OPEN BOOK EXAM Statistics for High-Dimensional Data, 27th May, 2014

May 20, 2014

Instructions:
- Write clearly. Scribbles will not be deciphered.
- Do no give ambiguous answers.
- Finish in time!

You are allowed to use background material, including slide hand-outs, etc. Computers are not allowed. Good luck!

1. FDR
   Let $f(p)$ be the density of p-values for a given study. Often, the following mixture model is used to model $f(p)$:
   $$f(p) = p_0 f_0(p) + p_1 f_1(p) \tag{1}$$

   (a) Explain the various quantities: $p_0, p_1, f_0(p)$ and $f_1(p)$.

   (b) Why is it reasonable to assume $f_1(1) = 0$ and how can this be used to estimate p_0?

   (c) Suppose the p-values result from a *discrete* test statistic (e.g. Wilcoxon two-sample rank test). To estimate bFDR $= p_0 F_0(p)/F(p)$, often $F_0(p) = p, 0 \leq p \leq 1$ is used. Why is this conservative (overestimation of FDR) when such a discrete test statistic is used and how can this be improved?

2. Ridge regression
   A researcher is interested in the post-transcriptional regulation of an mRNA by two microRNAs. Here microRNAs are small molecules, also part of the RNA, that may affect the expression of mRNA genes. The researcher has conducted a small experiment measuring the expression levels of these three entities. The data are given in the table below.

   (a) Write down the linear regression model that explains the expression levels of the mRNA by those of the two microRNAs. In this ignore the intercept and assume that the error has mean zero and unit variance.

   (b) Derive the loss function associated with ridge penalized maximum likelihood estimation for the model of part *a)* of this question.

   (c) Optimize the loss function of part *b)* of this question with respect to the regression coefficients. In this set the ridge penalty parameter λ_2 equal to 6.

Observation	mRNA	microRNA 1	microRNA 2
1	-1	1	1
2	2	-1	2
3	0	1	1
4	1	0	-2

(d) Instead of the traditional ridge penalty, now augment the maximum likelihood loss function with the following modified ridge penalty:

$$\frac{1}{2}\lambda_2(\boldsymbol{\beta} - \mathbf{1}_{2\times 1})^{\mathrm{T}}(\boldsymbol{\beta} - \mathbf{1}_{2\times 1}),$$

where $\boldsymbol{\beta}$ is the regression coefficient vector. What is the effect of this penalty? In particular, explain how it differs from the traditional ridge penalty considered above.

3. Classification
A researcher wants to build a classifier for recurrence of a tumor from microarray data. The final classifier should be implemented on a low-dimensional platform, which measures 20 genes maximally.

(a) The researcher uses a *single* 2/3 - 1/3 training-test split of the samples. Below you find the predicted probability of recurrence given the microarray data X, $p(Y = 1|X)$, and the actual event ($Y = 1$: yes, recurrence; $Y = 0$: no recurrence). Estimate the Brier score on the test samples.

Y	0	0	0	0	0	1	1	1	1	1
$p(Y = 1\|X)$	0.36	0.43	0.21	0.48	0.41	0.67	0.73	0.52	0.98	0.87

(b) Also estimate the accuracy of the classifier above, when cut-off 0.5 is used. The researcher is very happy with this accuracy and wants to publish it. Would you recommend to do so, and if not, what extra actions with the data at hand would you recommend before drawing conclusions?

(c) When selecting patients for the study the researcher deliberately balanced the number of recurrences and the number of non-recurrence (so, the same number of microarrays for both groups). In practice, however, it is known that 80% of patients show a recurrence. What is the advantage of the strategy chosen by the researcher with respect to a strategy which would have followed the population-based prevalence of recurrence? What is the disadvantage of this strategy when estimating accuracy and how should the researcher re-calibrate this estimate?

(d) Another researcher tries to reproduce the results using a new data set. She uses the same microarray technology, the same classification procedure and DNA material from patients of the same population. Yet, the set of features selected by the classification procedure (probes on the array) overlaps little with the original set. What could be the cause of that?

4. Zero-inflation (ShrinkBayes)
Consider

$$Y_j \sim \text{ZI-NB}(p_0, \mu_j, \phi) = p_0\delta(0) + (1 - p_0)\text{NB}(\mu_j, \phi), j = 1, \ldots, n$$

(a) What are the mean and variance of this random variable? Hint: condition on whether Y_j comes from the first component of the mixture (which happens with probability p_0) or not.

(b) Why would edgeR generally result in larger estimates of ϕ than ShrinkBayes for data rows (e.g. genes) with fairly many zeros?

5. Mixture priors, empirical Bayes
Suppose $Y_{ij} \sim F(\alpha_i, \boldsymbol{\beta}_i), i = 1, \ldots, p, j = 1, \ldots, n$. Here, α is the main parameter of interest and $\boldsymbol{\beta}_i$ are the other ones. In a Bayesian context one often uses a simple conjugate prior for α_i to obtain its posterior either analytically or via computationally efficient methods.

(a) Give two reasons why, in a high-dimensional context, a mixture prior with components that are of the same parametric form as the original conjugate prior is a very useful alternative.

(b) Suppose one would use such a mixture prior for α_i. Express the posterior of α_i in terms of the mixture proportions and the results obtained under each of the mixture components.

(c) An alternative is to use a completely non-parametric prior for α_i. Suppose sample size is small and in reality a very large proportion of α_i's are close to 0. Do you think it is wise to use a nonparametric prior in such a case?