Hierarchical models and shrinkage

Patrick Breheny

February 8

Introduction

- In this lecture, we will take a step back from how to assess significance in large-scale testing, and discuss an important consideration in performing the tests themselves
- Specifically, the collection of data concerning a large number of similar hypotheses allows for the possibility of borrowing information across tests
- This is the concept behind hierarchical modeling
- Certainly, the use of hierarchical models is not restricted to high-dimensional data, although as we will see, the concept comes up often in this setting

Basic question

- To illustrate the concepts, we will work with a different data set today, from microbiologists at the University of Iowa
- In a liquid, certain bacteria (*V. parahaemolyticus*) swim around seeking nutrients with the aid of a polar flagellum
- On a surface, however, the same bacteria will reorganize their cellular structure on a massive scale, growing large numbers of lateral flagella and allowing the bacteria to swarm over the surface
- The basic question of interest here is: how exactly do these bacteria initiate this “swarming” response?

Isolating surface sensing genes

- A simple way to address the question would be to compare "swimmers", growing in liquid, versus "swarmers", growing on a plate
- However, there are many changes between a liquid environment and a surface environment, and many of the differences between the cell types will have nothing to do with the swarming transformation specifically
- The novel innovation at work here is that the researchers discovered how to force the bacteria into swimmer and swarmer states – i.e., to grow swarmer cells in a liquid and swimmer cells on a plate

Isolating surface sensing genes (cont'd)

- Their study thus consisted of measuring gene expression under four experimental conditions: swimmer cells grown on a plate, swimmer cells grown in a liquid, swarmer cells grown on a plate, and swarmer cells grown in a liquid
- The goal is to find genes that are specifically turned on (or off) in response to a swarmer cell growing on a plate – not just growing on a plate or just the swarmer cell type, but when the two are combined
- From a statistical point of view, this is a two-way ANOVA and we are interesting in testing for an interaction between environment and cell type

Example

Here is an example of the kind of gene we're interested in, a flagellar-specific initiation factor called LafS:

	Plate	Liquid
Swarmer1	11.29	2.41
Swarmer2	11.43	2.37
Swimmer1	2.36	2.40
Swimmer2	2.36	2.34

Of course, most differentially expressed genes are not nearly as obvious as this one

Replications and expense

- As you can see from the previous slide, there are only two replicates per experimental condition
- Obviously, it would be nice to have more, but it tends to be expensive to measure the expression of thousands of genes; this often hinders the effort to collect larger sample sizes
- The main consequence we are interested in today is the fact that we have relatively few degrees of freedom with which to estimate the variance for any particular gene

Example: Outlying variance

- For example, consider gene ModA:

	Plate	Liquid
Swarmer1	5.61	5.97
Swarmer2	5.61	5.93
Swimmer1	5.60	6.16
Swimmer2	5.60	6.19

- This gene doesn't look particularly important, and yet the test for an interaction is highly significant: $p = 0.0004$

Remarks

- The primary factor driving this highly significant result is the fact that this gene has an extremely small sample variance
- As we have just mentioned, however, this sample variance is based on a mere four degrees of freedom, raising the question: is the true variance of this gene really that small, or is this just a coincidence?
- In particular, this gene has a much smaller variance than the vast majority of genes
- Perhaps, then, it would make sense to borrow information regarding the variance from the other genes for which we have data

Notation

- Let j index the features (here, genes), $\mathbf{X}$ denote the design matrix (here, an 8×4 matrix), and $\mathbf{y}_j$ denote the measurements for the jth feature
- Suppose we are interested in estimating $\theta_j = \lambda^\top \boldsymbol{\beta}_j$
- We then have

$$\begin{aligned} \mathbb{V}(\hat{\theta}) &= \lambda^\top (\mathbf{X}^\top \mathbf{X})^{-1} \lambda \sigma_j^2 \\ &= v\sigma_j^2, \end{aligned}$$

where $v = \lambda^\top (\mathbf{X}^\top \mathbf{X})^{-1} \lambda$, $\hat{\theta}_j = \lambda^\top \widehat{\boldsymbol{\beta}}_j$, and $\widehat{\boldsymbol{\beta}}_j$ is the usual least squares estimator

Distributional results

Under the usual distributional assumptions that y_{ji} is normally distributed with mean $\mathbf{x}_i^\top \boldsymbol{\beta}$ and variance σ_j^2, we have the following classical results:

$$
\begin{gathered}
\hat{\theta}|\theta_j, \sigma_j^2 \sim \mathrm{N}(\theta_j, v\sigma_j^2) \\
\hat{\sigma}_j^2|\sigma_j^2 \sim \frac{\sigma_j^2}{d}\chi_d^2 \\
\hat{\sigma}_j^2 \amalg \hat{\theta}_j|\theta_j, \sigma_j^2,
\end{gathered}
$$

where d denotes the residual degrees of freedom; here, $d = n - p = 4$

t-tests

The distributional results on the previous slide provide us with the following result for estimation and testing of coefficients and linear combinations or contrasts for linear models:

$$\frac{\hat{\theta}_j}{\hat{\sigma}_j \sqrt{v}} \sim t_d,$$

where t_d denotes a random variable following a t distribution with d degrees of freedom

Conjugate prior for σ^2

- To stabilize the estimate of variance and borrow information across genes, we will assume a prior distribution for σ_j^2; this allows the variance of each gene to differ, but assumes some degree of similarity across genes
- For many reasons, it is advantageous here to work with the following conjugate prior:

$$\frac{1}{\sigma_j^2} \sim \mathrm{Gamma}\left(\frac{d_0}{2}, \frac{d_0\sigma_0^2}{2}\right)$$

- **Result:**

$$\frac{1}{\sigma_j^2}\Big|\hat{\sigma}_j^2 \sim \mathrm{Gamma}\left(\frac{d_0+d}{2}, \frac{d_0\sigma_0^2 + d\hat{\sigma}_j^2}{2}\right)$$

Alternate form for prior

- **Homework:**

$$cX \sim \chi^2_\nu \implies X \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{c}{2}\right)$$

- This offers the somewhat cleaner way of writing our model:

Prior: $$\frac{1}{\sigma_j^2} \sim \frac{1}{d_0\sigma_0^2}\chi^2_{d_0}$$

Posterior: $$\frac{1}{\sigma_j^2}\Big|\hat{\sigma}_j^2 \sim \frac{1}{d_0\sigma_0^2 + d\hat{\sigma}_j^2}\chi^2_{d_0+d}$$

- Intuitively, we start with a prior d_0 observations of σ_j^2 centered on σ_0^2, then collect d additional observations centered on $\hat{\sigma}_j^2$

Shrinkage estimator for σ_j

- From the result on the previous slide, it is easy to see that the posterior mean for $1/\sigma_j^2$ is

$$\mathbb{E}(1/\sigma_j^2|\hat{\sigma}_j^2) = \frac{d_0 + d}{d_0\sigma_0^2 + d\hat{\sigma}_j^2}$$

- This implies the estimator

$$\tilde{\sigma}_j^2 = \frac{d_0\sigma_0^2 + d\hat{\sigma}_j^2}{d_0 + d}$$

- This estimate is a weighted average of the prior and sample means, with d_0 and d providing the weights

Moderated t-statistic

- This, in turn, implies the following test, a modified version of the classical t-test:

$$\frac{\hat{\theta}_j}{\tilde{\sigma}_j\sqrt{v_j}} \sim t_{d_0+d}$$

- Note that there are two changes here:
  - The variance has been shrunken towards a common variance σ_0^2
  - The degrees of freedom have increased from d to $d + d_0$

Estimation of hyperparameters

- One thing remains: we don't know σ_0 or d_0
- In Bayesian terminology, σ_0 and d_0 are called hyperparameters (parameters that govern the distribution of other parameters)
- A fully Bayesian approach would, of course, specify priors for σ_0 and d_0
- We will take an empirical Bayes approach today, calculating estimates for σ_0 and d_0 and plugging them where they are needed to perform the moderated t-test

Method of moments estimator

- A relatively simple estimator can be obtained using a method of moments approach on the log scale:

$$z_j = \log \hat{\sigma}_j^2$$

- The distribution of z_j is roughly normal with known (albeit slightly complicated) expressions for the mean and variance
- We'll skip the details, but the main idea is that
  - The mean of the z_j values allows us to estimate $\log(\sigma_0^2)$
  - The variance of the z_j values allows us to estimate d_0, with larger variance implying smaller d_0 and vice versa

R code

- The empirical Bayes approach we have laid out here is implemented in an R package called `limma`
- There are three main functions of interest to us:
  - `lmFit`: Fits the OLS models; here, the rows of `Y` represent features, and `X` is the design matrix

    ```
    fit <- lmFit(Y, X)
    ```

  - `eBayes`: Does all the shrinkage estimation and moderated t-tests

    ```
    eb <- eBayes(fit)
    ```

  - `topTable`: Provides a summary

    ```
    Tab <- topTable(eb)
    ```

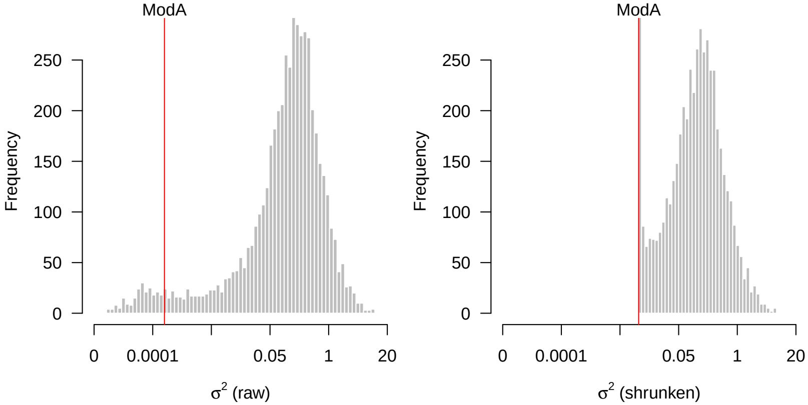
Hyperparameter estimates

- The method of moments approach described earlier results in an estimate of $\sigma_0 = 0.18$ for the prior standard deviation and $d_0 = 1.00$ for the prior degrees of freedom
- In other words, most genes have standard deviations of roughly 0.18, but there are enough differences among genes in terms of their variability that our estimate should give 80% of its weight to the observed variance and only 20% to the prior, or common, variance

Revisiting the example from earlier

- Revisiting ModA, our extremely low-variance gene from earlier, its sample standard deviation was 0.015
- Shrinking back towards the common variance by 20% results in a posterior standard deviation of 0.08 (still well under half the common standard deviation)
- Conventional t-test: $t = -10.99$, $p = 0.0004$, $q = 0.05$
- Moderated t-test: $t = -2.05$, $p = 0.095$, $q = 0.89$

Distribution of gene variances



Another example

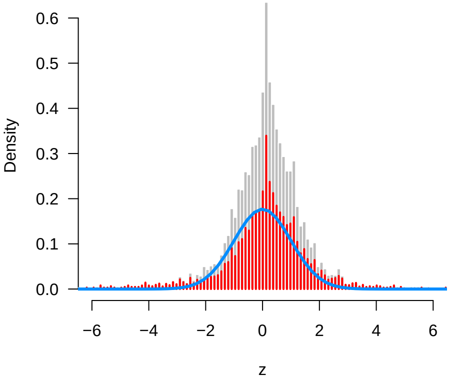
- The typical result, however, is that test results become more powerful, because we have additional degrees of freedom with which to estimate the residual variance
- For example, consider the following gene, another flagellar biosynthetic protein:

	Plate	Liquid
Swarmer1	7.12	2.90
Swarmer2	8.52	2.90
Swimmer1	2.93	2.90
Swimmer2	2.92	2.44

Another example (cont'd)

- The raw data seems fairly convincing, and yet the standard OLS test is not powerful enough to detect the interaction at a FDR of 10%: $t = -6.33$, $p = 0.003$, $q = 0.22$
- The gene is discovered, however, by the moderated t-test: $t = -6.98$, $p = 0.0009$, $q = 0.07$
- Overall, 43 genes can be identified at an FDR cutoff of 10% using the conventional test, compared to 72 genes for the empirical Bayes test

Local FDR (using ashr)



89 genes "turned on" (posterior mean > 4-fold) and only 5 genes "turned off"

Comments on lfdr

- Interestingly, local FDR estimates $\hat{\pi}_0 = 0.44$, despite the fact that only 217 genes (out of 5138) can be identified as differentially expressed (lfsr $< 10\%$)
- In general, this indicates something worth noting about power in high-dimensional testing: this experiment was certainly sufficiently powered to detect a number of interesting genes, but clearly not powered to detect all (or even a majority) of differentially expressed genes
- It is also worth noting that $\hat{\theta}_j \sim \mathrm{N}(\theta_j, s_j^2)$ probably does not hold in the case of moderated t-tests; the innovations in `limma` and `ashr` have not yet been combined

Remarks

- Prior to `limma`, a variety of ad hoc procedures were used to try to stabilize variance estimates, along with manually filtering out results that seemed like strange artifacts of unstable variance estimation
- The beauty of the empirical Bayes approach is that it provides a systematic, coherent, logical way of accomplishing all this with a minimal computational burden, since in the end, we're still performing t-tests
- Another option, of course, would be a fully Bayesian approach, although this tends to be fairly inconvenient in high dimensions, as MCMC procedures take a long time to run and lots of memory to store, and tends to give similar results

Sequencing and overdispersion

- Increasingly, many high-throughput molecular biology experiments use sequencing to measure things like gene expression, rather than microarrays
- From a statistical perspective, this means that y is now count data, and something like the Poisson or negative binomial distribution is more appropriate than the normal distributions we covered today
- All of the concepts we have talked about today still apply, though the details are more complicated – for this type of data, it is the estimation of gene-specific overdispersion parameters that is unstable and which requires borrowing information across genes