Communications in **M**athematical **A**nalysis
Special Volume in Honor of Prof. Peter Lax
Volume 8, Number 3, pp. 136–145 (2010)
ISSN 1938-9787
www.commun-math-anal.org

REGULARIZATION TECHNIQUES FOR MACHINE LEARNING ON GRAPHS AND NETWORKS WITH BIOLOGICAL APPLICATIONS

YUE FAN[*], SHINUK KIM[†], MARK KON[‡]
Department of Mathematics and Statistics
Bioinformatics and Systems Biology
Boston University
Boston, MA 02215, USA

LOUISE RAPHAEL[§]
Department of Mathematics
Howard University
Washington, DC 20059, USA

CHARLES DELISI[¶]
Bioinformatics and Systems Biology
Boston University
Boston, MA 02215, USA

(Communicated by Hitoshi Kitada)

Abstract

The representation of a high dimensional machine learning (ML) feature space F as a function space for the purpose of denoising data is introduced. We illustrate an application of such a representation of feature vectors by applying a local averaging denoising method for functions on Euclidean and metric spaces (together with its graph generalization) to the regularization of feature vectors in ML. We first discuss this technique for noisy functions on $\mathbb{R}$, and then extend it to functions defined on graphs and networks. This method exhibits a paradoxical property of the bias-variance problem in machine learning, namely, that as the scale over which averages are taken decreases, the error rate for classification first decreases and then increases. This approach is tested on two benchmark DNA microarray data sets used for classification of breast tumors based on predicted metastasis.

[*]E-mail address: yue@bu.edu
[†]E-mail address: kshinuk@bu.edu
[‡]E-mail address: mkon@bu.edu
[§]E-mail address: lraphael@howard.edu
[¶]E-mail address: delisi@bu.edu

AMS Subject Classification: 46N60, 68Q32, 92B05

Keywords: Bias-variance, computational biology, functional analysis, gene expression array, machine learning.

Dedicated To Professor Peter Lax – a remarkable mathematician as well as friend, with thanks for encouragement of our work on this problem.

1 Introduction

In this announcement we illustrate the transfer of some functional and numerical analysis tools to the solution of certain machine learning problems, and apply these to machine-learning based analysis of gene expression arrays (microarrays). Such biological array measurements produce feature vectors $\mathbf{z} = (z_1, \ldots, z_k)$ whose entries z_i are expression levels of genes q_i, yielding diagnostic and predictive information.

We view the feature vectors $\mathbf{z} = (f_1(q_1), \ldots, f_1(q_k))$ as (noisy) functions $f_1(q) = f(q) + e(q)$ on the space $G = \{q_1, \ldots, q_k\}$ of genes q_i, with f the underlying signal and $e(q) = \varepsilon g(q)$ an error term with $g(q)$ an independent $N(0,1)$ standard Gaussian distribution for each $q \in G$.

We will try to extract the underlying $f(q)$ from measurements $f_1(q)$ with function denoising approaches used in analysis on Euclidean and other metric spaces, and extend them to functions on G. The denoised functions f_{1t} will be more useful as approximations of f in further analysis, e.g., of gene expression arrays. This process requires a metric or network structure on G. Rappaort, Vert, et al. [11] have implicitly applied such an adapted denoising approach to regularize microarray feature vectors z using gene network structures on G. Yang and Kolaczyk [15] have studied network-based wavelets for denoising microarrays, again using networks on G, by extending analytic techniques such as wavelet denoising (e.g., Coifman and Donoho [2]). Chuang, Lee, et al. [1] have used local network-based averages based on protein-protein interactions (PPI), using local optimization methods to maximize discriminative ability of clusters. Clustering based on biological pathway membership of genes has also been investigated in [10, 8].

A gene network is a graph structure $\{G, w_{ij}\}$ with vertices G and edge weights w_{ij} (between genes i and j) which measure levels of interaction between pairs of genes. These might indicate for example their co-expression [16] (level of correlated gene expression), or existence of a protein-protein interaction (PPI) [9, 12] between their protein products.

Our discussion here is restricted to the adaptation of an analytic denoising tool which takes local function averages, adapted to microarray and other feature vectors which are functions on underlying networks or metric spaces. We denote this as martingale-based denoising, based on the notion of projection of noisy functions f_1 onto functions f_{1t} constant on increasingly refined partitions $\mathcal{A}_t = \{G_i^t\}_i$ $(t = 0, 1, 2, \ldots)$ of a base space $G = \bigcup_i G_i^t$. We point out that, as occurs in function denoising, accuracy of denoising based on such successively refined clustering initially increases and then decreases, as discussed in Theorems 2.1 and 3.3 below. Thus as $\mathcal{A}_t$ becomes less refined the decreasing variance due to microarray

This work was partially supported by NCI grant 2548-5, and NIH grant GM080625-01

noise is offset by a successively increasing bias due to local averaging, in another illustration of the bias-variance problem (Cuker and Smale [3], Geman et al. [5], Härdle et al. [6]).

Given an underlying expression function $f(q)$, the sequence of refined clusterings $\mathcal{A}_t$ ($t = 0, 1, \ldots, T$) of a graph or network G and its associated sequence of denoised (flattened) functions f_{1t} forms a martingale in t, i.e., a series of conditional expectations $f_{1t} = E(f_1|\mathcal{F}_t)$, where $\mathcal{F}_t$ is the field of sets generated by $\mathcal{A}_t$. Assuming the final refinement $\mathcal{A}_T = G$ consists of singleton sets, we note f_{1t} converges to f_1 (since in fact $f_{1T} = f_1$). As shown below, the error $||f_{1t} - f||$ typically decreases and then increases, being minimized at an intermediate value $t = t_0$. In applications we will seek an approximation to t_0, and use the regularized gene expression array $f_{1t_0}(q)$ for further inferences (e.g., biological predictions). We will illustrate this optimal averaging level by analyzing two gene expression datasets using two clustering methods.

An outline of the paper is as follows. The second section presents Theorem 2.1 on optimal martingale denoising for functions on the real line. Theorem 3.3 in section 3 extends this to an analogous denoising optimization theorem for functions on graphs and networks. Section 4 gives an application of Theorem 3.3, illustrating the existence of an intermediate level t_0 of refinement in a clustering which gives the best correspondence between denoised gene expression feature vectors f_{1t} and the ideal 'noise-free' vector f (as evidenced by the predictive accuracy of a denoising-based procedure). The networks giving our prior 'spatial' structure on the gene set G are gene co-expression [16] and protein-protein interaction (PPI) [9, 12] networks. The algorithm is then applied to two benchmark gene expression data sets used to test prediction of breast cancer metastases. These results are compared with classification results with no regularization, and also for regularization using random clustering. The numerical results imply that co-expression-based adjustment of PPI networks improves prediction scores. This validates the assumption that genes in the derived clusters should have similar expression behavior.

2 Optimized martingale denoising for functions on $\mathbb{R}$

We start in the Euclidean case by assuming an idealized machine learning feature vector in the form of a continuously differentiable function f on the interval $[0,1] \subset \mathbb{R}$. We assume that the measurement f_1 of f is noisy, i.e. that $f_1(q) = f(q) + e(q)$, is a measurement perturbed by a local error $e(q)$, normal and independent at each $q \in G$. More precisely, we assume that $e(q) = \varepsilon g(q)$, where $g(q)$ is a Gaussian noise distribution, of the form $g(q) = B'(q)$, the derivative of Brownian motion $B(q)$. To illustrate the continuous version of our network result, we try to recover f by averaging its measurement f_1 with respect to a hierarchical family (*filter*) of σ-algebras $\{\mathcal{F}_t\}_t$ of subsets of $G = [0,1]$. Specifically, for $t = 0, 1, 2, \ldots$ let $\mathcal{F}_t$ be the algebra of sets generated by the partition (clustering) $\mathcal{A}_t = \left\{ \left[\frac{i}{2^t}, \frac{i+1}{2^t} \right) : (i = 0, 1, \ldots, 2^t - 1) \right\}$. Averaging f over these clusters forms conditional expectations $f_t = E(f|\mathcal{F}_t)$, which are successively more accurate L^2 approximations to f. However, when f is perturbed by noise e as above, there is an optimal value t_0 for which the perturbed function f_1 has a conditional expectation $f_{1t} = E(f_1|\mathcal{F}_t)$ which minimizes the L^2 error, and after which (when $t > t_0$) the error begins to increase. We illustrate this bias-variance effect on the real line, and will then do so for functions on networks, with an illus-

tration involving genomic data sets. Throughout this paper all function norms $||\cdot|| = ||\cdot||_2$ are L^2 square integral norms.

Theorem 2.1. *Let the function f be defined on the unit interval be non-constant and continuously differentiable. Then for any $\delta > 0$, with probability greater than $1-\delta$, if the noise level ε is sufficiently small, the denoising approximation error $||f - f_{1t}||_2$ decreases monotonically for sufficiently small t, and increases monotonically for sufficiently large t.*

Sketch of Proof. The Gaussian form of the noise term $e(q) = \varepsilon g(q)$ together with a calculation of its square integral yields the identity

$$||f_{1t} - f||^2 = ||f_t + \varepsilon g_t(q) - f||^2 = ||f_t - f||^2 + \varepsilon^2 ||g_t||^2 = ||f_t - f||^2 + \varepsilon^2 \chi^2_{2^t} \qquad (2.1)$$

where the last term is a chi-square random variable with 2^t degrees of freedom. Clearly, if ε is sufficiently small and t is bounded, the second term is negligible, while the first term is decreasing, proving that $||f_{1t} - f||$ decreases in t for t small.

For large t, using (2.1) we see that changes in the error sequence $||f_{1t} - f||^2$ are bounded below by $-1 + \varepsilon^2 \left(\chi^2_{2^{t+1}} - \chi^2_{2^t}\right)$ (since $||f_t - f||^2$ approaches 0). Further, an estimate shows that $\sum_{t=1}^{\infty} P\left(\varepsilon^2 \left(\chi^2_{2^{t+1}} - \chi^2_{2^t}\right) - 1 < 0\right) < \infty$ for any $\varepsilon > 0$ (regardless of the dependence of the two χ^2 distributions) which together with the Borel-Cantelli lemma shows that $||f_{1t} - f||$ decreases only finitely often for $t = 0, 1, \ldots$. These two observations (for small and for large t) complete the argument. □

3 Graph Approximation Theorem Using Adaptive Martingales

Our main result is an extension of the above theorem to clustering-based function denoising on graphs. While a weighted graph G provides a measure of nearness, this does not necessarily lead to a metric on G, and the approach in the above theorem does not directly extend to graphs. However, if noisy perturbations of network-based data (i.e. a function $f_1(q) = f(q) + e(q)$ on the network) are given, then the associated graph structure together with some version of continuity for this function can similarly help eliminate noise. This is the basis for a graph-based theorem analogous to the one above. But first we define basic terms.

Definition 3.1. A *graph* (or *network*) $\{G, w\}$ consists of a collection G of elements (nodes), together with a function $w(i, j)$ $(i, j \in G)$ defining weights between each pair of nodes i and j.

Definition 3.2. A *filter* on a finite graph G is a sequence $\{\mathcal{F}_t\}_t$ of σ-fields (or equivalently, fields) of sets on G indexed by $t = 0, 1, 2, \ldots$, with the property $\mathcal{F}_{t+1} \supset \mathcal{F}_t$. We define the *clustering* $\mathcal{A}_t$ of G defined by $\mathcal{F}_t$ to be the most refined partition of G consisting of sets in $\mathcal{F}_t$. In some cases we will also use the sequence $\mathcal{A}_t$ of clusterings in place of the sequence of σ-fields $\mathcal{F}_t$.

If f is a function on G, we define the *f-martingale* on G with respect to $\{\mathcal{F}_t\}_t$ to be the sequence of conditional expectations $E(f|\mathcal{F}_t)$.

As in the continuous case, we assume measurements of the function f on G are subject to a noise $e(q) = \varepsilon g(q)$, with ε a (small) parameter and $g(q)$ an independent standard normal

random variable for each $q \in G$. We assume a filter $\{\mathcal{F}_t\}_t$ is defined for $t = 0, 1, \ldots, T$, with $\mathcal{F}_0 = \{G, \phi\}$ the trivial field of sets, and $\mathcal{F}_T = 2^G$ the full field of subsets of G. Now we approximate f from knowledge of $f_1 = f + \varepsilon g$ through the projection of the latter onto its conditional expectations $E(f_1|\mathcal{F}_t)$. Under proper values of the parameters, we show that the behavior exhibited earlier on the real line also occurs on a network.

Below all norms are L^2 norms defined on functions f on the graph G, i.e., $\|f\|^2 = \|f\|_2^2 = \sum_{q \in G} f(q)^2$, and $|A|$ denotes cardinality of a set A.

Theorem 3.3. *Let $K(t)$ $(t = 0, 1, 2, \ldots)$, denote a fixed positive function which is a lower bound for the change in error in $\|f_t - f\|_2$, i.e.,*

$$K(t) \leq \|f_t - f\| - \|f_{t+1} - f\|. \tag{3.1}$$

For a fixed choice of $C > 0$, let $\{\mathcal{A}_t\}_{1 \leq t \leq T}$ be any filter satisfying

$$|\mathcal{A}_{t+1}| \geq C\,|\mathcal{A}_t| \tag{3.2}$$

for some $C > 1$. Then with probability arbitrarily close to 1, the error $\|f_{1t} - f\|$ is decreasing for sufficiently small t and increasing for sufficiently large t, if the noise $\varepsilon \leq \varepsilon_0$ is small enough, and the graph size $|G| \geq n_0$ is large enough.

Note that the statement is uniform over all graphs G, all sequences $\{\mathcal{A}_t\}_{1 \leq t \leq T}$ of refinements satisfying (3.2), and all functions f on G satisfying (3.1), for fixed $K(t)$ and C. That is, within such a class of graphs, filters, and functions, the theorem holds for all of these with a single choice of ε_0, n_0.

Sketch of Proof. The proof relies on the identity

$$\begin{aligned} & \|e_{t+1}\| - |\|e_t\| - ||f_{t+1} - f|| - \|f_t - f\| \\ \leq\ & \left\|f_{1(t+1)} - f\right\| - \|f_{1t} - f\| \\ \leq\ & \|f_{t+1} - f\| - \|f_t - f\| + \|e_{t+1}\| + \|e_t\| \end{aligned} \tag{3.3}$$

recalling that $e(q) = \varepsilon g(q)$ $(q \in G)$, with $g(q)$ independent and identically distributed (iid) standard Gaussian noise. Note that $e_t(q) = \varepsilon g_t(q) = \varepsilon E(g(q)|\mathcal{F}_t)$, and it is easy to show that if $q \in a$ with $a \in \mathcal{A}_t$ a minimal set (cluster) in $\mathcal{F}_t$, then $E(g(q)|\mathcal{F}_t) = \frac{1}{\sqrt{|a|}} Z_a$ (with each Z_a an iid $N(0,1)$ variable), since the left side is simply an average of $|a|$ iid Gaussians; here $|a|$ denotes the size of the set a. Thus for ε small and if t is relatively small (meaning that cluster size $|a|$ is large), terms of the form $e_t(q) = \frac{\varepsilon}{\sqrt{a}} Z_a$ are negligible. Then the right side of the second inequality in (3.3) is (up to a small additive factor) bounded from above by $-K(t)$, which is negative and so implies a decreasing $||f_{1t} - f||$.

To show that $||f_{1t} - f||$ is increasing for t sufficiently large, we use the first inequality in (3.3). Since $||f_{t+1} - f|| - ||f_t - f||$ converges to 0 for large t, it suffices to show that $||e_{t+1}|| - ||e_t||$ becomes arbitrarily large. This in turn follows from the identity

$$\frac{1}{\varepsilon} E(||e_t||) = \frac{\sqrt{2}}{\Gamma(k_t/2)} \Gamma(k_t/2 + 1/2) = \sqrt{k_t} + O\left(1/\sqrt{k_t}\right), \tag{3.4}$$

with $k = |\mathcal{A}_t|$, and the second equality following from a standard Gamma function estimate. In addition, letting V denote variance,

$$\begin{aligned}\frac{1}{\varepsilon^2}V(||\varepsilon_t||) &= \frac{1}{\varepsilon^2}\left[E\left(||\varepsilon_t||^2\right) - E\left(||\varepsilon_t||\right)^2\right] = E\left(\sum_{a\in\mathcal{A}_t} Z_a^2\right) - \left(\sqrt{k_t} + O\left(1/\sqrt{k_t}\right)\right)^2 \\ &= k_t - \left(\sqrt{k_t} + O\left(1/\sqrt{k_t}\right)\right)^2 = O(1)\ (k_t \to \infty) \qquad (3.5)\end{aligned}$$

From (3.4) and (3.5) together with Chebyshev's inequality and the fact that $k_t = |\mathcal{A}_t| \geq Ck_{t-1}$, it follows that e_t grows sufficiently fast that $||e_{t+1}|| - ||e_t||$ and similarly the left side of (3.3) is eventually increasing with probability 1 (for a fixed choice of the sequence $|\mathcal{A}_t|$ and the lower bound $K(t)$), completing the argument. □

4 Application - Regularization of Gene Expression Data

We give an application of the network-based analytic denoising method in Theorem 3.3, to a machine learning analysis of two data sets of breast tumor gene expression arrays (microarrays) [14, 13], with respect to predictiveness of tumor metastasis. Biologically metastasis is the spreading of a disease, particularly cancer, to another part of the body. Our goal here is to design a test on microarrays that correctly predicts a breast patient's cancer tissue samples as either metastatic or non-metastatic. A prediction for a patient is calculated from measured gene expression levels for the genes $G = \{q_1, \ldots, q_k\}$.

Let the function $\mathbf{f}_i = (f_i(q_1), \ldots, f_i(q_k))$ denote the true gene expression values of the i^{th} patient in a data set D. As above, we assume the measured expression value (including error) is $f_{i1}(q) = f_i(q) + e_i(q)$ where the error consists of iid normal random variables. The measured gene expression feature vector for the i^{th} patient is $\mathbf{f}_{i1} = (f_{i1}(q_1), \ldots, f_{i1}(q_k))$. Thus the full data set is $D = \{\mathbf{f}_{1i}, y_i\}_{i=1}^n$, where for the i^{th} patient

$$y_i = \begin{cases} 1 & \text{if metastasis occurs} \\ -1 & \text{otherwise} \end{cases}$$

Following standard machine learning procedures, the data set of patient samples is separated into training and test sets. The training set is a subset of D that is used to discover predictive relationships between $\mathbf{f}_{1i}$ and y_i (here in training a machine learning algorithm). The accuracy of this predictive relationship is then tested on the remaining data set (the test set). Our goal in the test data set is to predict the outcome y_i from the measured feature vector $\mathbf{f}_{1i}$ (given we do not know the true underlying expression array $\mathbf{f}_i$). This estimate will be improved by using our cluster-based denoising technique of regularized expression values $\mathbf{f}_{i1t}$ (see below).

We tested the algorithm on two data sets, from Wang, et al. [14] and van de Vijver [13]. For each data set we first trained a classifier based on a support vector machine (SVM), a standard machine learning (ML) algorithm, using the training portion of the data set D. After this, predictions by the trained SVM were made on test data $\mathbf{f}_{i1} = (f_{i1}(q_1), \ldots, f_{i1}(q_k))$ from D, and the predictive accuracy of the algorithm was recorded. The cluster sets $\mathcal{A}_t = \left\{G_j^t\right\}_j$ (with $G = \bigcup_j G_j^t$) which were used were defined from proximity in a protein-protein

interaction (PPI) network as well as a gene co-expression network, with the a field $\mathcal{F}_t$ generated by $\mathcal{A}_t$.

We compared the accuracy of predicting y_i using the ML algorithm directly on feature data $\mathbf{f}_{1i}$, as compared to using the denoised (cluster-averaged) feature vectors $\mathbf{f}_{1it} = E(\mathbf{f}_{1i}|\mathcal{F}_t)$. The latter is constant on clusters G_k^t and so can be represented directly as a function on them, resulting in a k_t-dimensional vector $\mathbf{f}_{i1t} = \left(f_{i1t}(G_1^t), \ldots, f_{i1t}\left(G_{k_t}^t\right)\right)$, with $f_{i1t}\left(G_j^t\right) = \sum_{q \in G_j^t} f_{i1}(q) / \left|G_j^t\right|$, and $k_t = |\mathcal{A}_t|$. Practically, the premise is that genes in the same cluster should have similar expression behavior which is numerically reinforced and noise-cancelled by averaging.

We now describe the basis for our clustering procedure, whose primary basis is the so-called protein-protein interaction (PPI) gene network [9, 12]. This is a network of genes which in which two genes q_i and q_j are connected if the unique proteins they generate are involved in chemical interactions with each other. The PPI network is used to identify gene subgroups that have biologically related functions, and so are likely to have similar gene expression patterns. Grouping of genes based on PPI relatedness enables identification of potential macroscopic (group) biomarkers (e.g., indicators of metastasis) which should be relatively more robust than individual gene expression bio-markers. However, a potential disadvantage of such a network is that co-functioning proteins may not have similar gene expressions, since inverse expression patterns for pairs of genes in the same pathways (and so PPI clusters) is possible. For this reason we sub-clustered the PPI clusters according to a second network, the co-expression network. The co-expression network relates pairs of genes which have similar expression patterns among subjects in the training portion of the data set D. The resulting combined clustering is denoted as a *co-expression adjusted* PPI network $\{G, w_{ij}\}$. More specifically, the unweighted PPI network G was modified to include co-expression correlation by weighting edges (i, j) with weights $w_{ij} = exp\{-d_{ij}^2/\sigma^2\}$, with d_{ij} a distance defined by hierarchical co-expression clustering. We employed the *Graclus* software package (Dhillon et al. [4]) to perform graph clustering.

The PPI plus co-expression clustering-based conditional expectation $f_{1t} = E(f_1|\mathcal{F}_t)$ (the regularized gene expression array) was compared against use of unregularized feature vectors $\mathbf{f}_1$ as well as regularization using random clustering. This was tested on the above-mentioned benchmark breast cancer datasets from high throughput gene expression studies by Wang, et al. [14] and van de Vijver et al. [13], with the goal to distinguish metastatic patients from non-metastatic patients based on gene expression feature vectors. Wang's dataset contains 286 breast cancer patients of whom 93 had tumors which metastasized, while van de Vijver's dataset has 295 patients among which 79 were classified as metastatic.

The algorithm was tested by a 5-fold cross validation, reserving $1/5$ of patients in a data set D as test data, and building the classifier on the data of the remaining patients. The model was then used to predict metastasis on each test patient and to calculate accuracy, and the procedure was repeated 5 times until all patients were scored. Sensitivity (or recall) is the probability that a patient is predicted positive (for metastasis), given the patient is actually metastatic. Specificity is the probability that the patient is predicted negative (non-metastatic) given the patient is non-metastatic. Precision is the probability a patient is metastatic when predicted to be so.

Two measures are computed to compare the performance of the methods:

- the area under the ROC curve (AUROC) and
- the area under the Precision-Recall curve (AUPRC).

The ROC curve is the plot of sensitivity versus specificity. The precision-recall curve is the plot of precision versus recall (sensitivity). Both curves are obtained by varying decision thresholds for positivity, and plotting the two quantities of interest on the graph.

5 Numerical Results and Conclusion

We chose the numbers of gene clusters to be powers of 2, namely, $k_t = |\mathcal{A}_t|$ = 64, 128, 256, 512, 1024 and 2048. The predictive performances of support vector machine (SVM) machine learning algorithms on both datasets are improved after the regularization process, over use of unregularized individual gene feature vectors $\mathbf{z} = \mathbf{f} = (f(q_1), \ldots, f(q_k))$.

Specifically for AUROC, performance of co-expression adjusted PPI network clustering on Wang's dataset attained a value of 73% using $|\mathcal{A}_t| = 2048$ clusters. Using all the genes in SVM classification using unregularized expression arrays resulted in a reduction of 20% in the AUROC. The AUROC performance of the same regularized clustering on van de Vijver's dataset attains a value of 73% also for $k_t = |\mathcal{A}_t| = 2048$ clusters, which is decreased by 7% when using unregularized gene expression arrays. Our co-expression adjusted PPI clustering also improves the area under the precision recall curve (AUPRC) from 36.2% to 52.4% at $k_t = 2048$ for Wang's dataset, and from 34.6% to 43.0% at $k_t = 2048$ for the van de Vijver dataset.

In general as k_t increases, classification performance improves until it reaches an optimal value at a clustering level between $|\mathcal{A}_t| = 1024$ and $|\mathcal{A}_t| = 2048$ clusters, and then decreases, as suggested by Theorem 3.3. Indeed, the poorer unregularized performance can be interpreted as a performance at the most refined clustering level $\mathcal{A}_T = G$, i.e., with individual genes as clusters. Thus, the improvement of bias $||f_1 - f_{1t}||$ is offset by the increase variance $||e(t)||$ for larger values of t. Put even more briefly, averaging over a gene set which is too small does not sufficiently dampen out noise $e_i(t)$.

As a baseline we also considered random clustering. AUROC performance of random clustering on Wang's set was 69% with 512 clusters (as opposed to 53% for unregularized data), while van de Vijver's dataset attained values of 69% and 66%, respectively, with 512 cluster regularization and unregularized gene data, respectively, again using the SVM classifier. In fact for random clustering the best performances often ranged at cluster sizes from 256 to 512 genes. This apparently implies that the penalty for averaging over relatively large groups of randomly selected genes (whose expressions also can vary in either direction among metastatic and non-metastatic tissues) is overcome by the quenching of noise $e_i(t)$ attained by such averaging. In particular, this seems to imply that though most of the 512 random clusters in the Wang data set were not very informative, those few which happened to have co-expressed informative genes (together with the advantage of averaged out noise) were sufficient to overcome this disadvantage.

In addition, the strong performance of random clustering in the Wang data set (even though it did not quite match PPI clustering) implies that the data in that set were dominated by noise. The relatively worse performance of random clustering in the van de Vijver data

set indicates that this data set had generically less noise in it. This fact was validated by a study of standard deviations of accuracies of bootstrapped subsets of both data sets.

This behavior again emphasizes the fact that random noise is a large factor affecting predictive power of gene expression-based algorithms, and the need for further development of denoising methodologies for such networks.

Acknowledgments

The authors thank the editor for his initiation of this volume and his invitation to participate in it.

The authors thank the communicating editor for his careful reading of our manuscript and his insightful report.

References

[1] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, **3** (2007), pp 140-149.

[2] R. R. Coifman and D. L. Donoho, Translation invariant de-noising. *Wavelets and Statistics* A. Antoniadis and G. Oppenheim, Eds., Springer-Verlag Lecture Notes (1995).

[3] F. Cuker and S. Smale, Best Choices for Regularization Parameters in Learning Theory: On the Bias-Variance Problem. *Found. Comput. Math.*, **2** (2002), pp 413-428.

[4] I. S. Dhillon, Y. Guan, and B. Kulis, Weighted graph cuts without eigenvectors a multilevel approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **29** (Nov. 2007), pp. 1944-1957.

[5] S. Geman, E. Bienenstock and R. Doursat, Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **4** (1992), pp 1-58.

[6] W. Härdle, G. Kerkyacharian, D. Picard and A. Tsybakov, *Wavelets, Approximation and Statistical Applications* (1995) Springer-Verlag.

[7] P. D. Lax, *Functional Analysis* (2002) Wiley-Interscience.

[8] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*. **4**:e1000217 (2008).

[9] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein and P. D'Eustachio, Reactome knowledgebase of biological pathways and processes. *Nucleic Acids Res.* **3** (Nov. 2008).

[10] J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439** (2006), pp 353-357.

[11] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J-P. Vert, Classification of microarray data using gene networks. *BMC Bioinformatics*. **8**:35 (2007).

[12] S. Razick, G. Magklaras, and I. M. Donaldson, iRefIndex: A consolidated protein interaction database with provenance. *BMC Bioinformatics*. **9** (2008).

[13] M. J. van de Vijver, Y. D. He, L. J. van't Veer, A. A. Dai H, Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend et al, A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347** (2002) pp 1999-2009.

[14] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365** (2005) pp 671-679.

[15] S. Yang and E. D. Kolaczyk, Target detection via network filtering. *eprint arXiv:0902.3714* (2009).

[16] B. Zhang and S. Horvath, A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*. **4** (2005).