On the Number of Close-to-Optimal Feature Sets

Edward R. Dougherty\textsuperscript{1,2} and Marcel Brun\textsuperscript{2}

\textsuperscript{1}Department of Electrical and Computer Engineering, Texas A \& M University, College Station, TX. \textsuperscript{2}Computational Biology Division, Translational Genomics Research Institute, Phoenix, AZ.

Abstract: The issue of wide feature-set variability has recently been raised in the context of expression-based classification using microarray data. This paper addresses this concern by demonstrating the natural manner in which many feature sets of a certain size chosen from a large collection of potential features can be so close to being optimal that they are statistically indistinguishable. Feature-set optimality is inherently related to sample size because it only arises on account of the tendency for diminished classifier accuracy as the number of features grows too large for satisfactory design from the sample data. The paper considers optimal feature sets in the framework of a model in which the features are grouped in such a way that intra-group correlation is substantial whereas inter-group correlation is minimal, the intent being to model the situation in which there are groups of highly correlated co-regulated genes and there is little correlation between the co-regulated groups. This is accomplished by using a block model for the covariance matrix that reflects these conditions. Focusing on linear discriminant analysis, we demonstrate how these assumptions can lead to very large numbers of close-to-optimal feature sets.

Keywords: classification, covariance model, feature sets, gene expression.

Introduction

Concern has recently been expressed regarding the fact that different studies reveal different gene sets for microarray-based cancer classifiers meant to achieve a decision for the same characteristics. In particular, we refer to the concern regarding different gene sets for predicting prognosis for breast cancer (Ioannidis, 2005; Jenssen and Hovig, 2005). In one study, the issue is addressed by repeatedly randomly selecting a subset of the data, designing a classifier on the subset, estimating the error of the designed classifier on the remaining data points, and considering the distribution of the genes composing the various classifier feature sets (Michiels et al. 2005). Concern is expressed over the wide variability of the feature sets and the large number of genes contributing to the various feature sets. Reference is made to the sample size. While sample size is a factor, it is not the only one. A wide array of feature sets can occur with large samples owing to the relations among the genes. The potential for a large number of “best” feature sets is an inherent property of pattern recognition and is not particular to microarrays. In this paper we will explore this potential taking a model-based approach, specifically, modeling the covariance matrix of the features.

Before proceeding we must clarify the problem. Suppose we have some large number, $D$, of features from which to choose, say 20,000 gene-expression levels from a microarray. If we know the class conditional probability distributions, say the distributions of the genes for two phenotypes, then the collection of all $D$ features is optimal. Many may be redundant and can be removed with negligible effect on classification accuracy; nonetheless, all can be kept without doing harm to the classification accuracy. This, however, is abstract. In practice, feature sets and classifiers must be derived from sample data, and here there is the common situation of the error declining as the number of features grows to a certain point and then beginning to increase as the number of features increases beyond that point — the peaking phenomenon (Hughes, 1969; Kanal and Chandrasekaran, 1971; Jain and Waller, 1978) Peaking tends to occur later for larger samples. It occurs because too many features overfit the data. Hence, the issue of an optimal feature set is inherently one of classifier design from sample data. Given a sample, among all possible feature sets, which one results in a classifier with minimal error? Or more generally, what is the optimal number of features, in the sense that the expected error for samples of the optimal size is minimized relative to all possible sizes? The answer depends on the type of classifier being designed (Hua et al. 2005).

Since the actual distributions are unknown, evaluating feature sets requires error estimation. Different error estimators will yield different rankings of feature-set performance. Even for large samples,
error estimation suffers some imprecision. Thus, if there is a large number of close-to-optimal feature sets arising from a sample, then many of these may be statistically indistinguishable. Moreover, and this seems to be the issue concerning some investigators, different samples will yield different top-performing feature sets. But this is to be expected if each sample yields a list of feature sets (and classifiers) whose error rates are indistinguishable.

Putting aside error estimation for the moment, and focusing solely on classifier design, is it even conceivable to find an optimal feature set for a given sample? In the abstract, one could design a classifier for every possible feature set according to the desired classification rule, find the errors for these, and choose the best classifying feature set. But this program cannot be carried out because the number of possible feature sets to check is astronomical. Moreover, in the absence of prior distributional knowledge, which is the common situation in applications, a full search cannot be avoided if we wish to be assured of finding an optimal feature set (Cover and van Campenhout, 1977). This leads to the feature-selection problem, which is to apply some algorithm to select a less-than-optimal feature set that provides good classification. Depending on the feature-selection algorithm, or possible variant choices within the algorithm, different feature sets will be obtained for the same sample. There is the expectation that the more data one has, the better will be the selected feature set; nonetheless, outcome variability remains. Owing to the instability of feature-selection algorithms, one can surely expect slight changes in the data to result in different outputs. It is important to recognize that feature selection is part of classifier design. It determines the features that serve as variables for the classifier, which is a function on those variables. In doing so, it is part of the classification rule by which the classifier is constructed from the data. If we now incorporate error estimation into our reasoning, not only does it create imprecision in obtaining errors, but it also affects the feature-selection algorithm (Sima et al. 2005).

Given the preceding considerations, we will approach the optimal-feature-set problem in the following way: given the class conditional distributions and a positive integer \( d \), find the family of all feature sets of size \( d \) from among the potential features that possess minimal error, and hence are optimal. As posed, feature selection and error estimation do not play a role. Practically, error estimation is a must, so that, regarding the number of “best” feature sets, the issue is not so much one of simply finding the number optimal feature sets, but one of showing that there can be a very large number of feature sets of size \( d \) whose (true) errors are close to minimal, since these will be statistically indistinguishable. We note in this regard that increasing the sample size does not necessarily mitigate the problem of a large number of statistically indistinguishable feature sets, because as the sample size grows, so too does the size of the feature set producing minimal error, which means extraordinary growth in the number of feature sets of that size.

In the context of gene-expression classification, reflection on biological regulation indicates the enormous gene set variability that can be expected from a microarray study. A set of genes is useful for discriminating pathologies if its behavior, as a collection, varies according to the phenotypes. It may be that the genes involved have some primary regulatory function relating to the phenotype or that they are related, within the overall genome regulation, to other genes governing the phenotype. Given the complex regulatory connectivity within the genome—intricate feedback, massive redundancy, regulatory cascades, and tightly controlled co-regulation of gene cohorts—one should expect a large collection of gene sets having essentially equivalent capability for phenotype discrimination. Add to this the vast number of genes involved. Even if we throw away 80% of the genes on a 20,000-gene microarray, this leaves \( C(4000, d) \) feature sets of size \( d \), where \( C(D, d) \) denotes the number of subsets of size \( d \) that can be formed from \( D \) elements. Even for a modest size \( d \) this number is enormous. For instance, \( C(4000, 40) > 10^{80} \). Consequently, the errors corresponding to the feature sets can be expected to be extremely dense in the interval \([\varepsilon, 1]\), where \( \varepsilon \) is the Bayes error (minimal achievable error) for the classification problem. This observation alone should make it apparent that we can expect to find a large number of feature sets whose errors are statistically indistinguishable from the optimal one.

This paper considers a model in which the features are grouped in such a way that intra-group correlation is substantial whereas inter-group correlation is minimal, the intent being to model the situation in which there are groups of highly correlated co-regulated genes and there is little
correlation between the co-regulated groups. This will be accomplished by using a block model for the covariance matrix that reflects these conditions. Focusing on linear discriminant analysis, we demonstrate how these assumptions can lead to large numbers of close-to-optimal feature sets.

A Covariance Structure Leading to Many Close-to-Optimal Feature Sets

We consider classification for two Gaussian classes. Thus, the classes 0 and 1 possess normal conditional distributions, with mean vectors $\mu_k$ and covariance matrices $K_k$, for $k = 0, 1$. If we assume that the two classes are equally likely and the covariance matrices are equal, then the optimal classifier is determined by the discriminant $d_k(x) - d_0(x)$, where

$$d_k(x) = -(x - u_k)K^{-1}(x - u_k)$$

(1)

The discriminant is a linear function of $x$, produces hyperplane decision boundaries, and characterizes linear discriminant analysis (LDA). While it is unlikely that the covariance matrices are actually equal, it is this assumption that underlies the optimality of LDA. Moreover, in situations where LDA is applied as a classification rule and sample sizes are limited, such as is often the case with gene-expression data, the assumption of equal covariance matrices is a form of regularization, and is the simplest (and most used) of the various regularized covariance assumptions (Titterington, 1985; Friedman, 1988).

The decision boundary and error for LDA depend on the means, variances, and correlation coefficients. Obtaining direct insight concerning feature selection is not possible in the general case. A standard approach is to adopt a tractable covariance model that represents some assumptions on the relations among the features. For instance, a classic paper on finding the optimal number of features for LDA considers three covariance models (Jain and Waller, 1978). For all models it assumes a common variance $\sigma^2$ and a single base value $\rho$ from which the correlation coefficients are obtained. The models are defined by their correlation assumptions: (I) $\rho_{ij} = \rho$, (II) $\rho_{ij} = \rho^{|i-j|}$ and (III) $\rho_{ij} = \rho$ if $|i-j| = 1$ and $\rho_{ij} = 0$ if $|i-j| > 1$. The $3 \times 3$ forms of these models are given by the following covariance models:

\[
\begin{bmatrix}
1 & \rho & \rho \\
\rho & 1 & \rho \\
\rho & \rho & 1 \\
\end{bmatrix}
\]

(2)

In this paper we adopt a model suitable to the independent co-regulation of gene groups. Genes in different groups are uncorrelated and those within the same group are correlated. If we label these groups as $G_1, G_2, \ldots, G_m$, then the covariance matrix is blocked and takes the form

\[
B_k = \begin{pmatrix}
\sigma^2_{k1} & \sigma_{k1}\sigma_{k2}\rho_{k12} & \cdots & \sigma_{k1}\sigma_{kd}\rho_{kd} \\
\sigma_{k2}\sigma_{k1}\rho_{k21} & \sigma^2_{k2} & \cdots & \sigma_{k2}\sigma_{kd}\rho_{kd} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{kd}\sigma_{k1}\rho_{kd1} & \sigma_{kd}\sigma_{k2}\rho_{kd2} & \cdots & \sigma^2_{kd} \\
\end{pmatrix}
\]

(3)

where, for $k = 1, 2, \ldots, m$, $\sigma^2_{kj}$ is the variance of the $i$th feature in $G_k$ and $\rho_{ij}$ is the correlation coefficient for the $i$th and $j$th features in $G_k$. If $K$ has $d$ features split into $m$ equal-sized blocks, we write $K \sim \text{Cov}[d, m]$. Special cases of the model result if we assume the blocks take the form of one of the models I, II, or III. What is of interest to us here is that, depending on $\rho$ and the variances, it will often be the case that the best feature set of size $m$ has one feature from each block, namely, the best feature set is of the form $A = \{g_1, g_2, \ldots, g_m\}$, where $g_k \in G_k$ for $k = 1, 2, \ldots, m$. We define a feature set to be separated if the features come from different blocks.

We illustrate feature-set separation with several examples. Because the simulations required to construct the examples are extremely computational, we limit ourselves to relatively small covariance matrices. We first consider $K$ consisting of three blocks with three features in each, $K \sim \text{Cov}[9, 3]$. Fixing the value of $\rho$ throughout $K$, for each simulation the variances are selected from a uniform distribution. We consider two
cases: $\sigma_{ki} \sim U[0.3, 0.47]$ and $\sigma_{ji} \sim U[0.3, 1]$. The latter case allows a broader range of variances across the features. Fixing $\rho$ throughout $K$ gives a blockwise model corresponding to model I in (Jain and Waller, 1978) except that it is more realistic in allowing the variances to vary. In all cases 100,000 data points are used to train the best feature set containing 3 features, 50,000 per class, and another 100,000 data points are used for error estimation. The overall process is repeated 10,000 times at different values of $\rho$ uniformly distributed between 0 and 1. Figure 1(a) shows the percentage of separated optimal feature sets as a function of $\rho$ for $\sigma_{ij} \sim U[0.3, 0.47]$. Except for extreme values of $\rho$, the majority of optimal feature sets are separated, and for $0.3 \leq \rho \leq 0.9$, essentially 100% of the optimal feature sets are separated, indicating that this condition holds if there is at least modest correlation between features within the block and the correlation is not extreme. Note the steep fall off in separation percentage when $\rho$ exceeds 0.9. Figure 1(b), which corresponds to $\sigma_{ki} \sim U[0.3, 1]$ for the same covariance structure, shows that increased variability results in smaller percentages of separated optimal feature sets; nonetheless, for most of the range of $\rho$ there is a significant percentage of separated feature sets. For $0.2 \leq \rho \leq 0.8$, more than 60% of the optimal feature sets are separated.

To demonstrate the commonplace nature of the preceding results, as well as how they can vary depending on the model and the variability of the variance, we have done similar analyses in a number of circumstances. The results are shown in the remaining parts of Fig. 1: (c) $K \sim \text{Cov}[9,3]$, $\rho_{ij} = \rho^{ij}$ (corresponding to model II in (Jain and Waller, 1978)), and $\sigma_{ij} \sim U[0.3, 0.47]$; (d) $K \sim \text{Cov}[9,3]$, $\rho_{ij} = \rho^{ij}$ and $\sigma_{ij} \sim U[0.3, 1]$; (e) $K \sim \text{Cov}[10,2]$, $\rho_{ij} = \rho$, and $\sigma_{ij} \sim U[0.3, 0.47]$; (f) $K \sim \text{Cov}[10, 2]$, $\rho_{ij} = \rho$, and $\sigma_{ij} \sim U[0.3, 1]$; (g) $K \sim \text{Cov}[10, 2]$.

The percentage of separated sets is shown for different values of $\rho$ and $\sigma_{ij}$ in Figure 1. The figure illustrates how the separation of feature sets is affected by the correlation coefficient $\rho$ and the standard deviation $\sigma_{ij}$. The results are shown for several different covariance structures, including $K \sim \text{Cov}[9,3]$ and $K \sim \text{Cov}[10,2]$. The figure also shows how the separation of feature sets changes with the variability of the variance, as indicated by $\sigma_{ij}$. The results indicate that increased variability results in smaller percentages of separated optimal feature sets; nonetheless, for most of the range of $\rho$ there is a significant percentage of separated feature sets. For $0.2 \leq \rho \leq 0.8$, more than 60% of the optimal feature sets are separated.


\[ \rho_{i,j} = \rho^{ij} \text{ and } \sigma_{i,j} \sim U[0.3, 0.47]; \]  
\[ (h) \quad K \sim \text{Cov}[10, 2], \]  
\[ \rho_{i,j} = \rho, \text{ and } \sigma_{i,j} \sim U[0.3, 1]; \]  
\[ (i) \quad K \sim \text{Cov}[10, 5], \]  
\[ \rho_{i,j} = \rho, \text{ and } \sigma_{i,j} \sim U[0.3, 0.47]; \]  
\[ (j) \quad K \sim \text{Cov}[10, 5], \]  
\[ \rho_{i,j} = \rho^{ij}, \text{ and } \sigma_{i,j} \sim U[0.3, 1]. \]

Even though we have carried out very large simulations, there is still some wobble in the curves.

To look closer at the low separation percentage when \( \rho = 1 \), we consider two examples for the model having \( K \sim \text{Cov}[10, 2] \), \( \rho_{i,j} = \rho, \) and \( \sigma_{i,j} \sim U[0.3, 0.65] \). Let \( X_1, X_2, \ldots, X_10 \) denote the features, with \( X_1, X_2, \ldots, X_5 \) and \( X_6, X_7, \ldots, X_10 \) corresponding to the first and second block of \( K \), respectively. In the first example, we let \( \rho = 0.999 \) and generate one realization of \( K \) based on the randomly chosen variances. For the resulting \( K \) the best separated 2-feature set is \( \{X_1, X_7\} \), with an estimated error of 0.0253. There are nine non-separated feature sets possessing 0 error, \( \{X_1, X_3\}, \{X_1, X_4\}, \{X_1, X_5\}, \{X_6, X_8\}, \{X_6, X_9\}, \{X_6, X_{10}\}, \{X_7, X_8\}, \{X_7, X_9\}, \) and \( \{X_7, X_{10}\} \), where all error estimates have been obtained using 1,000,000 independent test sample points. Figures 2(a) and 2(b) show 1,000 sample points projected on to features \( X_1 \) and \( X_7 \), and on to features \( X_6 \) and \( X_8 \), respectively. The advantage of \( \{X_6, X_8\} \) over \( \{X_1, X_7\} \) is clear.

In the second example, we let \( \rho = 0.5 \) and generate one realization of \( K \) based on the randomly chosen variances. For the resulting \( K \) the best overall 2-feature set is the separated feature set \( \{X_1, X_7\} \), with an estimated error of 0.0257 (the difference with 0.0253 in the previous case resulting from the variability in error estimation). Figures 3(a) and 3(b) show 1,000 sample points projected on to the features \( X_1 \) and \( X_7 \), and on to the features \( X_6 \) and \( X_8 \), respectively. The advantage of \( \{X_1, X_7\} \) over \( \{X_6, X_8\} \), whose error is 0.0625, is clear.

Comparing Figures 2(b) and 3(b) for the feature set \( \{X_6, X_8\} \) shows what is happening, \( \rho \) is the common conditional correlation coefficient for the conditional densities \( f_{X|0} \) and \( f_{X|1} \). Since \( \rho = 1 \) in Fig. 2(b), the sample points for each class lie almost on a line, the lines being distinct because the class means are not equal. These two almost linear point masses are easily discriminated by a straight line between them. In Fig. 3(b), where \( \rho = 0.5 \), we see that the point masses are scattered around the two lines and extensively intermingle. Therefore they are not well discriminated.

### Analysis of Feature-Set Redundancy

Now that we know there are many separated feature sets resulting from the covariance model of Eq. 3, let \( A \) be such a feature set, \( \mu_{0i} \), \( \mu_{1i} \), \( \mu_{0m} \) and \( \mu_{1m} \) denote the means of the features in \( A \) for classes 0 and 1, respectively, and \( \sigma^2_1, \sigma^2_2, \ldots, \sigma^2_m \) denote their common variances. By a well-known formula (Duda, et al. 2001), the LDA error for \( A \) is given by

\[
\varepsilon_A = \frac{1}{\sqrt{2\pi}} \int_{\Delta(A)/2}^{\infty} e^{-u^2/2} du \tag{4}
\]

where \( \Delta(A) \) is the Mahalanobis distance. Since \( g_1, g_2, \ldots, g_m \) are independent, this distance is determined by

\[
\Delta(A)^2 = \sum_{j=1}^{m} \left( \frac{\mu_{0j} - \mu_{1j}}{\sigma_j} \right)^2 \tag{5}
\]
Each term of the sum gives the contribution of the feature in reducing the error. The contribution is increased for greater separation of the means and smaller variance. The error tends to 0 as $\Delta(A)$ tends to infinity. If, for any feature $g$ with variance $\sigma_g^2$ and class means $\mu_{0g}$ and $\mu_{1g}$, we let

$$\delta^s(g) = \left(\frac{\mu_{0g} - \mu_{1g}}{\sigma_g}ight)^2$$

(6)

then Eq. 5 can be rewritten as

$$\Delta(A)^2 = \sum_{j=1}^{m} \delta^s(g_j)$$

(7)

To consider feature-set redundancy, suppose $t$ is a very small positive number and that, for $k = 1, 2, ..., m$, the set $G_k$ has a subset $G_{k}^r = \{g_{k1}, g_{k2}, ..., g_{km}\}$ such that

$$\delta^s(g_k) - \delta^s(g_{k_1}) < t$$

(8)

for $i = 1, 2, ..., m_k$. We form a feature set $B$ by replacing $q$ features in $A$ by elements in $q$ of the sets $G_1^r, G_2^r, ..., G_m^r, q \leq m$. Then, according to Eq. 4, the increase in error resulting from using feature set $B$ instead of feature set $A$ is

$$\epsilon_B - \epsilon_A = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du$$

(9)

$$\Delta(A)^2 - \Delta(B)^2 < qt$$

(10)

Equations 7 and 8 yield

$$\Delta(B) > \Delta(A) - \frac{qt}{\Delta(A) + \sqrt{\Delta(A)^2 - qt}}$$

(11)

Some simple algebra shows that

Denoting the fraction on the right by $\eta(q, t)$, we conclude that

$$\epsilon_B - \epsilon_A = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} du$$

(12)

$$\Delta(A)^2 - \Delta(B)^2 < qt$$

(13)

For very small $t$, this difference is negligible, in which case the performances of the two feature sets are negligibly different. Since there are $m_k$ features in $G_k^r$ satisfying Eq. 8, the $q$ elements can be chosen in

$$v(q) = \sum_{1 \leq k_1 < k_2 < ... < k_q \leq m} m_{k_1} m_{k_2} ... m_{k_q}$$

(13)

ways. If we consider all feature sets formed by replacing 1, 2, ..., $m$ elements from different blocks,

Figure 3. Scatter plots for model $K \sim \text{Cov}[10, 2], \rho_{ij} = \rho, \sigma_{ki} \sim U[0.3, 0.65],$ with $\rho = 0.5$: (a) feature set $\{X_1, X_7\}$; (b) feature set $\{X_6, X_8\}$. 

![Figure 3](image-url)
then the total number of close-to-optimal feature sets formed in this manner is

\[ v = \sum_{q=1}^{m} v(q) \]  

(14)

If the blocks are large and there are several genes in some or all of the blocks satisfying Eq. 8, then this number can be very large.

The point is clear: if the blocks are large, which they may well be when they represent co-regulation of independent cohorts of genes, then there may be many features in a block possessing contributions close to the maximum. These can be interchanged with little effect on the error, thereby leading to close-to-optimal feature sets. In practice, feature sets whose errors are close to optimal will be indistinguishable owing to the need for error estimation.

We now illustrate feature-set redundancy with two examples generated from the model of Eq. 3 with \( K \sim \text{Cov} \{60, 5\} \) and \( \rho_{ij} = \rho \). In the first example, \( \rho = 0.8 \) and \( \sigma_{ij} \sim U[0.3, 1] \). The optimal 4-feature set is separated and has Bayes error 0.0485. Figure 4(a) shows the cumulative histogram of the errors for the superior 4-feature sets. Note that there are almost 700 feature sets whose errors are within 0.007 of the best. This is with 60 possible features. The number would likely be astronomically greater with 40,000 features. In the second example, \( \rho = 0.8 \) and \( \sigma_{ij} \sim U[0.3, 1] \). The optimal feature set is separated and has Bayes error 0.1642. Figure 4(b) shows the cumulative histogram of the errors for the superior 4-feature sets. With the larger error, feature-set errors are more spread out, but even in this case there are more than 80 features sets whose errors are within 0.02 of the best.

**Conclusion**

Feature-set variability should not be of concern in the context of finding expression-based diagnostic panels. It arises naturally from the relationships among the features, the vast number of feature sets when there is a large number of potential features, and the inherent natures of both classifier design and error estimation from sample data. Owing to the peaking phenomenon, larger sample sizes may not reduce the number of close-to-optimal feature sets because the number of feature sets rises dramatically as their size increases. The important issue is finding good feature sets that can achieve desirable classification rates. Based on what we have seen, it is not unlikely that two studies using comparable technologies and classification procedures will arrive at equally performing feature sets possessing small intersection—or, for that matter, no intersection. Rather than perceive this as a negative, one should see it as a positive. Given the incertitude of feature selection and the imprecision of error estimation, it can be beneficial to have many close-to-optimal feature sets.

Let us close by citing two previous empirical studies that have come to similar conclusions. In one, the authors consider a specific breast-cancer data set and classification with regard to good and...
bad prognosis (Ein-Dor et al. 2005). Their approach is to rank genes based on their correlation with survival, take feature sets consisting of the first through eighth sets of 70 genes in the list, and to show that classifiers based on each of these feature sets perform essentially the same. In another, it was shown that using an exhaustive search finds large numbers of indistinguishably performing small feature sets (Grate, 2005). By taking a model-based approach we have characterized this phenomenon for linear classification in terms of the discriminant and the actual error, and shown how intra- and inter-group correlation effect the phenomenon.

Reviewing Editor: James Lyons-Weiler.

References

Cover, T. and van. Campenout, J., 1977. On the possible orderings in the measurement selection problem. IEEE Trans. Systems, Man, and Cybernetics, 7:657–661.

Duda, R.O., Hart, P.E. and Stork, D.G., 2001. Pattern Classification, John Wiley, New York.

Ein-Dor, L., Kela, I., Getz, G., Givol, D. and Domany, E., 2005. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, 21:171–178.

Friedman, J.H., 1988. Regularized Discriminant Analysis. Journal of the American Statistical Association, 84:165–175.

Grate, L.R., 2005 Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. BMC Bioinformatics, 6:97.

Hua, J., Xiong, Z., Lowey, J., Suh, E. and Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21:1509–1515.

Hughes, G.F., 1969. Number of pattern classifier design samples per class. IEEE Trans. Information Theory, 15:615–618.

Ioannidis, J.P., 2005. Microarrays and molecular research: noise discovery? Lancet, 365:454–455.

Jain, A.K. and Waller, W.G., 1978. On the optimal number of features in the classification of multivariate Gaussian data. Pattern Recognition, 10:365–374.

Jenssen, T.K. and Hovig, E., 2005. Gene-expression profiling in breast cancer. Lancet, 365:634–635.

Kanal, L. and Chandrasekaran, B., 1971. On dimensionality and sample size in statistical pattern classification. Pattern Recognition, 3:225–234.

Michiels, S., Koscielny, S. and Hill, C., 2005. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 365:488–482.

Sima, C., Attoor, S., Braga-Neto, U., Lowey, J., Suh, E. and Dougherty, E.R., 2005. Impact of Error Estimation on Feature-Selection Algorithms. Pattern Recognition, 38:2472–2482.

Titterington, D.M., 1985. Common structure of soothing techniques in statistics. International Statistical Review, 53:141–170.