Badaoui *et al. J Big Data (2017) 4:32*
DOI 10.1186/s40537-017-0093-4

Journal of Big Data

METHODOLOGY Open Access

Dimensionality reduction and class prediction algorithm with application to microarray Big Data

Fadoua Badaoui[1*], Amine Amar[2], Laila Ait Hassou[3], Abdelhak Zoglat[3] and Cyrille Guei Okou[4]

*Correspondence:
fadoua.badaoui@gmail.com
[1] Département Statistique, Démographie et Actuariat, Institut National de Statistique et d'Economie Appliquée, Rabat, Morocco
Full list of author information is available at the end of the article

Abstract

The recent technology development in the concern of microarray experiments has provided many new potentialities in terms of simultaneous measurement. But new challenges have arisen from these massive quantities of information qualified as Big Data. The challenge consists to extract the main information containing the sense from the data. To this end researchers are using various techniques as "hierarchical clustering", "mutual information" and "self-organizing maps" to name a few. However, the management and analysis of the millions resulting dataset haven't yet reached a satisfactory level, and there is no clear consensus about the best method/methods revealing patterns of gene expression. Thus, many efforts are required to strengthen the methodologies for optimal analysis of Big Data. In this paper, we propose a new processing approach which is structured on feature extraction and selection. The feature extraction, is based on correlation and rank analysis and leads to a reduction of the number of variables. The feature selection, consists in eliminating redundant or irrelevant variables, using some adapted techniques of discriminant analysis. Our approach is tested on three type of cancer gene expression microarray and compared with concurrent other approaches. It performs well, in terms of prediction results, computation and processing time.

Keywords: Linear discriminant analysis, Tumor classification, Basis vectors, Kendall rank correlation, Cancer

Introduction

In recent years, technological innovations have led to a massive amount of data with relatively low cost. These massive and high-throughput data is commonly called Big Data. However, there is no universally agreed-upon definition of Big Data, but the more widely accepted explanations tend to describe it in terms of challenges that it presents. In terms of computational efficiency and time processing, Big Data motivate the development of new computational tools and data storage methods [17, 20, 27, 29]. Regarding this issue, [38] evokes three principal challenges which are related to dimensions of Big Data and which address volume, velocity and variety. Other authors have proposed additional dimensions such as veracity, validity or value [15]. The volume, one of the famous five Vs that characterize Big Data, is the main challenge that interests the statistician when

analyzing high dimension datasets. The other Big Data dimensions, interest particularly computer scientists and data investigators.

Besides challenges, Big Data give many opportunities in terms of results analysis and information extraction in different fields such as genomics and biology [30], climatology and water research [19], geosciences [44], neurology [18], spam detection and telecom [6, 13], Cyber-security [56], Software engineering [14, 40], social media analysis [37, 46], biomedical imaging [53], economics [21, 35], high frequency finance and marketing strategies [7]. The goal of using Big Data in the aforementioned fields is to develop accurate methods to predict the future, to gain insight into the relationship between the features and responses, to explore the hidden structures and to extract important common features across sub-populations.

The main problem with Big Data is still how to efficiently process it. To handle this challenge, we need new statistical thinking and computational methods. In fact, many statistical approaches that perform well for low dimension data, are inadequate when analyzing Big Data. Thus, to design effective statistical procedures for the exploration and prediction in this context, new needs will be identified, aside classical issues such as heterogeneity, noise accumulation, spurious correlations [23], incidental endogeneity [39], and [26], and sure independence screening [25], Hall and Miller [32, 33], and [12]. In terms of statistical accuracy, dimension reduction and variables selection play pivotal roles in analyzing high dimension data. For example, in high dimension classification, [48], and [22] showed that conventional classification rules using all features perform no better than random guess due to noise accumulation. This motivates new regularization methods [9, 10, 24, 54, 55].

The aim of dimension reduction procedures is to summarize the original p-dimensional data space in a form of a lower k-dimensional components subspace ($k \ll p$). To achieve this goal, statistical and mathematical theory provide many approaches. Based on frequency use, the most commonly applied methods are still principal component analysis (PCA) [2], and [34], partial least squares (PLS) [4, 5] and [45], linear discriminant analysis (LDA) [8], and sliced inverse regression (SIR) [3]. Rash Model (RM) is another recent efficient way for feature extraction which provides an appealing framework for handling high-dimensional datasets [36].

For all aforementioned considerations, and given the growing importance of alternative statistical approaches, we propose a new approach to reduce a dataset dimension, especially for classification purposes. The approach addresses the case where the number of variables p largely exceeds the sample size n ($p \gg n$), which is common in the Big Data context. To handle high dimension datasets in the prediction framework, we propose to proceed in five steps. The first three steps seek to reduce the number of variables using correlation arguments. The fourth and fifth steps consist in eliminating redundant or irrelevant variables, using adapted techniques of discriminant analysis. The performance of our approach is evaluated by measuring its accuracy of class prediction and processing time.

Before introducing a detailed description of our approach, it is worth to have a good understanding of state of the art in the concern of extraction and selection methodologies, especially for Big Data. Thus, the following section proposes to conduct a review of published studies to identify key trends with respect to the types of used methods.

Background and statistical review

The high dimension dataset can be represented by the following real-valued expression matrix

$$\begin{array}{c|ccc} \mathbb{Y} & \mathbb{X}_1 & \cdots & \mathbb{X}_p \\ \hline Y_1 & X_{11}^1 & \cdots & X_{1p}^1 \\ \vdots & \vdots & \cdots & \vdots \\ Y_1 & X_{n_1 1}^1 & \cdots & X_{n_1 p}^1 \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline Y_K & X_{11}^K & \cdots & X_{1p}^K \\ \vdots & \vdots & \cdots & \vdots \\ Y_K & X_{n_K 1}^K & \cdots & X_{n_K p}^K \end{array} \tag{1}$$

where individuals are scattered on K classes $C_1, \ldots, C_K$, n_k denotes the size of a kth class, for $k = 1, \ldots, K$ and $n = n_1 + \cdots + n_K$ is the global sample size. The objective is to explain the class membership defined by a categorical response $\mathbb{Y}$, using p variables $\mathbb{X}_1, \ldots, \mathbb{X}_p$, where X_{ij}^k is the ith value in the kth class of the variable $\mathbb{X}_j$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

For p smaller than n, classical methods of classification (LDA, PCA, ...) can be applied. In this work, we consider the case where p is much larger than n. This data structure has been used in special cases of gene expression data [11], to characterize different types of cancers [31] and the Lymphoma dataset [1].

The analysis of a high dimension dataset is primarily based on comparison of variables or observations, using a variety of similarity measures. The correlation, can be used as a measure of association between variables. To measure correlation between categorical and numerical variables, "the statistic η" can be used [28, 47] and [51, 52]. This statistic represents the ratio of variability between groups to the total variability.

In this paper, we elaborate a new approach to deal with the large dimension challenge presented by the Big Data framework. Our approach is summarized in an algorithm in five steps. The first three steps lead to the reduction of the number of columns (variables) in a dataset, the two others identify pertinent variables for building an accurate classifier. We apply our techniques to publicly available microarray datasets and compare our results with findings discussed in [36]. Our approach can clearly be used in many other areas (economy, finance, environment...etc.) where "high dimension" is a Big Data challenge.

A dimension reduction algorithm

Consider the dataset, represented by (1), of n observations and p variables with $p \gg n$. The following steps lead to a pertinent reduction of the dataset dimension p.

- `Step 1:` Calculate the correlation ratio between each variable $\mathbb{X}_j$ and nominal response ($\mathbb{Y}$) defined as:

$$\eta_j^2 = \frac{\sum_{k=1}^{K} n_k (\bar{X}_j^k - \bar{X}_j)^2}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{ij}^k - \bar{X}_j^k)^2} \tag{2}$$

where X_{ij}^k is the value of variable $\mathbb{X}_j$ measured on the ith individual belonging to the kth class, $\bar{X}_j^k$ is the mean of the restricted $\mathbb{X}_j$ to the kth class, and $\bar{X}_j$ is the (unrestricted) mean of $\mathbb{X}_j$.

- `Step 2`: For $j = 1, \ldots, p$, sort $\mathbb{X}_j$ in descending order according to η_j^2 values, and extract a basis of the first p' linearly independent variables, following the process of Gram-Schmidt ([49]).

This basis is optimal is the sense that it contains all the information about $\mathbb{Y}$ included in the p original variables. The linear independence condition reduces greatly the number of variables ($p' \leq n$).

- `Step 3`: For j and j' in $\{1, \ldots, p'\}$ with $j < j'$, calculate $\tau(\mathbb{X}_j, \mathbb{X}_{j'})$, the Kendall rank correlation coefficient between the $\mathbb{X}_j$ and $\mathbb{X}_{j'}$. If $\tau(\mathbb{X}_j, \mathbb{X}_{j'}) \geq 0.5$, eliminate $\mathbb{X}_{j'}$ (because $\eta_{j'}^2 < \eta_j^2$). Otherwise keep $\mathbb{X}_j$ and $\mathbb{X}_{j'}$.

At the end of this step, we are left with p'' linearly independent variables ($p'' \leq n$) ranked in descending order according to their correlations with $\mathbb{Y}$.

For classification purposes, it is desirable to further reduce the number of variables and keep only the most pertinent for building an accurate classifier. Numerous supervised classification methods can be used to achieve that. In our situation, we use the LDA [41, 42] and [50]. The objective is to explore the relationship between the numerical (independent) variables $\mathbb{X}_j$ and categorical (dependent) variable $\mathbb{Y}$, and use it to predict the value of the dependent variable.

The LDA is an implemented package in SPSS that leads to observations classification using scores, discriminant functions, and cross validation. For more details about implementation and output, we refer to SPSS guide users [43].

- `Step 4`: For ℓ ranging from 2 upto p'', perform the LDA to subsets, of the dataset resulting from `Step 3`, involving the ℓ first variables. For the classification purpose, retain the variables that maximize the cross validation percentage.

At this point, the retained variables could be considered the most reliable for predicting the dependent variable. The objective of the next step is an ultimate filter to discard variables that might be sensitive to the sample size.

- `Step 5`: Repeat the steps 1 to 4 with different sample sizes. The final set of retained variables contains those proven reliable predictors at least m% of the time (m may be set as 70%).

Application and results

In this section we consider the application of our approach to some real datasets recently used in cancer gene expression studies by several authors. The first dataset has been obtained from acute leukemia patients at the time of diagnosis [31]. This dataset comes from a study of gene expression in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data consist of 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML of $p = 3571$ human genes. The second dataset concerns the Prostate cancer that contains 52 prostate tumor observations and 50 non-tumor prostate observations of $p = 6033$ genes.

Both data sets are from Affymetrix high-density oligonucleotide microarrays and are publicly available [16].

Application to Leukemia dataset

The Leukemia dataset contains 72 observations. We randomly select 12 as a test sample. The 60 remaining are used as a training sample. We apply our approach to different sizes, and we retain the variables that maximize the cross validation percentage. These are the highly informative genes. Table 1 summarizes our results.

For a sample of 60 observations, the selected basis contains 60 vectors holding all the information from the 3571 initial genes. At this stage we get a dimension reduction of approximately 62%. In the next step (column 3), using the Kendall rank correlation we keep only 33 genes. The retained genes are the most highly correlated with the nominal response (cancer class). The Figure 1 represents the cross validation percentage against the number of genes (from `Step 4`). The 4th column contains the number of genes that maximize the cross validation percentage. The number of variables is reduced from 33 to a pertinent 3 genes which lead to a 98% correct classifications.

The steps described above are repeated for different sample sizes to ensure the model's stability, and we retain the variables which appear as reliable classifiers. Table 2 presents the genes occurrences with their cross validation percentages. The 11 retained genes have led to about 98% of good classification. The genes will be utilized to predict the classification of the 12 observations in the test sample. These prediction results, given in Table 3, show that our approach is highly accurate.

Kastrin and Peterlin [36] studied the potential of RM modeling using the same dataset. They demonstrate that the RM is as effective as the principal component analysis (PCA) with re-randomization scheme. Table 4 shows that our approach, applied to the Leukemia dataset, outperforms the RM.

Application to prostate cancer state

The prostate tumor dataset contains 102 observations. We randomly select 13 as a test sample. The 89 remaining are used as a training sample. We apply our approach to different sizes, and we retain the variables that maximize the cross validation percentage. These are the highly informative genes. Table 5 summarizes our results.

Table 6 presents the genes occurrences with their cross validation percentages. The 9 retained genes have led to about 95.5% of good classification. These genes are used to predict the classification of the 13 observations in the test sample. These prediction results, given in Table 7, show that our approach is highly accurate.

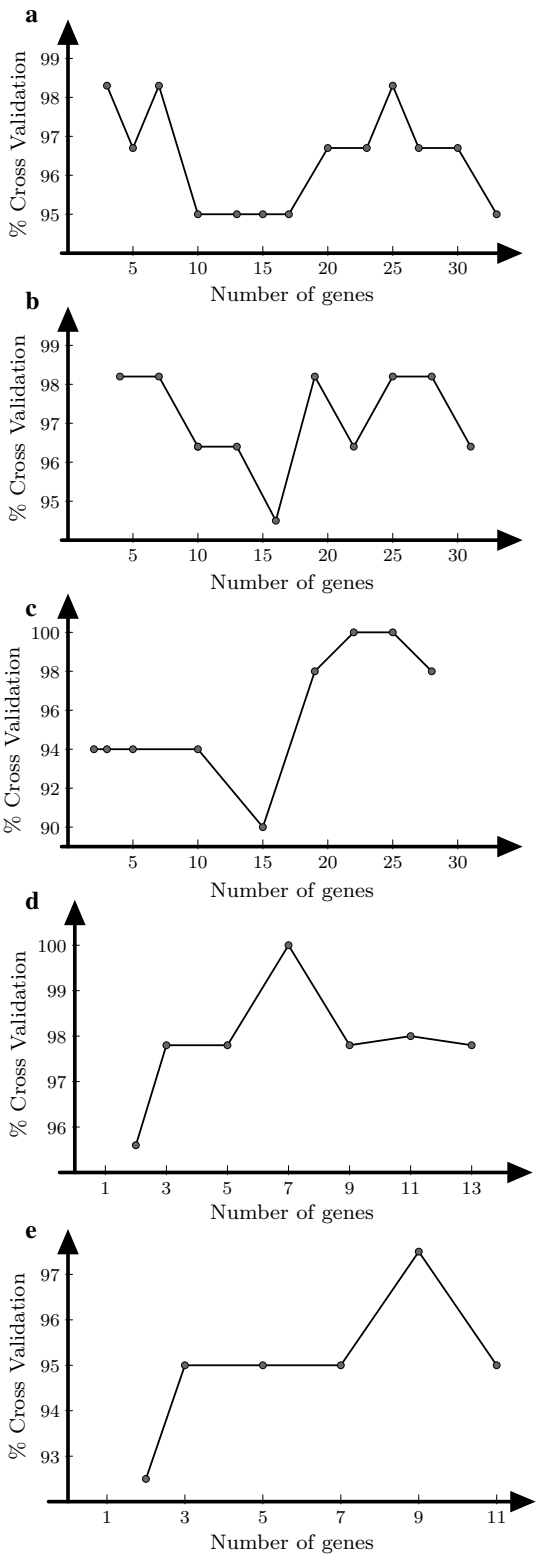

Fig. 1 Cross validation percentage against the number of genes for **a** 60, **b** 55, **c** 50, **d** 45 and **e** 40 samples of size

Table 1 Reduction of number of genes for different sample size in dataset

Sample size	Rank of the extracted basis (steps 1 and 2)	Nb of genes with Kendall rank correlation < 0.5 (step 3)	Number of final selected genes (step 5)
60	60	33	3
55	55	31	4
50	50	28	22
45	45	13	7
40	40	11	9

Table 2 Final retained classifiers

Frequency	Number of genes	Cross validation (%)	Final retained genes
At least 4 times	2 genes	91.7	gene376, gene456, gene626, gene672, gene874, gene907, gene918, gene951, gene956, gene979, gene1001
At least 3 times	4 genes	96.7	
At least 2 times	11 genes	98.3	

Table 3 Class prediction for the test sample

Observations	Observed class	Scores	Predicted class
1	0	1797	0
2	1	− 5044	1
3	1	− 4543	1
4	0	3743	0
5	0	0243	0
6	1	− 5924	1
7	1	− 5016	1
8	1	− 4070	1
9	0	− 0222	0
10	1	− 5733	1
11	1	− 4967	1
12	1	− 3395	1

Table 4 Performances comparison

RM-LDA			Our approach		
Number of selected genes	ER-random selection	ER-supervised selection	Number of selected genes	Random sample size	Error rate (%)
50	0.31	0.04	11	60	0
100	0.29	0.04	11	58	0
200	0.27	0.05	11	46	0

Table 5 Reduction of number of genes for different sample size in dataset

Sample size	Rank of the extracted basis (steps 1 and 2)	Nb of genes with Kendall rank correlation < 0.5 (step 3)	Final selected genes (step 5)
89	89	31	5
80	80	26	5
75	75	31	3
70	70	34	10
60	60	34	10
50	50	24	8

Table 6 Final retained classifiers

Frequency	Number of genes	Cross validation (%)	Final retained genes
At least 4 times	1 genes	92.1	gene2619, gene1495, gene2425, gene2746, gene4849, gene1788, gene1897, gene2848, gene4155
At least 3 times	5 genes	91	
At least 2 times	9 genes	95.5	

Table 7 Class prediction for testing sample

Observations	Observed class	Scores	Predicted class
1	0	− 0.857	0
2	0	− 1.064	0
3	0	− 0.614	0
4	0	− 2.846	0
5	0	− 1.593	0
6	0	− 1.933	0
7	1	1.035	1
8	1	2.149	1
9	1	2.806	1
10	1	2.751	1
11	1	0.584	1
12	1	0.722	1
13	1	0.048	1

Table 8 Performance comparison

RM-LDA			Our approach		
Number of selected genes	ER-random selection	ER-supervised selection	Number of selected genes	Random sample size	Error rate (%)
50	0.46	0.18	9	90	0
100	0.45	0.19	9	80	0
200	0.45	0.21	9	67	0

Table 8 shows that our approach, applied to the prostate tumor dataset, outperforms the RM.

It is worth noting that the use of the developed approach is not restricted to binary prediction problems. It can be extended to cover multiclass prediction. Indeed, we applied the approach on a third dataset which concerns the small blue cell tumors (SRBCTs) presented as a matrix of 2308 genes (columns) and 83 samples (rows), from a set of microarray experiments. The SRBCTs are 4 different childhood tumors classified into four major types: BL (Brkitt lymphoma), EWS (Ewings sarcoma), NB (neuroblastoma), and RMS (rhabdomyosarcoma). After applying the same approach described above for (2308×83) dataset, 8 genes are selected. Even if, we have 4 different classes, our approach performs well. It gives a mean accuracy rate of 90%.

Conclusions

Big Data is a highly topical issue of major importance in healthcare research. In fact, the role of Big Data in medicine consists to better build health profiles and predictive models around individual patients, so that we can better diagnose and treat disease. Big data comes into play an important role to overcome major challenges posed by cancer which represents an incredibly complex disease. The cancer disease is always changing, evolving, and adapting, where a single tumor can have more than 100 billion cells, and each cell can acquire mutations individually. To best understand evolution of cancer or to best distinguish tumor classes, we need advanced modeling by integrating Big Data. Different techniques are available, but it suffers from a lack of accuracy or processing complexity.

The purpose of this article is to present methods to reduce the number of variables and keep those that contain more information for reliable and informative classification. The article proposes methods for dimensionality reduction and classification, in several stages, using gene expression data from two recent studies. This way of proceeding, allows to retrieve the variables that contain most information for proper classification according to type of cancer. The retained model is the one that guarantees the best classification by cross-validation. The final model is then used to predict the class samples of the test set.

A comparative study was developed, for binary problems, between the results of our approach and that of the model developed by Rash [36]. The main conclusion is that our approach performs well the RM-LDA based approach with a null error rate and a 100% of accuracy.

It is worth to note that our approach can be compared with other multiclass prediction problems by integrating multiple ROC analysis and can be used to analyze other prediction problems in different fields such as, finance and banking, marketing and environment.

Authors' contributions
All mentioned authors contribute in the elaboration of the article. All authors read and approved the final manuscript.

Author details
[1] Département Statistique, Démographie et Actuariat, Institut National de Statistique et d'Economie Appliquée, Rabat, Morocco. [2] Moroccan Agency for Sustainable Energy, Rabat, Morocco. [3] Laboratoire de Mathématiques, Statistique et Applications, Département de Mathématiques, Faculté des sciences, Université Mohammed V de Rabat, Rabat, Morocco. [4] Unité de Formation et Recherche (UFR) Environnement, Université Jean Lorougnon Guédé de Daloa, Daloa, Côte d'Ivoire.

Acknowlegements
A particular acknowledgement is for the scientific and the editorial committee. Acknowledgement is also for the providers the used Data.

Competing interests
All authors confirm that there are no competing interests. The article is not under any other review process and is not subject of any other submission.

Availability of data and materials
All data used are publically available. Source of the used data is mentioned in the article.

Consent for publication
Authors approve the consent for publication.

Ethics approval and consent to participate
All authors confirmed the ethics approval and consent to participate.

Funding
No funding exists. We have ask for free charge processing and we have a confirmation for the Big data Journal, that we not need to pay the article processing charge, because we are based in a low-income country.

Publisher's Note

Received: 24 July 2017 Accepted: 26 September 2017
Published online: 10 October 2017

References

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403:503–11.
2. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. PNAS. 2000;97(18):10101–6.
3. Antoniadis A, Lambert-Lacroix S, Leblanc F. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics. 2003;19(5):563–70.
4. Boulesteix AL. PLS dimension reduction for classification with microarray data. Stat Appl Genet Mol Biol. 2004;3(1):1–30.
5. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high dimensional genomic data. Brief Bioinf. 2008;8:24–32.
6. Bughin J. Reaping the benefits of big data in telecom. J Big Data. 2016;3:14.
7. Casaca JA, da Gama AP. Marketing in the Era of Big data, human and social sciences at the common conference. 2013.
8. Cai T, Liu WD. A direct estimation approach to Sparse linear discriminant analysis. J Am Stat Assoc. 2011;106:1566–77.
9. Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat. 2005;35(6):23132351.
10. Chen S, Donoho D, Saunders M. Atomic decomposition by basis pursuit. SIAM J Sci Comput. 1998;20(1):3361.
11. Chiaromonte F, Martinelli J. Dimension reduction strategies for analyzing global gene expression data with a response. Math Biosci. 2002;176:123144.
12. Christopher G, Jiashun J, Wasserman L, Yao Z. A comparison of the lasso and marginal regression. J Mach Learn Res. 2011;13:21072143.
13. Crawford M, Khoshgoftaar M, Prusa D, Richter N, Al Najada H. Survey of review spam detection using machine learning techniques. J Big Data. 2015;2:23.
14. Depeige A, Doyencourt D. Actionable knowledge as a service (AKAAS): leveraging big data analytics in cloud computing environments. J Big Data. 2015;2:12.
15. Demchenko Y, Grosso P, de Laat C, & Membrey P. Addressing Big Data issues in scientific data infrastructure. Proceedings of the international conference on collaboration technologies and systems, May 20–24. San Diego: IEEE Xplore Press; 2013. p 48-5. DOI: 10.1109/CTS.2013.6567203.
16. Dettling M. BagBoosting for tumor classification with gene expression data. Bioinformatics. 2004;20(18):3583–93.
17. Donoho DL, Elad M. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. Proc Natl Acad Sci. 2013;100(5):2197–202.
18. Kondziolka Benjamin T C, Lunsford LD, Silverman J. Development, implementation, and use of a local and global clinical registry for neurosurgery. Big Data. 2015;3(2):80–9.
19. DongGuo H, Zhang L, WeiZhu L. Earth observation big data for climate change research. Adv Clim Change Res. 2015;6(2):108–17.
20. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2003;32:407451.
21. Einav L, Levin J. Economics in the age of big data. Science. 2014;346(6210):1243089.
22. Fan J, Fan Y. High dimensional classification using features annealed independence rules. Ann Stat. 2008;36:260537.
23. Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J R Stat Soc Ser B. 2012;74(1):3765.
24. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):13481360.
25. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with disussion). J R Stat Soc Ser B. 2007;70(5):849911.

26. Fan J, Liao Y. Endogeneity in ultrahigh dimension, technical report. New Jersey: Princeton University; 2014.
27. Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. J Mach Learn Res. 2009;10:20132038.
28. Fisher R. Statistical methods for research workers. ISBN 0-05-002170-2; 1926.
29. Friedman J, & Popescu B. Gradient directed regularization for linear regression and classification. Technical report. 2004.
30. Gesing S, Connor T, & Taylor I. Genomics and biological Big Data: facing current and future challenges around data and software sharing and reproducibility. Position paper at BDAC-15 (Big Data Analytics: Challenges and Opportunities), workshop in cooperation with ACM/IEEE SC15, Austin; 2015.
31. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531–7.
32. Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. J Comp Graph Stat. 2009;18(3):533550.
33. Hall P, Miller H. Modeling the variability of rankings. Ann Stat. 2010;38(20):2652–77.
34. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoro NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA. 2000;97:8409–14.
35. Husain S, Kalinin A, Truong A, Dinov D. SOCR data dashboard: an integrated big data archive mashing medicare, labor census and econometric information. J Big Data. 2015;2:13.
36. Kastrin A, Peterlin B. Rasch-based high-dimensionality data reduction and class prediction with applications to microarray gene expression data. Exp Syst Appl. 2010;37(7):5178–85.
37. Kramer A, Guillory J, Hancock J. Experimental evidence of massive scale emotional contagion through social networks. Proc Natl Acad Sci USA. 2014;111(24):8788–90.
38. Laney D. 3D Data management: controlling data volume, velocity and variety. 2001.
39. Liao Y, Jiang W. Posterior consistency of nonparametric conditional moment restricted models. Ann Stat. 2011;39(6):30033031.
40. Loureno JR, Cabral B, Carreiro P, Vieira M, Bernardino J. Choosing the right NoSQL database for the job : a quality attribute. J Big Data. 2015;2(1):1–26.
41. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. San Diego: Academic Press Inc; 1979.
42. McLachlan GJ. Discriminant analysis and statistical pattern recognition. New York: Wiley; 1992.
43. Meulman JJ, Heiser JW. IBM SPSS Categories 20. 2011. pp. 233–248
44. Narock TW, & Hitzler P. Crowdsourcing semantics for Big Data in geosciences applications. In: AAAI 2013 Fall symposium series, semantics for Big Data, November 15–17. Arlington; 2013.
45. Nguyen DV, Rocke DM. On partial least squares dimension reduction for microarray-based classification: a simulation study. Comput Stat Data Anal. 2004;46(3):407–25.
46. Pääkkönen P. Feasibility analysis of AsterixDB and spark streaming with Cassandra for stream-based processing. J Big Data. 2016;3:6. doi:10.1186/s40537-016-0041-8.
47. Pearson ES. Review of statistical methods for research workers (R. A. Fisher). Sci Prog. 1926;20:733–4.
48. Pittelkow PH, Ghosh M. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. J R Stat Soc B. 2008;70:15973.
49. Pursell L, Trimble SY. Gram-Schmidt orthogonalization by Gauss elimination. Am Math Month. 1991;98(6):544549. doi:10.2307/2324877.
50. Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.
51. Santos F. Le rapport de corrlation : mesurer la liaison entre une variable qualitative et une variable quantitative. CNRS, UMR 5199 PACEA. 2015.
52. Shaldehi AH. Using Eta (η) correlation ratio in analyzing strongly nonlinear relationship between two variables in practical researches. J Math Comput Sci. 2013;7(3):213–20.
53. Toga W, Dinov D. Sharing big biomedical data. J Big Data. 2015;2:7.
54. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58(1):267288.
55. Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894942.
56. Zuech R, Koshgoftaar M, Wald R. Intrusion detection and big heterogeneous data: a survey. J Big Data. 2015;2:3.