Research Statement
Noah Simon

1 Overview

Recent and somewhat less recent technological advances in high-throughput genomics and proteomics have given biomedical researchers a wealth of data. Unfortunately, classical tools are not well suited to handle these extensive, but often very noisy, data. My statistical interests revolve around the analysis of this data and its application to biomedical issues.

My interests in this area fall predominantly into two classes: I develop simple methodological tools which allow the less statistically inclined (medical doctors, biologists, etc.) to make honest inferences about potentially complex systems. I also think there is significant unrealized promise in personalized medicine (I’m sure that I’m not alone in this belief) which I pursue by building tools for prediction and classification, by collaborating with biomedical scientists to apply these tools to the development of biomarkers, and by developing clinical trial designs which take advantage of these biomarkers.

Honest inference is incredibly important in this age of high-throughput experiments and multiple comparisons — potentially more so than ever before. The community is plagued with reproducibility issues from poor inference, low power experiments, and poor experimental conditions. While the latter two problems cannot be “fixed” at the analysis stage, their effects are compounded by the first. In light of this, simple understandable analyses are ideal; complicated methods often obfuscate and trick scientists into “finding” signal where none truly exists. Our new wealth of data provides us with many new opportunities, but it’s important to tread with caution: a small step in the right direction is better than a large step in the wrong one.

As for personalized medicine, there have already been a number of successes: Herceptin for breast cancer [Bange et al., 2001], and Vemurafenib for melanoma [Bollag et al., 2010], along with a small but growing set of others. For many diseases we aren’t at the stage where we understand the biological pathways involved well enough to make predictions. However, there are some diseases for which we have candidate biomarkers that don’t seem completely outlandish (eg. for white blood cell cancers perhaps some serum pro-inflammatory cytokines are elevated). In these situations building biomarkers and honestly evaluating them in trials is within reach. In fact often this can be done at minimal cost through collaboration, as much of the data has already been collected (often at least enough for a rough sketch of the efficacy) but has not yet been analyzed with an eye towards clinical efficacy. For cases in which we lack candidate biomarkers, investigators could use genome, proteome, or metabolome-wide technologies to develop candidates.

2 Methodological Work

While my PhD research has been conducted with these ‘-omics’ applications in mind, I’ve worked on several different projects. I’ll somewhat artificially divide my past methodological research into 3 categories: my inferential work on interactions; my prediction work with penalized models; and my work on adaptive clinical trials.
2.1 Inference for Interactions

Pairwise interactions are of much interest in genomics: among other things they can give insight into how genes act in concert to cause/prevent disease; tell us about the effect of environment on the mechanisms of genetic conditions; and help us find predictive biomarkers indicative of future response to treatment. However their analysis is also complicated: in high dimensions there are a large number of potential interactions (both a computational and an inferential issue) and the interpretation of an interaction is often unclear (they are model specific, for correlated variables an interactions may be confounded with pure quadratic terms, etc.). We propose a simple but nonstandard approach to curb these issues and analyze these interactions in Simon and Tibshirani [2012b].

In this work we take data in two classes (e.g. diseased and healthy control) and look for marginal interactions in pairwise logistic models (i.e. considering models with only two variables at a time). There is a strong connection between marginal logistic interactions, and pairs of variables with a different correlation in one class than the other. In particular, the two are equivalent if the variables are Gaussian with the same variance within each class. We propose to test for marginal interactions by instead comparing marginal correlations between classes: those pairs with a changing marginal correlation, we call significant. Since we have reduced this to a problem of comparing correlations, we simply use as a statistic the difference of the Fisher transformed correlations. Also, by permuting the mean-centered, standardized variables we can estimate a null-distribution and false discovery rates.

We give basic asymptotic results to show that even if the number of observations is only on the order of the log of the number of features, under reasonable conditions we are not running a fishing expedition (we can still asymptotically find our signal). I’ve also been working more recently to better understand the statistical properties of our FDR estimates using tools from large deviation theory.

This approach remedies a number of issues from before: Calculating correlation matrices is quick and easily parallelizable; Fisher transformed correlation matrices are well-behaved, so small asymptotic $p$-values are believable (and further remedied by permutations); and these correlation-based interactions are very interpretable.

2.2 Penalized Models

Developments in biotechnology have made prediction more important. In particular many problems in personalized medicine can be viewed as prediction problems. However, prediction with genomic data also presents new challenges. Because the number of observations/patients $n$ is generally much smaller than the number of features/genes $p$, classical modeling with maximum likelihood estimation is impossible. Fortunately we often have outside information which can help with our estimates (sparsity patterns, grouping, etc.). One common way to codify this information is through penalized regression.

In penalized regression we modify standard maximum likelihood estimation by the addition of a penalty term:

$$\min_\beta -\ell (y, X, \beta) + \lambda P (\beta)$$

where $\ell (y, X, \beta)$ is the log-likelihood of the data for some parameter vector $\beta$, feature matrix $X$
and response vector \( y \), \( P(\cdot) \) is some penalty function attempting to curb the “complexity” of \( \beta \) (the type of “complexity” is specific to the choice of penalty), and \( \lambda > 0 \) is some fixed parameter that determines the tradeoff between goodness of fit (first term) and complexity (the second term).

A lot of my past and present work has dealt with these penalized regression models: Simon et al. [2012a] and Simon and Tibshirani [2011] propose new penalized models for regression and discriminant analysis and discuss their efficacy and numerical optimization.

Simon et al. [2012a] for example allows us to encode multiple types of sparsity. Often in biological phenomena, genes act in concert through known pathways (or groupings) of interest. When we estimate regression coefficients, we may believe that few genes are indicated, and also think that these genes belong to a similar, sparse set of pathways. Our work encodes this combination group-wise and overall sparsity into a penalty function, and solves the resulting optimization problem using first order methods.

Solving for \( \hat{\beta} \) in penalized models is also a challenge. As in standard regression, the problem is still a convex function minimization, however, often there are thousands (and sometimes millions) of variables. Standard interior-point methods involve matrix inversions and are slow past hundreds of variables and often infeasible past thousands.

Some of my other work (Simon et al. [2011], Tibshirani et al. [2012], Simon and Tibshirani [2012a], Witten et al. [2011], Simon et al. [2012b]) has been more focused on the numerical optimization — a number of these have contributed to the widely used \texttt{glmnet} package in R. Making good techniques computationally tractable in high dimensions with easy-to-use implementations is important — by and large people use methods that are straightforward to apply.

For example, work such as Simon et al. [2011] doesn’t bring new methodology to bear. Instead, it is a competitive algorithm for solving the Lasso-regularized Cox model, combining coordinate descent and Quasi-Newton iterations. We implemented this algorithm in the \texttt{R} package \texttt{glmnet} and it is commonly used in the biostatistics community. While this isn’t as glamorous as new methodological work, I think it was some of my more important work in terms of impact.

### 2.3 Adaptive Clinical Trials

There has also been a recent push to use ‘omics’ during clinical trials. Randomized trials are expensive and time consuming; to save time, money, and toxicity there have been attempts to combine the develop of treatments and companion predictive biomarkers (biomarkers which predict treatment response). Standard practice in the past has been either to restrict eligibility criteria based on a pre-determined biomarker or to broadly enroll patients and use some posthoc subset analysis to find a sub-population of interest.

In Simon and Simon [2012] we propose a method to adaptively change the eligibility criteria of a clinical trial in progress based on an adaptively updated biomarker while still controlling type 1 error under the global null of no treatment effect in any sub-population. In particular, one may predefine several accrual points at which to “update” the biomarker and criteria — this breaks the data up into blocks. By combining pivotal (or asymptotically pivotal) statistics from each block, one can control the type one error of the trial for a nearly arbitrary choice of updating function (based on the performance and random treatment assignments of all patients up to that point).
Examples of statistics are given in the paper for a number of common scenarios.

I also did work on controlling type 1 error in adaptive randomization trials by rerandomization [Simon and Simon, 2011]. This work doesn’t involve biomarkers, instead it applies when standard of care treatment is very poor, and, in the face of an apparently effective new treatment, randomizing too many patients to control may be unethical. In these trials, one chooses an adaptive randomization scheme which probabilistically assigns patients to treatment arms based on the outcome and assignment of previous patients (this scheme may randomize more patients to an arm if it appears to be effective). In our paper, we propose to condition on patient outcome, and calculate a null distribution by sequentially re-randomizing treatment assignment based on the original adaptive randomization scheme — this controls type 1 error even in finite samples and with a broad class of time trends in the characteristics of patients.

2a Optimization and Computing

Many of my statistical interests give rise to methods which are computationally intensive. To me, a method is only valuable in-so-far as it can be applied to, and is effective for, actual data. As such, part of my methodological work has been developing scalable algorithms and efficient implementations.

Many of my methods can be formulated as convex optimization problems. Much of the algorithmic work has been in developing, testing, and tuning first order optimization methods to solve these problems. Unlike standard interior point methods (which are generally very slow on these problems), there is often trickery and cunning required to develop a first order algorithm for some of the more complicated problems.

I have written R libraries implementing these algorithms. To be scalable, all of the “workhouse” optimization code is written in C (infrequently even calling fortran). I’m also working to build excel interfaces for these methods to make them accessible to the broader scientific community.

I’ve also recently been working on parallelized computing with GPUs. I’ve done some coding with NVIDIA’s CUDA framework and have been trying to find where its advantages are more and less pronounced (eg. memory contraints and load times really limit its efficacy for simple lasso problems). None of this code has reached “production” yet, but certain low memory, highly repetitive algorithms look to lend themselves well to this approach (eg. testing marginal interactions).

3 Collaborative Work

Collaboration is an important aspect of statistics: it’s exciting, it motivates meaningful methodological work, and even the fanciest statistical techniques alone aren’t curing diseases. Through collaboration I’ve been able to work more hands-on on biomarker development and better understand the opportunities and many challenges associated with it.

I worked with Dr. Lauren Harshman in the cancer center on developing predictive biomarkers for platinum and gemcitabine based chemotherapy in urothelial cancer patients (Harshman et al. [2011], Harshman et al. [2012]). We tested two DNA repair enzymes, ERCC1 and RRM1, to see if they were predictive of response to these chemotherapies — we retrospectively analyzed patients
from several treatment centers. Unfortunately, we were unable to find evidence that expression levels of these enzymes correlated with response to chemo. We did verify that, of the platinum agents, cisplatin appears to be more effective than carboplatin (though without randomization, this could have been the result of confounding).

I also have an ongoing collaboration with Professor Philip Tsao in cardiovascular medicine. We’re trying to develop a biomarker to screen for abdominal aortic aneurysms (AAA). We measured cytokines in serum using Luminex bead-based immunoassays — with this technology each patient has several hundred repeat measurements for each cytokine. We’ve been working to use this cytokine data to predict which patients have AAA. Part of the challenge here has been to effectively make use of the extensive replication within patient rather than just using the median as is standard practice by luminex. While results have been modest, we have found an interesting candidate protein and it has been a nice opportunity to familiarize myself with the science behind the Luminex technology, and cytokine expression in general.

Along with these longer term collaborations, I’ve also done some shorter term consulting work: I have been a frequent participant in (as well as often the instructor of) the Stanford Statistics Consulting Clinic — a free statistical consulting service offered by our department. Problems in the clinic are varied: ranging from analyzing the voting patterns of the UN security council, to examining the effect of body size and over-encephelization on the survival of species, to analyzing the effect of invasive rabbits on plant growth in Chile. It has been a unique experience: short term consulting has kept me on my toes, given me exposure to a scientific problems in many different fields (and different statistical techniques that I don’t generally come across in my methodological work), and my work as the instructor has given me the opportunity to mentor younger doctoral and master’s students.

4 Future Work

In the future I plan to continue both collaborating with scientists and developing methodology. I’m interested in continuing to attack useful scientific problems, and I think my past methodological and applied interests will only become more relevant as biotech tools become more advanced and medicine becomes more personalized, more precise, and more effective for individual patients.

4.1 Future Methodological Work

The relevant scientific issues of the future are unknown. Since I try to develop methodology relevant to important scientific questions, it is hard to say what my future methodological work will look like. That said, there are some questions I would like to work on. Though these are largely still in the planning phase, I will give details on two specific projects.

While all of our cells have the same DNA sequence (up to random mutations), our bodies have many different cell types which produce proteins at varying levels. These heritable cellular differences which don’t involve the nucleotide sequence itself are called epigenetic differences. Methylation is an important example of epigenetics. In our DNA, often when a cytosine nucleotide is followed by a guanine (a CpG site), a methyl group is attached to the cytosine. This CpG site is considered to be “methylated” and the expression of its corresponding gene is down-regulated.
Different cells (with the same DNA sequence) may be methylated differently which, in part, accounts for their different attributes. The pattern of methylation in cells has also been seen to play a role in cancer.

In general the probability that a CpG cite is methylated is a smooth function of its positional location — this smooth function is potentially different for different cell types and is seen to be altered in some locations for certain types of cancerous cells. The Irizarry lab recently developed a microarray based method to estimate methylation across the entire methylome (with significant preprocessing to map the base fluorescence intensities to methylation proportions: Irizarry et al. [2008], Aryee et al. [2011]). They use an empirical bayes approach with these proportions to compare diseased and healthy populations across the entire methylome and find regions of differential methylation (Irizarry et al. [2009]). I would also like to attack this problem using their preprocessing tools but with a different approach to the final analysis: I plan to find sites of differential methylation with penalized regression.

Various variational methods have been used to model smooth data: smoothing splines, trend filtering, etc.. I plan to use these methods with another lasso-like penalty to look for regions of differential methylation. One major attraction of this approach is that we can encode the entire problem as a penalized regression:

\[
\min_{\theta, \gamma} \ell(Y, \Theta) + \ell(Z, \Gamma) + \lambda_1 [P_v(\Theta) + P_v(\Gamma)] + \lambda_2 \|\theta - \Gamma\|_1
\]

where \(Y\) and \(Z\) are our vectors of methylation proportions by position from the two populations; \(\Gamma\) and \(\Theta\) are the corresponding population mean vectors for these counts; \(\ell(\cdot)\) is the log-likelihood of our data \(^1\); \(P_v(\cdot)\) is the variational penalty of choice, something like \(P_v(\cdot) = \|D\cdot\|\) with \(D\) a discrete \(k\)-th difference matrix; and \(\lambda_1 > 0\) and \(\lambda_2 > 0\) are parameters which determine the tradeoff between goodness of fit (first term), smoothness (Second term), and group shrinkage (third term). By combining these 3 terms we find two smooth functions which fit our data and have sparse differences. By varying \(\lambda_2\) from \(\infty\) to 0 one can also find a nice ordering of potentially differentially expressed regions. Also, because of the form of our regression (convex and nice), one can solve this in roughly linear time (in the number of positions).

Choosing two regularization parameters is always an issue. The simplest approach, an exhaus-tive search based on cross-validation, is computationally intensive and often highly variable. In this case, however, we can initially fit the two curves separately (ie. with \(\lambda_2 = 0\)) and use cross-validation, Cp or some either information criteria to choose \(\lambda_1\). Estimating FDR is also another important concern. In this case we might use a parametric bootstrap approach: fit a null/pooled curve (ie. fit with \(\lambda = \infty\)), resample counts, and refit our two curves to get a null distribution of supposedly differentially methylated regions.

I would also like to develop a method for approximating the false discovery rate of estimated-non-zero partial correlations in gene networks. In biological systems and processes we often believe (or assume for simplicity) that while most biological features are inter-dependent, the number of direct links (or conditional dependencies) are far fewer. For example colorblindness (a sex-linked disorder) and height are correlated, but the driving factor is sex; conditional on sex, the two are essentially independent. Given a set of biological features (often gene or protein expression), we are

\(^1\)possibly a weighted sum of squares as the technology only provides us with approximate proportions, not counts
increasingly interested in finding all non-zero partial correlations between these features (creating a network of partial dependencies). Sparse inverse covariance estimation is a standard technique for estimating these links [Friedman et al., 2008]. To date there are no reliable tools for estimating false discovery rates among the “significant” partial correlations. I believe one can attain an estimate by leveraging the duality between regression and covariance estimation — in the Gaussian setting, a partial correlation is non-zero iff the corresponding coefficient in a multiple regression is non-zero. This duality is mirrored in the relationship between lasso regression and sparse inverse covariance estimation [Friedman et al., 2008]. I propose to use this duality to estimate FDR of “significant” non-zero partial correlations.

For each pair of variables one can calculate partial residuals (using the implicit “lasso” regressions from our sparse precision matrix). We can then permute these residuals take their soft-thresholded inner-product and appropriately scale it to find a null distribution of these partial correlations (or perhaps just a null estimate of the number of non-zeros). By comparing our original partial-correlations to this null distribution we can find an FDR estimate for our “significant” links. Because the lasso is biased these FDR estimates will tend to be conservative. Hopefully this conservative bias is slight, however if not there are ways to de-bias the lasso estimate.

4.2 Future Collaborative Work

On the applied side, I’m looking forward to collaborating with new doctors and scientists (there’s no shortage of interesting problems). I also have a growing interest in biomarkers as applied to autoimmune diseases. For example, the (fairly) recent development of anti-tnf antibody treatments has revolutionized the long-term prognosis of many autoimmune diseases. Unfortunately these treatments are expensive, not everyone responds, and, because these proteins are not fully humanized they have a limited length of efficacy even among responders (patients develop auto-antibodies to the treatment). These issues give rise to two scientific questions.

First, can we develop a biomarker to predict treatment response? Or, even better, determine which anti-tnf agent is best for each patient? This is not a novel idea, but from my limited experience with autoimmune diseases it doesn’t look like a lot of high-throughput experimentation is attacking the problem (people seem to be looking at the most likely culprits, eg. tnf-α expression).

Second, can we develop a biomarker that will predict which patients are most at risk of developing auto-antibodies to the treatment? As of now, many patients are given an immunomodulator (azothioprine or methotrexate) while on anti-tnf medications in an attempt to subvert the immune system and slow the development of auto-antibodies: these medications are not without risk (they are low dose chemotherapy) and one would like to minimize the number of people on them. Furthermore, if we could develop this biomarker, those patients with more severe diseases that are predicted to be at very low risk of developing auto-antibodies could use anti-tnf drugs from an early stage, and significantly curb disease progression.

I’m excited about future collaborations both within and outside of the statistics community. I plan to continue pursing methodological work that is pertinent to applied problems, analyzing interesting data, and I hope to work with scientists as they develop and apply new technologies and data-aquisition tools.
References


