Questions Statistics for High-Dimensional Data

Mark van de Wiel and Wessel van Wieringen

1 Principle Component Analysis

1. We have 2xn measurement (e.g. measurements on two genes only).

   (a) First consider 2 uncorrelated measurements with covariance matrix

   $$\Sigma_0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

   What are the weights of the two genes in the first principle component for this data. No computations needed!

   (b) Next, consider covariance matrix:

   $$\Sigma = \begin{pmatrix} 2 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

   Find the weights of the two genes in the first principle component for this data.

   (c) Now change the covariate matrix to

   $$\Sigma' = \begin{pmatrix} 2 & 0.8 \\ 0.8 & 1 \end{pmatrix}.$$

   What do you expect to happen with the weights and verify your expectation?

2 Shrinkage (1)

Consider a high-throughput experiment in which a p-dimensional gene expression profile of 4 independent samples, equally distributed over two groups, has been determined using microarrays. Let the random variable Y_{ij} denote the expression level of gene j in sample i and X_i the group indicator for sample i. Assume the Y_{ij} (for all i and j) are independent and distributed as $Y_{ij} \,|\, X_i \sim \mathcal{N}(X_i, 1)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where the group levels are coded as `-1` and `1`.

In the remainder we consider fitting the linear regression model $Y_{ij} \,|\, X_i = \beta_j X_i + \varepsilon_{ij}$ for all genes. For this regression model the unbiased estimator is $\hat{\beta}_j = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}_j$, where $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ and $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, Y_{3j}, Y_{3j})^T$. As $\mathrm{Var}(\varepsilon_{ij}) = 1$, we have $\mathrm{Var}(\hat{\beta}_j) = (\mathbf{X}^T\mathbf{X})^{-1}$ for all j.

1. Show the Mean Squared Error (MSE) of $\hat{\beta}_j$ equals $\frac{1}{4}$.

2. Define the shrunken estimator: $\hat{\beta}_j(\lambda) = \hat{\beta}_j - \lambda(\hat{\beta}_j - \beta_{\text{target}})$, where $\lambda \in [0, 1]$ and $\beta_{\text{target}} = 1$. Calculate the expectation of the shrunken estimator $\hat{\beta}_j(\lambda)$.

3. Calculate the variance of the shrunken estimator $\hat{\beta}_j(\lambda)$.

4. Which λ minimizes the Mean Squared Error of the shrunken estimator $\hat{\beta}_j(\lambda)$? Explain your answer.

5. Instead of $\beta_{\text{target}} = 1$, we now set $\beta_{\text{target}} = \frac{1}{p}\sum_{j=1}^{p}\hat{\beta}_j$. Would this change your answer to Question 1d? Motivate your answer.

3 Analysis of high-dimensional count data

3.1 edgeR

1. What are the correct parametrizations $f(\mu, \phi)$ and $g(\mu, \phi)$ to arrive at the negative binomial distribution from the Poisson-Gamma? Tip: Equate the first 2 moments of the Poisson-Gamma to those of the Negative Binomial.

2. Show that $Z_i = \sum_{j=1}^{n} Y_{ij}$ is a sufficient statistic for μ when $Y_{ij} \sim \text{NB}(\mu, \phi)$.

3. Load the `d10000.Rdata` data set, available from: `http://www.few.vu.nl/~mavdwiel/HDDA/d10000.Rdata`
   Rename the feature names by entering:

   ```
   rownames(d10000$counts) <- sapply(1:10000,function(i) paste("Tag",i))
   d10000
   ```

   Group information is available in `d10000$sample`, enter:

   ```
   group <- d10000$samples$group group
   ```

   (a) Mimic the analysis from the demo to find the 50 most significant differentially expressed tags. Note that this is a simple 2-group study with no additional covariates. In this case you do not need to apply `estimateGLMCommonDisp()` (data were already normalized). Compare the results with those from the Wilcoxon analysis (`apply wilcox.test()`).

   (b) In the simple two-group setting (no covariates), the data can also be analysed using the functions `exactTest()` and `topTags()`, see `edgeR` manual, page 13 & 14. Use these functions to generate a top 50 of differential tags. In this case you do not need to apply `estimateCommonDisp()` (data were already normalized).

   (c) Which method generates smaller p-values? Compare the two lists in terms of intersection. Use intersect().

4 Shrinkage (2)

Consider a coin tossing experiment with many different coins. Many of the coins are (approximately) fair, some are not. We have a 1000 coins, each tossed 6 times. The number of heads is binomially distributed with $N = 6$ and some p_j for coin j. Moreover, assume $p_j \sim B(\alpha, \beta)$, where B denotes the beta-distribution.

1. Derive an Empirical Bayes moment estimate for α and β.

2. Let X_j be the number of heads for coin j. Let M_k be the number of coins for which $X_j = k$. Suppose we have $M_0 = 35, M_1 = 68, M_2 = 152, M_3 = 287, M_4 = 215, M_5 = 170, M_6 = 73$. What are the estimates for α and β?

3. We throw a coin 1001 and observe $X_{1001} = 1$. In reality the coin is fair. What is the improvement of the estimate of p_{1001} when applying the (empirical) Bayes estimate instead of the classical one?

5 ShrinkBayes

1. Show that the posterior of a parameter under a nonparametric prior f_{np} can be computed from the posterior $\pi(\theta|\mathbf{Y}_i)$ obtained under a parametric prior f_{p}

2. Verify the result for the posterior under the mixture prior, so compute $f(\theta|\mathbf{Y}), \theta \neq 0$ and $P(\theta = 0|\mathbf{Y})$.

3. For the CAGE data, we wish to test test $H_0 : \beta_{i,\text{group1}} = \ldots = \beta_{i,\text{group5}}$.

   (a) What is a suitable prior for this null-hypothesis?

   (b) Apply the ShrinkBayes software to test H_0

   Some tips:

   - the null-hypothesis implies a null-model without the group parameters.
   - use the argument excludefornull in the ShrinkSeq function
   - use the argument finalprior=TRUE in the FitAllShrink function
   - use BFupdatePosterior to compute posteriors under the mixture prior