei-Chun Chang, and Jeremy J W Chen. “Topology-based cancer classification and related pathway mining using microarray data.” *Nucleic Acids Research* 34, 14: (2006) 4069–4080.
- Ng, Andrew Y. “On Feature Selection: Learning with Exponentially many Irrelevant Features as Training Examples.” *Ieee Expert Intelligent Systems And Their Applications* 404–412.
- R Development Core Team. “R: A Language and Environment for Statistical Computing.”, 2010.
- Rubinstein, R. Y., and D. P. Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 48. Springer-Verlag, 2004.

- Sewell, M. "Feature Selection." <http://www.machinelearning.net/featureselection/feature-selection.pdf>, 2007.
- Somorjai, R. L., B. Dolenko, and R. Baumgartner. "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions." *Bioinformatics* 19, 12: (2003) 1484–1491.
- Statnikov, A., C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. "A comprehensive evaluation of multcategory classification methods for microarray gene expression cancer diagnosis." *Bioinformatics* 21, 5: (2004) 0.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A. Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* 102, 43: (2005) 15,545–15,550.
- Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B Methodological* 58, 1: (1996) 267–288.
- Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proceedings of the National Academy of Sciences of the United States of America* 99, 10: (2002) 6567–72.
- Xing, E. P., and R. M. Karp. "CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts." *Bioinformatics* 17 Suppl 1, Suppl 1: (2001) S306–S315.
- Yoon, H., K. Yang, and C. Shahabi. "Feature subset selection and feature ranking for multivariate time series." *IEEE Transactions on Knowledge and Data Engineering* 17, 9: (2005) 1186–1198.



# 4

## Cancer Microarray Feature Selection Using Support Vector Machines: Comparing Regularisation Techniques

This chapter constitutes an article published in the Proceedings of the 2009 Joint Statistical Meetings, Washington D.C. Wording has been changed slightly for this thesis by recommendation from examiners' comments. The citation is:

Peters, T., D. W. Bulger, T-H. Loi, J. Y. H. Yang, and D. Ma. 2009. Cancer Microarray Feature Selection Using Support Vector Machines: Comparing Regularization Techniques. In *JSM Proceedings*, Section on Statistical Learning and Data Mining. Alexandria, VA: American Statistical Association. 2951-2965.

## 4.1 Introduction

As discussed in Chapter 3, we realised that, despite the implementation of a sophisticated architecture ensuring all features were given a reasonable chance at being included in the final feature set, our chosen objective function (the  $F$ -statistic) was inferior to others. This is because the  $F$ -statistic is neither regularised, nor gives an accurate indication of how well classes are separated when the class distribution is non-Gaussian (Box, 1953). After extensive reading, we took three statistical learning methods that had both favourable reviews and frequent application in the field of transcriptomic data mining as candidates for the next phase of research: SVM (Cortes and Vapnik, 1995), Random Forests (Breiman, 2001) and Least Angle Regression (Efron et al., 2004). All have outstanding predictive ability and an intuitive approach to separating data samples, such that there is no consensus in the existing literature about a definitively superior dimension reduction method. Each has a substantially different approach to regularisation. SVMs regularise explicitly through a user-defined parameter, LAR relaxes model constraints by including only partial features, and Random Forest prunes trees that overfit after evaluating out-of-bag error, similar to the way a wrapper operates.

We implemented a standard cross-validation procedure on the training data, as well as an additional validation routine on unseen samples separate from those trained, to determine whether one of these candidate methods had a competitive advantage over the others. Three publicly available and well known transcriptomic data sets were used in addition to the in-house dataset used in Chapter 3. Given that the SVM does not have an inbuilt feature selection phase, we used a forward stepwise-style algorithm, similar to that of Zhu et al. (2004) to build a model path. We originally intended to use the Cross-Entropy method for feature selection for using SVMs, but were dissuaded by the poor convergence rates. We used receiver operating characteristics to assess our results, plotting true positive rate against false positive rate. In addition to evaluating the area-under-curve, we also plotted the model-defined decision boundary at the corresponding position on each curve, and measured the  $\ell^1$  distance from it to

$(0, 1)$ . The latter was done to assess the natural ability of each algorithm to classify unlabelled samples using its naturally generated decision boundary, as opposed to the one that corresponds to the position on the curve closest to  $(0, 1)$ .

The training set for the St Vincent's dataset is identical to the one used in the 'Level 1' analysis in Chapter 3. We found that our SVM algorithm performed better on this dataset than the other methods, and that all regularised methods improved on the F-statistic stepwise regression.





# CANCER MICROARRAY FEATURE SELECTION USING SUPPORT VECTOR MACHINES: COMPARING REGULARISATION TECHNIQUES

Tim Peters<sup>1,\*</sup>, David Bulger<sup>1,\*</sup>, To-ha Loi<sup>2,\*</sup>, Jean Y.H. Yang<sup>3,\*</sup>, David Ma<sup>2,\*</sup>

<sup>1</sup> Department of Statistics, Macquarie University, NSW 2109, Australia <sup>2</sup> Department of Haematology and Bone Marrow Transplant Unit, St Vincent's Hospital, Darlinghurst, Sydney, Australia <sup>3</sup> School of Mathematics and Statistics, University of Sydney, Australia

\* E-mail: t.peters@mq.edu.au

## Abstract

Microarray dataset dimensionality reduction is a prerequisite for avoiding overfitting, and hence developing diagnostic tools. Some previous work has selected features based, for example, on their individual Fisher discriminants ( $F$ -values), or ‘path-based’ training algorithms optimising the power of the resulting classifier. We show that a generic method, using a simple stepwise regression with the linear support vector machine penalised margin width as the objective function, subject to regularisation parameter grid-search, gives superior performance to three other feature-selection methods (least-angle regression, Random Forest, and stepwise regression on Fisher discriminants). We use a hierarchical validation method, applying leave-one-out cross-validation within the training subset, and applying the trained classifier to a separate test subset, on each of four two-class gene expression cancer datasets. The generic method shows superior results when classifying unseen samples, compared to three other feature selection methods, and a fixed regularisation value appears nearly optimal for all four datasets.

**Keywords:** feature selection, microarrays, support vector machines, path-based algorithms, regularisation.

## Background

Feature selection for supervised tissue classification from microarray data has become a field of considerable scrutiny over the past decade or so. Microarray datasets are afflicted by a large feature/sample ratio: only a small number of tissue samples (usually less than 100), but a feature set size in the tens of thousands, which makes the selection of significant gene groups both theoretically and computationally daunting (Somorjai et al., 2003). The evolution of feature selection algorithms in this field has been guided by a desire to find biologically significant genes or sets of genes with a view to building diagnostic tools (Tibshirani, 1996; Dettling, 2004). This has proven difficult, due in no small part to the large amount of biological and technical noise present in the datasets in question (Aris et al., 2004; Li et al., 2004), and features that appear redundant when considered individually are in fact useful for noise reduction and class separation when they are considered in complement (Guyon, 2003). The latter point is of critical significance, since the number of features needed for optimal discrimination varies from dataset to dataset, and is hence unknown when encountering a new dataset, rendering the problem of finding the optimal feature subset ill-posed.

A variety of feature selection methods has been employed with the aim of circumventing this noise and creating classifiers which are robust and insensitive to outliers (Somorjai et al., 2003). These selection methods can involve clustering algorithms (such as  $k$ -nearest neighbour, Dasarathy (2002), and PAMR, Tibshirani et al. (2002)), stochastic methods (such as simulated annealing (Kirkpatrick et al., 1983), genetic algorithms and Markov chain algorithms), and greedy hill-climbing methods. The success of these feature selection algorithms has been varied: no ‘gold standard’ has currently been accepted.

One of the most popular metaheuristics in the field of feature selection is the greedy hill climbing method (Cormen et al., 2001), also called ‘forward selection’ or ‘stepwise regression’. The first step of the process involves finding the most optimal feature by ranking all  $P$  features on the value of some objective function that informs its discriminatory power. This feature now constitutes the selected subset. The next

feature is selected based on the combined two-dimensional discriminatory power of the feature already within the subset, and one of the  $P - 1$  other features. Each of the  $P - 1$  two-feature models has its objective function calculated and ranked, before the gene most recently added to the most optimal model is then added permanently to the selected feature subset. The  $P - 2$  remaining features are then trialed one-by-one for the third addition, and so on.

Sets optimized by the greedy method can often find the global optimum, but this is not always guaranteed; sometimes only a local optimum is found. This is because the algorithm, in advancing an iteration, may overlook a set of features whose interactions confer a powerful discriminatory effect. Often a situation may occur where the algorithm has already advanced to an unrecoverable depth in another direction with a different feature subset, whose interactions with the overlooked set confer an unremarkable discriminatory power. Greedy feature selection therefore has a propensity to act too grossly in its quest for the optimal dataset. In terms of feature selection and subsequent sample classification, two recently developed algorithms have adapted the greedy algorithm architecture to advance iterations more cautiously when choosing the next feature to be added to the feature subset.

Least-angle regression (LARS) (Efron et al., 2004) is an improvement on the original LASSO algorithm, first described by Tibshirani (1996). The algorithm begins with a vector of regression coefficients size  $P$ , all set to zero. Firstly, the coefficient of the predictor (or feature) most correlated with the response variable is increased in the direction of the sign of this correlation, recalculating residuals along the way. This is done until some other feature is correlated with the residuals to the same degree as the original predictor is. The model is then shifted in the direction of the joint least-squares direction of these two predictors until its correlation with the residuals is equal to that of a third predictor, and so on. For microarray studies where the number of features  $P$  greatly exceeds the number of samples  $n$ , models typically contain up to  $n - 1$  variables, since any model can have no more than  $n - 1$  (mean centred) variables with non-zero co-efficients. LARS produces a continuous path of models, rather than a discrete sequence. Advantageously, the fit can produce a greater choice of models,

since they can be evaluated at any point along the path, even at fractions between feature-adding events.

Random Forest (Breiman, 2001), as its name suggests, contains a stochastic element to its training algorithm. Using a bootstrap aggregation method on randomly selected training and test sets with replacement, the algorithm also selects a random subset of features to train the model, and uses the best features to ‘split’ the node of each tree in the forest. The trees ‘vote’ on the predicted class for each test sample based on the computed proximities of each test case (that is, how often two samples occupy the same node), and are grown without pruning. Although disputed (Segal, 2004), Random Forest claims to avoid overfitting.

Both algorithms attempt to overcome the trap of classic forward selection becoming ‘too greedy’; least-angle regression by incorporating a continuous element into its architecture, and Random Forest by using the wrapper method of guarding against advancing the model in the direction of a classifier with a higher error rate at step  $k + 1$  than at step  $k$ . These are architectural modifications to the classic forward selection prototype, and have shown to be superior to the original (Efron et al., 2004; Breiman, 2001).

Feature selection can be seen as optimising an objective function defined on the set of feature subsets. Methods such as LARS and the Dantzig selector (Candes and Tao, 2007) incorporate regularisation into the optimisation method. However, there has been little discussion of incorporating regularisation into the objective function only, leaving a free choice of optimisation procedure. Traditionally, the objective function when performing stepwise regression is Fisher’s  $F$ -statistic (Fisher, 1936), but others such as Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) or Mallows’  $C_p$  are frequently used. These functions, with the exception of Mallows’  $C_p$ , have no inbuilt regularisation. We propose combining the support vector machine (Cortes and Vapnik, 1995); a highly robust and well-respected training algorithm used in classification, and more recently, feature selection, with a greedy optimisation process (Zhang et al., 2006; Guyon et al., 2002). The penalised margin-width in SVMs can be

expressed in the form:

$$\min \frac{1}{\|V\|} + \frac{\sum \xi_i}{\lambda} \quad (4.1)$$

where  $\|V\|$  is the distance between the two hyperplane margins,  $\xi_i$  the sum of  $i$  slack variables determined by the soft margin, and  $\lambda$  a regularisation trade-off parameter. Using the classic greedy forward selection model, we take the minimum value of the objective function for each feature in each successive ranking step, and add this value to the model. This is not a particularly original method; it is simply a combination of a heuristic and an objective function that are both well known. However, we propose that this approach confers a competitive, if not superior regularisation strategy for accurate class discrimination.

## Data and Methodology

Four cancer-related bipartite datasets were used in this study. One was generated in-house by the Department of Haematology and Bone Marrow Transplant Unit, St Vincent's Hospital, Darlinghurst, Sydney, Australia, and Golub et al. (1999), Alizadeh et al. (2000) and Armstrong et al. (2002) were downloaded from the Kent Ridge Biomedical Dataset Repository (Kent Ridge Biomedical Data Set Repository, 2008). These three datasets were chosen since all samples were from patients diagnosed with a form of haematological cancer, however they are all more homogeneous than our in-house dataset. An outline of the samples contained in each dataset can be found in Table 4.1.

For the dataset from St Vincent's Hospital, lymph node biopsies and mobilised normal peripheral blood stem cells (reference samples) were collected following written informed consent. The use of these specimens was approved by the Human Ethics and Research Committee of St Vincent's Hospital. The diagnosis of each biopsy was made based on standard histological, immunological and cytogenetic analysis. RNA was isolated using TRIzol reagent (Invitrogen, Victoria, Australia) followed by RNeasy micro column purification (Qiagen, Victoria, Australia). Labelled cDNA of lymph node and reference RNA were synthesised, labelled with either Cy3 or Cy5 fluorophores and

hybridised onto human 19K Compugen 70mer oligonucleotide microarrays (Adelaide Microarray Facility, Australia) using the 3DNA Array 900MPX labelling kit (Genisphere, Hatfield, PA) according to the manufacturer's protocol. Arrays were scanned using a GenePix 4000A scanner and fluorescent signals quantified using GenePix Pro 3.0 image analysis software (Molecular Devices, Sunnyvale CA). Data from GenePix results files were pre-processed without background subtraction followed by within-array print-tip loess normalization using algorithms from the Bioconductor packages (Gentleman et al., 2004). For all datasets, missing values were imputed using the GEPAS online preprocessing tool at <http://gepas.bioinfo.cipf.es/cgi-bin/preprocess> using  $k$ -nearest neighbour imputation with  $k=10$ .

	St. Vincent's Hospital	Golub (1999)	Alizadeh (2000)	Armstrong (2002)*
Sample Classes (Note: true positives listed first)	Cancer vs Re- active	Acute myeloid leukemia (AML) vs Acute lym- phoblastic leukemia (ALL)	Activated Dif- fuse Large B-Cell Lym- phoma (DL- BCL) vs Germinal DLBCL	Acute myeloid leukemia (AML) vs Acute lym- phoblastic leukemia (ALL)
Positives (Training set)	66	11	23	20
Negatives (Training set)	16	27	24	20
Positives (Testing set)	30	14	N/A	8
Negatives (Testing set)	8	20	N/A	4
Total number of features	18664	7129	4026	12582

TABLE 4.1: Attributes of the four datasets used in this study.

\*A third class, MLL, was removed from this dataset, since we are only working with bipartite splits.

Near-perfect discrimination between lower levels of lymphoma classes, such as follicular lymphoma vs. diffuse large B-cell lymphoma, was achieved in the St Vincent's dataset with rudimentary methods such as LDA, so we made the decision to analyse the top split: Cancer vs. Reactive.

Each dataset except (Alizadeh et al., 2000) contains a training set and a test set. For each training set, we trained the data using Linear SVM stepwise regression (SVMSR) with  $\lambda = 10$ , Least-Angle Regression (LARS), Random Forest (RF) and  $F$ -value stepwise regression (FSR). First, leave-one-out cross-validation (LOOCV) was performed on all four datasets' training sets using all four methods. A new subset of variables was selected for each of the  $n$  validation trials, and the error rate was calculated as the proportion of trials that conferred a false prediction. With the exception of the non-overfitting Random Forest, the model with the lowest error rate, taking the first reached in the event of ties, was selected for prediction of the unseen observation. For each method on each dataset, the accuracy, receiver operating characteristic (ROC), and the point along the algorithm path where the optimal model occurs were calculated. The crossover threshold that represents the decision boundary between classes (posterior probability = 0.5 for both stepwise regressions and Random Forest, and threshold = 0 for least-angle regression) was mapped onto each ROC.

We include pseudocode below (see next page) to clarify the aforementioned simulation experiment. The process shown was repeated for each of the four datasets in order to evaluate the behaviour of the SVMSR method. A similar procedure was used to evaluate each of the other three methods.

---

```

for  $\lambda \in \{100, 10, 1, 0.1, 0.01\}$  do
  for each sample  $n$  of the training set do
    for  $q$  (the number of genes in the model) = 1 to 20 do
      train SVMSR classifier on  $\{\text{training set}\} \setminus \{n\}$  using  $q$  and  $\lambda$ 
      classify left out sample  $n$  and record success or failure
    end for
  end for
end for
calculate sample classification accuracy for each  $q$  and  $\lambda$ 
 $(q_{\text{best}}, \lambda_{\text{best}}) \leftarrow (q, \lambda)$  with best sample classification accuracy
train SVMSR classifier on whole training set using  $q_{\text{best}}$  and  $\lambda_{\text{best}}$ 
classify test set and report accuracy

```

---

In the case of SVMSR, a grid search was performed using the values  $\lambda = \{0.01, 0.1, 1, 10, 100\}$  for SVMSR for all four LOOCV runs on the training samples,

and a parameter value of  $\lambda = 10$  was found to have not been bettered on classification performance from the resulting models. SVMs used the soft-margin objective function (4.1) to advance the algorithm, and the FSR used the  $F$ -value (Fisher, 1936) as the objective function. Care was taken for each implementation, to ensure each classifier trained had the weights of each (unbalanced) class taken into account. An upper bound of 20 features was used for the stepwise regression approaches, since the classification accuracy for both validation procedures did not seem to improve above this value, and also to keep the chosen model reasonably sparse. The software default of 500 trees was used for Random Forest, and the default number of features used to split each node of each tree:  $\sqrt{P}$  = total number of features. Since, for  $P \gg n$ , LARS has a deterministic stopping criterion of  $P_{LARS} > n$ , for standardization purposes the model was evaluated at 20 equally spaced points along the path length 1, at interval length 0.05. Following this, we trained each training set except (Alizadeh et al., 2000) using all four methods with the standards described above, and validated the trained model on the test set, calculating accuracies, ROCs, decision boundaries and optimal path locations. All analysis was carried out using R scripts, incorporating the following platforms: kernlab, lars, randomForest and MASS.

## Results and Discussion

SVMs outperform all other methods on the St. Vincent's dataset on both discrimination and classification, and it remains competitive on the other datasets analysed (Table 4.2). Least-angle regression performs well on both validation procedures on the Golub dataset, but does not gain an advantage over the other methods in other datasets, and noticeably struggles with the St. Vincent's dataset. As expected, the LARS and Random Forest algorithms outperformed the control:  $F$ -value stepwise regression, which has no regularisation.

With respect to class discrimination, Random Forest is able to outperform, or at least match, the other methods (Table 4.3). Considering, however, that software with an ability to classify unseen samples is needed for the development of diagnostic tools



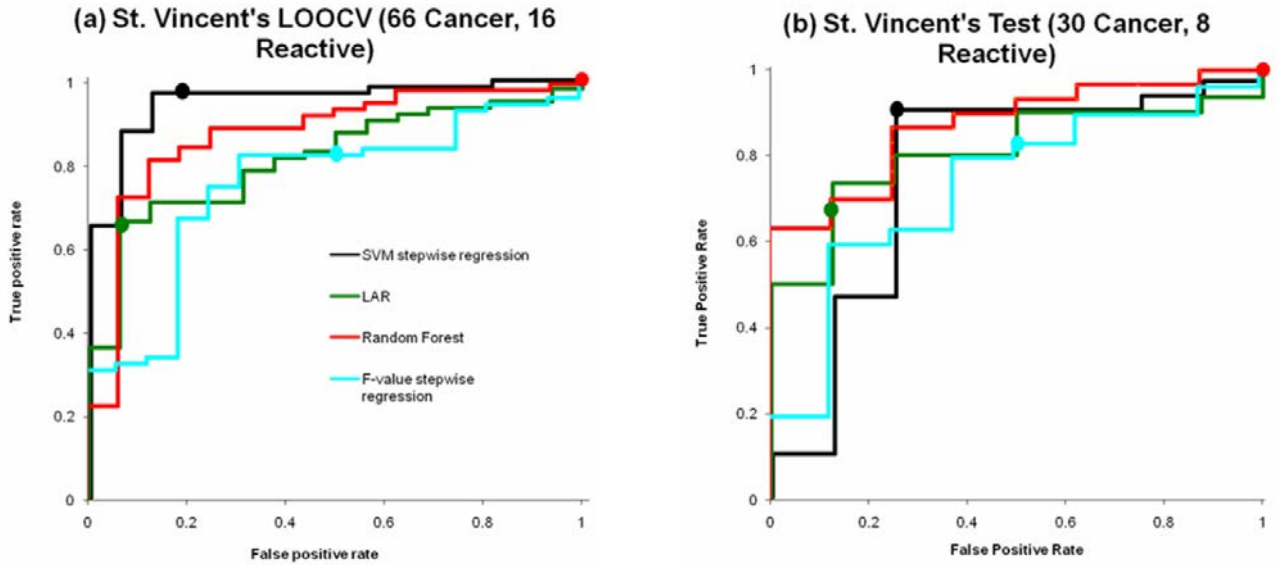


FIGURE 4.1: Receiver operating characteristics depicting the accuracy of each method on the optimized models from (a) the St. Vincent's training set and (b) St. Vincent's validation set. For this and Figures 4.2, 4.3 and 4.4, coloured dots represent the class decision boundary, i.e. class-weight adjusted posterior probability = 0.5 for both stepwise regressions and Random Forest, and response = 0 for least-angle regression. Each dataset represents the best performing model (with respect to classifier accuracy) from the regression path, with the exception of Random Forest, where the default of 500 trees was used. Index for colours used shown in (a). Exactly plotted curves will hide each other, so the curves in all figures have been slightly offset.

for use in medical practice, a heuristically-determined threshold alone is insufficient. For Random Forest, the user must perform trial-and-error runs to calibrate a suitable objective threshold for use in class prediction. This was certainly needed in the St Vincent's validation runs, where the trained forest classified incurred the maximum false positive rate using the default heuristic, hence rendering the prediction performance no better than random guessing (Figure 4.1 and Table 4.4). A comparison of this with the performance of SVMSR (Figure 4.1 and Tables 4.2 and 4.4) shows that SVMSR has a superior ability to draw a decision boundary that generates a minimal test error; this is shown by the fact that the threshold representing the decision boundary on the corresponding receiver operating characteristic (coloured large dots in Figure 4.1) is generally located at the point where the curve is most proximal to (0, 1). This shows

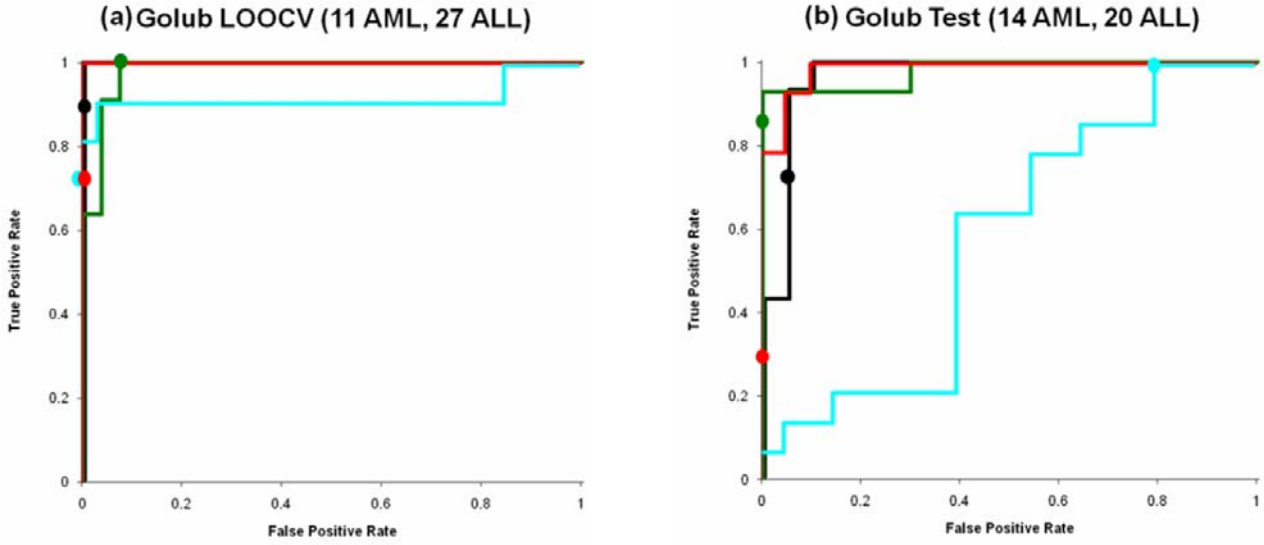


FIGURE 4.2: ROCs for (a) Golub training set and (b) Golub validation set.

that support vector machines are able to calibrate a more objective decision boundary through training than Random Forest, as well as being able to separate classes well. We also notice that the advantage gained from SVMs, especially in terms of classification, is most pronounced when the class sizes are unbalanced. The danger in training a classifier based on unbalanced classes is that the decision boundary may be drawn too close to the centroid of the larger class, heightening the risk of an increased test error. SVMs avert this problem by calculating the objective function using only a subset of observations (Cortes and Vapnik, 1995) within the vicinity of the decision boundary. The result is a more robust classifier.

LARS and LDA use all observations in model construction, and Random Forest uses iterative random sets which do not wittingly incorporate class weights into the algorithm; it needs user calibration. It is here that we see the fundamental difference between the regularisation approaches of SVMs, and Random Forest and LARS. The latter two incorporate regularisation into the algorithm architecture, by either making the model path continuous (LARS) or by eliminating stochastically generated paths that overfit too soon (Random Forest). These are valid and effective methods; however,

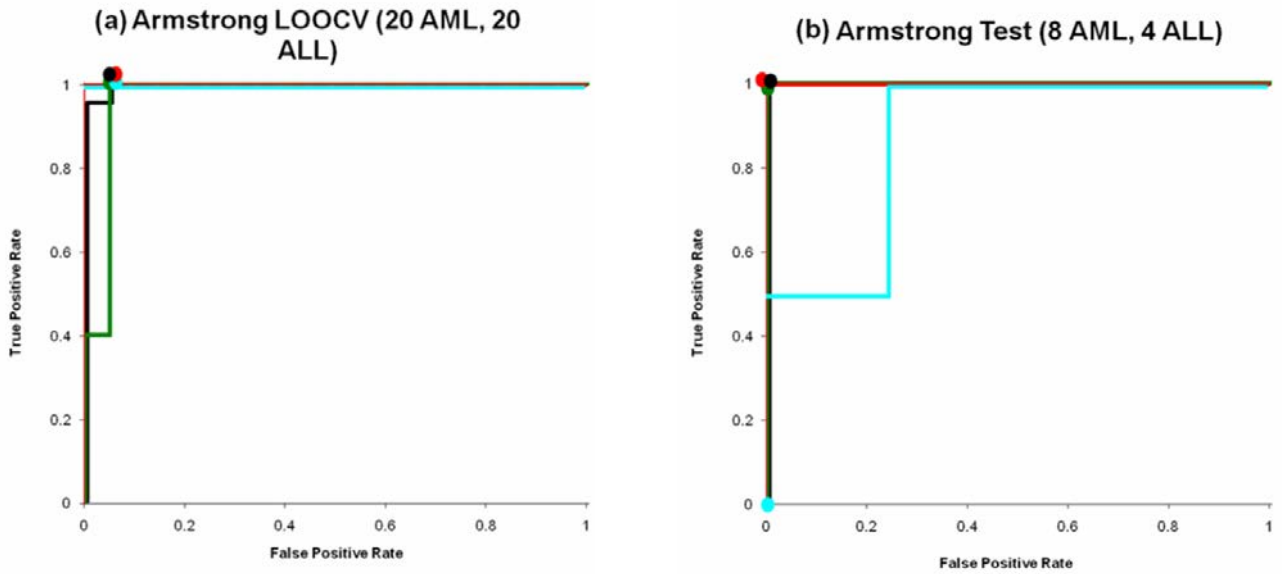


FIGURE 4.3: ROCs for (a) Armstrong training set and (b) Armstrong validation set.

SVMSR incorporates its regularisation through judicious selection of the most critical observations in the dataset, rather than changing the algorithm architecture. It is true that manual parameterization must be used with SVMs as well: a simple grid search is needed to find the regularisation parameter value that confers the most accurate classifier.

When training a SVM classifier, two distinct overfitting dangers need to be averted. The first concerns the number of dimensions present in the classification model, which, in the case of SVMSR, equals the number of features added. The more features are added, the less sparse the model becomes, which impairs the process of selecting important sets of genes for cancer diagnosis. Computational costs may also become prohibitive with a large number of features in the model. The second danger involves an inappropriate value of the regularisation parameter. In the case of too little regularisation, i.e.  $\lambda$  is too small, the model overfits to the misclassified training samples, shrinking the size of the margin, and compromising the robustness of the model. This can be seen in the fact that as  $\|V\|$  increases, the objective function, which we want to minimise, shrinks. This grid search was performed, and it was found that a value of

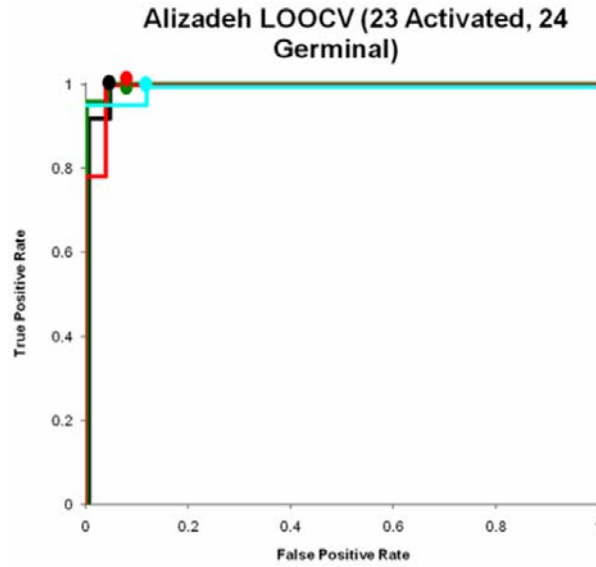


FIGURE 4.4: ROC for Alizadeh training set. Note that this study used a customized array, and the samples have a much greater homogeneity, since they are diagnosed as subclassifications from a single form of lymphoma.

$\lambda = 10$  within the set of  $\lambda = \{0.01, 0.1, 1, 10, 100\}$  could not be bettered with respect to class prediction when training across all datasets. As an afterthought, all values of  $\lambda$ , besides the calculated optimal value, were also used on the external validation sets, to check for bias. None performed better than the value chosen by cross-validation training. This promising consistency also supports the caveat that under-regularisation (i.e.  $\lambda \ll 1$ ) usually incurs a higher error rate on unseen samples (Hastie et al., 2004).

The sets of genes selected by all four methods generally overlapped heavily, in all four dataset cases. It is difficult to quantify the magnitude of overlap since the models are aggregated differently (and fractionally in the case of LARS), however, the overlap was noticeably less to the casual observer in the case of SVMSR. Whether the genes selected by SVMSR are more biologically meaningful than those selected by the other three methods is unclear; a weighted meta-analysis may be helpful.

In terms of SVMs, the circumvention of overfitting has been attempted using recursive backward feature selection, or SVM-RFE (Guyon et al., 2002). However, since we are interested in a small set of bellwether predictive genes for cancer diagnosis, this

% explained correctly	St. Vincent's		Golub		Alizadeh	Armstrong	
	LOOCV	Test Set	LOOCV	Test Set	LOOCV	LOOCV	Test Set
SVM linear stepwise regression	93.90	84.21	97.37	94.12	97.87	97.50	100.00
Least-angle regression	71.95	71.05	94.74	94.12	95.74	97.50	100.00
Random forest	80.49	78.95	92.11	70.59	95.74	97.50	100.00
F-value stepwise regression	76.83	78.95	92.11	79.41	93.62	97.50	66.67

TABLE 4.2: Classification accuracies for all models. The figure represents the highest accuracy achieved along each path for the regression algorithms, and the accuracy at 500 trees for Random Forest. Prediction is calculated from posterior probabilities for both stepwise regressions and Random Forest, and from signed responses for least-angle regression.

Area Under ROC Curve	St. Vincent's		Golub		Alizadeh	Armstrong	
	LOOCV	Test Set	LOOCV	Test Set	LOOCV	LOOCV	Test Set
SVM linear stepwise regression	0.953598	0.758333	1.000000	0.967857	0.996377	0.997500	1.000000
Least-angle regression	0.818182	0.808333	0.983165	0.978571	0.998188	0.970000	1.000000
Random forest	0.877841	0.870833	1.000000	0.985714	0.990942	1.000000	1.000000
F-value stepwise regression	0.758523	0.729167	0.919192	0.575000	0.994565	1.000000	0.875000

TABLE 4.3: Area under receiver operating characteristics shown in Figure 4.1

method is likely to only be of use in discarding non-informative features. Some studies (Subramanian et al., 2005; Liu et al., 2006) have approached the feature selection problem from the biological end, finding gene sets with constituents that share common biological function, chromosomal location, or regulation, although this method is self-limiting since its premises are restricted to current knowledge in the field of oncogenetics.

Furthermore, and perhaps most importantly, SVMsR is a simpler algorithm, with simpler classification rules, than LARS or Random Forest. Application of the Occam's Razor principle in machine learning and class prediction is highly valued and recommended (Pranckeviciene and Somorjai, 2006), and simpler decision rules should

$\ell^1$ norm	St. Vincent's		Golub		Alizadeh	Armstrong	
	LOOCV	Test Set	LOOCV	Test Set	LOOCV	LOOCV	Test Set
SVM linear stepwise regression	0.217803	0.350000	0.090909	0.335714	0.041667	0.050000	0
Least-angle regression	0.395833	0.458333	0.074074	0.142857	0.083333	0.05	0
Random forest	1.000000	1.000000	0.272727	0.714286	0.083333	0.050000	0
F-value stepwise regression	0.666667	0.666667	0.272727	0.800000	0.125000	0.050000	1.000000

TABLE 4.4:  $\ell^1$ -norm from decision boundary (coloured dots shown on ROC curves in Figures 4.1 - 4.4) to (0, 1). This represents the sum of the false positive and false negative rates.

be considered before more complex ones that may incur overfitting.

The advantages of SVMs over other forms of regularisation are clear, but come at a higher computational cost. The computation of an entire path (20 steps) of SVMs incurs a time cost approximately 20 times that of a full Random Forest path, and 100 times that of LARS. However, for the purposes of identifying gene sets that can be relied upon to provide accurate cancer class prediction in a diagnostic context, we opine that this is a worthwhile investment.

## Acknowledgements

Generation of the in-house microarray data was funded in part by Australian grants from The Sydney Foundation for Medical Research, St Vincent's Hospital Haematology Research Fund and the Arrow Bone Marrow Transplant Foundation. Thanks to Michael Baxter, for IT help.

## References

- Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever,

- J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.” *Nature* 403, 6769: (2000) 503–511.
- Aris, Virginie M., Michael J. Cody, Jeff Cheng, James J. Dermody, Patricia Soteropoulos, Michael Recce, and Peter P. Tolias. “Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer.” *BMC Bioinformatics* 5: (2004) 185.
- Armstrong, Scott A., Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. Den Boer, D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub, and Stanley J. Korsmeyer. “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.” *Sophia* 30, January: (2002) 41–47.
- Breiman, Leo. “Random Forests.” *Machine Learning* 45, 1: (2001) 5–32.
- Candes, Emmanuel, and Terence Tao. “The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ .” *The Annals of Statistics* 35, 6: (2007) 2313–2351.
- Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. “Greedy Algorithms.” In *Introduction to Algorithms*, The MIT Press, 2001, volume 56, chapter 16, 539–548.
- Cortes, Corinna, and Vladimir Vapnik. “Support-vector networks.” *Machine Learning* 20, 3: (1995) 273–297.
- Dasarathy, B. V. “Data mining tasks and methods: Classification: nearest-neighbor approaches.” *Handbook of data mining and knowledge discovery* .
- Dettling, M. “Finding predictive gene groups from microarray data.” *Journal of Multivariate Analysis* 90, 1: (2004) 106–131.
- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. “Least Angle Regression.” *Annals of Statistics* 32, 2: (2004) 407–499.
- Fisher, R. A. “The use of multiple measurements in taxonomic problems.” *Annals of Eugenics* 7, 2: (1936) 179–188.
- Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. “Bioconductor: open software development for computational biology and bioinformatics.” *Genome Biol* 5, 10: (2004) R80.
- Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.” *Science* 286, 5439: (1999) 531–7.

- Guyon, Isabelle. "An Introduction to Variable and Feature Selection 1 Introduction." *Journal of Machine Learning Research* 3: (2003) 1157–1182.
- Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene Selection for Cancer Classification using Support Vector Machines." *Machine Learning* 46, 19: (2002) 389–422.
- Hastie, Trevor, Saharon Rosset, Robert Tibshirani, and Ji Zhu. "The Entire Regularization Path for the Support Vector Machine." *Journal of Machine Learning Research* 5, 1: (2004) 1391–1415.
- Kent Ridge Biomedical Data Set Repository. "Kent Ridge Biomedical Data Set Repository." <http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html>. Accessed December 2008., 2008.
- Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. "Optimization by simulated annealing." *Science* 220, 4598: (1983) 671–680.
- Li, Tao, Chengliang Zhang, and Mitsunori Ogihara. "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression." *Bioinformatics* 20, 15: (2004) 2429–2437.
- Liu, Chun-Chi, Wen-Shyen E Chen, Chin-Chung Lin, Hsiang-Chuan Liu, Hsuan-Yu Chen, Pan-Chyr Yang, Pei-Chun Chang, and Jeremy J W Chen. "Topology-based cancer classification and related pathway mining using microarray data." *Nucleic Acids Research* 34, 14: (2006) 4069–4080.
- Pranckeviciene, Erinija, and Ray L. Somorjai. "On Classification Models of Gene Expression Microarrays: The Simpler the Better." In *IJCNN'06*. 2006, 3572–3579.
- Segal, Mark R. "Machine Learning Benchmarks and Random Forest Regression." *Biostatistics* 1–14.
- Somorjai, R. L., B. Dolenko, and R. Baumgartner. "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions." *Bioinformatics* 19, 12: (2003) 1484–1491.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A. Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* 102, 43: (2005) 15,545–15,550.
- Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B Methodological* 58, 1: (1996) 267–288.
- Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proceedings of the National Academy of Sciences of the United States of America* 99, 10: (2002) 6567–72.



- 
- Zhang, H. H., J. Ahn, X. Lin, and C. Park. “Gene selection using support vector machines with non-convex penalty.” *Bioinformatics* 22, 1: (2006) 88–95.



# 5

## Relieving Feature Selection AECS and Pains; a Consensus Approach to Identifying Biomarkers

This chapter constitutes an article in preparation for submission to a journal at the time of thesis submission.

This chapter forms the capstone of this thesis. The previous chapter tested three state-of-the-art learning algorithms against each other on their ability to find feature sets that correctly predict the class of unlabelled samples in a supervised setting. We chose improved versions of two of them in this paper: implementing a simulated annealing component to the SVM optimisation procedure to find deeper local minima, and substituting Generalised Path Seeking for Least Angle Regression. LAR is a very

good algorithm, but we decided that GPS provided a broader flexibility in terms of its constraints, and reflected the shrinkage metaheuristic more accurately.

However, despite both being closely linked, our focus in this paper is bioinformatic feature selection, not choosing the best classification algorithm. The validation procedure of the latter is straightforward, but how do we know that the features that comprise a classifier are truly a reflection of the biological phenomena they imply, or are just a statistical artifact? All algorithms used to find these genes are able to produce robust statistical models. Although SVMSR was shown to have an advantage over the others overall, this advantage was not clear in all datasets tested. We suspected that the use of only one data mining method would give unstable results, even with bootstrapping, and hence we wanted to ‘spread the net more widely’ in gathering our analyses. A review by Stolovitzky (2003) uses the well-known analogy of ‘the blind men and the elephant’ to argue that each feature selection algorithm only illuminates the data from a particular aspect, and that a synoptic view can only be obtained by multilateral sourcing of genetic significance; in other words, a meta-analysis. This inspired our desire to develop a feature selection method that selected features by concordance between ranked lists, where each list was generated by a different data mining algorithm. There is no clear favoured method of pooling the results from multiple ranked lists, and there are plenty to choose from (Boulesteix and Slawski, 2009), so we chose a method that was both simple and empirical: Zipf’s Law (Zipf, 1999). In keeping with the theme of biological significance, we validated our results on the degree to which they dovetailed with the current corpus of biological knowledge.

Note that the in-house dataset used in this study used the same 82 samples as the study in Chapter 3, plus the 34 used as the test set in Chapter 4, combined to produce a 116-sample dataset. This is because the extra 34 samples were not available at the time that the analyses for the study in Chapter 3 were performed. Further discussion on some aspects of the results in this chapter (such as the varying number of features common to all 3 dimension reduction methods among the different datasets) can be found in Chapter 6.

# RELIEVING FEATURE SELECTION AECS AND PAINS; A CONSENSUS APPROACH TO IDENTIFYING BIOMARKERS

Tim Peters<sup>1,\*</sup>, David Bulger<sup>1,\*</sup>, To-ha Loi<sup>2,\*</sup>, Jean Y.H. Yang<sup>3,\*</sup>, David Ma<sup>2,\*</sup>

<sup>1</sup> Department of Statistics, Macquarie University, NSW 2109, Australia <sup>2</sup> Department of Haematology and Bone Marrow Transplant Unit, St Vincent’s Hospital, Darlinghurst, Sydney, Australia <sup>3</sup> School of Mathematics and Statistics, University of Sydney, Australia

\* E-mail: t.peters@mq.edu.au

## Abstract

We consider a consensus approach to feature selection, based on the concordance of features between ranked lists, that includes a diverse group of robust data mining methods. We implement a suite of supervised learning techniques including Support Vector Machine, Random Forest and Generalised Path Seeking, and isolate key features from several transcriptomic datasets through a process of weighting, aggregation, ranking and concordance. We name this process Algorithm Ensemble Concordance Selection (AECS). Using AECS we give, through carefully observed dimension-reduction protocols and the avoidance of common feature selection pitfalls, further weight to the statistical and biological work conducted on well-known gene expression datasets for human lymphoma. AECS also demonstrates a high probability of identifying genes that act as *bellwethers*—genes that act as consistent and reliable biomarkers—for lymphoma subtypes, without the need for enrichment from biological data.

## Background

Interest in data mining has grown considerably in recent years, accompanied by an increase in the application of methods within this ambit to high-dimensional datasets. The development of finely-tuned methods, such as support vector machines (SVMs) and decision trees, has allowed statisticians to aid researchers in gleaning insights into

the latter's subjects of interest. Though also applicable to time-series and econometric datasets, the archetypal field where high-dimensional data-mining is found is gene expression profiling, of which cancer diagnostics is a large subset.

The genetic basis for lymphoma is understood to be undeniable (Skibola et al., 2007; Staudt and Dave, 2005), but its governing biological mechanisms remain enigmatic to the medical research community. Explication of these mechanisms has been found to show promise through bioinformatic approaches. Recently, the employment of computational statistics to analyse the transcriptomes of patients of diseases with underlying genetic components has been extensive. Most work is concerned with overcoming the curse of dimensionality (Somorjai et al., 2003), a phenomenon frequently found in gene expression datasets with few samples (often less than 100) but many variables (often exceeding 20000). Including all variables in the model frequently gives rise to the Hughes Effect (Hughes, 1968; Oommen et al., 2008), where the large number of variables causes the model to overfit, hence reducing its predictive power. This problem can be overcome by many approaches, but the preservation of core genetic components as a tool for elucidating disease aetiology remains elusive. Statistically speaking, the process can be thought of as dimension reduction, but in a bioinformatic context the term feature selection is more often used. The objective of feature selection is to identify a small subset of features that best differentiate between two or more labelled classes. The value of the application of feature selection is potentially enormous, since its biological implications herald promise of the identification of biomarkers which, although they may not reveal the complete dynamic behind the dysfunction, act as bellwethers for the benefit of diagnostic clinicians. Our definition of a bellwether gene set is one whose expression pattern typifies a larger and more complex biological phenomenon. Identification of these genes is crucial in developing diagnostic tools from gene expression data. Just as a small subset of selected features can predict a class label through training and validation, so can the expression pattern of a group of these features (genes) allow clinicians to make confident diagnoses.

A dimensionally-reduced dataset may not contain the complete set of features that allows one sample to be distinguished from another, but if the process is carried out

correctly then the remaining features have a high chance of being bellwethers for the disease class a given sample belongs to. A suite of quality control mechanisms must be implemented during the feature selection process in order to identify a subset of biologically meaningful genetic components. These include:

- Ensuring the risk (that is, the likelihood of an unseen sample being misclassified by the model) is minimised on each model built. Methodology used should be consistent with the principles of Structural Risk Minimization (Vapnik and Chervonenkis, 1974), where the risk is estimated from the data. An appropriate regularization parameter value must be chosen in order to avoid both underfitting and overfitting.
- Choosing an appropriate vector space norm to penalise model residuals.
- Making sure each gene or feature is considered in a setwise context.
- Ensuring the results are stable by bootstrapping.

A given feature selection method may be computationally more economical and elegant than others, but may also fail to confer results that are any less arbitrary. Comparison of feature selection methods has been performed (Caruana, 2006; Jeffery et al., 2006; Statnikov et al., 2008), and no single algorithm has emerged as a clear frontrunner. Caruana (2006) found Random Forests to be superior to Support Vector Machines, whilst Statnikov (2008) found the reverse was true. This study works on the premise that there is ‘no silver bullet’ in terms of a singular superior method. We propose an ensemble approach where a suite of state-of-the-art algorithms—one from each of three major statistical learning ‘strains’—is applied with all aforementioned quality control mechanisms.

The first strain of feature selection involves finding the feature set which best separates the supervised classes by plotting log-transformed gene expression values and hence explicitly resolving the decision boundary in the feature space. Beginning with discriminant analysis (Fisher, 1936), this method’s most widely recognised vanguard is the support vector machine (SVM, Cortes and Vapnik (1995)). The advantage of the

SVM lies in its ability to effectively ‘navigate’ the decision boundary through the regions where data points from different classes are proximal, as the equation parameters are tuned. Two hyperplane margins are drawn equidistant and parallel to the decision boundary. In the context of feature selection, the aim is to find a group of features that maximises the distance between them. However, when there is no hyperplane that can perfectly separate the samples, any samples found on the opposing side of its class’s hyperplane margin are penalised for that model. More formally, the value we want to find is reached by:

$$\min \frac{1}{\|V\|} + \frac{\sum \xi_i}{\lambda} \quad (5.1)$$

$\|V\|$  is the distance between the two hyperplane margins,  $\xi_i$  are the  $i$  slack variables determined by the soft margin, and  $\lambda$  is a regularization trade-off parameter. The trade off is between that of minimizing approximation error via a more generalized model (large  $\lambda$ ) and drawing a more precise decision boundary that may help in correctly classifying borderline samples (small  $\lambda$ ). The selection of the value of  $\lambda$  is critical in determining the minimal value of the criterion (Hastie et al., 2004), and its optimal value may change depending on the dataset. Although the value of the criterion can be seen as a proxy for the true risk of the model, there is no analytical method of finding an appropriate value of  $\lambda$ , and hence it must be estimated empirically (Vapnik, 2000, p. 155). The most rigorous metaheuristic available is via gradient descent methods (Chapelle et al., 2002; Platt, 2000; Bengio, 2000). This process, whilst superior in estimating the true risk of the model, is NP-hard when there is no hyperplane that can perfectly separate the samples (Feldman et al., 2009). In many cases a grid-search using pre-defined range of values of  $\lambda$  (Meinicke et al., 2003) will suffice. We propose that for the purposes of this study, a grid search is satisfactory given that we want to find a group of features whose separability is apparent from even coarse parameter selection, and hence fine-tuning is not likely to be needed. Nevertheless, a ‘ball-park’ value of  $\lambda$  is needed to select even the most obvious candidates. Note that additional optimization architecture (such as forward selection or Monte Carlo-style method) needs to be built



surrounding the SVM criterion value.

The second strain of machine learning is the Decision Tree family of algorithms, where the data  $(x, Y)$  is recursively partitioned into subsets  $(x_1, x_2, x_3 \dots x_k)$  until a subset is generated which confers an optimum predictive ability. Decision Tree learning dates from a method known as Automatic Interaction Detection (Morgan and Sonquist, 1963), and its formal extensions (Morgan and Messenger, 1973; Kass, 1980). The Decision Tree metaheuristic was formalised as a Classification and Regression Tree (CART, Breiman et al. (1984)) for both classification and regression models. Later, entropy minimisation was formally incorporated into this metaheuristic via the ID3 and C4.5 algorithms (Quinlan, 1986, 1993), and more interpretable regression models were also developed, such as MARS (Friedman, 1991). The most recent development in this area is the Random Forest algorithm (Breiman, 2001), whose main advantage is an inbuilt smoothing component achieved by bootstrap aggregation (Breiman, 1996), in order to prevent overfitting. This functionality combines with a stochastic element that randomly chooses subsets of features for each node, and then calculates the best split on this set based on an ‘out-of-bag’ error estimate. ‘Importance’ coefficients for each feature are calculated using the relative decreases in Gini impurity (Breiman, 2001) from a parent to child node. We have selected Random Forests as the representative from the decision tree family to use in this study.

The third and final strain concerns the algorithms derived from LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani (1996)). LASSO is itself derived from the least squares method, where the best model fit corresponds to the minimum sum of squared residuals. However, instead of calculating the sum of residuals, an upper bound is placed on the  $\ell^1$ -norm coefficient vector. This has the effect of shrinking some of the feature coefficients to exactly zero, hence giving a dimensionally reduced, or sparser, result. The value of this upper bound acts as a regularization parameter by which the risk on the model can be minimized, often by gradient descent (Kim and Kim, 2004; Meier et al., 2008; Friedman et al., 2010). LASSO was improved upon with the Least-Angle Regression (LARS) algorithm (Efron et al., 2004), which mimics a stepwise regression but for the coefficients increasing in a joint least squares direction

contingent upon their correlation with the residual. Although this improvement results in a more ‘democratic’ algorithm, it has been argued that the LASSO is limited by the fact that the maximum number of variables that it can hold in one model cannot exceed the sample size  $n$  (Zou and Hastie, 2005). In gene expression profiling, where sample sizes rarely exceed 100, and more often are in the vicinity of 50, it is likely that LASSO and LARS produce solutions which are often too sparse. In response to this Zou and Hastie (2005) developed the Elastic Net, which allows for an ‘interpolation’ of the  $\ell^\gamma$ -norm coefficient vector where  $1 \leq \gamma \leq 2$ . Although the Elastic Net uses a slightly different penalty function from LASSO, the method allows for the creation of a model with a mixture of LASSO-style ( $\ell^1$ -norm) and ridge regression ( $\ell^2$ -norm, Hoerl and Kennard (1970)) penalties, where an optimal value of  $\gamma$  can be found via a grid search.

On the other hand, there are problems that are better solved by an extension of LASSO into non-convex constraints, such as when features are highly independent. In these instances, the need for a sparse solution outweighs the advantages retaining the potential information gain from ancillary features (in other words, shrinking them to non-zero value), although the risk of multiple solutions is always present with non-convexity. We can see an analogy between the roles of  $\gamma$  and the regularization parameter  $\lambda$ , where  $\lambda$  calibrates the trade-off between bias and variance,  $\gamma$  allows for a trade-off between the desirable yet somewhat contradictory attributes of convexity and sparsity. Friedman (2008) extends Zou and Hastie (2005) to produce continuous path solutions for  $0 \leq \gamma \leq 2$ . This Generalized Path Seeking (GPS) algorithm serves as the representative from the LASSO family for this study.

Although each metaheuristic is powerful in its own way, each must be calibrated correctly in order to eliminate hidden biases. As mentioned before, support vector machines need an appropriate regularization parameter value for optimal results, and co-efficient shrinkage methods (LASSO, GPS) need both regularization and, in the case of GPS, the degree of shrinkage calibrated by the user. Random forests use a random subset of features at each node to provide a split; the size of the subset relative to the complete list is somewhat important (Breiman, 2001).

Two other important caveats will now briefly be discussed. It is important that the methods chosen analyse features contextually, that is, assessed in sets. Ranking features individually on their criterion value and then selecting an extreme percentile (called filtering) is a poor method of feature selection, since certain features on their own may have only a modest ability to separate two classes, but show a *gestalt* ability to achieve this in higher dimensions as members of a set (Guyon, 2003), (See Chapter 3). Embedded methods (Guyon, 2003), where the power and robustness of the learner are ratcheted up using a statistical criterion, and subsequent validation, as features are added (or removed) from the working feature set, are preferable to filtering. A gradient descent with an SVM margin width as the criterion, and GPS, are both embedded algorithms. The Random Forest method can be considered a ‘meta-wrapper’, a black box learner with an extra layer of validation, and although not strictly an embedded method, exploits the separation power of feature sets. The implementation of a robust embedded learning machine is effective at exploiting the setwise separation power of genes (Guyon, 2003; Saeys et al., 2007).

Finally, it is important for the final feature set to be stable. This means that each feature in the final set must have a high probability of being selected over a set of bootstrapped selection runs (Meinshausen and Bühlmann, 2010). Related to the previous point on feature context, a feature’s coefficient is not necessarily the best indicator of its worth as a biological bellwether, even for those features that have been selected contextually. The probability of a feature occurring in the final set (estimated from bootstrapping) is often a more reliable indicator (Meinshausen and Bühlmann, 2010). To reduce outlier sensitivity, implementing a bootstrapping mechanism is recommended.

The notion of a unified framework for feature selection—one that draws on varying methods contributing to a consensus model and produces end results using a scoring system—is not new to the field (Shaik and Yeasin, 2007; Yu et al., 2007). Other groups have attempted to extract biological insights by using a pre-defined set of genes, such as with Gene Set Enrichment Analysis (Subramanian et al., 2005). Our philosophy is to attack the problem purely from the statistical end, with a twofold justification:

1. It allows independent verification of shortlisted features that are associated with genes whose implication in the diseases studied have been proven in a laboratory environment.
2. It provides scope for novel bellwether gene detection, since all features are considered.

We have chosen three state-of-the-art algorithms to use, but welcome variations and embellishments on the group of algorithms chosen and in order to make the process even more rigorous. We call our method Algorithm Ensemble Concordance Selection (AECS). Our aim, using this ensemble approach on a number of well-studied publicly available datasets, and an in-house dataset, is to detect a subset of biologically significant genes using statistical methods only, and validate our findings in the existing biological literature.

## Data and Methodology

Four cancer-related datasets were used in this study. Alizadeh (2000) was downloaded from (Broad Institute of Medicine, 2009b), Golub (1999) and Shipp (2002) from (Broad Institute of Medicine, 2009a), and one was provided privately by the Department of Haematology and Bone Marrow Transplant Unit, St Vincent's Hospital, Darlinghurst, Sydney, Australia. The publicly available datasets are perhaps the most analysed publicly available transcriptome datasets from lymphoma patients. Alizadeh contrasts subtypes of diffuse large B-cell lymphoma: one expressing genes characteristic of germinal centre B cells (GC B-like) and the other expressing genes normally induced during in vitro activation of peripheral blood B cells (activated B-like). Golub compared acute myeloid (AML) and acute lymphoblastic leukaemia (ALL), and Shipp compared follicular (FL) and diffuse large B-cell lymphoma (DLBCL), a split we have also performed with our own samples.

The lymph node biopsies, RNA and microarray assays and the resulting data from St Vincent's Hospital (our in-house dataset) are identical to those described in (Loi

et al., 2011). Where applicable, missing values were imputed using the GEPAS online preprocessing tool at <http://gepas.bioinfo.cipf.es/cgi-bin/preprocess> using  $k$ -nearest neighbour imputation with  $k = 10$ .

Our 116-sample in-house dataset was split into seven subgroups, each with a binary response variable. An outline of the samples contained in each dataset can be found in Table 1. The term ‘Reactive’ refers to tissue samples that were initially suspected to have malignancy, but in fact were found to have inflammatory changes with no evidence of cancer. Lower subgroups were paired with the reactive in order to find characterising genes for these particular lymphoma subtypes.

Dataset Source	Biological Response Variable	Array Type	Number of samples ( $n$ )	Number of features ( $P$ )
Alizadeh et al. (2000)	DLBCL subtypes: GC B-like vs. Activated B-like	Two-colour	42 (21 Ac-tivated, 21 Germinal)	4029
Golub et al. (1999)	ALL vs. AML		38 (27 ALL, 11 AML)	7129
Shipp et al. (2002)	DLBCL vs. FL	Affymetrix	77 (58 DLBCL, 19 FL)	
In House	Hodgkin’s Lymphoma (HL) vs. Non-Hodgkin’s Lymphoma (NHL)	Two-colour	93 (19 HL, 74 NHL)	18661
	FL vs. DLBCL		54 (35 FL, 19 DLBCL)	
	HL vs. Reactive		42 (19 HL, 23 Reactive)	
	NHL vs. Reactive		97 (74 NHL, 23 Reactive)	

	FL vs. Reactive		58 (35 FL, 23 Reactive)
	DLBCL vs. Reactive		42 (19 DLBCL, 23 Reactive)

TABLE 5.1: Summary of the datasets used in this study. The structure of the response variable splits from the in-house dataset is described in Figure 5.1.

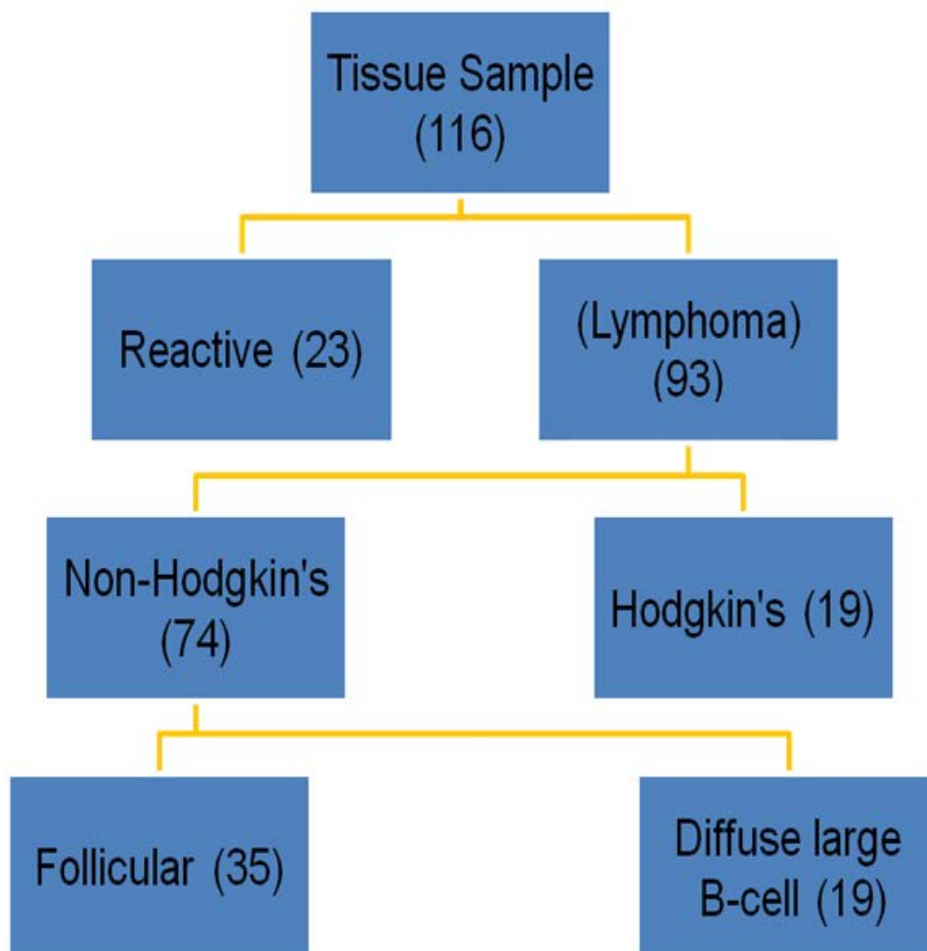


FIGURE 5.1: Hierarchical chart showing the biological classification of the response variables used in this study. ‘Reactive’ denotes a class of samples that were suspected to belong to patients with neoplasms, but received a negative diagnosis. Twenty remaining samples diagnosed as NHL were diagnosed as other lymphoma subclasses, and were left out of the lowest level split.

For all 10 data splits, we applied a triumvirate of dimension reduction methods (SVM forward selection with simulated annealing, Random Forest and Generalized Path Seeking). We bootstrapped all applications 150 times, with a randomised selection of  $n/2$  samples for each bootstrap. Odd totals were rounded up and class ratios were preserved. For each bootstrap, we ranked the final reduced group of features according to the degree of importance each feature had in contributing to the model; this depended on the method used and will be explained below for each method. We scored and aggregated features using Zipf’s Law (Zipf, 1999), an empirical observation which states that, for ranked data, the frequency with which an event occurs (for example, words in a corpus of text) is often approximately inversely proportional to the ranking it is given by that frequency. Hence for each of the 150 ranked lists of dimensionally reduced features, we gave the most important feature a score of 1, the second most important  $1/2$ , the third  $1/3$  and so on. We aggregated all totals for each feature, and a ‘master rank’ list for each data split was created based on these aggregates. This created thirty such lists.

To represent the SVM method of dimension reduction, we implemented a forward stepwise regression with an embedded simulating annealing (SA) routine. We used the penalised SVM soft-margin width, as described earlier, as the objective function. For each data split, an appropriate value of the regularisation parameter  $\lambda$  was obtained by way of a grid search; a leave-one-out (LOO) cross validation classification procedure was performed for  $\lambda = \{0.01, 0.1, 1, 10, 100\}$  and the value associated with the lowest LOO error was chosen for the selection runs proper. Features were added one by one to the model based on the feature whose addition conferred the greatest decrease in the objective function, with the SA step run after each added feature  $q$  for  $2 \leq q \leq 20$ . The SA routine ran through the entire current subset multiple times, one feature at a time, exhaustively substituting features that conferred a lower objective function (if found) for the existing members of the feature set until convergence was reached. Convergence was defined as a complete cycle of the current subset for which no feature substitutions occurred. To rank the importance of the 20 reduced features, the differences between the objective functions for the sets of 19 features with the feature in question removed

and the original objective function were calculated. Not surprisingly, this method was highly computationally intensive and was achieved in the order of weeks, but was performed for the sake of exhaustive search.

We implemented 150 Random Forest runs representing the decision tree metaheuristic. Although bootstrapping occurs within a single Random Forest application, we added another layer in the interests of standardisation. Random Forest estimates feature importance internally; the method is described in (Breiman, 2001). Importances for all features were ranked for each bootstrap and aggregated as described above.

We implemented 150 GPS runs representing the LASSO-style metaheuristic. As recommended by (Friedman, 2009a), we chose an appropriate penalty function, denoted by  $\gamma$ , by performing a validation grid search over  $\gamma = 0, 0.1, 0.2, 0.5, 1, 1.5, 2$  and choosing the value of  $\gamma$  associated with the lowest test error. Feature importance was determined by order of entry of each feature into the model, which, although not completely reliable, acts as a rule of thumb for predictor significance in LASSO-style results (Wu et al., 2009).

Both Random Forest and GPS had fast computational times; the entire bootstrapping run for each was completed in the order of minutes. All computations were performed using R Version 2.5 with additional packages *kernlab* and *randomForest*, and software from (Friedman, 2009b).

For each of our 10 bipartite splits, we produced 3 master lists of aggregated performance scores for each algorithm used. Using an arbitrary cutoff of the top 20 features in each ranked list, we selected the features that were common to all 3 master lists as components of our AECS-generated bellwether set. Features that were common to 2 of the 3 master lists were also earmarked as potential bellwethers. The sign of the  $t$ -statistic of each feature was used to determine whether the corresponding gene was up-regulated or down-regulated in terms of the response variable.



## Results and Discussion

The following table lists the genes corresponding to the features we isolated, using AECS, for all 10 data splits. The caption is at the end of the table.

Alizadeh et al. (2000)
<b>GC B-like</b> Clone=825217, UG Hs.169565 ESTs, Moderately similar to ALU SUBFAMILY SB [H.sapiens] Clone=1353041, linked to elongation factor 1-beta in humans AT5G19510 Clone=1334260, UG Hs.120716 ESTs, linked to Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 9 or SERPINA9 Clone=815539, JAW1=lymphoid-restricted membrane protein, linked to Lymphoid-restricted membrane protein LRMP Clone=1353015, linked to LRMP Clone=746300, UG Hs.136345 ESTs Clone=1338448, UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens] Clone=417502, JAW1=lymphoid-restricted membrane protein, linked to LRMP Clone=2005 Clone=1268870, UG Hs.120245 Homo sapiens mRNA for KIAA1039 protein, partial cds, linked to RAP1 GTPase activating protein 2 RAP1GAP2 Clone=825199 Clone=1358244, UG Hs.124922 ESTs, linked to LRMP Clone=1337653, UG Hs.124922 ESTs, linked to LRMP <b>Activated B-like</b> Clone=1355435, UG Hs.169081 ets variant gene 6 (TEL oncogene)
Golub et al. (1999)
<b>AML</b>

<p>Leukotriene C4 synthase LTC4S</p> <p>Leptin receptor overlapping transcript LEPROT</p> <p>Zyxin ZYX</p> <p>CD33 molecule CD33</p> <p>Fumarylacetoacetate FAH</p> <p>Cholinergic receptor, nicotinic, alpha 7 CHRNA7</p> <p>Complement factor properdin CFP</p>
<b>Shipp et al. (2002)</b>
<p><b>FL</b></p> <p>POU class 6 homeobox 1 POU6F1</p> <p><b>DLBCL</b></p> <p>lactate dehydrogenase A LDHA</p> <p>enolase 1, (alpha) ENO1</p>
<b>Lymphoma vs. Reactive (in house)</b>
<p><b>Reactive</b></p> <p>Immunoglobulin lambda variable 6-57 IGLV6-57</p> <p>Chemokine (C-C motif) ligand 21 CCL21</p> <p>Isolate RSV34L immunoglobulin light chain variable region</p> <p>IgG lambda light chain V-J-C region (clone Tgl11)</p> <p>Immunoglobulin kappa variable 3-20 IGKV3-20</p> <p>LUC7-like 3 (S. cerevisiae) LUC7L3</p> <p>Homo sapiens mRNA fragment, [Genbank:L10148]</p>
<b>HL vs. NHL (in house)</b>
<p><b>HL</b></p> <p>Chemokine (C-C motif) ligand 22 CCL22</p> <p>Killer cell lectin-like receptor subfamily B, member 1 KLRB1</p> <p>LENG10 mRNA, partial sequence [Genbank:AF211977]</p> <p>CD7 molecule CD7</p> <p>Chemokine (C-C motif) ligand 17 CCL17</p>

FL vs. DLBCL (in house)
<b>FL</b> Transmembrane protein 150C TMEM150C <b>DLBCL</b> CD163 molecule CD163 Phosphogluconate dehydrogenase PGD ATP-binding cassette, sub-family A (ABC1), member 1 ABCA1 Cell division cycle 25 homolog A (S. pombe) CDC25A GAR1 ribonucleoprotein homolog (yeast) GAR1
HL vs. Reactive (in house)
<b>HL</b> Chemokine (C-C motif) ligand 17 CCL17 Major vault protein MVP Glycerol kinase GK
NHL vs. Reactive (in house)
<b>Reactive</b> Isolate RSV34L immunoglobulin light chain variable region Immunoglobulin lambda variable 6-57 IGLV6-57 IgG lambda light chain V-J-C region (clone Tgl11) Homo sapiens mRNA fragment [Genbank:L10148] Oxidative stress induced growth inhibitor 1 OSGIN1 Similar to hCG2042717 [Genbank:AF035799] Immunoglobulin heavy variable 3-48 IGHV3-48 Cytoskeleton associated protein 2 CKAP2 Similar to hCG1642538 [Genbank:AF026929]
FL vs. Reactive (in house)
<b>Reactive</b> CD163 molecule CD163 X-box binding protein 1 XBP1

<p>Immunoglobulin lambda variable 6-57 IGLV6-57</p> <p>Ig gamma-1 chain C region-like</p> <p>Ig rearranged gamma-chain mRNA, subgroup VH2, V-D-J region</p> <p>Isolate RSV34L immunoglobulin light chain variable region</p> <p>IgG lambda light chain V-J-C region (clone Tgl11)</p> <p>Immunoglobulin lambda constant 1 (Mcg marker) IGLC1</p> <p>Accession NM.018395</p>
DLBCL vs. Reactive (in house)
<p><b>DLBCL</b></p> <p>Glycerol kinase GK</p> <p>RAD23 homolog B (S. cerevisiae) RAD23B</p> <p><b>Reactive</b></p> <p>Immunoglobulin lambda variable 6-57 IGLV6-57</p> <p>Chemokine (C-C motif) receptor 6 CCR6 (<b>Reactive</b>)</p> <p>Ubiquitin specific peptidase 53 USP53</p> <p>Placenta-specific 8 PLAC8</p> <p>Chromosome X open reading frame 48 CXorf48</p>

TABLE 5.2: Fully dimensionally-reduced feature lists from each of the 10 data splits analysed, using AECS. Known links to genes from expressed sequence tags (ESTs) are given. Features are grouped according to the class in which they show a higher degree of expression.

Features selected using AECS from all 3 of the publicly available datasets include genes that have an implication in these datasets’ diagnostic purview. Thirteen features were selected from Alizadeh (2000) and three of the five associated genes have an acknowledged lymphoma association. The presence of SERPINA9 is considered to be a defining molecular signature of DLBCL lymphomas with a germinal (GC B-like) centre phenotype (Paterson et al., 2008; Pan et al., 2003). LRMP/Jaw1 is also considered to be an associated marker of GC B-like DLBCL (Tedoldi et al., 2006).

Cytogenetic abnormalities on chromosome 12 involving TEL/ETV6 gene have been found in DLBCL tissue samples (Sevilla et al., 2009; Berger et al., 1997). TEL was the only gene identified to have up-regulation associated with activated B-like DLBCL; one sample described as having a complex karyotype in Sevilla et al. 2009 was found to have cytogenetic rearrangement on the TEL locus. Additionally, RAP1GAP2 plays an important role in platelet aggregation (Schultess et al., 2005).

In the Golub (1999) dataset, four of the seven features selected have implications in leukaemogenesis. CD33 has long been established as central to AML aetiology (Griffin et al., 1984; Dinndorf et al., 1986) and is the target of drugs used for AML treatment (Walter et al., 2007). Implications in AML have also been found for leptin (Hamed et al., 2003; Nakao et al., 1998) and zyxin (Wang et al., 2005), and a difference in bone marrow and peripheral blood expression was found for LCT4S (Sakhinia et al., 2006).

In the Shipp (2002) dataset, both features found to be up-regulated in DLBCL have literature supporting their involvement. Though implicated in FL, LDHA expression is variable within this phenotype, as compared to a more consistent presence of LDHA in DLBCL (Giatromanolaki et al., 2008), and its inhibition is found to slow tumour progression (Le et al., 2010). Enolase has also been suggested as playing a role in lymphoma differentiation (Mohammad et al., 1994). Peripherally, POU6F1 is found to have an important role in ovarian clear-cell carcinoma (Suzuki et al., 2010), but no role in FL has yet been published.

The studies that produced all 3 publicly available datasets performed their own feature selection analysis, ranking genes that best distinguished between their respective diagnostic classes. Both Shipp and Golub used an original supervised weighted class-voting method, and Alizadeh used a method identical to Golub. Of the 100 features selected as possessing the most significant class distinctions (50 for each class), all thirteen features selected by AECS were also in this list, including the top seven most significantly up-regulated in the GC B-like group and the most up-regulated feature in the activated B-like group. From the top 100 genes from the Shipp dataset, AECS selected the top ranked gene from FL with the top 2 from DLBCL as its bellwether set. Six of the seven genes found by AECS from the Golub dataset were found in the

top 50 of their own analysis, including the top 3 ranked in AML by Golub et al.

There is clearly a high degree of overlap of AECS-selected features with the most differentiated genes, as found by the publishers' own statistical analyses, in the 3 publicly available datasets we have studied. This, in addition to the success we have had in matching publicly available transcriptomes with supporting literature on their phenotypes, buttresses our belief that AECS is a discerning method that can not only convey valuable and potentially critical information to clinicians investigating lymphoma differentiation, but can be applied to any discipline that produces high-dimensional data. Particular areas of interest may include econometrics and other biological disciplines such as ecology.

Encouragingly, the expression patterns of the features selected by AECS from analysis of our in-house dataset support existing hypotheses about the nature of lymphoma. Firstly, and most strikingly, the group of features extracted by AECS when a broad spectrum of lymphoma samples was compared to reactive samples contained a large percentage of relatively down-regulated (in lymphoma) immunoglobulin (Ig) constituents. Four of the seven features from the lymphoma vs. reactive split and four of the nine from the NHL vs. reactive split were from Ig loci. Without exception, all were relatively down-regulated in the lymphoma tissues. This may be due to one or both of the following hypotheses. Firstly, the down-regulation may be the result of a preponderance of undifferentiated B-cells in the lymphoma samples, since Ig production does not occur until later in the maturation of B-cells (Alberts et al., 2008). Secondly, the up-regulation in the reactive samples may be explained by an immune response to an antigenic challenge, such as a foreign irritant, accelerating the production of mature B-cells. In addition, the mature B-cell marker chemokine receptor 6 (CCR6) was heavily down-regulated in DLBCL vs. Reactive, and is conspicuously absent in mediastinal large B-cell lymphoma (Rehm et al., 2009). Also, Xbox1 protein (XBP1), a master regulator of the secretory mechanism of plasma cells (Shaffer et al., 2004), and a likely driver of their immunoglobulin secretion (Staudt and Dave, 2005), is heavily down-regulated in FL vs. Reactive.

Other feature patterns also dovetailed with existing research, especially in Hodgkin’s lymphoma. Chemokine ligand 17 (CCL17) showed significant up-regulation in HL, which, along with CCL22 (Niens et al., 2008; Maggio et al., 2002) and CD7 molecule (Seegmiller et al., 2009), are biomarkers for that lymphoma subtype. Major vault protein (MVP), a gene selected when HL was split with the reactive samples, has been found to inhibit apoptosis in some instances (Ryu et al., 2008; Lloret et al., 2009), although no link has yet been published relating MVP to HL.

Comparing FL and DLBCL also yielded interesting feature patterns. Cell division cycle 25 homolog A (CDC25A) was relatively up-regulated in DLBCL, which is a marker of histologically aggressive B-cell NHLs (Moreira Júnior et al., 2003). Paradoxically, CD163 molecule, which was heavily down-regulated in FL in comparison to both reactive and DLBCL samples, is a biomarker for FL. However, its expression is complex, not observed in the lymphoma itself but restricted to the macrophage lineage (Nguyen et al., 2005) and the immediate sprouting environment (Clear et al., 2010). Further research on the nature of the relationship between CD163 and follicular lymphoma is advisable.

It has been our intention to create a feature selection tool that preserves the most critical features of a high-dimensional dataset while eliminating the incidental ones. While computation is no substitute for the judgement of the skeptical clinician or molecular biologist, reducing the feature set to a handful of key features is needed for clear direction in future cancer research. We have demonstrated the twofold objective of this project: to validate existing clinical research and to provide new directions for further study, the latter evidenced by our findings on the MVP/HL and CD163/FL relationships.

While we believe that we have used the most recent, state-of-the-art algorithms in this study, we welcome the substitution or addition of other machine learning methods into the AECS paradigm. A minimum of 3 algorithms is advisable for finding concordance between dimensionally-reduced lists, but an extensive application of heterogeneous algorithms may reduce the degree of concordance between them. An increase in the size of the top quantile of score-aggregated and ranked features may remedy this.

Ranking features by differential expression using the  $t$ -test may be useful for an extra level of quality control, but we recommend that  $t$ -test rankings should not be used as a selection method proper alongside the more robust decision tree and coefficient shrinkage methods. For example, our results evinced a feature that appeared in two of the three aggregated ‘master’ lists, up-regulated in HL (vs. Reactive), that was found to be ranked outside the top decile of up-regulated genes using the  $t$ -test. This re-enforces the need to always consider features as acting in consort with others statistically, as they do biologically, and to choose algorithms that exploit this statistical phenomenon. Attention to this caveat, along with careful implementation and tuning of regularisation, shrinkage and bootstrapping parameters described in this paper, is likely to ensure meaningful and valuable results.

## Acknowledgements

Generation of the in-house microarray data was funded in part by Australian grants from The Sydney Foundation for Medical Research, St Vincent’s Hospital Haematology Research Fund and the Arrow Bone Marrow Transplant Foundation.

## References

- Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, volume 54. Garland Press, 2008.
- Bengio, Yoshua. “Gradient-Based Optimization of Hyperparameters.” *Neural Comput.* 12, 8: (2000) 1889–1900.
- Berger, C., P. Brousset, C. McQuain, and H. Knecht. “Deletion variants within the NF-kappaB activation domain of the LMP1 oncogene in acquired immunodeficiency syndrome-related large cell lymphomas, in prelymphomas and atypical lymphoproliferations.” *Leukemia lymphoma* 26, 3-4: (1997) 239–250.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*, volume 19 of *Statistics/Probability Series*. Wadsworth, 1984.
- Breiman, Leo. “Bagging predictors.” *Machine Learning* 24, 2: (1996) 123–140.
- . “Random Forests.” *Machine Learning* 45, 1: (2001) 5–32.



- Broad Institute of Medicine. “GenePattern.” <http://www.broadinstitute.org/cancer/software/genepattern/datasets/>. Accessed November 2009., 2009a.
- . “Lymphoma/Leukemia Molecular Profiling Project.” <http://llmpp.nih.gov/lymphoma/>. Accessed November 2009, 2009b.
- Caruana, Rich. “An Empirical Comparison of Supervised Learning Algorithms.” *Advances* 161–168.
- Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee. “Choosing Multiple Parameters for Support Vector Machines.” *Machine Learning* 46, 1: (2002) 131–159.
- Clear, Andrew J., Abigail M. Lee, Maria Calaminici, Alan G. Ramsay, Kelly J. Morris, Simon Hallam, Gavin Kelly, Finlay MacDougall, T. Andrew Lister, and John G. Gribben. “Increased angiogenic sprouting in poor prognosis FL is associated with elevated numbers of CD163+ macrophages within the immediate sprouting microenvironment.” *Blood* 115, 24: (2010) 5053–5056.
- Cortes, Corinna, and Vladimir Vapnik. “Support-vector networks.” *Machine Learning* 20, 3: (1995) 273–297.
- Dinndorf, P. A., R. G. Andrews, D. Benjamin, D. Ridgway, L. Wolff, and I. D. Bernstein. “Expression of normal myeloid-associated antigens by acute leukemia cells.” *Blood* 67, 4: (1986) 1048–1053.
- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. “Least Angle Regression.” *Annals of Statistics* 32, 2: (2004) 407–499.
- Feldman, Vitaly, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. “Agnostic Learning of Monomials by Halfspaces Is Hard.” *Proc 50th IEEE Symp on Foundations of Comp Sci* 385–394.
- Fisher, R. A. “The use of multiple measurements in taxonomic problems.” *Annals of Eugenics* 7, 2: (1936) 179–188.
- Friedman, J. H. “Multivariate adaptive regression splines.” *Annals of Statistics* 19, 1: (1991) 1–67.
- . “GPS Help.” <http://www-stat.stanford.edu/~jhf/r-gps/GPShelp.html>, 2009a.
- . “GPS Software.” <http://www-stat.stanford.edu/~jhf/r-gps/>, 2009b.
- Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal Of Statistical Software* 33, 1: (2010) 1–22.

- Giatromanolaki, Alexandra, Michael I Koukourakis, Francesco Pezzella, Efthimios Sivridis, Helen Turley, Adrian L. Harris, and Kevin C. Gatter. "Lactate dehydrogenase 5 expression in non-Hodgkin B-cell lymphomas is associated with hypoxia regulated proteins." *Leukemia lymphoma* 49, 11: (2008) 2181–2186.
- Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286, 5439: (1999) 531–7.
- Griffin, J. D., D. Linch, K. Sabbath, P. Larcom, and S. F. Schlossman. "A monoclonal antibody reactive with normal and leukemic human myeloid progenitor cells." *Leukemia Research* 8, 4: (1984) 521–534.
- Guyon, Isabelle. "An Introduction to Variable and Feature Selection 1 Introduction." *Journal of Machine Learning Research* 3: (2003) 1157–1182.
- Hamed, Nahla A. M., Ola A. Sharaki, and Mohamed M. Zeidan. "Leptin in acute leukaemias: relationship to interleukin-6 and vascular endothelial growth factor." *The Egyptian journal of immunology Egyptian Association of Immunologists* 10, 1: (2003) 57–66.
- Hastie, Trevor, Saharon Rosset, Robert Tibshirani, and Ji Zhu. "The Entire Regularization Path for the Support Vector Machine." *Journal of Machine Learning Research* 5, 1: (2004) 1391–1415.
- Hoerl, A. E., and R. W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 1: (1970) 55–67.
- Hughes, G. "On the mean accuracy of statistical pattern recognizers." *IEEE Transactions on Information Theory* 14, 1: (1968) 55–63.
- Jeffery, Ian B., Desmond G. Higgins, and Aedín C. Culhane. "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data." *BMC bioinformatics* 7: (2006) 359.
- Kass, G. V. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics* 29, 2: (1980) 119–127.
- Kim, Yongdai, and Jinseog Kim. "Gradient LASSO for feature selection." *Twenty-first International Conference on Machine Learning ICML 04* 60.
- Le, Anne, Charles R. Cooper, Arvin M Gouw, Ramani Dinavahi, Anirban Maitra, Lorraine M. Deck, Robert E. Royer, David L. Vander Jagt, Gregg L. Semenza, and Chi V. Dang. "Inhibition of lactate dehydrogenase A induces oxidative stress and inhibits tumor progression." *Proceedings of the National Academy of Sciences of the United States of America* 107, 5: (2010) 2037–2042.

- Lloret, Marta, Pedro Carlos Lara, Elisa Bordón, Fausto Fontes, Agustin Rey, Beatriz Pinar, and Orlando Falcón. "Major vault protein may affect nonhomologous end-joining repair and apoptosis through Ku70/80 and bax downregulation in cervical carcinoma tumors." *International Journal of Radiation Oncology, Biology, Physics* 73, 4: (2009) 976–979.
- Loi, To Ha, Anna Campain, Adam Bryant, Tim J Molloy, Mark Lutherborrow, Jennifer Turner, Yee Hwa Jean Yang, and David D. F. Ma. "Discriminating lymphomas and reactive lymphadenopathy in lymph node biopsies by gene expression profiling." *BMC medical genomics* 4, 1: (2011) 27.
- Maggio, E. M., A. Van Den Berg, L. Visser, A. Diepstra, J. Kluiver, R. Emmens, and S. Poppema. "Common and differential chemokine expression patterns in rs cells of NLP, EBV positive and negative classical Hodgkin lymphomas." *International Journal of Cancer* 99, 5: (2002) 665–672.
- Meier, Lukas, Sara Van De Geer, and Peter Bühlmann. "The group lasso for logistic regression." *Society* 70, 1: (2008) 53–71.
- Meinicke, Peter, Thorsten Twellmann, and Helge Ritter. "Discriminative Densities from Maximum Contrast Estimation." In *Advances in Neural Information Processing Systems 15*, edited by S Thrun S Becker, and K Obermayer, MIT Press, 2003, 985–992.
- Meinshausen, Nicolai, and Peter Bühlmann. "Stability Selection (with discussion)." *Journal of the Royal Statistical Society Series B* .
- Mohammad, R. M., M. Y. Hamdan, and A. Al-Katib. "Induced expression of alpha-enolase in differentiated diffuse large cell lymphoma." *Enzyme protein* 48, 1: (1994) 37–44.
- Moreira Júnior, Gilberto, Gisele W. B. Colleoni, M. Giulia Cangi, Michael Murphy, Bradford Sherburne, José O Bordin, and Massimo Loda. "Reciprocal Cdc25A and p27 expression in B-cell non-Hodgkin lymphomas." *Diagnostic Molecular Pathology, The American Journal of Surgical Pathology Part B* 12, 3: (2003) 128–132.
- Morgan, James N., and Robert C. Messenger. *THAID, a sequential analysis program for the analysis of nominal scale dependent variables, by James N. Morgan [and] Robert C. Messenger*. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor,, 1973.
- Morgan, James N., and John A. Sonquist. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58, 302: (1963) 415–434.
- Nakao, T., M. Hino, T. Yamane, Y. Nishizawa, H. Morii, and N. Tatsumi. "Expression of the leptin receptor in human leukaemic blast cells." *British Journal of Haematology* 102, 3: (1998) 740–745.

- Nguyen, TuDung T., Erich J. Schwartz, Robert B. West, Roger A. Warnke, Daniel A. Arber, and Yasodha Natkunam. "Expression of CD163 (hemoglobin scavenger receptor) in normal tissues, lymphomas, carcinomas, and sarcomas is largely restricted to the monocyte/macrophage lineage." *The American Journal of Surgical Pathology* 29, 5: (2005) 617–624.
- Niens, Marijke, Lydia Visser, Ilja M. Nolte, Gerrit Van Der Steege, Arjan Diepstra, Pablo Cordano, Ruth F. Jarrett, Gerard J. Te Meerman, Sibrand Poppema, and Anke Van Den Berg. "Serum chemokine levels in Hodgkin lymphoma patients: highly increased levels of CCL17 and CCL22." *British Journal of Haematology* 140, 5: (2008) 527–536.
- Oommen, Thomas, Debasmita Misra, Navin K C Twarakavi, Anupma Prakash, Bhaskar Sahoo, and Sukumar Bandopadhyay. "An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing." *Mathematical Geosciences* 40, 4: (2008) 409–424.
- Pan, Zenggang, Yulei Shen, Cheng Du, Guimei Zhou, Andreas Rosenwald, Louis M. Staudt, Timothy C. Greiner, Timothy W. McKeithan, and Wing C. Chan. "Two newly characterized germinal center B-cell-associated genes, GCET1 and GCET2, have differential expression in normal and neoplastic B cells." *The American Journal of Pathology* 163, 1: (2003) 135–144.
- Paterson, Melinda A., Patrick S. Hosking, and Paul B. Coughlin. "Expression of the serpin centerin defines a germinal center phenotype in B-cell lymphomas." *American Journal of Clinical Pathology* 130, 1: (2008) 117–126.
- Platt, John. "Probabilities for support vector machines." In *Advances in Large Margin Classifiers*, edited by A Smola, P Bartlett, B Scholkopf, and D Schuurmans, MIT press, 2000.
- Quinlan, J. R. "Induction of decision trees." *Machine Learning* 1, 1: (1986) 81–106.
- . *C4.5: Programs for Machine Learning*, volume 240 of *Morgan Kaufmann series in Machine Learning*. Morgan Kaufmann, 1993.
- Rehm, Armin, Ioannis Anagnostopoulos, Kerstin Gerlach, Meike Broemer, Claus Scheidereit, Korinna Jöhrens, Michael Hübler, Roland Hetzer, Harald Stein, Martin Lipp, Bernd Dörken, and Uta E. Höpken. "Identification of a chemokine receptor profile characteristic for mediastinal large B-cell lymphoma." *International Journal of Cancer, Journal International du Cancer* 125, 10: (2009) 2367–2374.
- Ryu, S. J., H. J. An, Y. S. Oh, H. R. Choi, M. K. Ha, and S. C. Park. "On the role of major vault protein in the resistance of senescent human diploid fibroblasts to apoptosis." *Cell Death Differ* .
- Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *Bioinformatics (Oxford, England)* 23, 19: (2007) 2507–17.

- Sakhinia, E., M. Farahangpour, E. Tholouli, J. A. Liu Yin, J. A. Hoyland, and R. J. Byers. "Comparison of gene-expression profiles in parallel bone marrow and peripheral blood samples in acute myeloid leukaemia by real-time polymerase chain reaction." *Journal of Clinical Pathology* 59, 10: (2006) 1059–1065.
- Schultess, Jan, Oliver Danielewski, and Albert P. Smolenski. "Rap1GAP2 is a new GTPase-activating protein of Rap1 expressed in human platelets." *Blood* 105, 8: (2005) 3185–3192.
- Seegmiller, Adam C., Nitin J. Karandikar, Steven H. Kroft, Robert W. McKenna, and Yin Xu. "Overexpression of CD7 in classical Hodgkin lymphoma-infiltrating T lymphocytes." *Cytometry Part B Clinical cytometry* 76, 3: (2009) 169–174.
- Sevilla, D. W., S. V. Nandula, A. I. Colovai, S. Alexander, V. V. Murty, B. Alobeid, and G. Bhagat. "Diffuse large B-cell lymphoma with TEL/ETV6 translocation." *Human Pathology* 40, 1532-8392 (Electronic) PT - Case Reports: (2009) 588–593.
- Shaffer, A. L., Miriam Shapiro-Shelef, Neal N. Iwakoshi, Ann-Hwee Lee, Shu-Bing Qian, Hong Zhao, Xin Yu, Liming Yang, Bruce K. Tan, Andreas Rosenwald, Elaine M. Hurt, Emmanuel Petroulakis, Nahum Sonenberg, Jonathan W. Yewdell, Kathryn Calame, Laurie H. Glimcher, and Louis M. Staudt. "XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation." *Immunity* 21, 1: (2004) 81–93.
- Shaik, Jahangheer S., and Mohammed Yeasin. "A unified framework for finding differentially expressed genes from microarray experiments." *BMC Bioinformatics* 8, 7: (2007) 347.
- Shipp, Margaret A., Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C. T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub. "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Technical Report 1, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. [margaret\\_shipp@dfci.harvard.edu](mailto:margaret_shipp@dfci.harvard.edu), 2002.
- Skibola, Christine F., John D. Curry, and Alexandra Nieters. "Genetic susceptibility to lymphoma." *Haematologica* 92, 7: (2007) 960–969.
- Somorjai, R. L., B. Dolenko, and R. Baumgartner. "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions." *Bioinformatics* 19, 12: (2003) 1484–1491.
- Statnikov, Alexander, Lily Wang, and Constantin F. Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." *BMC Bioinformatics* 9, 1: (2008) 319.

- Staudt, Louis M., and Sandeep Dave. "The biology of human lymphoid malignancies revealed by gene expression profiling." *Advances in Immunology* 87, 05: (2005) 163–208.
- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A. Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* 102, 43: (2005) 15,545–15,550.
- Suzuki, Nao, Norihito Yoshioka, Atsushi Uekawa, Noriomi Matsumura, Akiko Tozawa, Jyunki Koike, Ikuo Konishi, Kazushige Kiguchi, and Bunpei Ishizuka. "Transcription factor POU6F1 is important for proliferation of clear cell adenocarcinoma of the ovary and is a potential new molecular target." *International journal of gynecological cancer official journal of the International Gynecological Cancer Society* 20, 2: (2010) 212–219.
- Tedoldi, S., J. C. Paterson, J. Cordell, S-Y. Tan, M. Jones, S. Manek, A. P. Dei Tos, H. Robertson, N. Masir, Y. Natkunam, S. A. Pileri, F. Facchetti, M-L. Hansmann, D. Y. Mason, and T. Marafioti. "Jaw1/LRMP, a germinal centre-associated marker for the immunohistological study of B-cell lymphomas." *The Journal of pathology* 209, 4: (2006) 454–463.
- Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B Methodological* 58, 1: (1996) 267–288.
- Vapnik, V. N. *The Nature of Statistical Learning Theory, 2nd ed.*, volume 8 of *Statistics for Engineering and Information Science*. Springer, 2000.
- Vapnik, V. N., and A. Y. Chervonenkis. "Ordered risk minimization." *Automation and Remote Control* 35: (1974) 1226–1235, 1403–1412.
- Walter, Roland B., Ted A. Gooley, Vincent H. J. Van Der Velden, Michael R. Loken, Jacques J. M. Van Dongen, David A. Flowers, Irwin D. Bernstein, and Frederick R. Appelbaum. "CD33 expression and P-glycoprotein-mediated drug efflux inversely correlate and predict clinical outcome in patients with acute myeloid leukemia treated with gemtuzumab ozogamicin monotherapy." *Blood* 109, 10: (2007) 4168–4170.
- Wang, Yu, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F. X. Mayer, and Hans W. Mewes. "Gene selection from microarray data for cancer classification—a machine learning approach." *Computational Biology and Chemistry* 29, 1: (2005) 37–46.
- Wu, Tong Tong, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics* 25, 6: (2009) 714–721.

- 
- Yu, Zhiwen, Hau-San Wong, and Hongqiang Wang. “Graph-based consensus clustering for class discovery from gene expression data.” *Bioinformatics* 23, 21: (2007) 2888–96.
- Zipf, G. K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. International Library of Psychology. Routledge, 1999.
- Zou, Hui, and Trevor Hastie. “Regularization and variable selection via the elastic net.” *Prostate, The* 301–320.





# 6

## Practical Considerations, Conclusions and Future Directions

In this chapter we detail some of the more important practical considerations a statistician must take into account when undertaking dimension reduction. Most of these issues can be solved by simple software engineering; they do not involve statistical learning theory. We will also present results that hint at future directions in the field, and summarise our findings.

### 6.1 Class Imbalance

In practice, when analysing two-class splits, there will usually be an unequal number of samples from each class. When comparing diseased tissue samples to healthy ones,

it is likely that there will be more of the former, since the clinician will not order a sample to be taken unless there is high suspicion of disease.

In statistical circles, the class imbalance problem is not trivial, and has generated a significant amount of discussion (Guo et al., 2008; Japkowicz and Stephen, 2002; Zheng et al., 2004). When the accuracy of the classification model is evaluated with unbalanced classes, it will be optimistically biased. For instance, if one class is more common than the other, even a simple classifier always choosing the majority class will clearly show more than 50% accuracy. It is our experience that a classifier built on poorly selected features will choose the more prolific class. One way to remedy this is to plot the performance of the classifier on a ROC and measure both the AUC and the distance from the plotted decision boundary to the point  $(0, 1)$ , representing perfect classification, instead of reporting the combined accuracy.

When performing feature selection, however, the influence of class imbalance cannot be so easily corrected. One solution is to oversample the minority class when bootstrapping so its frequency matches the majority class. This is an unsatisfactory method for two reasons: more frequent sampling of the minority class increases the risk of overfitting (Chawla et al., 2002), and it also increases the computational time needed to run the feature selection routine. Undersampling the majority class is also unsatisfactory since it means discarding a large amount of potentially useful data, which, as we have argued in Chapter 3, is antithetical to feature selection.

A third option for alleviation of this problem is cost modification. Unlike oversampling or undersampling, which takes place before a feature selection routine is run, the adjustment calculations are embedded within the routine. The calculations of the risk of the model are performed in such a way that each penalisation (such as the out-of-bag error in Random Forests, or residuals in shrinkage methods) is given a weight, depending on its mislabelled class. For example, a cost matrix may be set by the programmer such that each weight is proportional to the inverse of the class ratios (Guo et al., 2008). Cost modification does not discard data and, at the time of publication, we could not find any literature that raised concerns about overfitting with this technique.

We ought to finish this section with a remark about support vector machines. SVMs

do not need any cost modification routine built into them since the calculation of the objective function is done independently of any class ratios. The risk on an SVM model is calculated via the summation of sample overlaps within the soft margin only. This selectivity renders SVM models robust to the class imbalance problem, which has been shown empirically (Japkowicz and Stephen, 2002).

## 6.2 Computational Costs

All calculations were performed on a cluster server containing six Intel® Xeon® E5335 2 gigahertz CPUs, running GNU/Linux x86\_64. The computational time needed to complete one bootstrap selection varied greatly depending on the algorithm used. Generally speaking, methods with an incorporated stochastic element took much longer than deterministic methods. Building an SVM with 100 samples and only a handful of features takes only a fraction of a second, but a run of the stepwise regression feature selection with simulated annealing, complete with bootstrapping, took approximately six weeks for a complete set of results to be evaluated. This is because an SVM was created for each of the number of features  $P$ , which was over 18000 in the St. Vincent's datasets, every time a feature was either added or evaluated for replacement. So, for each of the 150 bootstraps, there were, at the very least (assuming no replacements), the 20th triangular number of sets of SVMs size  $P$  evaluated, which resulted in approximately  $10^8$  SVMs being created. Similarly, the Cross Entropy method was very slow to deliver satisfactory results. Depending on the dataset used, a single iteration took between 30 and 120 minutes to complete. The time needed to reach convergence was foremost dependent on, and inversely proportional to, the smoothing parameter  $\alpha$ . For example, using 81 samples and 18661 features, and a value of  $\alpha = 0.02$ , convergence (all values in  $V \geq 0.997$ ) took upwards of 300 iterations to occur, whereas using  $\alpha = 0.1$ , convergence usually occurred between 60 and 80 iterations.

Computational times for Random Forest, GPS and LAR were much better. Using 58 samples and 18661 features, where the subset of features taken for one tree was set at  $\sqrt{P}$ , one bootstrap run takes approximately 20 to 30 seconds. GPS and LAR were

even faster: as per the data dimensions used for Random Forests, one bootstrap took approximately 10 to 15 seconds, even when, in the case of GPS, 7 different penalty functions are considered.

When computational power or time is at a premium, we recommend Random Forest and shrinkage methods over SVM for feature selection. Other SVM feature selection methods, such as Recursive Feature Elimination (Guyon et al., 2002), have proven to be expensive, and methods to improve on them have been suggested (Ding and Wilkins, 2006; Zhou and Tuck, 2007). Computational speed was not our primary motivation to include an SVM feature selection, rather we aimed to find deep local optima that a faster algorithm may not have been able to find. In Chapter 5, however, the pooled master lists for our SVM algorithm were very similar to the GPS lists on almost all datasets studied, more so than Random Forest was to either, indicating that GPS may be the current choice algorithm for both fast and meaningful feature selection. In any case, any feature selection algorithm that allows feature rankings may be used on the AECS template at the discretion of the statistician.

## 6.3 What is the Optimal Number of Features a Dataset Should Be Reduced To?

Tuning parameters for optimal output in data mining is often an art, rather than a science. Each dataset has its own idiosyncrasies, and the statistician must often obtain a ‘feel’ for the data, by running a few experimental analyses. Often, he or she will stumble across the most significant aspect of the data by trial-and-error from intuition and natural curiosity, rather than systematically running an exhaustive suite of tests. An *ad hoc* approach may be used out of necessity when there is not enough time or computational power.

The regularisation parameter  $\lambda$  in a SVM and the feature cardinality (or dimensionality) of a model are both parameters whose global optimal value is difficult to

ascertain. There is evidence that very small numbers of features are able to form accurate classifiers (Grate, 2005), and hence it is worthwhile exploring how sparse we can make our final feature subset. The SVM framework makes an effort to account for training error by including soft margins, but it becomes a powerful tool for characterising datasets when it can (linearly or otherwise) separate labelled data. In this section, we show that it needs a certain lower bound on the number of features to do this effectively and recommend selecting a feature cardinality that is at or near this lower bound.

We used a fairly coarse method of  $\lambda$  optimisation in Chapters 4 and 5—a grid search, which we deemed satisfactory for our objectives—but we also, by way of exploratory analysis, performed a finer tuning of  $\lambda$  using software that is able to calculate the entire SVM regularisation path (Hastie et al., 2004). We performed a feature selection using a stepwise regression with simulated annealing with the SVM penalised margin width as the objective function to be minimised, identical to the algorithm we used in Chapter 5, and measured the optimal value of  $\lambda$  (that is, the one that conferred the minimum criterion value) at the end of each optimisation on a feature cardinality of  $q$ , where  $1 \geq q \geq 20$ . For comparison, we used the most regularised value of  $\lambda$  for which the training error was a minimum on the same path, and discovered a curious result when we compared the two. All seven of the in-house data splits used in Chapter 5 (the results from three of them are in Figure 6.1) showed the optimal value of  $\lambda$  preferring a high-variance model until it reaches the region of  $7 \leq q \leq 12$ , where optimal  $\lambda$  increases sharply, often (but not always) overtaking the most regularised path. Optimal  $\lambda$  then plateaus from approximately  $q = 15$  upwards; its value closely following the most regularised path. This trend, apparent over a number of different datasets, suggests a complex phenomenon.

At first glance, the separation power gained by adding further features to the SVM above  $q = 15$  is minimal, hence the model may be overfitting past this point. This hypothesis is further strengthened by the fact that the objective function (that is, the penalised margin width), in each case, decreased precipitously in the region corresponding to the jump in optimal  $\lambda$  when plotted against  $q$ , for each of the 7 datasets, and

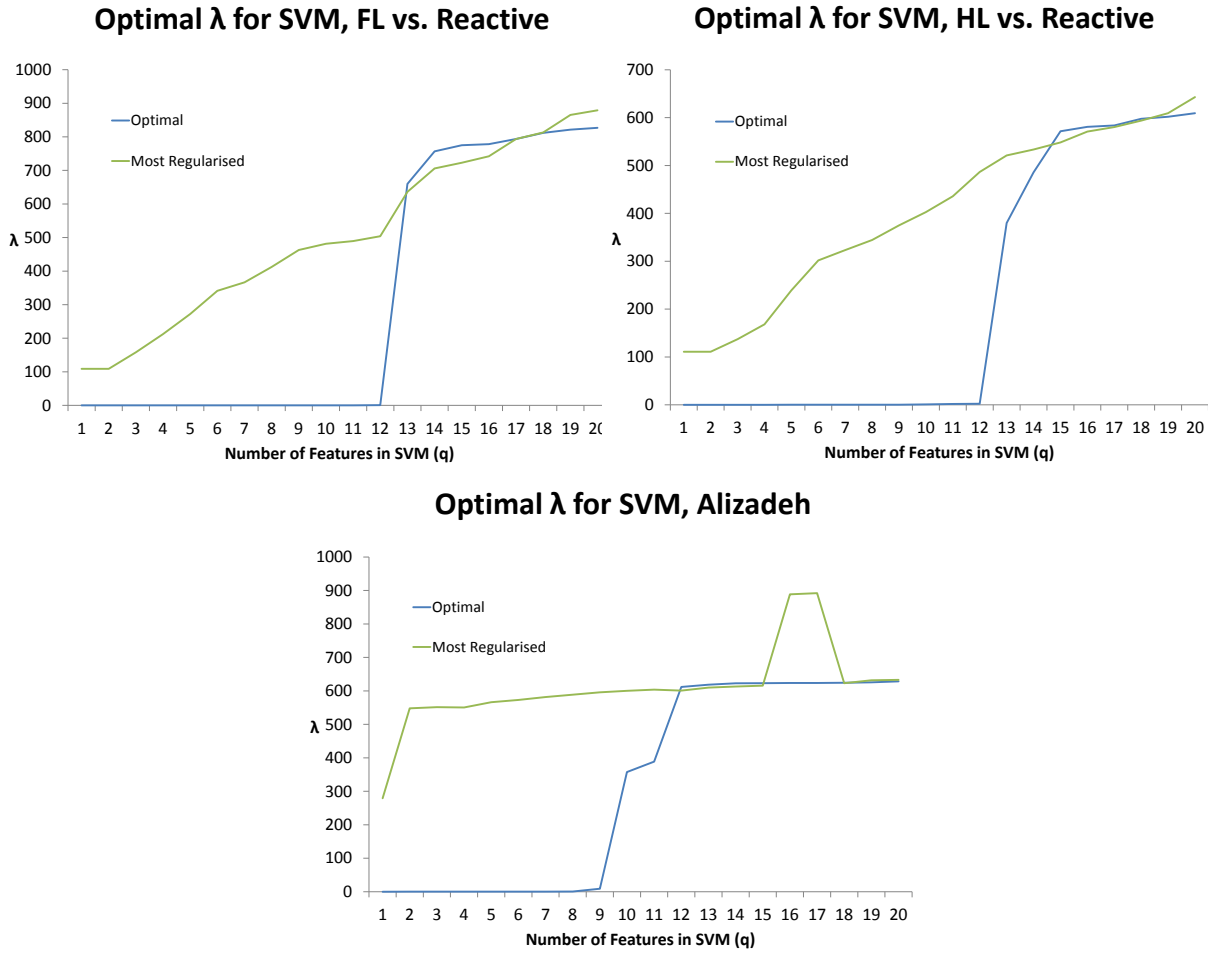


FIGURE 6.1: Plots of optimal value of  $\lambda$  against the most regularised value of  $\lambda$  conferring minimum training error along the same regularisation path, using SVM stepwise regression with simulated annealing, on three different datasets.

then also plateaued past  $q = 15$ . Since we have already optimised  $\lambda$ , then it is safe to assume that any further overfitting trend observed is due to some other parameterisation, such as the value of  $q$ . If we assume the differences in penalised margin width for differing values of  $q$  can act as a proxy for information gain as features are added to the model, then we may speculate that any features added to the model after  $q \approx 15$  may be superfluous in terms of the informative value they bring. Also supporting this hypothesis is the fact that the optimum models with  $q \leq 7$  display a high degree of variance. This suggests that the margin width of the SVMs in this region is forced to be kept very small, and optimising  $\lambda$  in this region essentially means minimising the residual. However, when we increase  $q$  past about 8, the SVM begins to function

as intended; the margin widens and  $\lambda$  increases sharply. It is therefore possible that the optimum number of features to include in the model is somewhere in the region of the  $\lambda$  ‘jump’. Whether this is a quirk of the SVM heuristic itself, or an underlying property of high-dimensional data remains to be seen, but it certainly warrants further investigation.

Sparsity is a highly desirable characteristic of a putatively informative model, and if the aforementioned trend is repeatable across a variety of synthetic and real datasets, and indeed other data mining algorithms, then it may serve as a valuable rule of thumb for the statistician tasked with producing a small but informative list of genes.

Also noteworthy is the existence of the regions in which the optimal  $\lambda$  is represented by a model along the path that has a training error *greater* than the minimum, reinforcing the point made in Chapter 2 that the minimisation of training error is a poor method of evaluating model robustness.

## 6.4 Differing Characteristics of Transcriptomic Datasets

The results from Chapter 5 can be further contextualised in a number of ways. Figure 6.2 shows how confident we can be about the results garnered from GPS. GPS performs an internal cross-validation procedure on the data, and calculates the *explainable risk* (or explained risk, ER), which is the proportional decrease in risk obtained by using a given subset of features, compared to the risk calculated on a null model. The risk of the latter is that on the expected distribution of the response variable if we randomly sample from the population (Korn and Simon, 1991). ER can be measured by mean squared error, or even just training error, but in this case it is the loss on the training data calculated by the shrinkage estimator. We can use the explained risk value as an indicator of how much confidence we can have in the biomarkers we selected being truly biologically influential.

When we average the ER of each optimised model across 150 bootstraps of the 10 different datasets in Chapter 5, we see that the degree of explication achieved by each dataset differs. This is likely explained by the biological homogeneity of each

comparison performed (see Figure 5.1). All three publicly available datasets have an ER of 95% or above, and the FL vs. DLBCL (the same biological diagnoses that Shipp used) from our in-house dataset shows an ER of 92%. The samples used by Golub are of a comparable level of homogeneity, since they compare different types of NHL, and the samples of Alizadeh are even more homogeneous, since they compare DLBCL subtypes. These splits are likely to have less biological noise than those that compare more heterogeneous samples, such as HL vs. NHL and the splits with non-cancerous samples.

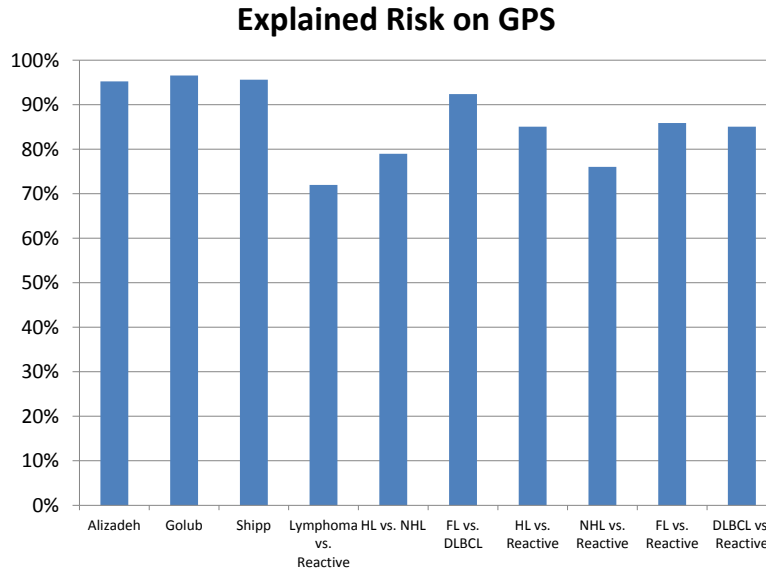


FIGURE 6.2: Explained risk (ER) for all datasets used in Chapter 5. Datasets with homogeneous samples tend to have a higher ER.

GPS chooses the best mode of residual shrinkage from a list of penalty functions that includes the LASSO ( $\gamma = 1$ ), ridge regression ( $\gamma = 2$ ) and all-subsets regression ( $\gamma = 0$ ), as well as other intermediate values of  $\gamma$ . It does this by creating shrinkage paths for each of several values of  $\gamma$  and selects the corresponding model that confers the maximum explained risk. As mentioned in Section 2.5.2, ridge regression tends to preserve groups of highly correlated features by shrinking their coefficients towards one another, but LASSO, with its tendency to produce sparser models, is more likely to shrink some of them to zero in a fashion that may be construed as arbitrary. Given the natural tension between wanting a sparse model and one that preserves enough



information to make biological inferences, we monitored the proportion of bootstraps chosen for which each of the available values of  $\gamma$  was optimal (Figure 6.3). Interestingly, the datasets with the fewest common features between the three ranked lists from each AECS algorithm (Shipp and HL vs. Reactive) had the highest proportion of bootstraps for which ridge regression was the optimal shrinkage method. Conversely, a dataset with a high number of common features, FL vs. Reactive, had the highest proportion of bootstraps for which the maximum sparsity penalty option ( $\gamma = 0$ ) was optimal. (The Alizadeh dataset, however, did not follow this trend.) This suggests that GPS attempts to compensate for the degree of correlation found in a dataset. Where there is low correlation, GPS feature selection will promote a penalty function that allows the coefficients of somewhat correlated features to be shrunk towards each other more easily, in order to create a ‘critical mass’ of predictor variables. On the other hand, when features are highly correlated, GPS will promote penalties which produce a sparser feature subset to guard against large number of highly correlated, but possibly artifactual predictors overwhelming the model. This powerful yet nuanced way of analysing microarray data allows the statistician to explore its properties more deeply.

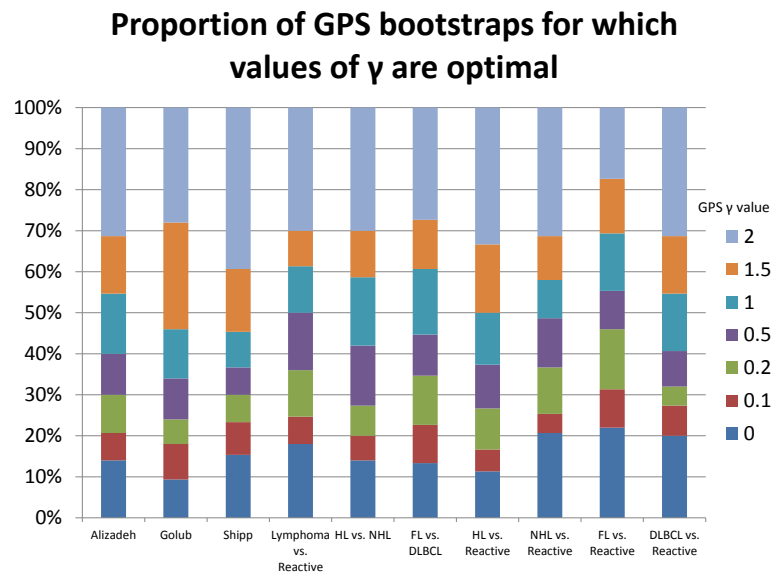


FIGURE 6.3: Proportions of bootstraps for which  $\gamma$  is optimal for all data splits used in Chapter 5. 150 bootstraps were used.

## 6.5 Future Directions and Conclusion

New microarray technologies are changing the nature of transcriptomics, and hence there are large ramifications for the statisticians who are involved in this field. The development most likely to influence the way gene expression profiles are analysed is the emergence of Whole Transcriptome Shotgun Sequencing (Wang et al., 2009; Morin et al., 2008), or more informally RNASeq, a next-generation high-throughput sequencing technique. RNASeq provides a deeper reading of transcriptomes than the tools used to read the data described in this thesis. This is because RNASeq can detect and measure the presence of many transcript isoforms from a single gene. Such isoforms may include allelic polymorphisms, post-transcriptional modifications and gene fusion events (Maher et al., 2009). The consequence for statisticians is a multifold increase in the number of features in a standard RNASeq gene expression profile from the more traditional two-colour and Affymetrix arrays, resulting in more noise and possible red herring features. The likely solution requires a redoubled effort to employ algorithms that value sparsity while maintaining adequate feature assessment coverage.

While an increase in the number of features in a standard expression profiling dataset is inevitable, sadly the same cannot be said for its sample size. One of the chronic problems in the field of gene expression profiling is the decentralisation of the work undertaken. Differing laboratory procedures, conditions and protocols lead to an inevitable non-standardisation of the worldwide corpus of samples obtained (Irizarry et al., 2005). Nevertheless, attempts have been made to overcome these problems, with varying success (Campain and Yang, 2010). What is considered a state-of-the-art sequencing technology one year may be obsolete the next, so it is difficult for a statistician to ‘hit a moving target’: establish a set of statistical protocols without them being superseded quickly. Nevertheless, (Fan et al., 2010) has demonstrated that features chosen from one platform can be exported to another platform with common transcripts to build a classifier, and that classifiers built on one platform can correctly predict samples prepared on another, with a 80 – 90% cross-platform prediction consistency in both cases, regardless of the learning method and common

features involved. Such consistency is helpful in building a collegial approach to solving large scale biomarker identification challenges, but if progress in the field is to be truly streamlined, the creation of a global umbrella network of transcriptomic analysts is needed. Such a network would include personnel responsible for microarray quality control and standardisation, as well as statisticians who ensure that the published work of the group maintains a high standard. Moves towards this ambitious goal have been initiated in the last decade (Brazma et al., 2001; Shi et al., 2010).

Central to the objectives of an integrated group of transcriptome analysts and affiliates is a set of statistical protocols and yardsticks which contributed work can both be informed by and measured against. Assuming the integrity of the microarray generation, normalisation and preprocessing steps, a rigorous set of data mining quality controls should be implemented by the bioinformatics team. Although by no means exhaustive, this thesis details what such a protocol list may contain. In summary, combining existing wisdom with that found in this thesis:

- The aim of most feature selection tasks is to find a small handful of discriminative features that define the differentially expressed classes. This means that any technique employed should be geared towards sparse results.
- When choosing an appropriate feature selection algorithm, or ensemble of algorithms, the user must take care to ensure each feature is assessed as fairly as the next.
- The procedure must be robust, and have high generalising capability.
- The procedure should have an inbuilt regularising function that is able to be tuned by the user.
- If shrinkage methods are used, a suite of penalty functions should be tested by the user in order to choose the most appropriate one.
- The parameter tuning for the previous two points must be executed in a supervised fashion, keeping test samples apart from training samples.

- When assessing the best model parameters, the risk on each model should be assessed by an appropriate loss function.
- If iterative continuous techniques are used, the smoothing parameter must also be tuned, and appropriate stopping criteria defined.
- Class imbalances must be addressed.
- Results must be stable; this can be achieved by aggregating the results of bootstrap resampling of the feature selection routines.
- When reporting the results to scientists, the direction in which features are regulated relative to the diagnostic class should be included.

While this list is not complete, it is our opinion that an implementation and mindfulness of these recommendations and caveats will yield meaningful results for both statisticians and clinicians alike. Bioinformatics is a discipline that requires a measure of insight and technical nous, in addition to a broad array of required knowledge. Clear-cut solutions to problems like feature selection are not readily obtainable, and progress is often made serendipitously. What this thesis cannot impart is the wit, perseverance and lateral thinking that come as welcome gifts to the talented researcher. Nonetheless, a marriage of these innate abilities and the necessary applied discipline described herein maximises the chances for success in this field.



# Appendix

## A.1 Cross-Entropy Code Example

Below is an example of a piece of R code used in Chapter 3. A separate script was written for each value of  $q$ ,  $2 \leq q \leq 20$ , and results from these were aggregated. This is the script where  $q = 10$

```
library (MASS)
```

```
gpdata <- as.matrix(read.delim( ‘numeric’ , header=FALSE)) #  
  numeric with “numeric” data , not factors  
preamblecols <- 9  
preamblerows <- 1  
genenames <- gpdata[,1]
```

---

```

groupA <- gpdata[,gpdata[2,]=='Reactive']
groupB <- gpdata[,gpdata[2,]=='Cancer']
gpdata <- cbind(groupA, groupB)
qAs <- ncol(groupA)
qBs <- ncol(groupB)
tissuenames <- gpdata[1,-1]
gpdata <- gpdata[-c(1:preamblecols),]
genenames <- genenames[-c(1:preamblecols)]
class <- factor(c(rep("Reactive", qAs), rep("Cancer", qBs)))

allgenes <- nrow(gpdata)
qgenes <- 10
alpha <- 0.1
weights <- vector(length = allgenes)
top <- as.integer(10)
popsize <- 100
alltally <- array(0, dim=c(1, allgenes))
toptally <- array(0, dim=c(1, allgenes))
iterations <- 10
results <- matrix(nrow=0, ncol=2*qgenes + 4)

transnum <- function(listinput){
  charmatrix <- t(listinput)
  numatrix <- array(as.numeric(charmatrix), c(nrow(
    charmatrix),ncol(charmatrix)))
  return(numatrix)
}

for(u in 1:ncol(gpdata)){
  vhat <- array(qgenes/allgenes, dim=c(1, allgenes))

```

---

```

leftout <- gpdata[, -u]
genesets <- array(0, dim=c(popsizes, allgenes+1))
for (x in 1:200){

  for (w in 1:popsizes){ #N = 100
    rand <- runif(allgenes)
    while(sum(as.numeric(rand < vhat))!=qgenes){rand
      <- runif(allgenes)}
    fitness <- lda(transnum(leftout[rand < vhat,]),
      class[-u])$svd #Calculate F-value
    genesets[w,] <- cbind(fitness, t(as.numeric(rand
      < vhat)))
  }

  ordered = order(genesets[,1], decreasing=TRUE)
  best = genesets[ordered[1:top], -1]
  alltally <- alltally + colSums(genesets[, -1])
  toptally <- toptally + colSums(best)
  vtilda <- colMeans(best)
  vhat <- alpha*vtilda + (1-alpha)*vhat #smoothing

}

save(alltally, file="F10alltally.results")
save(toptally, file="F10toptally.results")
optset <- sort(vhat, decreasing=TRUE, index.return=TRUE)$ix[1:
  qgenes]

classifier <- lda(transnum(leftout[optset,]), class[-u])

```

```

aprob <- predict(classifier , as.numeric(gpdata[optset,u]))$
  posterior[, "Reactive"]
bprob <- predict(classifier , as.numeric(gpdata[optset,u]))$
  posterior[, "Cancer"]
tissuresult <- c(tissuenames[u], levels(class)[class[u]],
  aprob, bprob, t(optset), t(sort(vhat, decreasing=TRUE))[1:
  qgenes])
results <- rbind(results, tissuresult)
save(results, file="F10output.results")
}

save(results, file="F10output.results")
save(alltally, file="F10alltally.results")
save(toptally, file="F10toptally.results")

```

## A.2 AECS Code Example

This code was an attempt to combine all sections of AECS into one script, complete with PubMed access for datasets appended with GenBank Accession Numbers. However, due to the SVMsR section requiring weeks of computation, this script was broken up into parts which were run separately, and then combined, to produce the results found in Chapter 5. Requires R packages *kernelab* and *randomForest* as well as Jerry Friedman's GPS package, which can be found at <http://www-stat.stanford.edu/~jhf/r-gps/>.

```

#arglist: inputfile, algotype, simann, regparam, numgenes,
  penalty, resamp, database

```



```

# Run in batch mode, for example <command> R CMD BATCH --no-
  save --no-restore --args inputfile="garvan.csv" algotype
  ="RF" simann=TRUE regparam=0.1 numgenes=20 resamp=20
  database='path' Miner.R test.out & </command>
#Don't forget to put the inputfile and algotype arguments in
  inverted commas, they will read as nonexistent objects
  instead of strings otherwise
##args is now a list of character vectors
## First check to see if arguments are passed.
## Then cycle through each element of the list and evaluate
  the expressions.

args=(commandArgs(TRUE))
if (length(args)==0){
  print("No arguments supplied.")
}else{
  for(i in 1:length(args)){
    eval(parse(text=args[[i]]))
  }
}
library(kernlab)
library(randomForest)

# Data entry file must have samples in rows and features in
  columns. Data begins on [2,2]
# with row 1 a binary vector of class names, and column 1 a
  vector of Genbank accession numbers. Cell (1,1) should be
  blank
# There should be no missing values, they can be imputed at
  http://gepas.bioinfo.cipf.es/cgi-bin/preprocess

```

```

# Preprocessing
gpdata <- as.matrix(read.csv(inputfile , header=FALSE)) #
  numeric with “numeric” data, not factors
genenames <- gpdata[-1,1]
groupA <- gpdata[,gpdata[1,]==unique(gpdata[1,-1])[1]]
groupB <- gpdata[,gpdata[1,]==unique(gpdata[1,-1])[2]]
gpdata <- cbind(groupA, groupB)
qAs <- ncol(groupA)
qBs <- ncol(groupB)

class <- factor(c(rep(unique(gpdata[1,-1])[1], qAs), rep(
  unique(gpdata[1,-1])[2], qBs)))
gpdata <- gpdata[-1,]
gpdata2 <- matrix(as.numeric(gpdata), nrow=nrow(gpdata), ncol=
  ncol(gpdata))
b <- modt.stat(t(gpdata2),class)

allgenes <- nrow(gpdata)

transnum <- function(listinput){
  charmatrix <- t(listinput)
  numatrix <- array(as.numeric(charmatrix), c(nrow(
    charmatrix),ncol(charmatrix)))
  return(numatrix)
}

```

---

```

noreplace <- function(samplenum, genesadded, minoflist){
count <- 1
while (count >= 0){

if(sum(samplenum == genesadded) == 0){count <- -1}else{count
  <- count + 1; samplenum = minoflist$ix[count]}
}
return(samplenum)
}

most.common <- function(x) {
  count <- sapply(unique(x), function(i) sum(x==i, na.rm=TRUE)
  )
  return(unique(x)[which(count==max(count))])
}

getcounts <- function(x, resultlist) {
  counts <- c()
  for (y in 1:length(x)){
    counts <- c(counts, sum(x[y] == resultlist))
  }
  return(counts)
}

if(algotype == "SVM" && as.logical(simann)){#SVMSR with
  simulated annealing and resamp

genesadded <- array(0, c(as.numeric(resamp),as.numeric(
  numgenes)))

```

```

objfunfs <- array(0, c(as.numeric(resamp), as.numeric(numgenes))
)
testresponse <- matrix(nrow=ncol(gpdata), ncol=as.numeric(
  numgenes))
testprobs <- array(0, c(ncol(gpdata), 2*as.numeric(numgenes)))
asweight <- qBs/(qAs + qBs)
bsweight <- -(qAs/(qAs + qBs))
classweights <- c(qBs, qAs)
names(classweights) <- c(levels(class)[1], levels(class)[2])
weights <- c(rep(1, qAs), rep(-1, qBs))

count <- 0
resultlist <- matrix(nrow=as.numeric(resamp), ncol=as.numeric(
  numgenes))
resultobjfunfs <- matrix(nrow=as.numeric(resamp), ncol=as.
  numeric(numgenes))
geneweights <- array(0, c(allgenes, 1))

for (x in 1:as.numeric(resamp)){

  if (as.numeric(resamp) != 1){sampchoose <- c(sample(1:qAs,
    ceiling(qAs/2)), sample((qAs+1):ncol(gpdata), floor(qBs/2))
  )} else {sampchoose=1:ncol(gpdata)}

  sampdata <- gpdata[, sampchoose]
  geneset <- array(0, c(ncol(sampdata), 0))

  for (qgenes in 1:as.numeric(numgenes)){

    currentobjfunfs <- array(0, c(allgenes, 1))

```

```

for (w in 1:allgenes){
  currentset <- cbind(geneset, as.numeric(
    sampdata[w,]))
  classifier <- ksvm(currentset, class[
    sampchoose], kernel="vanilladot", class.
    weights=classweights, C=as.numeric(regparam
  ))
  currentobjfuns[w] <- attributes(classifier)$
    obj
}

```

```

minoflist <- sort(currentobjfuns, decreasing=TRUE,
  index.return=TRUE)
genesadded[x, qgenes] <- noreplace(minoflist$ix[1],
  genesadded[x,], minoflist)
objfuns[x, qgenes] <- minoflist$x[which(minoflist$ix
  == genesadded[x, qgenes])]
geneset <- cbind(geneset, as.numeric(sampdata[
  genesadded[x, qgenes],]))
oldgeneset <- array(0, c(1,as.numeric(numgenes)))

```

```

while(sum((genesadded[x,]-oldgeneset)!=0)!=0) { #test
  for reaching local minimum
  if(qgenes==1){break}
  oldgeneset <- genesadded[x,]

```

---

```

for (repl in 1:qgenes){

  currentobjfuns <- array(0, c(allgenes ,
    1))

  for (y in 1:allgenes){
    #replace(a, ((index-1)*nrow(a)
      + 1):(index*nrow(a)), b)
    currentset <- replace(geneset ,
      ((repl-1)*nrow(geneset) +
        1):(repl*nrow(geneset)), as
      .numeric(sampdata[y,]))
    classifier <- ksvm(currentset ,
      class[sampchoose], kernel
      =“vanilladot”, class.
      weights=classweights , C=as.
      numeric(regparam))
    currentobjfuns[y] <-
      attributes(classifier)$obj
  }

  minoflist <- sort(currentobjfuns ,
    decreasing=TRUE, index.return=TRUE)
  genesadded[x, repl] <- noreplace(
    minoflist$ix[1], genesadded[x,-repl
    ], minoflist)

```

```

        objfun[x, qgenes] <- minoflist$x[
            which(minoflist$ix == genesadded[x,
                repl])]
        geneset <- replace(geneset, ((repl-1)*
            nrow(geneset) + 1):(repl*nrow(
                geneset)), as.numeric(sampdata[
                    genesadded[x, repl],]))

    }

    write.csv(cbind(genesadded[x,], objfun[x,]), file='
        svmoutput.csv')
}

}

resultlist[x,] <- genesadded[x,]
resultobjfun[x,] <- objfun[x,]
proporder <- rbind(resultlist[x,], resultobjfun[x,])

for (z in 1:numgenes){
weightedorder <- proporder[1,][sort(proporder[2,], decreasing=
    FALSE, index.return=TRUE)$ix)]
for (y in 1:numgenes){
geneweights[as.numeric(weightedorder[y])] <- geneweights[as.
    numeric(weightedorder[y])] + 1/y
}
}

```

```

}
finallist <- sort(geneweights, index.return=TRUE, decreasing=
  TRUE)$ix[1:numgenes]

}

if(algotype == "SVM" && !(as.logical(simann))) {#normal SVMsR
  resamp

  genesadded <- array(0, c(as.numeric(resamp), as.numeric(
    numgenes)))
  objfuns <- array(0, c(as.numeric(resamp), as.numeric(numgenes))
    )
  testresponse <- matrix(nrow=ncol(gpdata), ncol=as.numeric(
    numgenes))
  testprobs <- array(0, c(ncol(gpdata), 2*as.numeric(numgenes)))
  asweight <- qBs/(qAs + qBs)
  bsweight <- -(qAs/(qAs + qBs))
  classweights <- c(qBs, qAs)
  names(classweights) <- c(levels(class)[1], levels(class)[2])
  weights <- c(rep(1, qAs), rep(-1, qBs))

  count <- 0
  resultlist <- matrix(nrow=as.numeric(resamp), ncol=as.numeric(
    numgenes))
  resultobjfuns <- matrix(nrow=as.numeric(resamp), ncol=as.
    numeric(numgenes))
  geneweights <- array(0, c(allgenes, 1))

  for (x in 1:as.numeric(resamp)){

```



---

```

genesadded <- array(0, c(as.numeric(resamp), as.numeric(
  numgenes)))

if (as.numeric(resamp) != 1) { sampchoose <- c(sample(1:qAs,
  ceiling(qAs/2)), sample((qAs+1):ncol(gpdata), floor(qBs/2))
)} else { sampchoose = 1:ncol(gpdata) }

sampdata <- gpdata[, sampchoose]
geneset <- array(0, c(ncol(sampdata), 0))

for (qgenes in 1:as.numeric(numgenes)) {

currentobjfuns <- array(0, c(allgenes, 1))

  for (w in 1:allgenes) {
    currentset <- cbind(geneset, as.numeric(
      sampdata[w,]))
    classifier <- ksvm(currentset, class[
      sampchoose], kernel="vanilladot", class.
      weights=classweights[sampchoose], C=as.
      numeric(regparam))
    currentobjfuns[w] <- attributes(classifier)$
      obj

  }
}

```

---

```

    minoflist <- sort(currentobjfuns , decreasing=TRUE,
      index.return=TRUE)
    genesadded[x, qgenes] <- noreplace(minoflist$ix[1] ,
      genesadded[x,] , minoflist)
    objfuns[x, qgenes] <- minoflist$x[which(minoflist$ix
      == genesadded[x, qgenes])]
    geneset <- cbind(geneset , as.numeric(sampdata[
      genesadded[x, qgenes] ,]))
write.csv(cbind(genesadded[x,] , objfuns[x,]) , file='svmoutput.
  csv ')
}
resultlist[x,] <- genesadded[x,]
resultobjfuns[x,] <- objfuns[x,]
for (y in 1:numgenes){
  geneweights[as.numeric(resultlist[x,y])] <- geneweights[as.
    numeric(resultlist[x,y])] + 1/y
}

}

finallist <- sort(geneweights , index.return=TRUE, decreasing=
  TRUE)$ix[1:numgenes]
print(t(genenames[finallist]))

}

if(algotype == "RF"){# Random Forest
gpdata <- t(gpdata)
genesadded <- array(0 , c(as.numeric(resamp) , as.numeric(
  numgenes)))

```

```

count <- 0
resultlist <- matrix(nrow=as.numeric(resamp), ncol=as.numeric(
  numgenes))
importancelist <- matrix(nrow=as.numeric(resamp), ncol=as.
  numeric(numgenes))
geneweights <- array(0, c(allgenes, 1))
classweights <- c(qBs, qAs)
names(classweights) <- c(levels(class)[1], levels(class)[2])

for (x in 1:as.numeric(resamp)){
genesadded <- array(0, c(as.numeric(resamp), as.numeric(
  numgenes)))

if (as.numeric(resamp) != 1){sampchoose <- c(sample(1:qAs,
  ceiling(qAs/2)), sample((qAs+1):nrow(gpdata), floor(qBs/2))
)}else{sampchoose=1:nrow(gpdata)}

sampdata <- gpdata[sampchoose,]
geneset <- array(0, c(ncol(sampdata), 0))

forest <- randomForest(sampdata, class[sampchoose], classwt=
  classweights)
resultlist[x,] <- sort(forest$importance, decreasing=TRUE,
  index.return=TRUE)$ix[1:numgenes]
geneweights <- geneweights + forest$importance
print(sum(forest$importance))
}

```

---

```

finallist <- sort(geneweights, index.return=TRUE, decreasing=
  TRUE)$ix[1:numgenes]

}

if(algotype == "GPS"){#GPS resamp
platform = "PLATFORM"
gpshome = getwd()
source("GPS.r")
gpdata <- t(gpdata)
genesadded <- array(0, c(as.numeric(resamp), as.numeric(
  numgenes)))

count <- 0
resultlist <- matrix(nrow=as.numeric(resamp), ncol=as.numeric(
  numgenes))
weights <- c(rep(1, qAs), rep(-1, qBs))

penalties <- array(0, c(1, numgenes*resamp))
penaltyindex <-0

for (x in 1:as.numeric(resamp)){

if(as.numeric(resamp)!=1){sampchoose <- c(sample(1:qAs,
  ceiling(qAs/2)), sample((qAs+1):nrow(gpdata), floor(qBs/2))
)}else{sampchoose=1:nrow(gpdata)}
sampdata <- gpdata[sampchoose,]

```

---

```

sampdata <- array(as.numeric(sampdata), c(nrow(sampdata), ncol
      (sampdata)))
geneset <- array(0, c(ncol(gpdata), 0))
model <- gpsbridge(x=sampdata, y=weights[sampchoose], pens=c
      (0,0.1,0.2,0.5,1.0,1.5,2.0))
soln <- gpssoln(model, vars=1:as.numeric(numgenes), ord='entry
      ')

resultlist[x,] <- soln$order
}
finallist <- array(0, c(1, as.numeric(numgenes)))
for (a in 1:as.numeric(numgenes)){
      athgene <- most.common(resultlist[,a])
      if(length(athgene)!=1){athgene <- athgene[which(
            getcounts(athgene, resultlist) == max(getcounts(
            athgene, resultlist)))]}
      athgene <- athgene[1]
      finallist[a] <- athgene
}

}

unique <- genenames[unique(as.numeric(finallist))]
lengthunique <- length(genenames[unique(as.numeric(finallist))
      ])
print(paste(lengthunique, 'unique genes found.'))
for (b in 1:lengthunique){
      print(paste(b, ': ', genenames[unique(as.numeric(
            finallist))][b]))
}

```

```

}
exit <- FALSE
while(exit==FALSE){

print('How many of these genes would you like to search the
      database for? Note: Searching smaller gene sets, or genes
      individually, will more likely yield a positive result.')
searchquant <- scan('', n=1)

print('Please enter, one by one, the indices of the genes you
      want to search for.')
generequest <- scan('', n=searchquant)
searchlist <- unique[generequest]
#termlist <- paste(searchlist, sep=',', collapse=',')
termlist<-paste(searchlist, sep='', collapse='')
print(paste('http://www.genome.jp/dbget-bin/www_bfind_sub?
            dbkey=refseq&keywords=', termlist, '&max_hit=1000&mode=bget ',
            sep=''))
print('Would you like to do another search? (y/n)')
renew=readline()
if(renew=='n'){exit==TRUE}
}

```

## List of Symbols

- $n$  Total number of samples in a high-dimensional dataset
- $P$  Total number of features in a high-dimensional dataset
- $X$  Data matrix containing gene expression values from all features
- $Y$  Binary response variable. Can be used for classification (categorical) or regression (numerical)
- $B$  Vector of feature coefficients in a linear model
- $B_0$  The intercept of  $B$
- $\epsilon$  General variable describing residuals on fitting  $X$  to  $Y$
- $\beta$  Coefficient vector for all features in  $X$
- $\mu$  Population mean
- s.e.* Standard Error
- $t_k$   $t$ -statistic for a given feature  $k$
- $S_o$  Constant in the denominator of the modified  $t$ -statistic, guarding against  $k$ s with a small fold change being included in the top echelon of a ranked feature list
- $C$  Covariance matrix, used in PCA

$k$  When not subscripted, the number of folds in a  $k$ -fold cross validation. When subscripted, usually a symbol for a generic feature.

$\sigma$  Variance

$Cov$  Covariance

$T$  (superscripted) Transpose (of a matrix or vector)

$\mathbb{R}^q$  The vector space with dimension  $q$ , usually describing a model with  $q$  features

$\|V\|$  Normed vector  $V$  representing the distance between margins in a SVM

$\xi_i$  Slack variables representing distances by which trained data points overlap the margin in a SVM

$\lambda$  Regularisation parameter that regulates a trade-off between the importance a model gives to how biased its function is, and its variance by way of residuals

$Q$  The dimensionally-reduced feature set

$q$  User-defined constant representing an upper bound on the amount of information present in  $Q$ . In the Cross-Entropy Method, the user-defined size (in features) of  $Q$

$\gamma$  Penalty parameter  $0 \leq \gamma \leq 2$ , where the  $\ell^\gamma$  norm is used to penalise residuals in shrinkage regression.

$V$  (not normed) The vector of probabilities for inclusion in  $Q$ .  $\sum V = q$

$\alpha$  User-defined parameter controlling the degree of smoothing as the CE Method updates



---

$\rho$  User-defined parameter controlling the proportion of top feature sets retained after random sampling and evaluation in the Cross-Entropy Method



## References

- Aizerman, M. A., E. M. Braverman, and L. I. Rozonoer. “Theoretical foundations of the potential function method in pattern recognition.” *Automation and Remote Control* 25: (1964) 821–837.
- Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*, volume 54. Garland Press, 2008.
- Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.” *Nature* 403, 6769: (2000) 503–511.
- Ambroise, Christophe, and Geoffrey J. McLachlan. “Selection bias in gene extraction on the basis of microarray gene-expression data.” *Proceedings of the National Academy of Sciences* 99, 10: (2002) 6562–6566.
- Aris, Virginie M., Michael J. Cody, Jeff Cheng, James J. Dermody, Patricia Soteropoulos, Michael Recce, and Peter P. Tolias. “Noise filtering and nonparametric analysis of microarray data underscores discriminating markers of oral, prostate, lung, ovarian and breast cancer.” *BMC Bioinformatics* 5: (2004) 185.
- Armstrong, Scott A., Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. Den Boer, D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub, and Stanley J. Korsmeyer. “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.” *Sophia* 30, January: (2002) 41–47.
- Bengio, Yoshua. “Gradient-Based Optimization of Hyperparameters.” *Neural Comput.* 12, 8: (2000) 1889–1900.
- Berger, C., P. Brousset, C. McQuain, and H. Knecht. “Deletion variants within the NF-kappaB activation domain of the LMP1 oncogene in acquired immunodeficiency syndrome-related large cell lymphomas, in prelymphomas and atypical lymphoproliferations.” *Leukemia lymphoma* 26, 3-4: (1997) 239–250.
- Berrar, Daniel, Ian Bradbury, and Werner Dubitzky. “Avoiding model selection bias in small-sample genomic datasets.” *Bioinformatics* 22, 10: (2006) 1245–1250.

- Boulesteix, Anne-Laure, and Martin Slawski. "Stability and aggregation of ranked gene lists." *Briefings in Bioinformatics* 10, 5: (2009) 556–568.
- Box, G. E. P. "Non-normality and tests on variances." *Biometrika* 40, 3-4: (1953) 318–335.
- Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." *Nature Genetics* 29, 4: (2001) 365–371.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*, volume 19 of *Statistics/Probability Series*. Wadsworth, 1984.
- Breiman, Leo. "Bagging predictors." *Machine Learning* 24, 2: (1996) 123–140.
- . "Random Forests." *Machine Learning* 45, 1: (2001) 5–32.
- Broad Institute of Medicine. "GenePattern." <http://www.broadinstitute.org/cancer/software/genepattern/datasets/>. Accessed November 2009., 2009a.
- . "Lymphoma/Leukemia Molecular Profiling Project." <http://llmpp.nih.gov/lymphoma/>. Accessed November 2009, 2009b.
- Campaign, Anna, and Yee Hwa Yang. "Comparison study of microarray meta-analysis methods." *BMC Bioinformatics* 11, 1: (2010) 408.
- Candes, Emmanuel, and Terence Tao. "The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ ." *The Annals of Statistics* 35, 6: (2007) 2313–2351.
- Caruana, Rich. "An Empirical Comparison of Supervised Learning Algorithms." *Proceedings of the 23rd International Conference on Machine Learning* 161–168.
- Chapelle, O., V. Vapnik, O. Bousquet, and S. Mukherjee. "Choosing Multiple Parameters for Support Vector Machines." *Machine Learning* 46, 1: (2002) 131–159.
- Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research* 16, 1: (2002) 321–357.
- Clear, Andrew J., Abigail M. Lee, Maria Calaminici, Alan G. Ramsay, Kelly J. Morris, Simon Hallam, Gavin Kelly, Finlay MacDougall, T. Andrew Lister, and John G. Gribben. "Increased angiogenic sprouting in poor prognosis FL is associated with elevated numbers of CD163+ macrophages within the immediate sprouting microenvironment." *Blood* 115, 24: (2010) 5053–5056.

- Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. "Greedy Algorithms." In *Introduction to Algorithms*, The MIT Press, 2001a, volume 56, chapter 16, 539–548.
- . *Introduction to Algorithms*, volume 7 of *The MIT Electrical Engineering and computer Science Series*. MIT Press, 2001b.
- Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine Learning* 20, 3: (1995) 273–297.
- Crick, F. "Central dogma of molecular biology." *Nature* 227, 5258: (1970) 561–563.
- Dasarathy, B. V. "Data mining tasks and methods: Classification: nearest-neighbor approaches." *Handbook of data mining and knowledge discovery* .
- Dash, M., and H. Liu. "Feature selection for classification." *Intelligent Data Analysis* 1, 3: (1997) 131–156.
- Detting, M. "Finding predictive gene groups from microarray data." *Journal of Multivariate Analysis* 90, 1: (2004) 106–131.
- Devijver, P. A., and J. Kittler. *Pattern recognition: A statistical approach*. Prentice Hall, 1982.
- Ding, Yuanyuan, and Dawn Wilkins. "Improving the Performance of SVM-RFE to Select Genes in Microarray Data." *BMC Bioinformatics* 7, Suppl 2: (2006) S12.
- Dinndorf, P. A., R. G. Andrews, D. Benjamin, D. Ridgway, L. Wolff, and I. D. Bernstein. "Expression of normal myeloid-associated antigens by acute leukemia cells." *Blood* 67, 4: (1986) 1048–1053.
- Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern Classification*, volume 2 of *Pattern Classification and Scene Analysis: Pattern Classification*. Wiley, 2001.
- Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association* 97, 457: (2002) 77–87.
- Efron, B. "Bootstrap methods: another look at the jackknife." *Annals of Statistics* 7, 1: (1979) 1–26.
- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. "Least Angle Regression." *Annals of Statistics* 32, 2: (2004) 407–499.
- Efroymson, M. A. "Multiple regression analysis." *Mathematical methods for digital computers* 1: (1960) 191–203.

- Fan, X., E. K. Lobenhofer, M. Chen, W. Shi, J. Huang, J. Luo, J. Zhang, S. J. Walker, T.-M. Chu, L. Li, R. Wolfinger, W. Bao, R. S. Paules, P. R. Bushel, J. Li, T. Shi, T. Nikolskaya, Y. Nikolsky, H. Hong, Y. Deng, Y. Cheng, H. Fang, L. Shi, and W. Tong. "Consistency of predictive signature genes and classifiers generated using different microarray platforms." *The pharmacogenomics journal* 10, 4: (2010) 247–257.
- Feldman, Vitaly, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. "Agnostic Learning of Monomials by Halfspaces Is Hard." *Proc 50th IEEE Symp on Foundations of Comp Sci* 385–394.
- Fisher, R. A. "The use of multiple measurements in taxonomic problems." *Annals of Eugenics* 7, 2: (1936) 179–188.
- Friedman, J. H. "Multivariate adaptive regression splines." *Annals of Statistics* 19, 1: (1991) 1–67.
- . "GPS Help." <http://www-stat.stanford.edu/~jhf/r-gps/GPShelp.html>, 2009a.
- . "GPS Software." <http://www-stat.stanford.edu/~jhf/r-gps/>, 2009b.
- Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal Of Statistical Software* 33, 1: (2010) 1–22.
- Friedman, Jerome H. "Fast sparse regression and classification." In *Proceedings of the 23rd International Workshop on Statistical Modelling*, edited by Paul Eilers. Statistical Modelling Society, 2008, 27–57.
- Fushiki, Tadayoshi. "Estimation of prediction error by using K-fold cross-validation." *Statistics and Computing* 21, September: (2009) 137–146.
- Garey, M. R., and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, volume 44 of *Books in the Mathematical Sciences*. W. H. Freeman, 1979.
- Geisser, S. *Predictive Inference*, volume 24. Chapman and Hall, 1993.
- Geladi, P. "Partial least-squares regression: a tutorial." *Analytica Chimica Acta* 185, 1: (1986) 1–17.
- Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biol* 5, 10: (2004) R80.

- Giatromanolaki, Alexandra, Michael I Koukourakis, Francesco Pezzella, Efthimios Sivridis, Helen Turley, Adrian L. Harris, and Kevin C. Gatter. "Lactate dehydrogenase 5 expression in non-Hodgkin B-cell lymphomas is associated with hypoxia regulated proteins." *Leukemia lymphoma* 49, 11: (2008) 2181–2186.
- Goffin, J. L. "The Relaxation Method for Solving Systems of Linear Inequalities." *Mathematics of Operations Research* 5, 3: (1980) 388–414.
- Golub, G. H., and C. F. Van Loan. *Matrix Computations*, volume 10 of *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins University Press, 1996.
- Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286, 5439: (1999) 531–7.
- Grate, Leslie R. "Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery." *BMC Bioinformatics* 6: (2005) 97.
- Griffin, J. D., D. Linch, K. Sabbath, P. Larcom, and S. F. Schlossman. "A monoclonal antibody reactive with normal and leukemic human myeloid progenitor cells." *Leukemia Research* 8, 4: (1984) 521–534.
- Grinstein, G, M Trutschl, and U Cvek. "High-Dimensional Visualizations.", 2001.
- Guo, Xinjian, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. "On the Class Imbalance Problem." *2008 Fourth International Conference on Natural Computation* 1: (2008) 192–201.
- Guyon, Isabelle. "An Introduction to Variable and Feature Selection 1 Introduction." *Journal of Machine Learning Research* 3: (2003) 1157–1182.
- Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. "Gene Selection for Cancer Classification using Support Vector Machines." *Machine Learning* 46, 19: (2002) 389–422.
- Hamed, Nahla A. M., Ola A. Sharaki, and Mohamed M. Zeidan. "Leptin in acute leukaemias: relationship to interleukin-6 and vascular endothelial growth factor." *The Egyptian journal of immunology Egyptian Association of Immunologists* 10, 1: (2003) 57–66.
- Hand, David J., and Keming Yu. "Idiot's Bayes: Not So Stupid after All?" *International Statistical Review* 69, 3: (2001) 385–398.
- Hastie, T., R. Tibshirani, and J. Friedman. *Elements of Statistical Learning (2nd Ed.)*. Springer, 2009.
- Hastie, Trevor, Saharon Rosset, Robert Tibshirani, and Ji Zhu. "The Entire Regularization Path for the Support Vector Machine." *Journal of Machine Learning Research* 5, 1: (2004) 1391–1415.

- Hoerl, A. E., and R. W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12, 1: (1970) 55–67.
- Horner, M. J., L. A. G. Ries, M. Krapcho, N. Neyman, R. Aminou, N. Howlader, S. F. Altekruse, E. J. Feuer, L. Huang, A. Mariotto, B. A. Miller, Lewis D. R., M. P. Eisner, D. G. Stinchcomb, and B. K. Edwards. "SEER Cancer Statistics Review, 1975-2006." Technical report, National Cancer Institute. Bethesda, MD, 2008.
- Hosmer, David W., and Stanley Lemeshow. *Applied logistic regression*, volume 2nd of *Wiley series in probability and statistics: Texts and references section*. Wiley, 2000.
- Hughes, G. "On the mean accuracy of statistical pattern recognizers." *IEEE Transactions on Information Theory* 14, 1: (1968) 55–63.
- Irizarry, Rafael A., Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal, Bryan C. Frank, Edward Gabrielson, Joe G. N. Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C. Hilmer, Eric Hoffman, Anne E. Jedlicka, Ernest Kawasaki, Francisco Mart, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, and Michael Wilson. "Multiple-laboratory comparison of microarray platforms." *Nature Methods* 2, 5: (2005) 1–6.
- Japkowicz, N., and S. Stephen. "The class imbalance problem: A systematic study." *Intelligent Data Analysis* 6, 5: (2002) 429–449.
- Jeffery, Ian B., Desmond G. Higgins, and Aedín C. Culhane. "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data." *BMC Bioinformatics* 7: (2006) 359.
- Jolliffe, Ian T. "A Note on the Use of Principal Components in Regression." *Applied Statistics* 31, 3: (1982) 300.
- Kass, G. V. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics* 29, 2: (1980) 119–127.
- Kelley, James, Bernard de Bono, and John Trowsdale. "IRIS: a database surveying known human immune system genes." *Genomics* 85, 4: (2005) 503–511.
- Kent Ridge Biomedical Data Set Repository. "Kent Ridge Biomedical Data Set Repository." <http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html>. Accessed December 2008., 2008.
- Kevil, C. G., L. Walsh, F. S. Laroux, T. Kalogeris, M. B. Grisham, and J. S. Alexander. "An improved, rapid Northern protocol." *Biochem Biophys Res Commun* 238, 2: (1997) 277–279.
- Kim, Yongdai, and Jinseog Kim. "Gradient LASSO for feature selection." *Twenty-first International Conference on Machine Learning ICML 04* 60.



- Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. "Optimization by simulated annealing." *Science* 220, 4598: (1983) 671–680.
- Kohavi, Ron. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence* 14, 12: (1995) 1137–1143.
- Kohavi, Ron, and George H John. "Wrappers for feature subset selection." *Artificial Intelligence* 97, 1-2: (1997) 273–324.
- Kohavi, Ron, and Dan Sommerfield. "Feature Subset Selection Using the Wrapper Model: Overfitting and Dynamic Search Space Topology." In *Proc of KDD*. 1995, 192–197.
- Korn, Edward L., and Richard Simon. "Explained Residual Variation, Explained Risk, and Goodness of Fit." *The American Statistician* 45, 3.
- Lachenbruch, Peter A, and M Ray Mickey. "Estimation of error rates in discriminant analysis." *Technometrics* 10, 1: (1968) 1–11.
- Le, Anne, Charles R. Cooper, Arvin M Gouw, Ramani Dinavahi, Anirban Maitra, Lorraine M. Deck, Robert E. Royer, David L. Vander Jagt, Gregg L. Semenza, and Chi V. Dang. "Inhibition of lactate dehydrogenase A induces oxidative stress and inhibits tumor progression." *Proceedings of the National Academy of Sciences of the United States of America* 107, 5: (2010) 2037–2042.
- Li, Tao, Chengliang Zhang, and Mitsunori Ogihara. "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression." *Bioinformatics* 20, 15: (2004) 2429–2437.
- Liu, Chun-Chi, Wen-Shyen E Chen, Chin-Chung Lin, Hsiang-Chuan Liu, Hsuan-Yu Chen, Pan-Chyr Yang, Pei-Chun Chang, and Jeremy J W Chen. "Topology-based cancer classification and related pathway mining using microarray data." *Nucleic Acids Research* 34, 14: (2006) 4069–4080.
- Lloret, Marta, Pedro Carlos Lara, Elisa Bordón, Fausto Fontes, Agustin Rey, Beatriz Pinar, and Orlando Falcón. "Major vault protein may affect nonhomologous end-joining repair and apoptosis through Ku70/80 and bax downregulation in cervical carcinoma tumors." *International Journal of Radiation Oncology, Biology, Physics* 73, 4: (2009) 976–979.
- Loi, To Ha, Anna Campaign, Adam Bryant, Tim J Molloy, Mark Lutherborrow, Jennifer Turner, Yee Hwa Jean Yang, and David D. F. Ma. "Discriminating lymphomas and reactive lymphadenopathy in lymph node biopsies by gene expression profiling." *BMC medical genomics* 4, 1: (2011) 27.
- Lönnstedt, Ingrid, and Terry Speed. "Replicated microarray data." *Statistica Sinica* 12, 1: (2002) 31–46.

- Loughrey, J., and P. Cunningham. “Overfitting in Wrapper-Based Feature Subset Selection: The harder you try the worse it gets.” *Search* 33–43.
- Lukes, R. J., and J. J. Butler. “The pathology and nomenclature of Hodgkin’s disease.” *Cancer Res* 26, 6: (1966) 1063–1083.
- Maggio, E. M., A. Van Den Berg, L. Visser, A. Diepstra, J. Kluiver, R. Emmens, and S. Poppema. “Common and differential chemokine expression patterns in rs cells of NLP, EBV positive and negative classical Hodgkin lymphomas.” *International Journal of Cancer* 99, 5: (2002) 665–672.
- Maher, Christopher A., Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M. Chinnaiyan. “Transcriptome sequencing to detect gene fusions in cancer.” *Nature* 458, 7234: (2009) 97–101.
- Mallows, C. L. “Some comments on C p.” *Technometrics* 15, 4: (1973) 661–675.
- McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition*, volume 156 of *Wiley Series in Probability and Statistics*. Wiley, 1992.
- Meier, Lukas, Sara Van De Geer, and Peter Bühlmann. “The group lasso for logistic regression.” *Society* 70, 1: (2008) 53–71.
- Meinicke, Peter, Thorsten Twellmann, and Helge Ritter. “Discriminative Densities from Maximum Contrast Estimation.” In *Advances in Neural Information Processing Systems 15*, edited by S Thrun S Becker, and K Obermayer, MIT Press, 2003, 985–992.
- Meinshausen, Nicolai, and Peter Bühlmann. “Stability Selection (with discussion).” *Journal of the Royal Statistical Society Series B* .
- Mitchell, Tom M. “Chapter 1: Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression.” In *Machine Learning*, 2010, 1–17.
- Mohammad, R. M., M. Y. Hamdan, and A. Al-Katib. “Induced expression of alpha-enolase in differentiated diffuse large cell lymphoma.” *Enzyme protein* 48, 1: (1994) 37–44.
- Monti, Stefano, Kerry J. Savage, Jeffery L. Kutok, Friedrich Feuerhake, Paul Kurtin, Martin Mihm, Bingyan Wu, Laura Pasqualucci, Donna Neuberg, Ricardo C. T. Aguiar, Paola Dal Cin, Christine Ladd, Geraldine S. Pinkus, Gilles Salles, Nancy Lee Harris, Riccardo Dalla-Favera, Thomas M. Habermann, Jon C. Aster, Todd R. Golub, and Margaret A. Shipp. “Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response.” *Blood* 105, 5: (2005) 1851–1861.

- Moreira Júnior, Gilberto, Gisele W. B. Colleoni, M. Giulia Cangi, Michael Murphy, Bradford Sherburne, José O Bordin, and Massimo Loda. "Reciprocal Cdc25A and p27 expression in B-cell non-Hodgkin lymphomas." *Diagnostic Molecular Pathology, The American Journal of Surgical Pathology Part B* 12, 3: (2003) 128–132.
- Morgan, James N., and Robert C. Messenger. *THAID, a sequential analysis program for the analysis of nominal scale dependent variables, by James N. Morgan [and] Robert C. Messenger*. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor,, 1973.
- Morgan, James N., and John A. Sonquist. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58, 302: (1963) 415–434.
- Morin, Ryan, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor Pugh, Helen McDonald, Richard Varhol, Steven Jones, and Marco Marra. "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." *Biotechniques* 45, 1: (2008) 81–94.
- Nakao, T., M. Hino, T. Yamane, Y. Nishizawa, H. Morii, and N. Tatsumi. "Expression of the leptin receptor in human leukaemic blast cells." *British Journal of Haematology* 102, 3: (1998) 740–745.
- NCBI. "MICROARRAYS: CHIPPING AWAY AT THE MYSTERIES OF SCIENCE AND MEDICINE." Available as url: <http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html>, 2007.
- Ng, Andrew Y. "On Feature Selection: Learning with Exponentially many Irrelevant Features as Training Examples." *Ieee Expert Intelligent Systems And Their Applications* 404–412.
- Nguyen, TuDung T., Erich J. Schwartz, Robert B. West, Roger A. Warnke, Daniel A. Arber, and Yasodha Natkunam. "Expression of CD163 (hemoglobin scavenger receptor) in normal tissues, lymphomas, carcinomas, and sarcomas is largely restricted to the monocyte/macrophage lineage." *The American Journal of Surgical Pathology* 29, 5: (2005) 617–624.
- Niens, Marijke, Lydia Visser, Ilja M. Nolte, Gerrit Van Der Steege, Arjan Diepstra, Pablo Cordano, Ruth F. Jarrett, Gerard J. Te Meerman, Sibrand Poppema, and Anke Van Den Berg. "Serum chemokine levels in Hodgkin lymphoma patients: highly increased levels of CCL17 and CCL22." *British Journal of Haematology* 140, 5: (2008) 527–536.
- Oommen, Thomas, Debasmita Misra, Navin K C Twarakavi, Anupma Prakash, Bhaskar Sahoo, and Sukumar Bandopadhyay. "An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing." *Mathematical Geosciences* 40, 4: (2008) 409–424.

- Pan, Zenggang, Yulei Shen, Cheng Du, Guimei Zhou, Andreas Rosenwald, Louis M. Staudt, Timothy C. Greiner, Timothy W. McKeithan, and Wing C. Chan. “Two newly characterized germinal center B-cell-associated genes, GCET1 and GCET2, have differential expression in normal and neoplastic B cells.” *The American Journal of Pathology* 163, 1: (2003) 135–144.
- Parmigiani, Giovanni, Elizabeth S. Garrett, Rafael A. Irizarry, Scott L. Zeger, Yulei Shen, Cheng Du, Guimei Zhou, Andreas Rosenwald, Louis M. Staudt, Timothy C. Greiner, Timothy W. McKeithan, and Wing C. Chan. *The analysis of gene expression data: methods and software*, volume 14. Springer Verlag, 2003.
- Parvin, Hamid, Hosein Alizadeh, and Behrouz Minaei-Bidgoli. “MKNN : Modified K-Nearest Neighbor.” *WCECS* 2173: (2008) 22–25.
- Paterson, Melinda A., Patrick S. Hosking, and Paul B. Coughlin. “Expression of the serpin centerin defines a germinal center phenotype in B-cell lymphomas.” *American Journal of Clinical Pathology* 130, 1: (2008) 117–126.
- Pearson, K. “On lines and planes of closest fit to systems of points in space.” *Philosophical Magazine* 2, 6: (1901) 559–572.
- Platt, John. “Probabilities for support vector machines.” In *Advances in Large Margin Classifiers*, edited by A Smola, P Bartlett, B Scholkopf, and D Schuurmans, MIT press, 2000.
- Pranckeviciene, Erinija, and Ray L. Somorjai. “On Classification Models of Gene Expression Microarrays: The Simpler the Better.” In *IJCNN’06*. 2006, 3572–3579.
- Quinlan, J. R. “Induction of decision trees.” *Machine Learning* 1, 1: (1986) 81–106.
- . *C4.5: Programs for Machine Learning*, volume 240 of *Morgan Kaufmann series in Machine Learning*. Morgan Kaufmann, 1993.
- R Development Core Team. “R: A Language and Environment for Statistical Computing.”, 2010.
- Rehm, Armin, Ioannis Anagnostopoulos, Kerstin Gerlach, Meike Broemer, Claus Scheidereit, Korinna Jöhrens, Michael Hübler, Roland Hetzer, Harald Stein, Martin Lipp, Bernd Dörken, and Uta E. Höpken. “Identification of a chemokine receptor profile characteristic for mediastinal large B-cell lymphoma.” *International Journal of Cancer, Journal International du Cancer* 125, 10: (2009) 2367–2374.
- Rennie, Jason D. M., Lawrence Shih, Jaime Teevan, and David R. Karger. “Tackling the Poor Assumptions of Naïve Bayes Text Classifiers.” *Proceedings of the 20th International Conference on Machine Learning* .
- Reunanen, Juha. “Overfitting in Making Comparisons Between Variable Selection Methods.” *Journal of Machine Learning Research* 3, 7-8: (2003) 1371–1382.

- Riddihough, Guy. "In the Forests of RNA Dark Matter." *Science* 309, 5740: (2005) 1507.
- Ripley, Brian D. *Stochastic Simulation*. Number Mc in Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, 1987.
- Rosenblatt, F. "The perceptron: a perceiving and recognizing automaton (Technical Report 85-460-1).", 1957.
- Roweis, Sam. "EM Algorithms for PCA and SPCA." *Advances in neural information processing systems* 10: (1998) 626–632.
- Rubinstein, R. Y., and D. P. Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 48. Springer-Verlag, 2004.
- Ryu, S. J., H. J. An, Y. S. Oh, H. R. Choi, M. K. Ha, and S. C. Park. "On the role of major vault protein in the resistance of senescent human diploid fibroblasts to apoptosis." *Cell Death and Differentiation* 15, 5: (2008) 1020–1028.
- Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *Bioinformatics (Oxford, England)* 23, 19: (2007) 2507–17.
- Sakhinia, E., M. Farahangpour, E. Tholouli, J. A. Liu Yin, J. A. Hoyland, and R. J. Byers. "Comparison of gene-expression profiles in parallel bone marrow and peripheral blood samples in acute myeloid leukaemia by real-time polymerase chain reaction." *Journal of Clinical Pathology* 59, 10: (2006) 1059–1065.
- Schäfer, Juliane, and Korbinian Strimmer. "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." *Statistical Applications in Genetics and Molecular Biology* 4, 1: (2005) Article32.
- Schultess, Jan, Oliver Danielewski, and Albert P. Smolenski. "Rap1GAP2 is a new GTPase-activating protein of Rap1 expressed in human platelets." *Blood* 105, 8: (2005) 3185–3192.
- Seegmiller, Adam C., Nitin J. Karandikar, Steven H. Kroft, Robert W. McKenna, and Yin Xu. "Overexpression of CD7 in classical Hodgkin lymphoma-infiltrating T lymphocytes." *Cytometry Part B Clinical cytometry* 76, 3: (2009) 169–174.
- Segal, Mark R. "Machine Learning Benchmarks and Random Forest Regression." *Biostatistics* 1–14.
- Sevilla, D. W., S. V. Nandula, A. I. Colovai, S. Alexander, V. V. Murty, B. Alobeid, and G. Bhagat. "Diffuse large B-cell lymphoma with TEL/ETV6 translocation." *Human Pathology* 40, 1532-8392 (Electronic) PT - Case Reports: (2009) 588–593.
- Sewell, M. "Feature Selection." <http://www.machinelearning.net/featureselection/feature-selection.pdf>, 2007.

- Shaffer, A. L., Miriam Shapiro-Shelef, Neal N. Iwakoshi, Ann-Hwee Lee, Shu-Bing Qian, Hong Zhao, Xin Yu, Liming Yang, Bruce K. Tan, Andreas Rosenwald, Elaine M. Hurt, Emmanuel Petroulakis, Nahum Sonenberg, Jonathan W. Yewdell, Kathryn Calame, Laurie H. Glimcher, and Louis M. Staudt. "XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation." *Immunity* 21, 1: (2004) 81–93.
- Shaik, Jahangheer S., and Mohammed Yeasin. "A unified framework for finding differentially expressed genes from microarray experiments." *BMC Bioinformatics* 8, 7: (2007) 347.
- Shi, Leming, Gregory Campbell, Wendell D. Jones, Fabien Campagne, Zhining Wen, Stephen J. Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M. Goodsaid, Lajos Pusztai, John D. Shaughnessy, André Oberthuer, Russell S. Thomas, Richard S. Paules, Mark Fielden, Bart Barlogie, Weijie Chen, Pan Du, Matthias Fischer, Cesare Furlanello, Brandon D. Gallas, Xijin Ge, Dalila B. Megherbi, W. Fraser Symmans, May D. Wang, John Zhang, Hans Bitter, Benedikt Brors, Pierre R. Bushel, Max Bylesjo, Minjun Chen, Jie Cheng, Jing Cheng, Jeff Chou, Timothy S. Davison, Mauro Delorenzi, Youping Deng, Viswanath Devanarayan, David J. Dix, Joaquin Dopazo, Kevin C. Dorff, Fathi Elloumi, Jianqing Fan, Shicai Fan, Xiaohui Fan, Hong Fang, Nina Gonzaludo, Kenneth R Hess, Huixiao Hong, Jun Huan, Rafael A. Irizarry, Richard Judson, Dilafruz Juraeva, Samir Lababidi, Christophe G. Lambert, Li Li, Yanen Li, Zhen Li, Simon M. Lin, Guozhen Liu, Edward K. Lobenhofer, Jun Luo, Wen Luo, Matthew N. McCall, Yuri Nikolsky, Gene A. Pennello, Roger G. Perkins, Reena Philip, Vlad Popovici, Nathan D. Price, Feng Qian, Andreas Scherer, Tielu Shi, Weiwei Shi, Jaeyun Sung, Danielle Thierry-Mieg, Jean Thierry-Mieg, Venkata Thodima, Johan Trygg, Lakshmi Vishnuvajjala, Sue Jane Wang, Jianping Wu, Yichao Wu, Qian Xie, Waleed A Yousef, Liang Zhang, Xuegong Zhang, Sheng Zhong, Yiming Zhou, Sheng Zhu, Dhivya Arasappan, Wenjun Bao, Anne Bergstrom Lucas, Frank Berthold, Richard J. Brennan, Andreas Bunes, Jennifer G. Catalano, Chang Chang, Rong Chen, Yiyu Cheng, Jian Cui, Wendy Czika, Francesca Demichelis, Xutao Deng, Damir Dosymbekov, Roland Eils, Yang Feng, Jennifer Fostel, Stephanie Fulmer-Smentek, James C Fuscoe, Laurent Gatto, Weigong Ge, Darlene R. Goldstein, Li Guo, Donald N. Halbert, Jing Han, Stephen C. Harris, Christos Hatzis, Damir Herman, Jianping Huang, Roderick V. Jensen, Rui Jiang, Charles D. Johnson, Giuseppe Jurman, Yvonne Kahlert, Sadik A. Khuder, Matthias Kohl, Jianying Li, Menglong Li, Quan-Zhen Li, Shao Li, Zhiguang Li, Jie Liu, Ying Liu, Zhichao Liu, Lu Meng, Manuel Madera, Francisco Martinez-Murillo, Ignacio Medina, Joseph Meehan, Kelci Miclaus, Richard A Moffitt, David Montaner, Piali Mukherjee, George J Mulligan, Padraic Neville, Tatiana Nikolskaya, Baitang Ning, Grier P. Page, Joel Parker, R. Mitchell Parry, Xuejun Peng, Ron L. Peterson, John H Phan, Brian Quanz, Yi Ren, Samantha Riccadonna, Alan H. Roter, Frank W. Samuelson, Martin M. Schumacher, Joseph D. Shambaugh, Qiang Shi, Richard Shippy, Shengzhu Si, Aaron Smalter, Christos Sotiriou, Mat Soukup, Frank

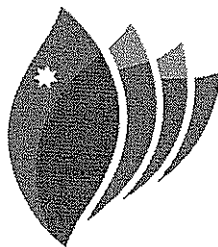
- Staedtler, Guido Steiner, Todd H. Stokes, Qinglan Sun, Pei-Yi Tan, Rong Tang, Zivana Tezak, Brett Thorn, Marina Tsyganova, Yaron Turpaz, Silvia C. Vega, Roberto Visintainer, Juergen Von Frese, Charles Wang, Eric Wang, Junwei Wang, Wei Wang, Frank Westermann, James C. Willey, Matthew Woods, Shujian Wu, Nianqing Xiao, Joshua Xu, Lei Xu, Lun Yang, Xiao Zeng, Jialu Zhang, Li Zhang, Min Zhang, Chen Zhao, Raj K. Puri, Uwe Scherf, Weida Tong, and Russell D. Wolfinger. "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models." *Nature Biotechnology* 28, 8: (2010) 827–838.
- Shipp, Margaret A., Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C. T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S Lander, Jon C. Aster, and Todd R. Golub. "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." Technical Report 1, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. margaret\_shipp@dfci.harvard.edu, 2002.
- Skibola, Christine F., John D. Curry, and Alexandra Nieters. "Genetic susceptibility to lymphoma." *Haematologica* 92, 7: (2007) 960–969.
- Smola, Alexander J., Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans. "Introduction to Large Margin Classifiers." *Advances in Large Margin Classifiers* 1–30.
- Smyth, Gordon K. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical applications in genetics and molecular biology* 3, 1: (2004) Article3.
- Somorjai, R. L., B. Dolenko, and R. Baumgartner. "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions." *Bioinformatics* 19, 12: (2003) 1484–1491.
- Statnikov, A., C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. "A comprehensive evaluation of multcategory classification methods for microarray gene expression cancer diagnosis." *Bioinformatics* 21, 5: (2004) 0.
- Statnikov, Alexander, Lily Wang, and Constantin F. Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." *BMC Bioinformatics* 9, 1: (2008) 319.
- Staudt, Louis M., and Sandeep Dave. "The biology of human lymphoid malignancies revealed by gene expression profiling." *Advances in Immunology* 87, 05: (2005) 163–208.
- Stolovitzky, Gustavo. "Gene selection in microarray data: the elephant, the blind men and our algorithms." *Current Opinion in Structural Biology* 13, 3: (2003) 370–376.

- Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A. Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” *Proc Natl Acad Sci U S A* 102, 43: (2005) 15,545–15,550.
- Suzuki, Nao, Norihito Yoshioka, Atsushi Uekawa, Noriomi Matsumura, Akiko Tozawa, Jyunki Koike, Ikuo Konishi, Kazushige Kiguchi, and Bunpei Ishizuka. “Transcription factor POU6F1 is important for proliferation of clear cell adenocarcinoma of the ovary and is a potential new molecular target.” *International journal of gynecological cancer official journal of the International Gynecological Cancer Society* 20, 2: (2010) 212–219.
- Tedoldi, S., J. C. Paterson, J. Cordell, S-Y. Tan, M. Jones, S. Manek, A. P. Dei Tos, H. Robertson, N. Masir, Y. Natkunam, S. A. Pileri, F. Facchetti, M-L. Hansmann, D. Y. Mason, and T. Marafioti. “Jaw1/LRMP, a germinal centre-associated marker for the immunohistological study of B-cell lymphomas.” *The Journal of pathology* 209, 4: (2006) 454–463.
- Tibshirani, R. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society Series B Methodological* 58, 1: (1996a) 267–288.
- Tibshirani, Robert. “A comparison of some error estimates for neural network models.” *Neural Computation* 8, 1: (1996b) 152–163.
- . “A comparison of fold-change and the t-statistic for microarray data analysis.” *Analysis* 1: (2007) 2–3.
- Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. “Diagnosis of multiple cancer types by shrunken centroids of gene expression.” *Proceedings of the National Academy of Sciences of the United States of America* 99, 10: (2002) 6567–72.
- Tibshirani, Robert, and Larry Wasserman. “Correlation-sharing for detection of differential gene expression.” *Arxiv preprint math0608061* .
- Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. “Missing value estimation methods for DNA microarrays.” *Bioinformatics* 17, 6: (2001) 520–5.
- Tusher, V. G., R. Tibshirani, and G. Chu. “Significance analysis of microarrays applied to the ionizing radiation response.” *Proc Natl Acad Sci U S A* 98, 9: (2001) 5116–5121.
- Vapnik, V. N. *The Nature of Statistical Learning Theory, 2nd ed.*, volume 8 of *Statistics for Engineering and Information Science*. Springer, 2000.
- Vapnik, V. N., and A. Y. Chervonenkis. “Ordered risk minimization.” *Automation and Remote Control* 35: (1974) 1226–1235, 1403–1412.



- Walter, Roland B., Ted A. Gooley, Vincent H. J. Van Der Velden, Michael R. Loken, Jacques J. M. Van Dongen, David A. Flowers, Irwin D. Bernstein, and Frederick R. Appelbaum. "CD33 expression and P-glycoprotein-mediated drug efflux inversely correlate and predict clinical outcome in patients with acute myeloid leukemia treated with gemtuzumab ozogamicin monotherapy." *Blood* 109, 10: (2007) 4168–4170.
- Wang, Li, Michael Gordon, and Ji Zhu. "Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning." *Sixth International Conference on Data Mining ICDM06* 690–700.
- Wang, Yu, Igor V. Tetko, Mark A. Hall, Eibe Frank, Axel Facius, Klaus F. X. Mayer, and Hans W. Mewes. "Gene selection from microarray data for cancer classification—a machine learning approach." *Computational Biology and Chemistry* 29, 1: (2005) 37–46.
- Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10, 1: (2009) 57–63.
- Webb, Geoffrey I., Janice R Boughton, and Zhihai Wang. "Not So Naive Bayes: Aggregating One-Dependence Estimators." *Machine Learning* 58, 1: (2005) 5–24.
- Weisberg, Sanford. *Applied Linear Regression*. Wiley, 1980.
- WHO. "National Cancer Institute sponsored study of classifications of non-Hodgkin's lymphomas: summary and description of a working formulation for clinical usage. The Non-Hodgkin's Lymphoma Pathologic Classification Project." *Cancer* 49, 10: (1982) 2112–2135.
- . *WHO classification of tumours of haematopoietic and lymphoid tissues*. Lyon, France: International Agency for Research on Cancer, 2008.
- Wright, Margaret H. "The interior-point revolution in optimization: History, recent developments, and lasting consequences." *Bulletin of the American Mathematical Society* 42, 01: (2004) 39–57.
- Wu, Tong Tong, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics* 25, 6: (2009) 714–721.
- Xing, E. P., and R. M. Karp. "CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts." *Bioinformatics* 17 Suppl 1, Suppl 1: (2001) S306–S315.
- Xu, Ronghui, and Xiaochun Li. "A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data." *Bioinformatics* 19, 10: (2003) 1284–1289.

- Yoon, H., K. Yang, and C. Shahabi. "Feature subset selection and feature ranking for multivariate time series." *IEEE Transactions on Knowledge and Data Engineering* 17, 9: (2005) 1186–1198.
- Yu, Zhiwen, Hau-San Wong, and Hongqiang Wang. "Graph-based consensus clustering for class discovery from gene expression data." *Bioinformatics* 23, 21: (2007) 2888–96.
- Zhang, Cun-hui. "Penalized linear unbiased selection." *Ann Statist to appear* 1–43.
- Zhang, H. H., J. Ahn, X. Lin, and C. Park. "Gene selection using support vector machines with non-convex penalty." *Bioinformatics* 22, 1: (2006) 88–95.
- Zheng, Zhaohui, Xiaoyun Wu, and Rohini Srihari. "Feature selection for text categorization on imbalanced data." *ACM SIGKDD Explorations Newsletter* 6, 1: (2004) 80.
- Zhou, Xin, and David P. Tuck. "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data." *Bioinformatics* 23, 9: (2007) 1106–1114.
- Zhu, Ji, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. "1-norm Support Vector Machines." *Statistics* 16, x.
- Zipf, G. K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. International Library of Psychology. Routledge, 1999.
- Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Prostate, The* 301–320.
- Zuber, Verena, and Korbinian Strimmer. "Gene ranking and biomarker discovery under correlation." *Bioinformatics* 25, 20: (2009) 2700–2707.



4 June 2009

Mr Tim Peters  
Department of Statistics  
Faculty of Science  
Macquarie University

**Reference: HE29MAY2009-D06614**

Dear Mr Peters,

## **FINAL APPROVAL**

### **Title of project: Selection of Microarray Elements for Optimal Linear Discrimination**

The above application was reviewed by the Ethics Review Committee (Human Research) at its meeting on 29 May 2009. Approval of the above application is granted, effective 29 May 2009 and you may now proceed with your research.

Please note the following standard requirements of approval:

1. Approval will be for a period of twelve (12) months. At the end of this period, if the project has been completed, abandoned, discontinued or not commenced for any reason, you are required to submit a Final Report on the project. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. The Final Report is available at: [http://www.research.mq.edu.au/researchers/ethics/human\\_ethics/forms](http://www.research.mq.edu.au/researchers/ethics/human_ethics/forms)
2. However, at the end of the 12 month period if the project is still current you should instead submit an application for renewal of the approval if the project has run for less than five (5) years. This form is available at [http://www.research.mq.edu.au/researchers/ethics/human\\_ethics/forms](http://www.research.mq.edu.au/researchers/ethics/human_ethics/forms). If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report (see Point 1 above) and submit a new application for the project. (The five year limit on renewal of approvals allows the Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).
3. Please remember the Committee must be notified of any alteration to the project.
4. You must notify the Committee immediately in the event of any adverse effects on participants or of any unforeseen events that might affect continued ethical acceptability of the project.
5. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University [http://www.research.mq.edu.au/researchers/ethics/human\\_ethics/policy](http://www.research.mq.edu.au/researchers/ethics/human_ethics/policy)

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide Macquarie University's Research Grants Officer with a copy of this letter as soon as possible. The Research Grants Officer will not inform external funding agencies that you have final approval for your project and funds will not be released until the Research Grants Officer has received a copy of this final approval letter.

Yours sincerely



P.R

Ms Karolyn White  
Director of Research Ethics  
Chair, Ethics Review Committee (Human Research)

Cc: Dr David Bulger, Department of Statistics

---

ETHICS REVIEW COMMITTEE (HUMAN RESEARCH)  
MACQUARIE UNIVERSITY

[http://www.research.mq.edu.au/researchers/ethics/human\\_ethics](http://www.research.mq.edu.au/researchers/ethics/human_ethics)

[www.mq.edu.au](http://www.mq.edu.au)