LECTURE #1

LECTURER: BARBARA ENGELHARDT SCRIBE: KIRAN VODRAHALLI

Contents

1 Introduction

I am a professor in Computer Science and in Center for Stastics and Machine Learning This is joint work with Ryan P Adams from Harvard, who is on leave of absence at Twitter.

I will talk about this from the perspective of genotypes and stuff, which is what I tend to be interested in.

Definition 1.1. Genotypes are single nucleotide polymorphisms (SNPs) are changes at one position in a genome.

Definition 1.2. An SNP (single nucleotide polymorphisms) are the places where genomes are different. Humans are diploids meaning they have two copies of each genome, one from mom and one from dad.

You can do the full genome sequence of a full individual for a few thousand dollars now.

Definition 1.3. Phenotypes are characteristics of traits of individual or samples, including disease states.

Our goal is to identify associations between SNPs and phenotypes.

We are going to be talking about expression quantitative trait loci. eQTLs are highly enriched in the set of SNPs that have been found to be associated with complex traits or disease.

Each sample has hundreds of samples (n), tens of thousands of gene expression levels, tens of millions of SNPs (p).

We can target specific genes for diseases we want to control. Now we start paralleling neuroscience problems. We have so many predictors 10s of millions, but only hundreds of samples: so we need to highly regularize.

How does this relate to fMRI: SNPs are voxels, and gene expression is a task. SNPs are often well-correlated, but voxels are also well-correlated. It may even be easier because there is spatial locality in the brain, but this plays less of a role in genotypes (because of the way you inherit). The locality effects in fMRI are often measurement error rather than the spatial structure of brain activity. A lot of the interesting temporal correlations are more long range, but they can be obscured by local patterns. There is an analogy with the gene.

In any case, we are in the setting $p >> n$: When you have this situation, statistics and machine learning are not really good at handling these problems. We only recently started talking about these kinds of problems. This is why Bayesian approach may be effective and is a center of connection between neuroscience and genomics.

The question is: **How can we leverage known biological structure to gain statistical power?**

2 Bayesian Model Selection

The question is: We want to quantify association. There is a SNP not associated with a phenotype since the means of the regression model have a flat slope: Not correlated. There can be a slope as well. The assumptions we make: There is a dosage effect (number of different alleles): One dose of a minor allele gives a change of $-\alpha$, and a second one gives -2α, etc. This is our linear model assumption. Thus we can use the simplest possible linear model. We just fit standard linear regression here.

Now we will give the sparse regression notation, introduce a Bayesian model, and look at the different results for whole-genome association.

2.1 Standard linear regression

We have $y \in \mathbb{R}^n$ as a quantitative trait. $X \in \mathbb{R}^{n \times p}$ where each genotype is encoded as $0, 1, 2$. Then we try to predict $y|X, \beta, \nu \sim \mathcal{N}\left(X\beta, \nu^{-1} I_n\right)$, where $\beta \in \mathbb{R}^{p \times 1}$. We would like to shrink β since we only want to find a few things associated with our trait. $\nu > 0$ is the precision of the residual. We are looking for non-zero β values. Recoding the zeros, ones, and twos can gives you different encodings (Mendelian etc.).

2.2 Bayesian sparse regression

We want to induce sparsity: $\nu \sim G_A(a, b)$, $\beta_j \sim \mathcal{N}(0, \sigma_j^2)$, $\sigma_j^{-2} \sim G_A(a_\sigma, b_\sigma)$. This is the automatic relevance detection (ARD) prior. This is the l_1 type regularization on the β_js. The intution is if you integrate out the σ_j, you get something that looks like Student-t which gives you l_1 like behavior. This is computationally very fast since all these terms are conjugate (closed form integrals).

2.3 Two-group priors

We might want to integrate two-group sparsity, which says we will introduce sparsity with a mixture model. Each of these β_j will be assigned to one of two groups. Either it will be noise or it will be normal: $\beta_j|w, \tau^2 \sim w\delta(\beta_j) + (1-w)\mathcal{N}(\beta_j|0, \tau^2)$. You explicitly model each predictor as either noise or signal (spike-and-slab prior). Now we can pull out the posterior probability on SNP inclusion. This is desirable: In Lasso, you choose a threshold on β after you estimate. In reality, you can have true associations with very small effect sizes. If you're able to model inclusion probability separately from the effect size, you have a huge benefit. You have many more variables included at the equivalent false discover rate. Bradley Efron talked about this in 2008 and called it the zero-assumption. You're going to be able to estimate effect sizes that looks approximately zero. It's really hard to work in a hypothesis space with 2^p things.

2.4 One-group priors

The Laplace distribution (l_1) penalty, automatic relevance determination, Horseshoe prior, Generalized Double Pareto. All of these have been proposed as priors on effect sizes to be able to induce sparsity in a computationally efficient way: Closed form or variational methods to do parameter estimates very quickly.

The one-group prior has heavy tails allowing signals to escape aggressive shrinkage. We use these at the expense of not getting posterior probabilities.

2.5 Our contribution

We have straightforward Bayesian linear regression from before, but then we modify the spike-and-slab prior: $\beta \sim \mathcal{N}(0, (\nu\lambda)^{-1}\Gamma)$, where Γ is diagonal with $\Gamma_{ij} = \mathbf{1}\{\gamma_j > \gamma_0\}$, where $\gamma \sim \mathcal{N}(0, \Sigma)$, $\gamma_0 \sim \mathcal{N}(\mu_\gamma, \nu_\gamma)$, $\lambda \sim G_A(a_\lambda, b_\lambda)$.

Each axis has one gamma term. Anything below γ_0 is going to be 0. Anything above is going to be included. This is a probit prior on inclusion/ exclusion parameters.

There are other sparse regression models: Lasso, forward stepwise regression (algorithmic simplicity): Until your BIC score is lowest (Bayesian Information Criteria - model complexity tradeoff), including each excluded predictor one at a time, include the predictor which minimizes BIC score, and then refit the model. It's just a greedy algorithm. Then only include non-zero β_j.

Another idea is projection pursuit, which has a relationship to ICA. We want to find the predictor most incorporated with the residual.

3 Linkeage disequilibrium

Linkeage disequilibrium induces correlation in local SNPs. Your grandparents and parents will give you crossovers between maternal and paternal chromosomes. If you think about what this means, if there are two three recombinations, you've crossed over a handful of times. When you're nearby in this chromosomes, mutations will probably be inherited together. You're not just inheriting one, you're inheriting two. This is two haplotypes. This can cause some trouble.

The SNPs are observed variables. The fact that there is correlations in the way they're generated, we're not actually modeling the correlation. So why if we know all of the SNPs and we are trying to predict phenotype, why does the way they were mattered generated.

First of all, we are missing SNPS - they do not necessarily have the causal SNP! We care because other sparse regression models assume independence. What are we trying to do: share strength across predictors. We would like all the predictors to support the predictiveness of the model.

We also tend to think about things locally for SNPs.

3.1 Group sparsity

Let's group predictors to share penalty terms to capture the correlation structure we just described. There are definitely groups of correlated p-values, and within SNPs, they will be very highly correlated. Given a disjoint clustering (block out certain regions where SNPs are very well-correlated). Bayesian group Lasso (Normal-Gamma) is a another new thing.

The SNP $\times$ SNP correlation matrix correlation matrix reveals some blocks of correlation. We would like to encourage sparsity patterns in proportion to their correlation, not just do group-wise sparsity. We would like to move beyond disjoint groups. We would like to disambiguate between association and effect size: Some idea of the posterior probability of inclusion, not conflate small β_j and lack of association. We would like to include structural information into the model a priori to increase statistical power of these models. We would like to directly compute posterior estimates of inclusion probability.

Before we have Bayesian linear model with modified spike-and-slab model, with a specific modification: Make Σ PSD, then this looks suspiciously like a Gaussian process! Now the axis of predictors gives a threshold via γ_0. You are going to regularize the inclusion: nearby snips are included together with a probability and excluded together with a probability. Nearby here means in terms of SIMILARITY, according to the covariance matrix. We use the Pearson's correlation as a kernel. The correlation data between SNPs comes from an external reference. How you get your kernel - are you going to double-dip? Go to an external reference to compute covariance matrix and then use on the data, which theoretically should be identical. On average there are about 10000 SNPs per linear regression. So there is a 1000 genome database, and that's enough to give you $10^4 \times 10^4$ covariance matrices. You just compute Pearson correlation for the two vectors which are over $0, 1, 2$, which represent the things.

If we have smooshed our covariance matrix down to a line; then local snips (in terms of covariance matrix) are going to be included or excluded together, and we get sparse blocks.

To do parameter estimation in the model is that you integrate out the effect sizes and don't estimate those, since we don't care what they are: We care if a parameter (SNP) is included or excluded in the model.

3.2 Parameter estimation

We use elliptical slice sampling (ESS). Use the covariance matrix ellipse to identify an acceptable move with no free parameters. The idea is that when you take a step in this space, the correlations are controlled for so it mixes fast. Everything else is just Gibbs sampling.

4 Other multi-SNP approaches

The general idea is sparse regression and model averaging to identify QTLs. There are Bayesian approaches: MISA model (do MCMC very quickly, integrate out effect sizes). But for $p >> n$, so things don't work as well as people would like.

Do people ever fit interaction terms in these models $(i.e. xy)$: No. Sometimes they do conditional analysis. But if you do interaction terms, you get huge numbers of parameters: You get no statistical power after squared effects (it's not even computational here).

There's a 2014 paper with heteroskedastic terms so assume different variances are different parts.

To summarize the model is

$y|X, \beta, \nu \sim \mathcal{N}(X\beta, \nu^{-1} I_n)$, $\nu \sim G_A(a, b)$, $\beta \sim \mathcal{N}(0, (\nu\lambda)^{-1}\Gamma)$, Γ diagonal where G is a gamma prior.

5 Extending this work for fMRI analysis

We want to do faster parameter inference, Brain kernels/time kernels for locality informative similarity of voxels in the brain. How to do dimension reduction methods using structured sparsity? Connect dimension reduction methods with regression and prediction. Find meaningful low-dimensional representations by using dimension reduction on the sparse structured data.

Bayesian CCA from Barbara - you can do something with CCA and think of things in terms of different modalities. Find the covariance specific to the genotypes, find gene expression, then find structure common to both matrices (genotypes and gene expression data). This is where the real excitement is: explore a sparse space where genotypes are regulating very well: Think about networks, pleotropy, and figure out which signals are related to each other using these structures. Shared latent structure in canonical correlation with sparsity priors on top: fMRI, fMRI + EEG, etc.

CCA seems a lot like RSA (Representation Similarity Analysis).

The idea of interactions: What kind of non-additive interactions go on in fMRI? Some work using a mind coefficient: nonlinear relationships, not use Pearson correlation; look for spatial structure in the scatter plot.

Impose grids of different granularity, look for topological features of different scales; find voxel relationships that don't show up in Pearson correlation. There are very simple cases where you go back to main effect and interaction point: Find correlations as the basic unit of analysis instead of nesting correlation analysis; pull out main thing.

You're looking at second order things. When you do CCA, you get all of that: If you integrate over factors you can get, you can describe regularized correlation matrix. You get either shared covariation or modality specific correlation. Think about factor analysis and Bayesian CCA in terms of a regularized covariance estimate.

Some work has been done by looking at over the squared voxel set - they quickly ran into a problem where we had too much p and too little n, so we could not discriminate between how much we were overfitting and how much was actually signal. Performance tended to stay constant. There are ways to estimate low-rank approximations of covariance matrices.

One difference between Pillow and Engelhardt: You don't have to pre-specify groups in this, but you do in Group Lasso. Not imposing that a priori is a great strength: You are using an extra set of co-occurrence of genes to impose a prior structure: Which snips are likely to be in or out together. We have used no additional information; two voxels are either likely to be in the model or out. 3D distance does not reflect likeilhood to be within same functional unit. Learning about group structure; a lot of people try to flatten brains. There's the anatomy, spatial auto-correlation in the anatomy. Also there is a functional spread because of the BOLD response, measurement error, and so on.

A big project would be a brain kernel: It's a fundamental thing.

Finally, connecting back with dimension reduction methods in regression and prediction. Can we do dimension reduction in much higher dimensional setting: reduced rank regression problems. How do we use structured high dimensional data in a nice, well powered way.