33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Is the Reduction of Dimensionality to a Small Number of Features Always Necessary in Constructing Predictive Models for Analysis of Complex Diseases or Behaviours?

Amin Zollanvari, Nancy L. Saccone, Laura J. Bierut, Marco F. Ramoni, and Gil Alterovitz

Abstract— **Gene expression and genome wide association data have provided researchers the opportunity to study many complex traits and diseases. When designing prognostic and predictive models capable of phenotypic classification in this area, significant reduction of dimensionality through stringent filtering and/or feature selection is often deemed imperative. Here, this work challenges this presumption through both theoretical and empirical analysis. This work demonstrates that by a proper compromise between structure of the selected model and the number of features, one is able to achieve better performance even in large dimensionality. The inclusion of many genes/variants in the classification rules can help shed new light on the analysis of complex traitstraits that are typically determined by many causal variants with small effect size.**

I. INTRODUCTION

DNA microarray technology, which made possible measuring the expression of thousands of genes simultaneously, has found many applications in biomedical research and made studying the variation of population through genetic markers possible. They have been widely used in medicine to help researchers in better understanding the etiology of diseases by discovering new biomarkers that correlate well with progression of a disease or by finding new drugs through studying the differences in gene expressions in cells exposed to different doses. In recent years, with the advent of genome wide association data many causal variants with strong evidence of association to complex disease or behaviors have been identified; yet, they perform poorly in a *predictive* setting [1], [2]. This is due to the fact that many complex traits, have a phenotypic response determined by interactions between numerous environmental and genetic factors and therefore, each individual disease locus has a small effect size [3]. Few studies have tried to capture the polygenic nature of complex traits such as overt stroke in sickle cell anemia [2], Coronary Artery calcification in atherosclerosis [4], and nicotine dependence [1] through constructing prognostic models capable of dissecting the complex web of interactions between causal variants; yet, they have all considered limited number of single nucleotide polymorphisms (SNPs) in the proposed prognostic model. The limited number of features considered in the model have not been limited to the aforementioned genome wide association studies (GWAS) and SNP based models; many gene expression based studies have been following similar design machinery [5]–[8].

The commonly employed procedure to design the classifiers in these studies have been based on the presumption of reducing the number of dimensions from thousands of features to a small number of features, usually based on filtering and some statistical tests such as t-test, ANOVA or their variants. Then, the designer tries to find a best set of features in this lower dimension based on some search methods such as exhaustive search, best first search, ranker, and other methods to construct the classifier in a space of much lower dimensionality, such as two or three commonly in gene expressions based studies [5], [7], [8], and commonly less than one hundred SNPs in GWAS [1], [2], [4]. Eventually, assessing the performance of the classifier is performed by different error estimation techniques.

The rationale behind reducing the dimensionality involves small-sample situations and the well-known phenomenon, *curse of dimensionality* [14], [15]. Devroye, one of the pioneers in statistical pattern recognition, mentions that: "Just recall the curse of dimensionality that we often faced: to get good error rates, the number of training samples should be exponentially large in the number of components. Also, computational and storage limitations may prohibit us from working with many components". In today's world, thanks to advanced technology, the latter problem, namely storage limitations, has been alleviated to some extend. However, the salient concluding point we try to make in this paper is that in order to achieve a classifier of reasonable performance, one does not need to presumably reduce the number of features for the classifier or the predictive model; a compromise between complexity of the model (for example measured by VC dimension [14]) and the number of features can result in substantially better performance. This needs to be investigated more in future as a continuum of the very few classical works on determining the optimal number of features in very specific scenarios [16]. Unfortunately due to the lack of proper measures of such a compromise, many researchers regardless of the scenarios that call for the exponentially

Amin Zollanvari and Marco F Ramoni are with Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, and Partners Healthcare Center for Personalized Genetic Medicine, Boston, MA. Gil Alterovitz is with Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Partners Healthcare Center for Personalized Genetic Medicine, and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. Nancy L Saccone is with Department of Genetics, Washington University School of Medicine, St. Louis, MO. Laura J Bierut is with Department of Psychiatry, Washington University School of Medicine St. Louis, MO. Correspondence to: Amin.Zollanvari@childrens.harvard.edu, Ga@alum.MIT.edu

higher number of samples than features, try to reduce the number of dimensions to a small number in a hope to find well-behaved predictors. This procedure not only eliminates many important features from the scope of further analysis, but also results in predictors with limited accuracy rate. For many biostatisticians and bioinformaticians, the presumption of dimensionality reduction can be rooted back to the dawn of microarray technology in which the technology was still merging; in addition, the number of volunteer patients for genetic profiling was limited. The data in many studies from that time, and even yet, are characterized by many features but very limited number of samples [5], [9], [10], and are commonly known as "tall data" matrices [11]. In [11], the authors have pointed out several common "fads" and "fallacies" regarding the classification problem using microarray data; however, many hints and tips mentioned there are valid for analysis of tall data matrices. In recent years, with the advanced recent technology, which facilitates genetic profiling of the patients, and by accumulating the data over time, we are witnessing the emergence of many studies in which the number of features and the samples can be considered as being "comparable" (at least after initial filtration or quality controls); yet, many classification rules pertinent to tall data matrices are still being employed intact. To cite just a few works with comparable numbers of samples and features, consider [19] where 3,713 SNPs were genotyped for 1,929 samples; [2] in which 235 SNPs were genotyped for 1,398 samples, and [4] with 2,882 SNPs and 712 samples. In this work, we focus on the scenario in which sample size and number of features are comparable, and using different synthetic and practical examples we demonstrate the efficacy of incorporating a large number of features in relatively simple classifiers, or in general, predictive or prognostic models.

II. Significant dimensionality reduction can be avoided where sample size is comparable to number of features

A. An Analytic Example

The scenario in which the number of features is comparable to the sample size can be analytically studied under a specific asymptotic assumption, namely double asymptotics, which is commonly credited to S. Raudys and A. Kolmogorov [17], [18] and was extended by us in [20], [21]. Intuitively, the behavior of a statistic is studied as both sample size and dimensionality (or generally the number of parameters) increase to infinity in a controlled fashion, where the ratio between sample size and dimensionality converges to a finite constant [18]. Denoting the Mahalanobis distance between classes by $\delta_p^2 = (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)$, the number of samples in each class n_i, $i = 0, 1$, and the number of features, p, the double asymptotic conditions can be represented as $n_0 \to \infty, n_1 \to \infty, p \to \infty, \frac{p}{n_0} \to J_0 < \infty, \frac{p}{n_1} \to J_1 < \infty, \delta_p^2 \to c$ in which c and J_i, $i = 0, 1$ are all constants. These conditions can be used to analyze different statistics of interests in situations where the number of samples is comparable to the number of features [20].

In order to make a simple example, we assume a binary classification setting where n_i, $i = 0, 1$, are comparable to the number of features, p, and both are large e.g. $n_i = 2,000$ and $p = 1,000$. Furthermore, we assume that the sampling distributions of both classes are multivariate normal distributions with common and known covariance matrix Σ and $\delta_p^2 = 4$. We further assume that Fisher linear classifier, commonly known as Linear Discriminant Analysis (LDA) [22], is chosen to execute the classification task. LDA is given by:

$$\psi(x) = \begin{cases} 1\,, & \text{if } W(x) < 0 \\ 0\,, & \text{if } W(x) \geq 0 \end{cases}, \tag{1}$$

in which

$$W(X) = \left(x - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} \right)^T \Sigma^{-1} \left(\hat{\mu}_0 - \hat{\mu}_1 \right). \tag{2}$$

where μ_i, i= 0, 1, are the sample estimates of class conditional densities. Under double asymptotic conditions, it can be shown that LDA constructed using all the features has on average (over sample space) a true error equal to $\epsilon_p = \alpha_0 \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_1 - J_0}{\sqrt{\delta^2 + J_0 + J_1}}\right) + \alpha_1 \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_0 - J_1}{\sqrt{\delta^2 + J_0 + J_1}}\right)$, in which α_i's are just prior probability of classes [17], [20], [21] . Under aforementioned double asymptotic conditions and certain other regularity conditions, the assumption of Gaussianity can be alleviated; however, this is not the focus of this paper and interested readers are referred to [18] for more information. Assuming equal prior probabilities and substituting $J_i = \frac{p}{n_i} = \frac{1,000}{2,000} = \frac{1}{2}$, we observe that Fisher linear discriminant using all the features has on average an error of 0.185, which is reasonably close to the Bayes error, denoted by ϵ^*. The Bayes error is the error of the optimal classifier that is often unknown in practice. Here ϵ^* can be computed since we have the parameters of the distributions, and furthermore, we know that the optimal linear classifier for a multivariate normal distribution of classes has $\epsilon^* = \Phi\left(-\frac{1}{2}\delta\right) = 0.158$. Closeness of the average error of linear classifiers to the optimal linear classifier shows that, in this example, reduction of dimensionality is not a critical stage of classification rule. This point becomes even more clear by noticing the fact that reduction of dimensionality to a lower space, say $p' < p$, can even diminish the performance of the classifier. This is because δ_p^2 is changed to $\delta_{p'}^2$ and assuming that the common covariance matrix of classes is identity, then, it can be shown that $\delta_{p'}^2 < \delta^2$ and $\epsilon_{p'} > \epsilon_p$. We have depicted the average true error versus dimension for this example in Figure 1-a, where in order to have $\delta_{1,000}^2 = 4$, the means of the normal distributions need to be $\mu_0 = -\mu_1 = \mathbf{0.031}_p$. Clearly, the true error is a decreasing function of dimensionality confirming the fact that the best performance is achieved by considering all the features in the classifier.

B. A Simulation Example

Here, in contrast to the previous example, we consider a more realistic situation in which the covariance matrix of classes that appears in the discriminant $W(X)$ given by (2) is estimated from the data using a regularized estimation of

sample covariance matrix. In order to simulate the situation under study in this paper (comparable sample size and dimension), we have generated 1,000 samples for each class taken from two multivariate Gaussian distributions of 900 dimensions with a common covariance matrix having 1 as diagonal and 0.2 as off diagonal elements and the means adjusted such that $\delta^2_{900} = 4$. Since the performance of regularized-LDA depends on regularization parameter, L, three values of L were selected to design three different LDAs in this scenario ($L = 0.1, 1, 10$). In order to estimate the error of these three designed LDAs, we have generated 10,000 additional samples from the aforementioned multivariate distributions as a test set and found the rate of misclassification of these test samples by each LDA. On the other hand, in order to compare the performance of the three LDAs designed using the full 900 dimensions to classifiers designed on a much lower dimensionality, we have chosen 10,000 different randomly selected combinations of two-feature sets out of 900 features and designed LDA classifiers using each set (hence 10,000 LDAs designed in two dimensions), and estimated the error of each LDA on the test samples. The histogram of the error of these 10,000 LDAs is plotted in Figure 1-b. The errors of the three regularized-LDAs constructed by considering all the features and different choice of L, are shown as vertical dashed and dotted lines in Figure 1-b. As we can see the histogram of the error rates of 10,000 classifiers designed on two dimensionality spaces is on the right of the error rate of all three regularized-LDAs designed on 900 dimensions. This clearly shows the advantage of using all the features in this scenario.

C. A Practical Situation

Nicotine dependence has a strong genetic component. Twin studies have demonstrated the heritability of a large proportion of phenotypic variance ranging from 40-75% [22]. In order to identify novel causal genetic factors for nicotine dependence, several GWAS carried out using the case-control experimental design [19], [26]. To address the issues related to standard statistical methods such as logistic regression in high dimensionality space [24], and building a predictive model able to simultaneously capture interactions between causal loci, Ramoni et al. [1] considered 73 SNPs previously reported by Bierut et al. and Saccone et al. in [19], [26] as SNPs associated with nicotine dependence. Ramoni et al. utilized a multivariate probabilistic model, namely Bayesian network, to predict the nicotine dependence with up to 75% accuracy, measured by the area under the receiver operator characteristic curve (AUROC) on the fitted data. The data we use here are a subset of data used in the Collaborative Genetic Study of Nicotine Dependence (COGEND) (for more information about the data the reader is referred to [25]). The data set we have considered is a cohort of 2,062 European Americans. We randomly split the data into 1,857 training and 205 data. After controlling for SNPs with high genotype call, and removing those with minimum allele frequency, $MAF < 0.01$, the number of 1,642 initial SNPs was reduced to 1,501 SNPs. After training the predictive Bayesian model, namely naive Bayesian network, we utilized the model to predict the risk of individuals' nicotine dependence in the validation data set containing 205 samples. To train each classifier, we first ranked the SNPs using Cochran-Armitage trend test of association [27]. Then, naive Bayesian network was employed to construct the model on the SNPs selected from the top of the list. The selected number of SNPs was increased until all 1,501 SNPs in the list were considered in the network. Figure 2 shows that the best possible performance, measured by AUROC, corresponds to the case where all SNPs are considered in the model with AUROC=0.772% on an independent data set. The AUROC of our model on the training data set (as considered in [1]) is 0.861%. Therefore, considering all 1,501 SNPs and a simple classifier, namely naive Bayesian network, we easily outperformed the classifier constructed in [1].

III. Conclusion

A proper compromise between the complexity of the classifier and the number of features selected to be involved in the model is a critical step in achieving the best possible performance. Often researchers select a structure and regardless of how simple the structure is, employ stringent filtration or significant dimensionality reduction through feature selection methods. Here, three relatively simple classifiers have been designed to demonstrate the efficacy of considering large number of features in situations where the number of samples is comparable. This work shows the necessity of continuing very few classical works on determining the optimal number of features in very specific scenarios. Double asymptotics that was presented in this paper can be a promising analytical tool to accomplish this goal.

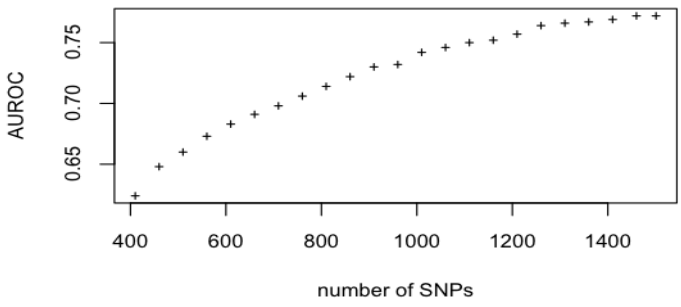

Fig. 2. True error of naive Bayesian network versus number of SNPs evaluated on 205 test samples.

IV. Acknowledgments

Lead investigators directing data collection are Laura Bierut, Naomi Breslau, Dorothy Hatsukami, and Eric Johnson. The authors thank Heidi Kromrei and Tracey Richmond for their assistance in data collection. This work was supported by the NIH grants P01CA89392 (L. Bierut) from the National Cancer Institute, K02DA021237 (L. Bierut) from the National Institute on Drug Abuse, 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01(G. Alterovitz), and 4R00LM00982603 (G. Alterovitz).

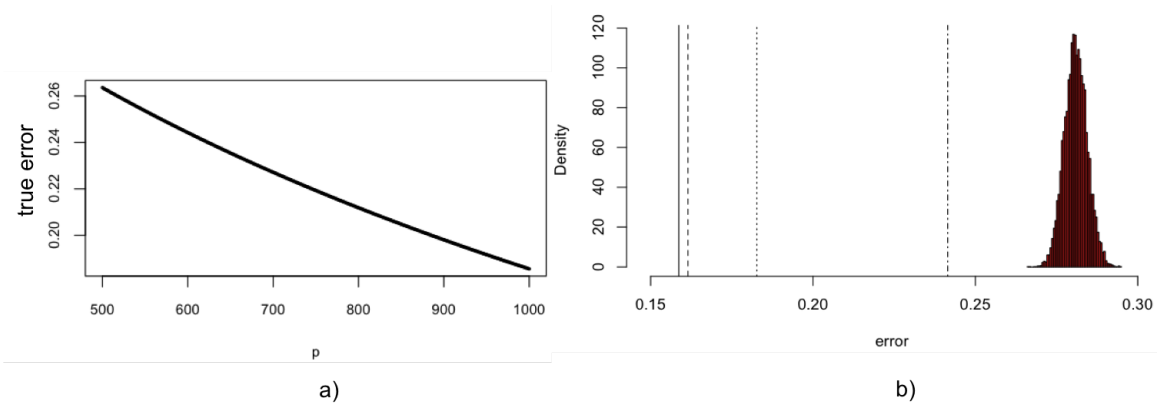

Fig. 1. a) True error vs. dimension for the analytic example: p and n are both large and comparable. Clearly, considering all the features is substantially better than reducing the dimensionality for the choice of classifier considered here. b) Comparison between error of two-feature LDA classifiers vs. regularized-LDAs constructed by 900 features with different choice of L. The histogram on the right shows the error of 10,000 LDAs designed on randomly selected sets of two features from the original 900 features. The solid vertical line is the Bayes error. Other vertical lines are the error of regularized-LDAs for different choice of L and constructed by considering all the features: solid-dashed: $L = 0.1$, dotted: $L = 1$, and dashed line: $L = 10$. Figure shows the error rates of all two-feature designed LDAs are substantially larger than the error rates of regularized-LDAs constructed by 900 features.

REFERENCES

[1] Ramoni, R. B., Saccone, N. L., Hatsukami, D. K., Bierut, L. J., Ramoni, M. F., "A Testable Prognostic Model of Nicotine Dependence," . *Journal of Neurogenetics*, vol. 23, 283-292, 2009.

[2] Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T., "Steinberg, M.H. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia," *Nature genetics*, vol. 37, 435-40, 2005.

[3] Zondervan, K. T., Cardon, L. R., "The complex interplay among factors that influence allelic association", *Nature reviews. Genetics*, vol. 5, No. 2, 89-100, 2004.

[4] McGeachie, M., Ramoni, R. L., Mychaleckyj, J. C., Furie, K. L., Dreyfuss, J. M., Liu, Y., Herrington, D., Guo, X., Lima, J. A., Post W., Rotter, J. I., Rich, S., Sale, M., Ramoni, M. F., "Integrative predictive model of coronary artery calcification in atherosclerosis", *Circulation*, vol. 120, No. 24, 2448-2454, 2009.

[5] S. Kim, E. R. Dougherty, I. Shmulevich, K. R. Hess, S. R. Hamilton, J. M. Trent, G. N. Fuller, and W. Zhang, "Identification of Combination Gene Sets for Glioma Classification," *Molecular Cancer Therapeutics*, vol. 1, No. 13, 1229-1236, 2002.

[6] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa, "Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival, *Blood*, vol. 103, no. 7, pp. 27712778, 2004.

[7] H. Somura, N. Iizuka, T. Tamesa, K. Sakamoto, T. Hamaguchi, R. Tsunedomi, H. Yamada-Okabe, M. Sawamura, M. Eramoto, T. Miyamoto, Y. Hamamoto, and M. Oka, "A three-gene predictor for early intrahepatic recurrence of hepatocellular carcinoma after curative hepatectomy", *Oncol. Rep.*, 19, 489-95, 2008.

[8] E.J.M. Nascimento, U.M. Braga-Neto, C. Calvazara, A.L. Gomes, F. Abath, B. Acioli, C.A.A. Brito, M.T. Cordeiro, A.M. Silva, C. Magalhaes, R. Andrade, L.H.V.G. Gil and E.T.A. Marques, Jr., "Gene Expression Profiling During Acute Stage of Dengue Infection", *PLoS ONE*, 4, 2009.

[9] van de Vijver M.J., He Y.D., van't Veer L.J., Dai H., Hart A.A., Voskuil D.W., Schreiber G.J., Peterse J.L., Roberts C., Marton M.J., Parrish M., Atsma D., Witteveen A., Glas A., Delahaye L., van der Velde T., Bartelink H., Rodenhuis S., Rutgers E.T., Friend S.H., Bernards R. "A gene-expression signature as a predictor of survival in breast cancer,"*N Engl J Med*, vol. 347, No. 25, 1999-2009, 2002.

[10] Bittner, M., Meltzer, P., Chen, Y., Jiang, Y. et al. "Molecular classification of cutaneous malignant melanoma by gene expression profiling", *Nature*, vol. 406, No. 6795, 536-540, 2000.

[11] Braga-Neto U.M. , "Fads and Fallacies in the Name of Small-Sample Microarray Classification," *IEEE Signal Processing Magazine, Special Issue on Signal Processing Methods in Genomics and Proteomics*, vol. 24, No. 1, 91-99, 2007.

[12] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", *Nature*, vol. 403, No. 6769, 503-511, 2000.

[13] L.J. Beirut, et. al. "A genome-wide association study of alcohol dependence", *Proc. Natl. Acad. Sci.*. 16, 2010.

[14] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* New York: Springer, 1996.

[15] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers", *IEEE Transactions on Information Theory*, 14, 1968.

[16] S. J. Raudys and A. K. Jain, "Small Sample Size Effects in Statistical Pattern. Recognition: Recommendations for Practitioners", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, 1991.

[17] S. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former soviet union literature," *Journal of Multivariate Analysis*, vol. 89, pp. 1-35, 2004.

[18] V. Serdobolskii, *Multivariate Statistical Analysis: A High-Dimensional Approach*. Springer, 2000.

[19] Saccone, S.F. et al., "Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs", *Human molecular genetics* vol. 16, 36-49, 2007.

[20] A. Zollanvari, U. M. Braga-Neto, E. R. Dougherty, "Analytic Study of Performance of Error Estimators for Linear Discriminant Analysis", *accepted for publication in IEEE Transactions on Singal Processing,* doi: 10.1109/TSP.2011.2159210.

[21] A. Zollanvari, M. G. Genton, "On Kolmogorov Asymptotics of Misclassification Error Rates in Linear Discrimianat Analysis", *submitted*.

[22] Vink, J.M., Willemsen, G., Boomsma, D.I., "Heritability of smoking initiation and nicotine dependence", *Behavior genetics* vol. 35, 397-406, 2005.

[23] A. Zollanvari, U. M. Braga-Neto, E. R. Dougherty, "Joint Sampling Distribution Between Actual and Estimated Classification Errors for Linear Discriminant Analysis", *IEEE Transactions on Information Theory*, vol. 56, 784-804, 2010.

[24] Heidema, A. G., Boer, J. M., Nagelkerke, N. Mariman, E. C., van der, A. Dl, Feskens, E. J., "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases", *BMC genetics*, vol. 23, 2006.

[25] Saccone, N. L. ,Saccone, S. F., Hinrichs, A. L., Stitzel, J. A. et al. , "Multiple distinct risk loci for nicotine dependence identified by dense coverage of the complete family of nicotinic receptor subunit (CHRN) genes", *American journal of medical genetics. Part B, Neuropsychiatric genetics*, vol. 150B, 453-466, 2009.

[26] Bierut, L.J. et al. "Novel genes identified in a high-density genome wide association study for nicotine dependence", *Human molecular genetics*, vol. 16, 24-35, 2007.

[27] Agresti, A. *Categorical data analysis*. Wiley, New York, 2002.