Expert Systems With Applications 167 (2021) 114200

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A new local covariance matrix estimation for the classification of gene expression profiles in high dimensional RNA-Seq data

Necla Kochan [a,*], G. Yazgı Tütüncü [b,c], Göknur Giner [d,e]

[a] Izmir Biomedicine and Genome Center, Izmir, Turkey
[b] Izmir University of Economics, Department of Mathematics, Izmir, Turkey
[c] IESEG School of Management CNRS, LEM, Lille, France
[d] Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, 3052, Australia
[e] Department of Medical Biology, University of Melbourne, Melbourne, 3010, Australia

ARTICLE INFO

Keywords:
RNA-seq
Gene expression
Local Covariance matrix
Classification
Quadratic Discriminant Analysis

ABSTRACT

Recent developments in the next-generation sequencing based on RNA-sequencing (RNA-Seq) allow researchers to measure the expression levels of thousands of genes for multiple samples simultaneously. In order to analyze these kinds of data sets, many classification models have been proposed in the literature. Most of the existing classifiers assume that genes are independent; however, this is not a realistic approach for real RNA-Seq classification problems. For this reason, some other classification methods, which incorporates the dependence structure between genes into a model, are proposed. Quantile transformed Quadratic Discriminant Analysis (qtQDA) proposed recently is one of those classifiers, which estimates covariance matrix by Maximum Likelihood Estimator. However, MLE may not reflect the real dependence between genes. For this reason, we propose a new approach based on local dependence function to estimate the covariance matrix to be used in the qtQDA classification model. This new approach assumes the dependencies between genes are locally defined rather than complete dependency. The performances of qtQDA classifier based on two different covariance matrix estimates are compared over two real RNA-Seq data sets, in terms of classification error rates. The results show that using local dependence function approach yields a better estimate of covariance matrix and increases the performance of qtQDA classifier.

1. Introduction

Dependence relation between random variables is one of the most commonly studied subjects in statistical data analysis. It is an important task to figure out the dependence structure of a data set and incorporate it into a statistical model in data analysis field. Generally, one can incorporate the dependence structure via covariance matrices which play an important role in multivariate statistical models, data classification, image processing, etc. A simple way to estimate the covariance matrix is to use Maximum Likelihood Estimator. However, this simple estimator may not reflect the complex dependence structures in medical and biological sciences due to the high dependence between the variables (attributes) in data sets. Hence, there have been a few recent approaches proposed for improving the covariance matrix estimation (Matteoli et al., 2010; Velasco-Forero et al., 2015) in the literature.

Caefer and Rotman (2009) developed a quasi-local covariance matrix estimation to be applied on spectral data analysis. Instead of estimating the whole covariance matrix they use the variance of neighbors surrounding the reference point and they define dependence areas. That is, the points in highly variable areas will have higher variances and the points in low variable areas will have less variances, accordingly. Similar to the approach given in Caefer and Rotman (2009), Oruc and Ucer (2009) proposed a new methodology to construct local dependence map which can identify three regions: positive, negative and zero dependence. They applied it on real medical data sets and showed that local dependence is much more informative in some instances.

Since it is known that RNA-Seq data sets are composed of many genes which are highly correlated with a high dependence degree, we claim that new samples will have an individual impact on the estimation of the covariance matrix while classifying the new samples. For this purpose, in this study, we propose a new type of covariance matrix estimate, which is called local covariance matrix, that can be implemented in qtQDA classifier. Integrating this new local covariance matrix into the qtQDA classifier improves the performance of the classifier. In this study, since the local covariance is updated for each

* Corresponding author.
E-mail addresses: necla.kochan@msfr.ibg.edu.tr (N. Kochan), yazgi.tutuncu@ieu.edu.tr (G. Yazgı Tütüncü), giner.g@wehi.edu.au (G. Giner).

https://doi.org/10.1016/j.eswa.2020.114200
Received 24 November 2019; Received in revised form 28 September 2020; Accepted 29 October 2020
Available online 2 November 2020
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

new sample observation with a newly proposed method, the classifier, qtQDA, becomes an adaptive algorithm and we call it Local-quantile transformed Quadratic Discriminant Analysis (L-qtQDA). The source code implementing the method is available on https://github.com/Necla/LocalDependence.

2. Methodology

Classification of gene expression data has become an important research area in the last decade (Algamal & Lee, 2015; Bielza et al., 2011; Huang et al., 2012). Particularly in cancer research, true classification of the sub-type of a patient with a particular cancer, leads a better predictive and a customized treatment for that patient. Therefore, classification of a patient to a cancer sub-type at gene expression level has a crucial importance. Due to the discrete structure of RNA-Seq data, classification of these kind of data is not as simple as other classification models that are proposed for microarray gene expression data sets.

There are certain number of classifiers proposed especially for RNA-Seq data in the literature (Goksuluk et al., 2019). The most recent one is qtQDA classifier proposed by Koçhan et al. (2019). Since qtQDA incorporates the dependence structure into the model, we apply qtQDA in order to compare a differently estimated covariance matrix, local covariance matrix, with the simple one used in qtQDA model. In the following section we explain the qtQDA classifier in details.

2.1. Negative binomial marginals

Suppose that we have k distinct classes and want to classify new samples into one of those k classes on the basis of m genes. Let $\mathbf{X}^{(k)} = [X_1^{(k)}, X_2^{(k)}, \dots, X_m^{(k)}]^T$ be a gene expression data matrix from kth class where $X_i^{(k)}$ is the number of reads (counts) for gene i. Assume that counts are marginally negative binomial distributed, i.e.

$$X_i^{(k)} \sim \mathrm{NB}(\mu_i^{(k)}, \phi_i^{(k)}), \tag{1}$$

where $\mu_i^{(k)} = E[X_i^{(k)}]$ and $\phi_i^{(k)}$ is the dispersion for gene i. It can be easily calculated that

$$\mathrm{Var}(X_i^{(k)}) = \mu_i^{(k)} + \phi_i^{(k)}(\mu_i^{(k)})^2.$$

If $\phi_i^{(k)}$ is different than zero then

$$\mathrm{Var}(X_i^{(k)}) = \mu_i^{(k)} + \phi_i^{(k)}(\mu_i^{(k)})^2 > \mu_i^{(k)}$$

which is consistent with known properties of RNA-seq data when there are biological replicates (readers are referred to McCarthy et al. (2012) for more details).

2.2. Quantile transformation

In order to incorporate the dependence into the model, a quantile transformation process is applied :

1. Let $\mathbf{Z}^{(k)}$ be an m-vector from a multivariate normal distribution: $\mathbf{Z}^{(k)} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}^{(k)})$, where $Z_i^{(k)} \sim \mathrm{N}(0,1)$.
2. Then transform ith component of $\mathbf{Z}^{(k)}$ into the ith component of $\mathbf{X}^{(k)}$

$$X_i^{(k)} = F_k^{-1}\{\Phi(Z_i^{(k)})\}, \tag{2}$$

where Φ is the standard normal distribution function, $\mathbf{X}^{(k)}$ is the transformed random variable and F_k is the $\mathrm{NB}(\mu_i^{(k)}, \phi_i^{(k)})$ distribution function.

Note here that each class has its own different covariance matrix which is expected to increase the performance of the classification.

2.3. Classification

Suppose we observe a new sample $\mathbf{x}^* = [x_1^*, x_2^*, \dots, x_m^*]^T$ from unknown class y^*, where $y^* \in \{1, 2, \dots, K\}$ is the class label. Using inverse of quantile transformation, we transform components of the new sample $\mathbf{x}^*$ to a new vector $\mathbf{z}^{*(k)}$ which is multivariate normally distributed with parameters $\mu = \mathbf{0}$ and $\Sigma = \mathbf{\Sigma}^{(k)}$. It is obvious that this transformation is applied for each class separately. Then, by Bayes theorem, posterior probability of $\mathbf{x}^*$ belonging to the kth class is given as

$$P\left(y^* = k|\mathbf{x}^*\right) \propto f_k\left(\mathbf{z}^{*(k)}\right)\pi_k, \tag{3}$$

where π_k is the prior probability and f_k is the density

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}\left|\Sigma^{(k)}\right|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\left(\Sigma^{(k)}\right)^{-1}\mathbf{x}\right\}, \tag{4}$$

Using Eqs. (3) and (4), the quadratic discriminant score for qtQDA can be defined as follows:

$$\delta_k(\mathbf{x}^*) = -\frac{1}{2}\left(\mathbf{z}^{*(k)}\right)^T \mathbf{\Sigma}^{(k)^{-1}}\mathbf{z}^{*(k)} + \log\pi_k \tag{5}$$

Thus we classify new sample $\mathbf{x}^*$ into one of k distinct classes which maximizes the Eq. (5).

2.4. Parameter estimation

In order to apply the model in practice, there exist some parameters to be estimated in the classification model. Now, we explain how they are estimated.

- *Negative Binomial Parameters(mean and dispersion)*. Like qtQDA, we use **estimateDisp** function in the R package **edgeR**. This function estimates mean using maximum likelihood and calculates a matrix of likelihoods for each gene at a set of dispersion grid points. Then weighted likelihood empirical Bayes method is applied to obtain posterior dispersion estimates for each gene (Chen et al., 2014; McCarthy et al., 2012).
- *Covariance Matrix*. In order to estimate the class specific covariance matrices, we apply inverse quantile transformation given in Koçhan et al. (2019). Unlike Koçhan et al. (2019), in this study we use local dependence function explained in Section 3 in order to improve the covariance matrix estimation and we call this estimation as local covariance matrix. Note here that similar to Koçhan et al. (2019), we use the R package "corpcor" to guarantee that local covariance matrix is symmetric and positive definite for downstream analysis.
- Classification Error Rate (CER). To assess the performance of the classifiers, we used Classification Error Rate, which is defined as follows:

$$\mathrm{CER} = \frac{\text{the number of misclassified samples}}{\text{the total number of samples}}.$$

3. Local dependence function

Let (X, Y) be a continuous bivariate random variable with joint cumulative distribution function $F(x, y)$ and with joint probability density function $f(x, y)$. Then the Pearson correlation coefficient between X, Y is given as

$$\rho(X,Y) = \frac{E(X - EX)(Y - EY)}{\sqrt{E(X - EX)^2}\sqrt{E(Y - EY)^2}} \tag{6}$$

Indeed, Eq. (6) is a way of measuring linear dependence between two random variables and in some researches it is called measure of association (Bairamov & Kotz, 2000; Bairamov et al., 2003). But in some cases, this strength of association between two random variables

can vary locally. In order to define a local measure of the association between two random variables (Bairamov & Kotz, 2000) proposed a new local dependency function which replaces the expectations EX and EY by conditional expectations $E(X|Y=y)$ and $E(Y|X=x)$, respectively. The Bairamov & Kotz local dependence function (Bairamov & Kotz, 2000) is given as follows:

$$L(x,y)=\frac{E(X-E(X|Y=y))(Y-E(Y|X=x))}{\sqrt{E(X-E(X|Y=y))^2}\sqrt{E(Y-E(Y|X=x))^2}} \tag{7}$$

Let $\varepsilon_X(y)=EX-E(X|Y=y)$ and $\varepsilon_Y(x)=EY-E(Y|X=x)$. Then

$$L(x,y)=\frac{\nu+\varepsilon_X(y)\varepsilon_Y(x)}{\sqrt{\sigma_X+\varepsilon_X^2(y)}\sqrt{\sigma_Y+\varepsilon_Y^2(x)}} \tag{8}$$

where $\nu=Cov(X,Y)$.

Thus, local dependence function $L(x,y)$ which represents the dependence between X and Y at any specific point (x,y) is more robust and accurate if there exists a dependence in the model.

In order to estimate the covariance matrix from the data available we need to estimate the local dependence function from the data. Therefore, Nadaraya (1964) and Watson (1964) proposed the following estimates for the regression functions $E(X|Y=y)$ and $E(Y|X=x)$:

$$A_X^{(n)}(y)=\frac{\sum_{i=1}^n X_i K\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{y-Y_i}{h_n}\right)} \quad \text{and} \quad A_Y^{(n)}(x)=\frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \tag{9}$$

where K is an integrable kernel function with short tails and $h_n\to 0$ is a width sequence tending zero at approximate rates.

Since it is given in Silverman (1986) that the optimal choice for h is

$$h_n=\left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}}\approx 1.06\hat{\sigma}n^{-1/5} \tag{10}$$

where $\hat{\sigma}$ is the standard deviation of the samples, we use Eq. (10) in order to estimate the conditional expectations. Moreover, we use triangular kernel function which is given as follows:

$$K(u)=1-|u|,\ |u|\le 1 \tag{11}$$

Using those estimates given in Eq. (9), we suggest the following estimate for local dependence function

$$L^n(x,y)=\frac{\nu^{(n)}+\left(\overline{X}-A_X^{(n)}(y)\right)\left(\overline{Y}-A_Y^{(n)}(x)\right)}{\sqrt{1+\frac{\left(\overline{X}-A_X^{(n)}(y)\right)^2}{s_X^2}}\sqrt{1+\frac{\left(\overline{Y}-A_Y^{(n)}(x)\right)^2}{s_Y^2}}} \tag{12}$$

$$=s_x s_y H^{(n)}(x,y)$$

where

$$\nu^{(n)}=\mathrm{Cov}(X,Y),$$

$$\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i,$$

$$\overline{Y}=\frac{1}{n}\sum_{i=1}^n Y_i,$$

$$s_X^2=\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2$$

$$s_Y^2=\frac{1}{n-1}\sum_{i=1}^n\left(Y_i-\overline{Y}\right)^2.$$

$H^{(n)}(x,y)$: local dependence function suggested in Bairamov and Kotz (2000).

Note here that X and Y are any two genes across samples.

4. Results

4.1. Application on real data sets

In this section, we implement qtQDA based on two different estimates of covariance matrix. For qtQDA classifier, we use R package "qtQDA" which is available at https://github.com/goknurginer/qtQDA. For the discriminant function in the qtQDA package, We use trended dispersion estimate. Then, we compare the classification error rates using two real RNA-Seq data sets. These are not only publicly available data sets but also commonly used data sets in order to test the performance of RNA-Seq classification methods.

The first data is cervical cancer data (see Witten et al., 2010). The cervical cancer data is composed of 714 microRNAs and 58 samples where 29 samples are tumor and 29 samples are non-tumor.

The second data is HapMap data (see Montgomery et al., 2010; Pickrell et al., 2010). Similar to cervical cancer data, the HapMap data also includes two groups of samples; CEU and YRI where CEU represents Utah residents with Northern and Western European Ancestry and YRI represents Yoruba in Ibadan and Nigeria, respectively. There are 91 CEU samples and 89 YRI samples with a total number of 52,580 genes.

It is known that RNA-Seq technology measures the expression levels of thousands of genes for multiple samples. However, not all genes are relevant and informative. Therefore, a gene selection technique is required not only to reduce the computing time but also to improve the classification performance. We apply edgeR pipeline to select informative genes which will be used in the classification algorithm. Basically, a likelihood ratio test (LRT) is performed in edgeR to detect differentially expressed (DE) genes between groups. After that, DE genes are sorted according to the value of LRT statistic and finally, the top m genes are used for the classification process. In our study, the top $20, 50, 100, 200, 300, 500$ DE genes are selected for both cervical cancer data and HapMap data.

After conducting gene selection procedure, we randomly split the data set into two sets: training set and test set. 70% of the data set is randomly assigned to the training and the rest 30% of the data set is assigned to the test set. Training set is used to train the classifiers and test set is used to measure the classification error rate. The whole procedure is repeated 300 times for different number of genes and the average classification error rate is computed.

4.2. Performance comparison of two different approaches

In this section, we compare and analysis the results. It is obvious to see from Table 1 that improving the covariance matrix estimate, i.e using local dependence function to estimate the covariance matrix, leads generally better results. Interestingly, for both data sets, qtQDA performs better then L-qtQDA. However, for the cervical cancer data, we obtain better performances except for 20, 60 and 200 genes selected in gene selection process. For HapMap data, we obtain better performances except for 200 and 500 genes selected in gene selection process. Overall we can conclude that L-qtQDA performs generally better than qtQDA.

In order to show the difference between the error rates of L-qtQDA and qtQDA, we extended our experiment, run the algorithm for different number of genes such as n = $10, 30, 40, 60, 70, 80, 90, 120$ and combine these results with the results given previously in the submitted manuscript. Then in order to show the significance of our new method we applied pairwise Wilcoxon Rank Sum test. The test results had been used to collect evidence whether the classification error rates are smaller for the new method than the qtQDA method. Since the p-values (p-value = 0.04538 for cervical cancer and p-value = 0.003324 for HapMap) are less than 0.05, we can conclude that the error rates for L-qtQDA and qtQDA are different and there is a significant evidence that the error rates for L-qtQDA are less than qtQDA.

Table 1
Classification error rates for cervical cancer and HapMap data sets.

Data	# of genes	qtQDA	L-qtQDA
Cervical	10	0.0822	**0.0787**
	20	**0.0367**	0.0372
	30	0.0309	**0.0294**
	40	0.0276	**0.0265**
	50	0.0280	**0.0265**
	60	**0.0206**	0.0207
	70	0.0185	0.0185
	80	0.0244	**0.0243**
	90	0.0204	0.0204
	100	0.0126	**0.0124**
	120	0.0132	0.0132
	200	**0.0117**	0.0122
	300	0.0161	**0.0159**
	500	0.0189	**0.0170**
HapMap	10	0.0146	0.0146
	20	0.0172	**0.0166**
	30	0.0079	**0.0078**
	40	0.0194	**0.0190**
	50	0.0064	**0.0057**
	60	0.0483	**0.0456**
	70	0.0542	**0.0514**
	80	0.0640	**0.0635**
	90	0.0615	**0.0611**
	100	0.0448	**0.0434**
	120	0.0430	**0.0426**
	200	**0.0120**	0.0116
	300	0.0074	**0.0073**
	500	**0.0106**	0.0109

5. Conclusion

Incorporating the true/accurate covariance matrix into the classification model is an important and crucial step particularly for cancer prediction. In this study we investigated the impact of covariance matrix estimated with the help of local dependence function on RNA-Seq data classification. This new approach assumes the dependencies between genes are locally defined rather than complete dependency. We have shown that locally estimated covariance matrix decreases the classification error rate which can have a significant impact on patients' survival. Therefore, one should take the estimation of the covariance matrix into account when it comes to classification of real RNA-Seq data sets.

In qtQDA method the classification has been done using the Pearson calculations for class specific covariance matrices whereas in L-qtQDA correlation calculations are done by using local dependence function. In Pearson calculations the output was a scalar that has been used to estimate the covariance matrices, which means that if two genes are correlated this correlation is a constant for any expression levels. On the contrary, in the new method, the output was a function that has been used to estimate the local covariance matrix, which means that the correlation between two genes is a function that variates for different expression levels. Therefore, incorporating local covariance matrix yields better covariance estimates, which can improve the classification performance.

Although the improvement in error rate is small (for qtQDA 0.019 and for L-qtQDA 0.017 when the number of genes is 500 in Cervical cancer data), in real life situations, this difference can play a crucial role, such as potentially increasing the survival of a patient for 0.2%. Considering the accuracy of the classification of a cancer patient impacts, this encourages to and embraces any improvement could have an impact on patient's survival. Moreover, in some cases, such as breast cancer, there are a number of treatment options available for patients and the effectiveness of these treatments relies on the accurate classification of a patient into one of the breast cancer subtypes. Therefore, we consider any increment as significant when it comes to classification.

Since we only used triangular kernel function and Gaussian bandwidth in local dependency calculation, we note here that different kernel functions and a different optimal bandwidth selection can also be implemented and may improve the classification performances. The only disadvantage of the L-qtQDA is that the algorithm is computationally intensive due to the estimation of the local covariance matrix. Nevertheless, we believe that this new estimation technique will be useful for classification of RNA-Seq profiles or other genomic studies.

CRediT authorship contribution statement

Necla Kochan: Conceptualization, Methodology, Software, Resources, Formal analysis, Writing - original draft, Writing- review & editing. **G. Yazgı Tütüncü:** Conceptualization, Methodology, Writing-review & editing. **Göknur Giner:** Conceptualization, Methodology, Resources, Software.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Prof. Dr. Gordon K. Smyth, Prof. Dr. Terry Speed and Luke C. Gandolfo for their support and suggestions and WEHI Bioinformatics division for using their resources. This work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK $2214/A-1059B141601270$) and by the Australian National Health and Medical Research Council (Program Grant 1054618 and Fellowship 1154970 to Gordon K. Smyth), the Cancer Therapeutics CRC, Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIIS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Algamal, Z. Y., & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, *42*, 9326–9332.

Bairamov, I., & Kotz, S. (2000). On local dependence function for multivariate distributions. *New Trends in Probability and Statistics*, *5*, 27–44.

Bairamov, I., Kotz, S., & Kozubowski, T. J. (2003). A new measure on linear local dependence. *Statistics*, *37*(3), 243–258.

Bielza, C., Robles, V., & Larrañaga, P. (2011). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, *38*, 5110–5118.

Caefer, C. E., & Rotman, S. R. (2009). Local covariance matrices for improved target detection performance. In *Proceedings of the 1st workshop on hyperspectral image and signal processing: Evolution in emote sensing (WHISPERS '09)* (pp. 1–4).

Chen, Y., Lun, A. T., & Smyth, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edger. In S. Datta, & D. Nettleton (Eds.), *Statistical Analysis of Next Generation Sequencing Data* (pp. 51–74). Springer.

Goksuluk, D., Zararsiz, G., Korkmaz, S., Eldem, V., Zararsiz, G. E., Ozcetin, E., Ozturk, A., & Karaagaoglu, A. E. (2019). Mlseq: Machine learning interface for RNA-sequencing data. *Computer Methods and Programs in Biomedicine*, *175*, 223–231.

Huang, H., Li, J., & Liu, J. (2012). Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis. *Expert Systems with Applications*, *39*(3), 2314–2320.

Koçhan, N., Tütüncü, G. Y., Smyth, G. K., Gandolfo, L. C., & Giner, G. (2019). Qtqda: quantile transformed quadratic discriminant analysis for high-dimensional RNA-seq data. *PeerJ*, *7*, Article e8260.

Matteoli, S., Diani, M., & Corsini, G. (2010). Improved estimation of local background covariance matrix for anomaly detection in hyperspectral images. *Optimization and Engineering*, *49*(4), 1–16.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., & Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, *464*(7289), 773–777.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, *9*(1), 141–142.

Oruc, O. E., & Ucer, B. (2009). A new method for local dependence map and its applications. *Turkiye Klinikleri Journal Biosta*, *1*(1), 1–8.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall/CRC.

Velasco-Forero, S., Chen, M., Goh, A., & Pang, S. K. (2015). Comparative analysis of covariance matrix estimation for anomaly detection in hyperspectral images. *IEEE Journal of Selected Topics in Signal Processing*, *9*(6), 1061–1073.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya A*, *26*(4), 359–372.

Witten, D., Tibshirani, R., Gu, S. G., Fire, A., & Lui, W.-O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology*, *8*(58), 1–14.