STATEMENT OF RECENT WORK – HAIM BAR

Abstract. My research interests include statistical modeling for high-throughput data, variable selection, and network analysis. I develop statistical methods for the so-called 'Omics data, which includes microarray data, RNA-seq, metabolomics, and proteomics. New techniques which enabled the sequencing of the genome (or proteome, or microbiome, etc.) introduced fantastic opportunities in medical research. However, the abundance of data at the gene, protein, or metabolite level created multiple challenges from the statistical point of view. For example, quite often the number of response variables (e.g. gene expression levels) is very large, and analyzing each one separately in the traditional way (e.g. via t-test or linear regression) may result in insufficient statistical power, since it is essential to account for (possibly massive) multiple testing. In addition, although the cost of experiments involving sequencing is decreasing, obtaining data from a large sample is still not affordable to most organizations. Typical experiments involve fairly small sample sizes, which again adversely affects the statistical power. Another challenge in these applications is that response variables are not likely to be independent. Many traditional, and even modern statistical methods rely on the convenient, but unrealistic assumption that response variables are independent. I develop powerful and computationally efficient methods to address these challenges, and I collaborate extensively with multiple research groups. Generally, my approach to addressing the aforementioned challenges has been to develop hierarchical mixture models with random effects. In my models I assume that the data are drawn from a parametric distribution in which the number of parameters is not only small, but remains constant as the dimensionality of the problem increases. Using this approach yields parsimonious models and 'shrinkage estimation', which not only increases power by borrowing information across responses, but also increases computational efficiency and scalability.

1. High Throughput Data Analysis – 'Omics Applications

My work in the area of high throughput data analysis has been motivated by problems in genetics and bioinformatics. My earlier contribution was to develop a powerful method to detect differentially expressed genes in microarray data, in the two group setting [3]. The power of the method stems from the parsimonious nature of the mathematical model. More specifically, when comparing gene expression levels in two populations, I assume that genes may belong to one of three components, namely, genes which are not differentially expressed, genes which are more expressed in population 1, and genes which are more expressed in population 2. This is implemented as a three component mixture model such that each component is modeled to be normally distributed with different means and variances. This mixture model random-effect approach leads to 'borrowing information' across all genes in a component, thus increasing the power to detect differentially expressed genes, even when the number of genes is very large (tens of thousands) and the sample sizes are small. I implemented it as an R package [2] using an empirical Bayes approach. I also implemented it using a fully Bayesian approach [4]. The method was applied in a number of collaborations with biologists. For example, in [5] and [6] we were interested in finding which genes in the genome of the Shewanella oneidensis bacteria are associated with an increased ability of the bacteria to convert waste water to electricity. In another collaboration, with the Caudill group at Cornell University, we investigated the association between pregnancy status, choline intake levels, and gene expression [7, 8, 9, 10].

This work led to the interesting observation that in some situations the difference between the populations may be in the variance of the gene expression levels, rather than the mean. I developed a model for parallel testing for unequal variances [11]. This approach has proven to be quite robust to deviations from the normality assumption, and it substantially increases the power to detect differences in variance over the more traditional approaches. I applied this method in [12] where we compared metabolic levels across different strains of Staphylococcus aureus, in order to understand their Intermediate-Type Vancomycin Resistance (VISA).

Recently, I developed a bivariate model in which it is possible to test for differential dispersion and differential expression, simultaneously [13]. I have also developed an extension to the model to the k groups scenario, and to account for additional fixed-effects in the model. I am currently in the final stages of the implementation of the extended model as a free and user-friendly software (using the 'Shiny' package in R).

I am currently involved in a number of collaborative projects in which my bivariate method is applied to other 'omics applications. For example, with the Pyle lab at Yale University we have analyzed RNA-seq data to investigate molecular pathways involved in human oocyte senescence (under review.) With the Smyth lab at the University of Connecticut, we are analyzing proteomics data in order to understand the necrotic enteritis disease mechanisms in poultry, by detecting which proteins are associated with increased or decreased intensity levels across four disease-producing strains.

2. Model Selection in the 'large p, small n' Scenario

I have been working on variable selection in linear regression, where the number of predictors is very large, possibly much larger than the number of responses (the so-called 'large p, small n' problem). To understand the source of the problem, consider the normal setting, where the response vector, Y, is a linear function of p predictors with a mean vector, $\boldsymbol{\beta}$. If X is an $n \times p$ design matrix, and $\boldsymbol{\epsilon}$ is a vector of normally distributed errors, then we write $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To obtain estimates for the mean effect of the predictors, we have to invert the matrix $X'X$, whose dimensions are $p \times p$. When p is large this matrix becomes singular, and estimates can not be obtained. The analysis of quantitative trait loci (QTL) falls into this category. In a QTL analysis, the response is some quantitative trait (e.g., height, wing-span), and the putative variables are SNPs (single-nucleotide polymorphism) in the genome. The objective is to find which of the (potentially millions) of SNPs is associated with the trait. I developed a variable selection method which is suitable for situations in which the number of predictors is very large [14]. It uses the empirical Bayes approach, and a fast implementation of the Expectation Maximization (EM) algorithm [1]. In addition, to further improve its efficiency it uses mathematical tricks that exploit the sparse structure of the variance-covariance matrix. More recently, I improved the method by accounting for correlations between variables (*A Scalable Empirical Bayes Approach to Variable Selection*, with Prof. Booth and Prof. Wells, from Corenll University. Submitted.) We are working on extensions to the method, including to the Generalized Linear Models (GLM) framework, and to survival analysis.

3. Analyzing Gene Networks

Although much has been learned over the years from analyzing data at the individual gene level, it is clear that in many cases groups of genes are associated with certain conditions. I have been developing methods to better understand gene pathways via co-expression analysis. One important objective of this line of research is to reduce the dimensionality of the data and to obtain a set of genes that represent a small set of gene clusters. These eigengenes are then used to perform differential expression analysis. I developed a gene network analysis method which addresses four important issues that arise in existing approaches. Note that a gene network corresponds to a very large adjacency matrix in which element (m, n) represents the co-expression levels of genes m and n. The first of these four issues is that a large number of gene pairs are truly uncorrelated, but their sample correlation is non-zero. Thus, the complete graph includes a very large number of spurious correlations, which may lead to incorrect module assignments. Second, the adjacency matrix of the complete graph is not sparse, which increases the computational complexity of the network analysis. Third, co-expression measures, such as the correlation coefficient, can only encode direct relationships between pairs of genes, while indirect relationships through other genes are ignored. Fourth, approaches focusing on modules and eigengenes or hub genes are suitable for modular or hierarchical networks where nodes within a module are highly connected but connections across modules are relatively rare. However, biological networks such as protein-protein interaction and gene interaction networks may have other network features where modules cannot be clearly partitioned. I use a random effect mixture model in order to identify spurious correlations and remove the corresponding edges from the graph. The model yields shrinkage estimation, which leads to increased power to detect significant connections between pairs of genes, and to a sparse adjacency matrix. The truncated matrix is used to compute the 'average commuting times' as distances, which account for indirect connections between genes. It also allows to use random walk models in order to analyze graph-theoretic properties of nodes and of the network as a whole, as well as to test for structural differences between networks obtained from different populations. A paper with a former student (Seojin Bang) has been submitted to the Annals of Applied Statistics.

References

[1] A. P. Dempster and N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, **Journal Of The Royal Statistical Society, Series B**, 39:1–38, 1977.

[2] H. Bar and E. Schifano, *lemma: Laplace approximated EM Microarray Analysis*, **http://CRAN.R-project.org/package=lemma**.

[3] H. Bar and J. Booth and E. Schifano and M. T. Wells, *Laplace Approximated EM Microarray Analysis: An Empirical Bayes Approach for Comparative Microarray Experiments*, **Statistical Science**, 25:388–407, 2010.

[4] H. Bar and E. Schifano, *Empirical and fully Bayesian approaches for random effects models in microarray data analysis*, **Statistical Modelling**, 11:71–78, 2011.

[5] M. A. Rosenbaum and H. Bar and W. K. Beq and D. Segré and J. Booth and M. A. Cotta and L. T. Angenent, *Shewanella oneidensis in a lactate-fed pure-culture and a glucose-fed co-culture with Lactococcus lactis with an electrode as electron acceptor*, **Bioresource Technology**, 102(3):2623–2628, 2010.

[6] M. A. Rosenbaum and H. Bar and W. K. Beq and D. Segré and J. Booth and M. A. Cotta and L. T. Angenent, *Transcriptional analysis of Shewanella oneidensis MR-1 with an electrode compared to soluble Fe(III) or oxygen as terminal electron acceptor*, **PLoS ONE**, 7(2):e30827, 2012.

[7] Jiang, Xinyin; Yan, Jian; West, Allyson; Perry, Cydne; Malysheva, Olga; Bar, Haim; Wells, Mar- tin; Devapatla, Srisatish; Pressman, Eva; Caudill, Marie. *A higher maternal choline intake favorably alters placental gene expression of biological pathways related to disease risk*, **The FASEB Journal**, Vol. 25:599.5, 2011.

[8] Jiang, Xinyin; Bar, Haim Y.; Yan, Jian; West, Allyson A.; Perry, Cydne A.; Malysheva, Olga V.; Devapatla, Srisatish; Pressman, Eva; Vermeylen, Francoise M.; Wells, Martin T.; Caudill, Marie A. *Pregnancy Induces Transcriptional Activation of the Peripheral Innate Immune System and Increases Oxidative DNA Damage among Healthy Third Trimester Pregnant Women*, **PLoS ONE** 7(11): e46736. doi:10.1371/journal.pone.0046736, 2012.

[9] Jiang, Xinyin; Yan, Jian; West, Allyson; Perry, Cydne; Malysheva, Olga; Bar, Haim; Wells, Martin; Devapatla, Srisatish; Pressman, Eva; Caudill, Marie. *Pregnancy status and choline intake alter DNA integrity, epigenetic marks and gene expression.* **The FASEB Journal**, Vol. 26:116.1, 2012.

[10] Jiang, Xinyin; Bar, Haim Y.; Yan, Jian; Jones, Sara; Brannon, Patsy M.; West, Allyson A.; Perry, Cydne A.; Ganti, Anita; Pressman, Eva; Devapatla, Srisatish; Vermeylen, Francoise; Wells, Martin T.; Caudill, Marie A, *A higher maternal choline intake among third-trimester pregnant women lowers placental and circulating concentrations of the antiangiogenic factor fms-like tyrosine kinase- 1 (sFLT1)*, **The FASEB Journal**, Vol. 27, No. 3, pages 1245-1253, 2013.

[11] H. Bar and J. Booth and M. T. Wells, *A Mixture-Model Approach for Parallel Testing for Unequal Variances*, **Statistical Applications in Genetics and Molecular Biology** Vol. 11, Iss. 1, Article 8, 2012.

[12] Elizabeth L. Alexander; Susana Gardete; Haim Y. Bar; Martin T. Wells; Alexander Tomasz; Kyu Y. Rhee, *Intermediate-Type Vancomycin Resistance (VISA) in Genetically-Distinct Staphylococcus aureus Isolates is Linked to Specific, Reversible Metabolic Alterations*, **PLOS ONE**, Vol. 9, No. 5, pages e97137, 2014

[13] Bar, Haim Y.; Booth, James G.; Wells, Martin T. *A Bivariate Model for Simultaneous Testing in Bioinformatics Data* **Journal of the American Statistical Association**, June 2014, Vol. 109, No. 506, Applications and Case Studies, 2014.

[14] H. Bar and J. Booth and M. T. Wells, *An Empirical Bayes Approach to Variable Selection and QTL Analysis*, **In the Proceedings of the 25th International Workshop on Statistical Modelling, Glasgow, Scotland**, 2010.