

qtQDA: Quantile Transformed Quadratic Discriminant Analysis for High-dimensional RNA-seq Data

Göknur Giner

September 25, 2019

Background

This work has been conducted during Necla Koçhan's visit to Smyth Lab as a PhD student. Necla is currently finalising her PhD in Izmir University of Economics in Turkey.



Background

This is a joint project with Luke Gandolfo and Gordon Smyth from WEHI Bioinformatics Division.

We also received invaluable theoretical advice and support from Terry Speed.

Classification

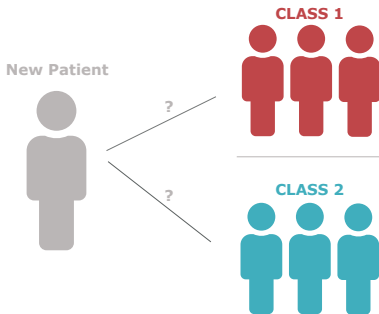
“Classification is an everyday instinct as well as a full-fledged scientific discipline” [Song et al., 2015]

Classification

“In practice, it was regarded more often as an art than a science, as there is no axiomatic classification theory that applies to all problems” [Song et al., 2015]

A note on terminology

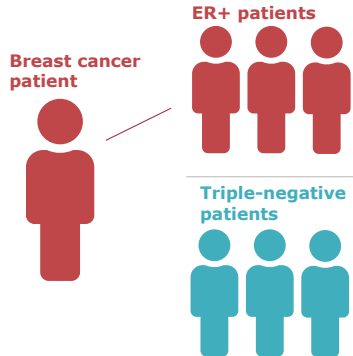
The term classification refers to **supervised** sample assignment, using features identified in a **training** set containing samples with known class labels. Sample to be assigned is **test** sample.



Classification in medicine

Accurate classification of a patient could

- help identify the subtype of the disease for patient
- suggest treatment options that are more accurately matched to the patient
- slow down the disease progression



General classification methods applied to RNA-seq data

During this project we explored several different approaches

machine learning approaches

- support vector machines (**SVMs**)
- k-Nearest Neighbour (**kNN**)

regression modelling approaches

- regularized logistic regression (**Glmnet**)[Friedman et al., 2010]

Methods developed specifically for RNA-seq data

count-based approaches

- **PLDA** (Poisson linear discriminant analysis), which models the counts using the Poisson distribution.[Witten 2011]
- **NBLDA** (Negative binomial linear discriminant analysis), which instead models the counts using the negative binomial distribution.[Dong et al., 2016]

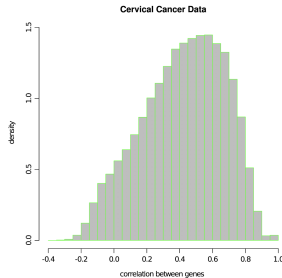
log-transformed count-based approaches

- **voomDLDA** (voom-based diagonal linear discriminant analysis), which models the log transformed counts.[Zararsiz et al., 2017]

Methods above assume the genes are statistically INDEPENDENT!

Correlation between genes is unavoidable

However, [Zhang 2017] showed that there is fairly strong (above 0.5) correlation between every pair of genes in the cervical cancer data (log-transformed), indicating that the independence assumption in PLDA, NBLDA and voomDLDA is violated.



Methods that incorporated **dependence** between genes

- **[Zhang 2017]** modeled the counts directly and developed a Bayesian approach where the covariance structure is modelled using a (multivariate) Gaussian copula. **Has no publicly available implementation and computationally intensive!**
- **SQDA** (Sparse Quadratic Discriminant Analysis)
[Sun and Zha 2015] models log-transformed counts with the multivariate normal distribution using regularized estimates of covariance matrices. **Computationally intensive!**

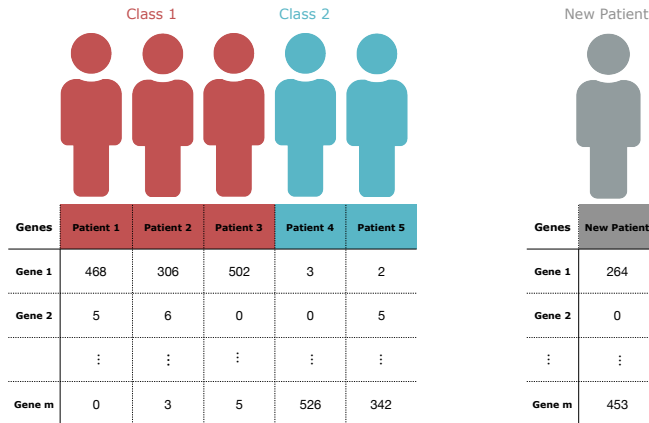
qtQDA: quantile transformed quadratic discriminant analysis

- based on counts that are negative binomial but **dependent**.
- utilizes **multivariate normal distribution** for classification.

qtQDA has two key differences from existing approaches:

- instead of modelling log-transformed counts, qtQDA models **quantile transformed counts**.
- a novel application of a powerful **regularization** technique for covariance matrix estimation.

Gene expression data to classify samples



qtQDA model assumptions

Let $\mathbf{X}^{(k)} = [X_1^{(k)}, X_2^{(k)}, \dots, X_m^{(k)}]^T$ be a random vector from the k th class where $X_i^{(k)}$ denotes the count for gene i .

We assume the counts are marginally negative binomial, i.e.

$$X_i^{(k)} \sim \text{NB}(\mu_i^{(k)}, \phi_i^{(k)}),$$

where $\mu_i^{(k)}$ and $\phi_i^{(k)}$ are the mean and dispersion for gene i .

qt: Quantile transformation

We suppose that $\mathbf{X}^{(k)}$ is generated by the following process:

- 1 Let $\mathbf{Z}^{(k)}$ be an m -vector from a multivariate normal distribution: $\mathbf{Z}^{(k)} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_k)$, where $Z_i^{(k)} \sim N(0, 1)$.
- 2 Then let the i th component of $\mathbf{X}^{(k)}$ be the transformed random variable

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

where Φ is the standard normal distribution function and F_k is the $\text{NB}(\mu_i^{(k)}, \phi_i^{(k)})$ distribution function.

Frame Title

This is a consequence of the following elementary fact from probability theory: if F and G are distribution functions, and X has distribution function F , then the transformed variable $G^{-1}\{F(X)\}$ has distribution function G (see [Lange 2010, p. 432])

Inverse quantile transformation

Suppose we observe $\mathbf{x}^* = [x_1^*, x_2^*, \dots, x_m^*]^T$ from unknown class y^* , where $y^* \in \{1, 2, \dots, K\}$.

For each class we apply the inverse of the quantile transformation to the components of \mathbf{x}^* to produce a new vector $\mathbf{z}^{*(k)}$, i.e. where

$$z_i^{*(k)} = \Phi^{-1}\{F_k(x_i^*)\}$$

The transformation from x_i^* to $z_i^{*(k)}$ is implemented by the `zscoreNBinom` function in the `edgeR` [Chen et al., 2014].

Posterior probabilities with Bayes theorem

If \mathbf{x}^* is from the k th class then $\mathbf{z}^{*(k)}$ is an observation from the $\text{MVN}(\mathbf{0}, \Sigma_k)$ distribution. The posterior probability that \mathbf{x}^* belongs to the k th class is

$$\Pr(y^* = k | \mathbf{x}^*) \propto f_k(\mathbf{z}^{*(k)}) \pi_k,$$

where π_k is the prior probability that $\Pr(y^* = k)$, and f_k is the density

$$f_k(\mathbf{v}) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} \mathbf{v}^T \Sigma_k^{-1} \mathbf{v} \right\}$$

evaluated at $\mathbf{z}^{*(k)}$.

QDA: Quadratic discriminant analysis

Since maximizing $\Pr(y^* = k | \mathbf{x}^*)$ is equivalent to maximizing $\log \Pr(y^* = k | \mathbf{x}^*)$, this classification rule entails the following (quadratic) discriminant function:

$$\delta_k(\mathbf{x}^*) = -\frac{1}{2} \mathbf{u}_k^T \mathbf{u}_k + \log \pi_k,$$

where $\mathbf{u}_k = \Sigma_k^{-1/2} \mathbf{z}^*(k)$.

Parameter estimation

We first need to **train** qtQDA to estimate three necessary parameters for this algorithm to work:

- the negative binomial parameters $\mu_i^{(k)}$ and $\phi_i^{(k)}$ to parametrize the quantile transformation.

For this purpose, we use `glmQLFit` to estimate the mean and `estimateDisp` to estimate the dispersions.

Both of these functions are implemented in `edgeR`.

Parameter estimation

The last parameter to estimate is:

- the covariance matrix Σ_k of the transformed variables, so QDA can perform with posterior probability.

However, standard estimation of the covariance matrix performs poorly [Tong et al., 2014].

The reason: the number of genes used for classification is greater or equal than the number of samples (i.e. $m \approx n$ or $m > n$).

Covariance regularization

We regularize the standard covariance estimate using the R package `corpcor` ([Strimmer 2008]).

`corpcor` is:

- 1 always positive definite and the invertible
- 2 guaranteed to have minimum mean squared error
- 3 computationally very fast
- 4 does not require any “tuning” parameters

Feature selection

Lastly, we select the more informative subset of genes using edgeR with the strategy below:

- 1 Filter genes with low expression across all samples
- 2 For each remaining gene, perform a likelihood ratio test (LRT) to test for genes differentially expressed between groups
- 3 Sort the list of genes by LRT statistic
- 4 Select the top m genes from this list

Data sets to assess the performance of qtQDA

	Source	Classes	Features
Cervical	Witten 2010	Cancer (29 samples) Non-cancer (29 samples)	714 microRNAs
Prostate	Kannan 2011	Cancer (20 samples) Non-cancer (10 samples)	Whole transcriptome
HapMap	Montgomery 2010 Pickrell 2010	CEU (60 samples) YRI (69 samples)	Whole transcriptome

Methods compared with qtQDA

machine learning approaches

- SVM (e1071)
- kNN (e1071)

regression modelling approaches

- Regularized logistic regression (glmnet)

count-based approaches

- PLDA (PoiClaClu)
- NBLDA (<http://www.comp.hkbu.edu.hk/xwan/NBLDA.R>)

log-transformed count-based approaches

- voomDLDA (MLSeq)

methods that incorporated dependence between genes

- SQDA (SQDA)

We used bootstrap to estimate the true error rate

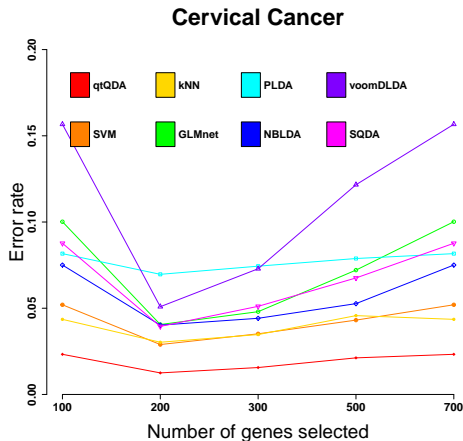
Bootstrap procedure:

- 1 Samples of each data set are randomly divided into two parts, where 70% used for training the model and 30% used to be tested.
- 2 This is repeated 1,000 times.
- 3 We estimated the error rates for $m = 100, 200, 300, 500, 700$ genes.

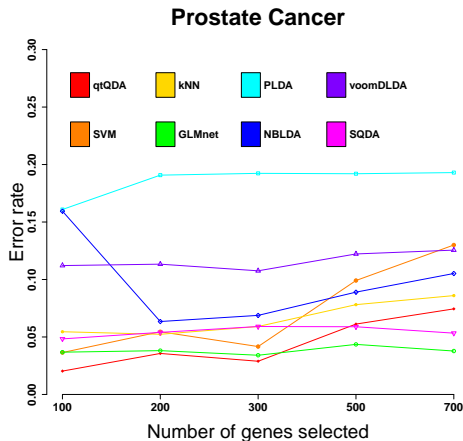
Minimum error rates

Method	Cervical cancer	Prostate cancer	HapMap
qtQDA	0.0125 (200)	0.0203 (100)	0.0018 (300)
SVM	0.0276 (100)	0.0364 (100)	0.0014 (500)
kNN	0.0277 (100)	0.0523 (200)	0.0009 (200)
GLMnet	0.0406 (200)	0.0341 (300)	0.0009 (500)
PLDA	0.0608 (100)	0.1609 (100)	0.0123 (100)
NBLDA	0.0402 (200)	0.0634 (200)	0.0058 (100)
voomDLDA	0.0425 (100)	0.1076 (300)	0.0029 (100)
SQDA	0.0318 (100)	0.0483 (100)	0.0046 (300)

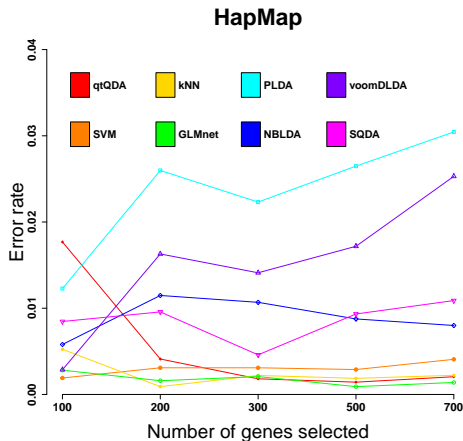
Cervical cancer data



Prostate cancer data



HapMap data



Summary

- 1 Data is negative binomial but **dependent**.
- 2 qtQDA produces multivariate normal variables performing **quantile transformation** then applies Gaussian **quadratic discriminant analysis** to estimate the posterior probabilities.
- 3 To perform the second step, qtQDA is trained by using the sophisticated **edgeR** methodology for negative binomial parameter estimation and the powerful **corpcor** methodology for regularized covariance matrix estimation.

Advantages of qtQDA

- 1 does not have any “tuning” parameters
- 2 outputs the posterior probabilities
- 3 is computationally much faster than the approaches that take gene dependence into account
- 4 is implemented as a publicly available R package

qtQDA paper and R package



<https://www.biorxiv.org/content/10.1101/751370v1>



<https://github.com/goknurginer/qtQDA>



@goknurginer




Future work

The feature selection method can be improved and this may potentially lead to even better disease classifications.




We aim at developing a sparse version of qtQDA, involving some level of regularization for features, i.e. identifying less informative features and reducing their influence to zero (e.g. like the GLMnet logistic regression classifier).

Sparse qtQDA may help identify “marker” genes that characterize different disease subclasses, i.e. variable selection.




References

-  Chen Y., Lun ATL. and Smyth G. K., *Differential expression analysis of complex RNA-seq experiments using edgeR*, In: Statistical Analysis of Next Generation Sequence Data, Somnath Datta and Daniel S. Nettleton (eds), Springer, New York, (2014) 51–74.
-  Dong K., Zhao H., Tong T. and Wan X., *NBLDA: negative binomial linear discriminat analysis for RNA-Seq data*, BMC Bioinformatics, **17** (2016), no. 369, 1–10.
-  Friedman, J., Hastie, T., and Tibshirani, R. *Regularization paths for generalized linear models via coordinate descent*. Journal of statistical software, **33(1)** (2010), 1–22.




References

-  Lange, K., *Numerical analysis for statisticians*. Springer (2010).
-  Opgen-Rhein, R. and Strimmer, K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology*, 6(1), (2007).
-  Schafer J. and Strimmer K., *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, *Stat. Appl. Genet. Mol. Biol*, **4** (2005), no. 1, 1–30.

References

-  Song, Q. and Merajver, S. D. and Li, J. Z., *Cancer classification in the genomic era: five contemporary problems*, Human genomics, **9** (2015), 1, 27.
-  Strimmer, K. (2008). Comments on: Augmenting the bootstrap to analyze high dimensional genomic data. Test, 17(1):25–7.
-  Sun, J. and Zhao, H. The application of sparse estimation of covariance matrix to quadratic discriminant analysis. BMC bioinformatics, 16(1), (2015).

References

-  Tong, T., Wang, C., and Wang, Y. (2014). Estimation of variances and covariances for high-dimensional data: a selective review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4):255–64.
-  Witten D. M., *Classification and clustering of sequencing data using a Poisson model*, *Annals Appl Stat.*, **5** (2011), no. 2, 493–518.
-  Zararsiz, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G. E., Duru, I. P., and Ozturk, A. (2017). A comprehensive simulation study on classification of RNA-Seq data. *PloS one*, 12(8).

References



Zhang Q., *Classification of RNA-Seq data via Gaussian copulas*, The ISI's Journal for the Rapid Dissemination of Statistics Research, **6** (2017), 171–183.

Acknowledgement



Necla Koçhan

Gordon Smyth

Luke Gandolfo

G. Yazgı Tütüncü

Terry Speed

Marie Trussart & Ramyar
Molania

Thanks for your time and attention! Any questions?