Exam: high-dimensional data analysis
February 28, 2014

Instructions:
- Write clearly. Scribbles will not be deciphered.
- Answer each main question (not the subquestions) on a separate piece of paper.
- Finish in time!

Good luck!

Question 1
A medical researchers studies the effect of microRNAs A and B on mRNA Z in terms of their expression levels. Expression data of A, B and Z are available for 100 samples. Below the result of the regression analysis (the mRNA is regressed on the microRNAs, without an intercept):

```
Coefficients:
             Estimate  Std.Error  t-value   Pr(>|t|)
microRNA A   -0.24071    0.03872   -6.217  1.24e-08  ***
microRNA B    0.72330    0.03904   18.526   < 2e-16  ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  0.6742 on 98 degrees of freedom
Multiple R-squared:  0.7779, Adjusted R-squared:  0.7734
F-statistic:  171.6 on 2 and 98 DF, p-value:  < 2.2e-16
```

Question 1a
Assume the expression levels of the microRNAs form an orthonormal design matrix. Now consider fitting the same linear model by means of ridge regression. How do the ridge estimates (corresponding to any positive penalty parameter) of the regression parameter relate to those of the unpenalized fit above? A motivated qualitative statement on the relation between the two estimates is demanded (in which the formula for the ridge estimator may be used for the motivation).

Question 1b
Still assume the orthonormality of the design matrix. How does the coefficient of determination (R^2) change with the ridge penalty parameter?

Question 1c
How would your answer to Question 1a change if the ridge penalty is replaced by the lasso penalty?

Question 1d
Relax the orthonormality assumption and allow for correlation between the expression levels of the two microRNAs. How does this change your answer to Question 1a?

Question 2
The expression levels, denoted by Y, of the genes comprising a 3-gene pathway follow a trivariate normal distribution:

$$\begin{pmatrix} Y_{1,i} \\ Y_{2,i} \\ Y_{3,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 14 & 12 & 3 \\ 12 & 16 & 4 \\ 3 & 4 & 3\frac{1}{2} \end{pmatrix} \right)$$

This model has an equivalent formulation as a system of regression equations. Calculate the values of the regression parameters of the first equation of this system, that is of the equation: $Y_{1,i} = \beta_2 Y_{2,i} + \beta_3 Y_{3,i} + \varepsilon_i$. *Hint:* recall that the explicit expression for the inverse of a 3×3 matrix

A is given by:

$$\begin{aligned}\mathbf{A}^{-1} &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}^{-1} \\ &= [\det(\mathbf{A})]^{-1} \begin{pmatrix} a_{33}a_{22} - a_{32}a_{23} & -(a_{33}a_{12} - a_{32}a_{13}) & a_{23}a_{12} - a_{22}a_{13} \\ -(a_{33}a_{21} - a_{31}a_{23}) & a_{33}a_{11} - a_{31}a_{13} & -(a_{23}a_{11} - a_{21}a_{13}) \\ a_{32}a_{21} - a_{31}a_{22} & -(a_{32}a_{11} - a_{31}a_{12}) & a_{22}a_{11} - a_{21}a_{12} \end{pmatrix}\end{aligned}$$

with $\det(\mathbf{A}) = a_{11}(a_{33}a_{22} - a_{32}a_{23}) - a_{21}(a_{33}a_{12} - a_{32}a_{13}) + a_{31}(a_{23}a_{12} - a_{22}a_{13})$.

Question 3
Consider the following list of p-values:

$$\text{plist} = (0.001, 0.003, 0.006, 0.01, 0.02, 0.15, 0.25, 0.46, 0.68, 0.79).$$

The p-values result from a PCR experiment, which is typically used after selecting genes from a microarray experiment as validation for gene expression.

Question 3a
Would you advise to use FDR or FWER here?

Question 3b
Do you think the null-hypotheses are true for all 10 genes? A qualitative argument is sufficient (no calculations).

Question 3c
Compute Bonferroni and Holm-adjusted p-values for plist. Does it help to apply Holm instead of Bonferroni?

Question 3d
Let us focus now on the 4th p-value, $p_4 = 0.01$. What is 5% lower-quantile of $T_k = \min_{i=k}^{10} p_{0i}$, for $k = 4$, where p_{0i} are independent p-values following the null-hypothesis? Hint: note that the null-distribution of T_4 is the same as that of $T' = \min_{i=1}^{7} p_{0i}$. Definition $\alpha * 100\%$ lower quantile q of continuous random variable X: value of q such that $P(X \leq q) = 0.05$.

Question 3e
The correlation between the 10 items corresponding to the 10 p-values is quite high (and positive). What may be the consequence of this when using the Westfall & Young permutation Holm-equivalent procedure instead of the ordinary Holm procedure?

Question 4
A researcher is interested in differentially expressed genes between normal tissue and pre-cursor (Ned: 'voorloper') lesion (which is a lesion that *could* lead to cervical cancer). Genomic differences between these two conditions are likely to be subtle. A sequencing experiment is designed to detect genomic differences between the two conditions.

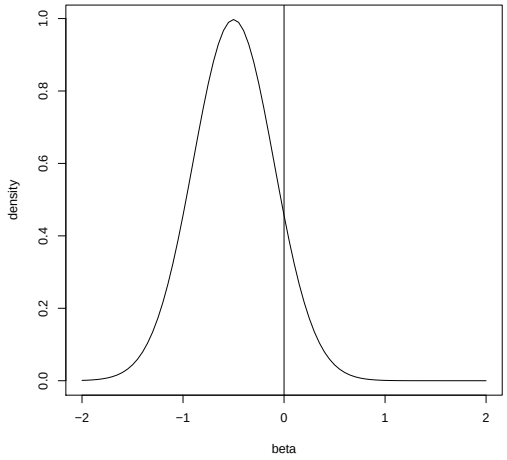

Figure 1: Prior density of log-fold change parameter β between two conditions

Question 4a
The researchers sequences 15 individuals from a retrospective cohort twice: once the healthy, normal cervical tissue (which was stored in the freezer) and once the pre-cursor lesion of the same individual. At this point no other covariates are considered important. Formulate the model for analyzing this data.

Question 4b
Figure 1 shows the prior found by ShrinkBayes for the (log-fold) difference between the two conditions. Observing this prior the researcher doubts whether the analysis has been performed correctly. Why?

Question 4c
The experiment was performed in two batches. Batch 1 contains sequencing experiments of 10 normal tissues and 5 precursor lesions and batch 2 contains sequencing experiments of 5 normal tissues and 10 precursor lesions. Give a simple estimate of the mean batch effect (mean over all genes) using Figure 1 assuming that the mean effect between conditions equals 0.

Answer to question 1
Question 1a
The ridge estimate of the regression coefficient is always smaller than that of the unpenalized fit. Hereto observe that the loss function is the sum of two loss functions: one for each parameter. Each loss function is sum of sum-of-squares and penalty. Optimization of this loss function (for any positive penalty parameter) balances the increase of the sum-of-squares by deviation from the OLS estimator and the decrease of the penalty by reduction of the regression parameter. In particular, with an orthogonal design it decreases monotonically to zero. All this may be also deduced from the explicit expression of the ridge estimator.

Question 1b
Smaller regression coefficients, less variance explained. But the total variance in the response is unaffected by the fitting method. Hence, the coefficient of determination decreases with an increasing lambda.

Question 1c
No, the argument is unaffected.

Question 1d
A monotonuous decrease of each regression parameter estimate in lambda is no longer warranted.

Answer to question 2
For starters: $\det(\mathbf{\Sigma}) = 200$ and:

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 0.20 & -0.15 & 0.00 \\ -0.15 & 0.20 & -0.10 \\ 0.00 & -0.10 & 0.40 \end{pmatrix}$$

This means directly means $\beta_3 = 0$. Furthermore:

$$\mathrm{Var}(Y_1, \,|\, Y_2, Y_3) = \mathrm{Var}(Y_1, \,|\, Y_2) = \mathrm{Var}(Y_2, \,|\, Y_1) = \mathrm{Var}(Y_2, \,|\, Y_1, Y_3).$$

Then,

$$\begin{aligned} \beta_1 &= [\mathrm{Var}(Y_1, \ , \,|\, Y_2, Y_3)/\mathrm{Var}(Y_1, \,|\, Y_2, Y_3)]^{1/2} \rho(Y_1, Y_2 \,|\, Y_3) \\ &= \rho(Y_1, Y_2 \,|\, Y_3). \end{aligned}$$

Rests to obtain the partial correlation from the partial correlation matrix, which equals:

$$\begin{pmatrix} 1.00 & 0.75 & 0.00 \\ 0.75 & 1 & * \\ 0.00 & * & 1 \end{pmatrix}.$$

Thus, $\beta_1 = 0.75$.

Answers to Question 3
The answers to this question are available in the handout containing the solutions for the multiple

testing exercises.

Answers to Question 4

Answer to Question 4a

$$Y_{ijk} \sim \text{ZI-NB}(\mu_{ijk}, \phi_i, w_i),$$
$$\log(\mu_{ijk}) = \beta_{i0} + \beta_{i1} X_{1jk} + \gamma_{ij},$$
$$\gamma_{ij} \sim N(0, \tau_i^2)$$

Here, i : gene, j: individual, $k = 1, 2$: distinguishes the two measurements on each individual (say $k = 1$ corresponds to normal tissue). Moreover, ϕ_i: overdispersion, w_i: zero-inflation (optional). Finally, X_{1jk} : group indicator, equals 0 for $k = 1$, and 1 for $k = 2$. Between individual random effect is modeled using γ_{ij} and Gaussian prior. It accounts for the pairing.

Answer to Question 4b

The prior is not centered around zero. This might indicate lack of appropriate normalization or the presence of a batch effect not accounted for (unless one expects a a large global difference across features between the two groups).

Answer to Question 4c

First observe that

$$\text{mean}_i\{\text{mean}_j[\log(\mu_{ij2}) - \log(\mu_{ij1})]\} \approx \hat{\bar{\beta}}_1 = -0.5,$$

as observed from the Figure.

In a model with batch effects: add $\beta_{i,B} X_{2jk}$ to the model where $X_{2jk} = 0$ when measurement (j, k) belongs to batch 1 and 1 for batch 2. Then we have:

$$\text{mean}_i\{\text{mean}_j[\log(\mu_{ij2}) - \log(\mu_{ij1})]\} = \text{mean}_i\{(10\beta_{i,B} - 5\beta_{i,B})/15\},$$

because $\bar{\beta}_1 = 0$ and β_{i0} and γ_{ij} cancel as well. Then, $1/3\hat{\bar{\beta}}_B = -0.5$, so $\hat{\bar{\beta}}_B = -1.5$.