Exam: high-dimensional data analysis
January 20, 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 researcher is interested in the post-transcriptional regulation of an mRNA by two microRNAs. The researcher has conducted a small experiment measuring the expression levels of these three entities. The data are given in the following table:

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

Question 1a
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.

Question 1b
Give the loss function associated with ridge penalized maximum likelihood estimation of the regression coefficients for the model of part *a)* of this question.

Question 1c
Optimize this loss function with respect to the regression coefficients. In this set the ridge penalty parameter λ_2 equal to 6.

Question 1d
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.

Question 1e
Replace in the loss function of part *b)* of this equation the traditional ridge penalty by the modified one of part *d)* of this question. Now find the 'modified ridge' penalized maximum likelihood estimate of $\boldsymbol{\beta}$ when $\lambda_2 = 6$.

Question 1f
Write down lasso analogue of the loss function employed in part *e)* of this question, and find its optimum for $\lambda_1 = 6$.

Question 2
Consider a p-gene pathway. Gene expression data on each gene in the pathway are available from

an observational study involving n samples. Assume these data from the pathway can be modeled by a multivariate normal distribution.

Question 2a
When uncovering the conditional independence graph underlying the pathway, one may exploit the link between regression coefficients and partial correlations. Hence, irrespective of whether one directly estimates the (inverse of the) covariance matrix or uses a linear regression approach, the same conclusion is reached. Discuss whether you think the two approaches still give identical results in a high-dimensional setting ($p > n$).

Question 2b
The topology of the conditional independence graph is known: it is known which genes in the pathway interact. For convenience, you may now assume $n > p$. Explain (in words) how you would estimate the covariance matrix of the expression levels of the p genes, taking into account the known structure of the conditional independence graph.

Question 3
A biologist detected 1000 significant genomic features in a 10 vs 10 comparison using a Benjamini-Hochberg FDR threshold $t = 0.1$. Now she aims at validating those results. She uses a new type of high-resolution microarray which measures 3 times as many genomic features as the the one used for the original experiment, including all the genomic features of the original one. She selects 20 (10+10) new samples from exactly the same population as the 20 original ones, follows exactly the same laboratory protocols and performs the same analysis as before. She is surprised to find only 300 significant features.

Question 3a
Argue what could have caused this apparent loss of power.

Question 3b
Assume now that the new experiment is only a partial confirmation experiment: she has means to run a third, final validation experiment with larger sample sizes on all features that are significant according to this 2nd experiment. Advise her on how to perform (p-value based) multiple testing correction on this 2nd experiment.

Question 3c
Now assume that the 2nd experiment is really the final validation experiment. Advise her on how to perform the (p-value based) multiple testing correction in this case.

Question 4
A researcher has performed an RNAseq experiment and now wishes to analyze the data. The following factors have to be accounted for: I) the main comparison of interest is between 2 groups of 5 individuals; II) the experiment contains four repeated measurements for 10 individuals; these repeats are all within one group; III) the experiment is done in two batches, equally spread over groups and individuals.

Question 4a

Write down the full model that you would like to use to analyse this data.

Question 4b
For the parameter that codes for the group difference, say β_i^g (i denotes genomic feature), the researcher wishes to test $H_{0i} : |\beta_{ig}| \leq 0.5$. He decides to use a Gaussian mixture prior with three components for this instead of a simple Gaussian. Why is this is a good choice?

Question 4c
With INLA/ShrinkBayes we obtain fits under each of the three Gaussian components separately, and hence also posterior probabilities like: $\pi(|\beta_i^g| > 0.5|Y_i, C_k)$ which denotes the posterior tail-probability under the kth component of the mixture prior. Show how we can compute the posterior probability of interest, $\pi(|\beta_i^g| > 0.5|Y_i)$ from these probabilities, the prior and other results from INLA.

Question 4d
It turns out that the RNAseq experiment contains a lot of noisy genomic features that are biologically not interesting and which can a priori be excluded. Is this a valid and wise approach and what consequences would it have for the Gaussian mixture prior?

Question 4e
The INLA computations take a lot of time. The researcher has all the results from the original prior. Show how to recalculate the posteriors without applying INLA again.

Answer to question 1

Answer to question 1a

Let Y_i, $X_{i,1}$, $X_{i,1}$ and εi be random variables representing the expression levels of the mRNA, microRNA 1, microRNA 2, and the error in sample i. The linear regression model is

$$Y_i \quad = \quad \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i \qquad \text{for } i = 1, \ldots, n,$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\text{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = 0$ if $i_1 \neq i_2$.

Answer to question 1b

The ridge loss function is:

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{2}\lambda\|\boldsymbol{\beta}\|_2^2 \quad = \quad \sum_{i=1}^{n}(Y_i - \beta_1 X_{i,1} - \beta_2 X_{i,2})^2 - \frac{1}{2}\lambda(\beta_1^2 + \beta_2^2).$$

Answer to question 1c

Equate the derivative w.r.t. $\boldsymbol{\beta}$ to zero and obtain the estimating equation:

$$-\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} \quad = \quad \mathbf{0},$$

which has the solution:

$$\boldsymbol{\beta}(\lambda) \quad = \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

In this:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} \quad = \quad \begin{pmatrix} 3 & 0 \\ 0 & 10 \end{pmatrix} \qquad \text{and } \mathbf{X}^{\mathrm{T}}\mathbf{Y} \quad = \quad \begin{pmatrix} -3 \\ 1 \end{pmatrix}.$$

Thus:

$$\hat{\boldsymbol{\beta}}(\lambda) \quad = \quad \begin{pmatrix} 9 & 0 \\ 0 & 16 \end{pmatrix}^{-1} \begin{pmatrix} -3 \\ 1 \end{pmatrix} \quad = \quad \begin{pmatrix} -1/3 \\ 1/16 \end{pmatrix}.$$

Answer to question 1d

It shrinks the regression coefficient to 1 as $\lambda \to \infty$. The penalty now includes a target other than zero.

Answer to question 1e

The estimating equation now changes to:

$$-\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta} - \lambda\mathbf{1}_{2\times 1} \quad = \quad \mathbf{0}_{2\times 1}.$$

Solving for $\boldsymbol{\beta}$ yields:

$$\boldsymbol{\beta}(\lambda) \quad = \quad (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y} + \lambda\mathbf{1}_{2\times 1}).$$

Answer to question 1f

First note $\|\boldsymbol{\beta} - \mathbf{1}_{2\times 1}\|_1 = \lambda_1|\beta_1 - 1| + \lambda_1|\beta_2 - 1|$. Apply the transformation-of-variables: $\gamma_1 = \beta_1 - 1$ and $\gamma_2 = \beta_2 - 1$. The loss function then becomes:

$$\|\tilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 - \lambda_1\|\boldsymbol{\gamma}\|_1,$$

where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{1}_{2\times 1}$. As the design matrix $\mathbf{X}$ is orthogonal, the loss function may be optimized w.r.t. γ_1 and γ_2 separately. The rest is analogous to the lecture notes.

Answer to question 2

Answer to question 2a

Both penalized estimates (regression coefficient and covariance) are baised. As loss functions and penalties are different, their bias may also be rather different. It is thus unclear whether this will uphold.

Answer to question 2b

Estimate the (inverse) covariance matrix by means of maximization of the log-likelihood augmented with a lasso-type penalty. In this penalty each parameter of the precision matrix has its own penalty parameter. This parameter equals zero is the corresponding edge is in the topology, and infinity if it is not.

Answers to Question 3

Answer to Question 3a

The new features included may contain much less differential signal than the existing ones. When using FDR, the significance of each single gene depends also on the signal of the others. Even when the p-values of the 1000 would exactly reproduce the FDR in the new experiment at the threshold used for the first experiment might be higher because V, the number of false discoveries increases proportionally with the number of tests, whereas R, the total number of discoveries, increases much less than proportionally. Hence, a smaller p-value threshold is required which potentially leads to fewer findings.

Answer to Question 3b

It would suffice to only test the 1000 genes that were detected in the first experiment, and apply BH-FDR, because further validation will be available.

Answer to Question 3c

Apply an FWER criterion. Either Bonferroni (or Holm) for simplicitly, or Westfall-Young permutation when high correlations are expected.

Answers to Question 4

Answer to Question 4a

$$Y_{ijk} \sim \text{ZI-NB}(\mu_{ijk}, \phi_i, w_i),$$
$$\log(\mu_{ijk}) = \alpha_{i0} + \alpha_{i1}X_{1j} + \alpha_{i2}X_{2jk} + \beta_{ij},$$
$$\beta_{ij} \sim N(0, \tau_i^2)$$

Here, i : gene, j: individual, k : repeat. Moreover, ϕ_i: overdispersion, w_i: zero-inflation (optional). Finally, X_{1j} : group indicator, equals 0 if individual j belongs to group 1, and 1 otherwise. X_{2jk} : indicator for batch, 0 for batch 1, 1 for batch 2. Between individual random effect is modeled using

β_{ij} and Gaussian prior.

Answer to Question 4b
Because such a mixture may better discern the negatively and positively expressed genes from the non-expressed ones. Hence, power may increase. In addition, a mixture allows for asymmetry between negatively and positively expressed genes (while maintaining the mean differential expression at 0), a simple Gaussian does not.

Answer to Question 4c
(Note that β_i^g corresponds to α_{i1} in the model above)
First:

$$\pi(|\beta_i^g| > 0.5|Y_i) = \sum_{k=1}^{3} P(C_k|Y_i)\pi(|\beta_i^g| > 0.5|Y_i, C_k).$$

Then:

$$P(C_k|Y_i) = P(Y_i|C_k)P(C_k)/\sum_k P(Y_i|C_k)P(C_k),$$

which are all available; $P(Y_i|C_k)$ as marginal likelihood from the fits under the separate model components and $P(C_k)$ from the prior.

Answer to Question 4d
Yes, as long as the removal is done a priori. It is wise because it may lead to a prior with better separated mixture components (smaller standard deviations because the noisy features were removed), which may improve power for the relevant features.

Answer to Question 4e
Compute it from the old posterior, the new prior and the old prior:

$$\pi_{\text{new}}(\beta_i^g|Y_i) \propto \pi_{\text{old}}(\beta_i^g|Y_i)\pi_{\text{new}}(\beta_i^g)/\pi_{\text{old}}(\beta_i^g).$$