Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 22, No. 3, 2024 (New Series), pp 171–197
http://www.ssca.org.in/journal

Model-Free Data Cleaning for Raw Data: An Eigen-Structure Approach

Ravindra Khattree
Department of Mathematics and Statistics
Center for Data Science and Big Data Analytics
Center for Biomedical Research
Oakland University, Rochester, MI, 48309, USA

Received: 13 June 2024; Revised: 10 July 2024; Accepted: 11 July 2024

Abstract

Preprocessing of data at the initial stages before assuming any model for the data is a necessary requirement for observational data. With preliminary data cleaning of raw data in mind, we introduce a model-free approach based on the eigen-structure of the data matrix to assess if a particular observation induces multicollinearity or is excessively outlying within the data. Specifically we look at the eigenvalues and antieigenvalues obtained from the singular value decomposition of the data matrix or a function thereof. We also study detection of the outlier induced multicollinearity or outlier induced masking of multicollinearity present in the data. Usefulness of our approach is illustrated via several examples describing a variety of situations and for several classical data sets. Emphasis is on data matrix of variables rather than model matrix, although these approaches can be later used in model based contexts as well.

Key words: Antieigenvalue; Condition indexes; Eccentricity; Emphasis measure; Multicollinearity; Outlying observations.

AMS Subject Classifications: 62J05, 62P99

1. Introduction

Prof. C. R. Rao was my PhD adviser at the University of Pittsburgh. His teaching and research both have continued to have a lasting impact on my own academic career. Throughout his classes, there was always an implicit but very definite message that any research in statistics should have a definite purpose of understanding and solving some meaningful and practical problem. Reflecting on this philosophy of Prof. Rao, this article on a very fundamental first step of data processing is written as a personal tribute to Prof. Rao and with a purpose of honoring his legacy and place in the world of statistics and science.

Preprocessing of data and data cleaning are essential steps in observational studies and may involve the steps of detecting freak values, identification of outlying observations,

Corresponding Author: Ravindra Khattree
Email: khattree@oakland.edu

choosing the meaningful variables and understanding the underlying dependence among various variables. Inference can be greatly distorted due to some or all of the above issues and a substantial body of work has been done in the past to remedy such problems. See for example, Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Belsley (1991), Khattree and Naik (1999), Seber and Lee (2003) and Khattree (2019). All of the above referenced work, except that by Khattree (2019), deal with various model based approaches. While these are perfectly valid approaches for diagnostics, all being model based, they however, tend to not look into the very basic structure of the raw data. We think that in order to gain full understanding of and more insight into data, we must also look at the anomalies in the data at the very fundamental level, the fundamental level being the patterns and differences at the level of **X**'**X** matrix, where **X** is our data matrix of all variables under consideration. This is a fundamental issue in data science, taking a precedence over any subsequent statistical modeling. This article is a step towards addressing this problem by looking at the eigen-structure of the data itself.

In principle and in general, **X** may be either a data matrix or a model matrix (in which case, we will use the notation $\mathbf{X}^*$ to distinguish it from the raw data matrix). We here assume all variables to be quantitative. Since the data cleaning at the preprocessing stages must assume no specific model and no specific prespecified choice of a few selected variables, data matrix is a more appropriate context for our work. Therefore it is meaningful that we rely more on the mathematical structure of the data matrix **X** than on statistical evaluation of the model to be fitted. This is especially relevant because for large data sets at preprocessing stages, data cleaning is equally important for the explanatory variables as well as response variables and our data matrix may contain both types of variables. This is the approach adopted by Wang and Nyquist (1991) and Khattree (2019). Other approaches not exclusively based on matrix structure or model but based on various other tentative techniques such as aggregate queries are given by Chu *et. al* (2016), Chu and Ilyas (2016) and Ilyas and Chu (2019).

As indicated, our approach here will rely on an evaluation of the eigen-structures of the matrices which are closely related to the singular value decomposition of the whole or parts of the data matrix. We will focus on the evaluation of multicollinearity and the detection of outlying observations by evaluating the changes or deviations in these eigen-structures by the use of eigenvalues and antieigenvalues. While theory of eigenvalues is well established, recent discussions of antieigenvalues along with various applications thereof are available in Khattree (2001, 2002, 2003, 2006, 2010, 2014, 2019) and in Tran and Khattree (2024). Applications have also been presented in Khattree and Bahuguna (2019), Cuntoor and Chellappa (2006) and Guo *et al.* (2018).

We must emphasize that data matrix consists of raw data on variables and not on the mathematical functions thereof. To press that point, although we will not rely on it, suppose the framework was the standard linear model, namely,

$$\mathbf{y}_{n\times 1} = \mathbf{X}^*_{n\times p^*}\beta^*_{p^*\times 1} + \epsilon_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \mathbf{Z}_{n\times q}\gamma_{q\times 1} + \epsilon_{n\times 1},$$

and suppose the data cleaning was confined to only data on the explanatory variables. The matrix **X** may then represent the raw data matrix and **Z** contains columns corresponding to other $q(= p^* - p)$ terms in the model such as intercept, and specific mathematical transformations of columns of $\mathbf{X}$ *e.g.* polynomial or cross products terms. When the type of true model (in terms of which variables, which degree of polynomial or which cross product terms) and/or its dimension are unknown, our interest should be exclusively in the response variable and in the matrix of raw data on explanatory variables namely, $\mathbf{X}$ and what specific model will subsequently be used will be a secondary consideration for a later time. In that sense, we are interested in measuring the multicollinearity and outlying nature of observations within the raw data and we do not concern ourselves as to what specific model is being assumed. The problem of *model induced outlyingness* due to outliers will involve the *model* matrix $\mathbf{X}^*$ $= [\mathbf{X} : \mathbf{Z}]$ and the corresponding response variable. A version of this latter problem, albeit with a very different approach has been discussed in Mason and Gunst (1985).

Clearly, at initial stages, from data cleaning point of view only the former context is relevant and should always take precedence over modeling and model selection. Further, in big data context, one encounters a very large number of observations and a large number of variables, both of which are, presumably, to be used in several different modeling problems in the future. It is thus imperative to have a clean data where "cleanliness" of the data may refer to general robustness of the data, with respect to specific observations and/or specific variables. This is the situation, where we believe our approach has the most currency.

In Section 2, we motivate antieigenvalues of a positive definite matrix as a way to look into the eigen-structure and as the measures of eccentricities of an ellipsoid for various cross-sections which in turn provide us a way to measure the interdependences between a set of variables or multicollinearity.

Section 3 is about identification of outlying observations via antieigenvalues. Towards the end of this section, we also discuss the issues pertaining to *collinearity* − *outlyingness* where one of these two may cause the other. We provide illustrations of our approaches using several data sets. Admittedly, to make the understanding of the approach more accessible, we must use data sets which are not excessively large and are readily available. However, that does not disqualify our approach for bigger data sets. In fact, those are the situations where once computationally implemented, these methods will have most utility. Thus in Section 4, we also consider a relatively larger data set consisting of 1599 observations on the quality of red wine. We apply our procedure on this data to make a point that procedure is effective even when we have large data sets and that our approach is able to successfully pick out the observations which are not easy to identify otherwise but whose presence excessively corrupts the data and subsequently also affects the modeling steps. Section 5 provides some concluding remarks.

2. Eccentricities and measurement of multicollinearity

Let, as earlier, $\mathbf{X}_{n\times p}$ be a data matrix. Assume $rank(\mathbf{X}_{n\times p}) = p$ and let $\mathbf{A} = \mathbf{X}\text{'}\mathbf{X}$. Consider the quadratic surface, $\mathbf{u}\text{'}\mathbf{A}^{-1}\mathbf{u} = c$ where c is a known constant in a $p-$ dimensional space. Since $\mathbf{A}$ is positive definite, this represents an ellipsoid and with an appropriate orthogonal rotation $\mathbf{v} = \mathbf{P}\text{'}\mathbf{u}$ where $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}\text{'}$ is the spectral decomposition of $\mathbf{A}$, the surface can be represented as,

$$\mathbf{v}\text{'}\Lambda^{-1}\mathbf{v} = c \text{ with } \Lambda = diagonal(\lambda_1, \lambda_2, ..., \lambda_p)$$

or

$$\frac{v_1^2}{\lambda_1} + \frac{v_2^2}{\lambda_2} + ... + \frac{v_p^2}{\lambda_p} = c \text{ where } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0.$$

The eccentricities of certain two dimensional elliptical cross-sections of this ellipsoid can be quantified in decreasing order as $\sqrt{\frac{\lambda_1}{\lambda_p}} \geq \sqrt{\frac{\lambda_2}{\lambda_{p-1}}} \geq \sqrt{\frac{\lambda_3}{\lambda_{p-2}}} \geq \cdots$. The quantity $e_1 = \sqrt{\frac{\lambda_1}{\lambda_p}}$ is the eccentricity measured respectively via the two most elongated and most compressed directions and hence measures the extreme eccentricity. The next quantity $e_2 = \sqrt{\frac{\lambda_2}{\lambda_{p-1}}}$ represents the comparison of the next two most elongated and most compressed directions and similar comparisons continue for $r = [p/2]$ pairs where $[p/2]$ is the integer part of $p/2$. Clearly, with λ_i considerably larger than $\lambda_{p-i+1}, i = 1, 2, \cdots, r$, e_i will also be large, indicating a particular cross section of the ellipsoid highly elongated thereby indicating the high multicollinearity. A one-to-one monotonically decreasing function of $e_i = \sqrt{\frac{\lambda_i}{\lambda_{p-i+1}}}$ is the i^{th} *antieigenvalue* of the matrix **X**'**X** namely,

$$\eta_i = \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \frac{2}{e_i + e_i^{-1}}, i = 1, 2, ..., r = [p/2]. \tag{1}$$

It can be shown that $0 < \eta_1 \leq \eta_2 \leq ... \leq 1$ are ordered by their magnitudes. Being a monotonic function of $e_i, i = 1, 2, \cdots, r$, these also measure the eccentricities and hence the multicollinearity in the data. To connect this unfamiliar quantity to a familiar context, values close to zero for at least one of the antieigenvalues indicate high multicollinearity while higher values (close to 1) of all $\eta_i, i = 1, 2, \cdots, r$ indicate a lack of multicollinearity. Also, the most ideal situation namely, $\eta_1 = 1$, implies that the matrix $\mathbf{A} = \mathbf{X}'\mathbf{X}$ is orthogonal and then there is absolutely no multicollinearity among the columns of **X**. Further, greater the number of η_i that are close to zero, higher is the number of linear near-dependencies that may exist.

Note that **X**'**X** and $(\mathbf{X}'\mathbf{X})^{-1}$ share the same set of antieigenvalues and hence these r measures of multicollinearity for the two matrices are equal. It is, in some way, a reasonable and desirable property in that, multicollinearity, being synonymous to ill-conditioning of a given matrix, indicates a *computational difficulty* in obtaining an accurate inverse matrix. Intuitively, this computational difficulty should be same for the matrix **A** as well as for its *true* inverse $\mathbf{A}^{-1}$ because their eigen-structures are directly related (i^{th} ordered eigenvalue of inverse of a matrix is the reciprocal of the $(p-i+1)^{th}$ ordered eigenvalue of the original matrix).

A single index of multicollinearity combining all antieigenvalues defined in Equation (1) can be defined as the generalized antieigenvalue (See (Khattree, 2002, 2003)),

$$\Delta = \prod_{i=1}^{r} \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}} = \prod_{i=1}^{r} \eta_i, \text{where } r = [p/2], \tag{2}$$

which is a function of all antieigenvalues and can be interpreted as an overall measure of eccentricity. Clearly Δ is also same for **X**'**X** and $(\mathbf{X}'\mathbf{X})^{-1}$. One may alternatively use the r^{th} root of Δ which would then be the geometric mean of all antieigenvalues.

Belsley, Kuh and Welsch (1980) and Belsley (1991) suggest to look at the *condition number* $\psi = \sqrt{\frac{\lambda_1}{\lambda_p}}$. They also look at the *condition indexes* $\psi_2, \psi_3, \cdots, \psi_p$, where $\psi_i = \sqrt{\frac{\lambda_1}{\lambda_i}}, i = 2, 3, \cdots, p$. Larger values are indicative of possible multicollinearity. Note that $\psi_2, \psi_3, .., \psi_p (= \psi$, the condition number) are all greater than or equal to 1 with no upper bound specified. There is apparently no way to decide what constitutes a large condition index/number. That aside, unlike the sets of antieigenvalues, the two sets of condition indexes, – for $\mathbf{X}'\mathbf{X}$ and for $(\mathbf{X}'\mathbf{X})^{-1}$, – are different from each others. Specifically these are $\{\sqrt{\lambda_1/\lambda_2}] \leq \sqrt{\lambda_1/\lambda_3} \leq ... \leq \sqrt{\lambda_1/\lambda_p}\}$ and $\{\sqrt{\lambda_{p-1}/\lambda_p}] \leq \sqrt{\lambda_{p-2}/\lambda_p} \leq ... \leq \sqrt{\lambda_1/\lambda_p}\}$ respectively.

How useful and practical are the indexes defined in Equations (1) and (2)? This can be best explained and demonstrated by applying them on real data sets. We will thus illustrate the utility of these indexes by first applying them to four data sets of varying sizes, with different number of variables and of varying features. Specifically, we consider,

(*i*) A data set on properties of soil given by Kendall (1975) with $p = 4, n = 20$. The data collected here is a set of 20 samples of soil for each of which salt content (x_1) clay content (x_2), organic matter (x_3) and acidity on pH scale (x_4) are measured.

(*ii*) A data set by Daniel and Wood (1980) on clinkers with $p = 5, n = 14$. This data set on clinker compounds was collected with a purpose to study their effects on amount and rate at which heat evolves during cement hardening. The independent variables are weight percent of $SiO_2(x_1)$, Al_2O_3 (x_2), Fe_2O_3 (x_3), CaO (x_4) and $MgO(x_5)$. The data are compositional in that ideally their sum should add to hundred percent. However possibly due to impurities and also due to round off errors, these variables do not add to hundred percent (in many cases, in fact, they add to much more than hundred percent and thus cannot be fully explained by round off errors). This data set has been extensively analyzed by Chatterjee and Hadi (1988).

(*iii*) A data set due to Chatterjee, Hadi and Price (2006) with $p = 6, n = 40$. This data set with six predictors (x_1 through x_6) is given in Chatterjee, Hadi and Price (2006) as Table 4.8 (p. 128) and as part of Exercises 4.12-4.14. No detailed description is available. However, data set was used to illustrate the strong presence of multicollinearity.

(*iv*) A data set due to Rao (1948) on cork deposits with $p = 4, n = 28$. This classic data, more easily available in Khattree and Naik (1999), pertains to the cork deposits in four directions (North, East, South and West), denoted respectively by x_1 through x_4 on twenty eight trees in the Himalayan range. These latter authors have extensively studied this data set in various contexts including the detection of outlying observations.

Later in Section 4, we consider a very large dataset as well. The above four data sets, being of manageable size to include here, are given in the appropriate columns of Tables 1-4. In each case, the question is, how well behaved, with respect to multicollinearity, the particular data set is. Thus, we calculate the antieigenvalues η_i as well as generalized antieigenvalue Δ in each case. We will work with r^{th} root of the generalized antieigenvalue as it is essential to bring this measure on equal footing when comparing multicollinearities of various data sets with different number of variables and this measure, as the geometric mean of all antieigenvalues, does so. To make results more readable, all values of antieigenvalues and generalized antieigenvalue in various tables are multiplied by 100. Smaller values indicate more severe presence of multicollinearity.

Table 1: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Kendall's data, $r = 2$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**4.58**	**93.42**	**4.28**	**20.70**
1	13.0	9.7	1.5	6.4	4.56	95.16	4.34	20.82
2	10.0	7.5	1.5	6.5	4.47	94.49	4.22	20.55
3	20.6	12.5	2.3	7.0	4.68	94.39	4.42	21.02
4	**33.8**	**19.0**	**2.8**	**5.8**	**4.15**	**87.77**	**3.64**	**19.07**
5	20.5	14.2	1.9	6.9	4.69	94.68	4.44	21.06
6	10.0	6.7	2.2	7.0	4.58	93.18	4.26	20.65
7	12.7	5.7	2.9	6.7	4.61	91.43	4.22	20.53
8	36.5	15.7	2.3	7.2	4.78	92.81	4.44	21.07
9	37.1	14.3	2.1	7.2	4.58	93.91	4.30	20.74
10	25.5	12.9	1.9	7.3	4.58	93.54	4.29	20.70
11	26.5	14.9	2.4	6.7	4.72	93.07	4.39	20.95
12	22.3	8.4	4.0	7.0	4.39	92.18	4.05	20.12
13	30.8	7.4	2.7	6.4	4.59	95.04	4.36	20.88
14	**25.3**	**7.0**	**4.8**	**7.3**	**4.11**	**90.56**	**3.72**	**19.28**
15	31.2	11.6	2.4	6.5	4.72	94.03	4.44	21.08
16	22.7	10.1	3.3	6.2	4.48	93.34	4.19	20.46
17	31.2	9.6	2.4	6.0	4.69	95.13	4.46	21.11
18	13.2	6.6	2.0	5.8	4.61	93.12	4.29	20.72
19	11.1	6.7	2.2	7.2	4.54	92.75	4.22	20.53
20	20.7	9.6	3.1	5.9	4.48	93.38	4.19	20.46

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 5 presents the values of all antieigenvalues along with r^{th} root of generalized antieigenvalue. Based on first antieigenvalue as well as on the r^{th} root of generalized antieigenvalue, the Rao's data largely seems to be relatively well behaved. The Chatterjee, Hadi and Price' data set appears to be suffering from very severe multicollinearity issues. Other two data sets fall in between. The data set by Daniel and Wood does exhibit a certain degree of multicollinearity and reasons for its presence are extensively discussed in Chatterjee and Hadi (1988).

What if the data sets were standardized prior to fitting the model? Needless to say that eigenvalues and hence the antieigenvalues will change. Does that in any way distort the picture in terms of multicollinearity? There is no reason to expect an answer one way or the other since standardization eliminates the differences among variables in terms of degree of variability in relative terms. See Naik and Khattree (1996), Timm (2002) and Johnson and

Wichern (2014) for extensive discussions on this aspect of the data. The standardization would certainly affect the eccentricities of the ellipsoid. Thus, we suggest that one should also analyze the standardized version of **X**'**X** matrix. Table 6 presents the antieigenvalues for the four data sets in this case. While the conclusions are more or less same as those for original unstandardized data, we do notice that in each case corresponding antieigenvalues are larger except in case of cork data and for η_2. Understandably, scaling makes the **X**'**X** matrix more "spherical" compared to what it was for the unscaled data. Thus, standardization seems to help in the sense that standardized data seem to exhibit less multicollinearity.

Note that in this case the most "well behaved" data set among the four is that by Kendall. First and second antieigenvalues as well as the generalized antieigenvalue are all highest for this data set. Rao's cork data has next highest first antieigenvalue as well as the generalized antieigenvalue. As earlier, the data set by Chatterjee, Hadi and Price exhibits a very severe case of multicollinearity as seen by small values.

Table 2: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Daniel and Wood's data, $r = 2$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**.**	**1.95**	**60.77**	**1.18**	**10.88**
1	27.68	3.76	1.98	64.97	2.48	2.02	61.66	1.25	11.17
2	25.96	3.48	5.06	63.15	2.32	2.02	64.30	1.30	11.40
3	**21.86**	**5.75**	**2.77**	**65.02**	**5.04**	**0.25**	**61.92**	**0.16**	**3.96**
4	24.60	5.85	2.80	64.18	2.40	2.02	58.48	1.18	10.86
5	25.04	3.86	2.11	66.57	2.36	1.98	55.75	1.10	10.51
6	22.32	6.17	2.85	66.47	2.43	2.01	62.00	1.24	11.16
7	20.93	4.64	5.74	66.26	2.08	1.95	56.69	1.10	10.51
8	23.54	4.83	7.21	62.03	2.24	2.01	60.36	1.22	11.03
9	21.96	4.65	6.06	64.07	2.32	2.02	60.32	1.22	11.04
10	**21.44**	**8.81**	**1.19**	**66.64**	**2.48**	**1.64**	**56.22**	**0.92**	**9.60**
11	22.48	5.00	7.46	62.72	2.24	2.02	60.39	1.22	11.04
12	21.34	6.07	2.93	67.03	2.56	2.01	62.79	1.26	11.25
13	21.94	5.57	2.68	67.71	2.44	1.99	60.79	1.21	11.01
14	25.72	4.12	6.06	61.05	2.08	2.02	63.92	1.29	11.35

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

3. Detection of outlying observations

One of the major investigations during the preprocessing and data cleaning is to identify outlying observations. In a model based approach, this problem is usually dealt by calculating the *leverage values* (= $\mathbf{x}_i$'$(\mathbf{X}$'$\mathbf{X})^{-1}\mathbf{x}_i$ where $\mathbf{x}_i$' is the i^{th} observation, $i = 1, 2, \cdots, n$) of each of the n observations. Investigations in Wang and Nyquist (1991) and Khattree (2019) look at the problem in terms of effect of outlyingness of the observation on the eigen-structure of the data matrix in the sense how it affects the eigen-structure of the

Table 3: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [Chatterjee, Hadi and Price's data, $r = 3$.]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	x_6	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**.**	**.**	**0.36**	**42.61**	**56.48**	**0.09**	**9.56**
1	49	79	76	8	15	205	0.37	41.81	57.24	0.09	9.57
2	27	70	31	6	6	129	0.36	43.67	56.82	0.09	9.66
3	115	92	130	0	9	339	0.37	40.03	53.74	0.08	9.25
4	92	62	92	5	8	247	0.37	42.65	56.19	0.09	9.60
5	67	42	94	16	3	202	0.34	42.34	56.62	0.08	9.38
6	31	54	34	14	11	119	0.36	42.86	54.86	0.08	9.47
7	105	60	47	5	10	212	0.36	42.02	59.85	0.09	9.67
8	114	85	84	17	20	285	0.37	42.14	55.16	0.09	9.50
9	98	72	71	12	-1	242	0.37	41.34	56.88	0.09	9.53
10	15	59	99	15	11	174	0.36	42.65	62.72	0.10	9.92
11	62	62	81	9	1	207	0.36	42.31	55.97	0.08	9.46
12	25	11	7	9	9	45	0.35	42.61	54.53	0.08	9.32
13	45	65	84	19	13	195	0.37	42.37	56.33	0.09	9.56
14	92	75	63	9	20	232	0.36	40.48	56.10	0.08	9.38
15	27	26	82	4	17	134	0.35	40.93	58.02	0.08	9.37
16	111	52	93	11	13	256	0.36	43.29	57.51	0.09	9.64
17	78	102	84	5	7	266	0.36	42.65	55.44	0.09	9.50
18	106	87	82	18	7	276	0.37	41.85	56.95	0.09	9.59
19	97	98	71	12	8	266	0.36	43.00	56.86	0.09	9.59
20	67	65	62	13	12	196	0.36	42.62	55.82	0.09	9.50
21	38	26	44	10	8	110	0.35	42.72	55.83	0.08	9.44
22	56	32	99	16	8	188	0.37	43.83	56.94	0.09	9.70
23	54	100	50	11	15	205	0.37	43.93	56.16	0.09	9.67
24	53	55	60	8	0	170	0.35	42.20	56.09	0.08	9.43
25	61	53	79	6	5	193	0.36	42.80	56.23	0.09	9.56
26	60	108	104	17	8	273	0.37	42.33	58.79	0.09	9.73
27	83	78	71	11	8	233	0.37	42.65	56.59	0.09	9.61
28	74	125	66	16	4	265	0.36	44.08	56.54	0.09	9.64
29	89	121	71	8	8	283	0.37	44.22	55.92	0.09	9.67
30	64	30	81	10	10	176	0.37	43.82	56.24	0.09	9.66
31	34	44	65	7	9	143	0.36	42.56	57.09	0.09	9.59
32	71	34	56	8	9	162	0.36	42.83	56.76	0.09	9.61
33	88	30	87	13	0	207	0.36	42.86	57.01	0.09	9.56
34	112	105	123	5	12	340	0.36	41.17	55.36	0.08	9.39
35	57	69	72	5	4	200	0.36	42.60	55.92	0.08	9.47
36	61	35	55	13	0	152	0.36	41.40	56.64	0.09	9.49
37	29	45	47	13	13	123	0.35	42.61	55.07	0.08	9.40
38	82	105	81	20	9	268	0.36	42.32	56.29	0.09	9.49
39	80	55	61	11	1	197	0.37	41.95	56.90	0.09	9.56
40	82	88	54	14	7	225	0.37	42.96	56.76	0.09	9.64

Remark: Most-outlying (none for this data) observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 4: Detecting multicollinearity, raw data and antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$ of $\mathbf{X}_{(-j)}$'$\mathbf{X}_{(-j)}$ [C. R. Rao's Cork data]

Deleted Obs. (j)	x_1	x_2	x_3	x_4	η_1	η_2	Δ	$\Delta^{1/r}$
none	**.**	**.**	**.**	**.**	**8.83**	**90.24**	**7.97**	**28.23**
1	72	66	76	77	8.69	90.68	7.88	28.08
2	60	53	66	63	8.94	91.05	8.14	28.53
3	56	57	64	58	8.76	90.05	7.89	28.09
4	41	29	36	38	8.89	89.38	7.94	28.18
5	32	32	35	36	8.70	90.30	7.86	28.03
6	30	35	34	26	8.85	89.65	7.94	28.17
7	39	39	31	27	8.89	91.35	8.12	28.49
8	42	43	31	25	8.82	92.63	8.17	28.58
9	37	40	31	25	8.89	91.62	8.15	28.54
10	33	29	27	36	8.60	88.95	7.65	27.66
11	32	30	34	28	8.89	90.11	8.01	28.30
12	63	45	74	63	8.95	93.76	8.39	28.96
13	54	46	60	52	9.01	90.69	8.17	28.58
14	47	51	52	43	8.88	89.60	7.96	28.21
15	91	79	100	75	8.53	88.95	7.59	27.55
16	**56**	**68**	**47**	**50**	**8.09**	**93.28**	**7.54**	**27.47**
17	79	65	70	61	8.76	90.17	7.90	28.10
18	**81**	**80**	**68**	**58**	**9.06**	**93.89**	**8.51**	**29.17**
19	**78**	**55**	**67**	**60**	**8.05**	**89.64**	**7.21**	**26.86**
20	46	38	37	38	8.91	89.41	7.96	28.22
21	39	35	34	37	8.87	89.67	7.95	28.20
22	32	30	30	32	8.82	90.04	7.94	28.18
23	60	50	67	54	8.96	90.37	8.09	28.45
24	35	37	48	39	8.73	89.43	7.81	27.94
25	39	36	39	31	8.88	90.15	8.01	28.30
26	50	34	37	40	8.55	88.24	7.54	27.46
27	43	37	39	50	8.26	89.09	7.36	27.14
28	48	54	57	43	8.85	88.07	7.79	27.92

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion. All originally calculated statistics are multiplied by 100.*

Table 5: Antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$: Unstandardized (complete) data. All originally calculated statistics are multiplied by 100.

Data Set	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
Kendall	4.58	93.42	.	4.28	20.70
Daniel and Wood	1.95	60.77	.	1.18	10.88
Chatterjee, Hadi, Price	0.36	42.61	56.48	0.09	9.56
Rao	8.83	90.24	.	7.97	28.23

Table 6: Antieigenvalues η_i values, generalized antieigenvalue Δ and $\Delta^{1/r}$: Standardized (complete) data. All originally calculated statistics are multiplied by 100.

Data Set	η_1	η_2	η_3	Δ	$\Delta^{1/r}$
Kendall	63.84	99.61	.	63.59	79.74
Daniel and Wood	11.27	90.58	.	10.21	31.96
Chatterjee, Hadi, Price	0.00	59.36	93.36	0.00	0.13
Rao	27.25	85.42	.	23.28	48.25

data matrix. We here consider the problem in terms of the sensitivities of the antieigenvalues of **X**'**X** matrix to a particular observation. The premise is that when an observation is outlying, it may be due to considerable changes in the values of one or more variables (or combinations thereof) that will affect the usual pattern among the variables. This will in turn show up in the diagonal and non-diagonal elements of the **X**'**X** matrix. Such changes will then have an effect on the eccentricities of the corresponding ellipsoid. Accordingly, by reverse logic, if the i^{th} observation is not outlying then if $\mathbf{X}_{(-i)}$ is the corresponding $(n-1) \times p$ data matrix obtained by discarding the i^{th} observation from **X** matrix, the $p \times p$ matrices **X**'**X** and $\mathbf{X}_{(-i)}$'$\mathbf{X}_{(-i)}$ must not be too different from each other in terms of their eccentricities. Therefore we compare the antieigenvalues of these two matrices for every i and identify the observations to which these eccentricities are very sensitive. It must be emphasized that depending on the situation, the effect of an outlying observation may manifest on different antieigenvalues and hence one must ideally consider all antieigenvalues as well as the generalized antieigenvalue.

To illustrate the procedure, we return to our four data sets discussed earlier. Raw unscaled data are used in each case. For the sake of easy comparison, Tables 1-4 each present the original data along with the antieigenvalues when the particular observation has been deleted. As earlier, all antieigenvalues and the appropriate root of the generalized antieigenvalue have been multiplied by 100.

For the Kendall's data (Table 1), we observe that deleting the observation number 4 results in both antieigenvalues becoming unusually small. A closer look at the particular observation shows that corresponding (x_1, x_2) values are both unusually large for this data point. Another observation which stands out is the observation number 14 for which the first antieigenvalue is the smallest and the second antieigenvalue is second smallest. For this data point x_3 is unusually large while x_4 is among the several larger values. Clearly, both of these observations have substantial effects on the eigen-structure and eccentricities of the corresponding $\mathbf{X}'\mathbf{X}$ matrix.

Daniel and Wood's data set is relatively smaller in size. Yet, it still has two observations which are deemed outlying. These are observations 3 and 10 respectively and they stand out essentially due to relatively large x_5 value (for 3^{rd} observation) and very large x_2 value (for 10^{th} observation) respectively. See Table 2.

In case of Rao's cork data, the first antieigenvalues are somewhat different when the 19^{th} or 16^{th} observations are discarded. When the 18^{th} or 16^{th} observations are removed, the second antieigenvalue shows a considerable change. See Table 4. These observations were identified by Khattree and Naik (1999) as outlying using other techniques. The reasons for them being outlying are also explained there. However the effect is rather mild.

The data set of Chatterjee, Hadi and Price does not seem to contain any outlying observation because for all the antieigenvalues in Table 3, corresponding values are not very different when an observation has been deleted.

To explore further and perhaps in a more definite way than the previous approach where the relative closeness of antieigenvalues was visually assessed in a table, we may yet adopt another criterion (Also, see Khattree, 2019) and directly look at the eigen or antieigenstructures of the matrix

$$\mathbf{G}_i = \mathbf{U}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_i'$$

where $\mathbf{U}_i$ is the upper triangular square root matrix of $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, defined by $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)} = \mathbf{U}_i'\mathbf{U}_i$.

Ideally, if an observation was not outlying then we must expect $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, to be nearly the same. Thus, $\mathbf{G}_i$ in above equation must be "close" to an identity matrix, for which all eigenvalues and hence all antieigenvalues are 1. Therefore, we may argue that an observation is outlying if there is considerable departure of these quantities from unity. It turns out (See Theorem 1 in Appendix) that except possibly for the smallest eigenvalue, all other eigenvalues of $\mathbf{G}_i$ must always be equal to 1. Thus it suffices to look at the departure of the smallest eigenvalue $\delta_{i,p}$ or equivalently the smallest antieigenvalue $\eta_{i,1}$ of $\mathbf{G}_i$ from unity.

All four data sets are subjected to this criterion as well. This criterion results in the identification of all outlying observations found previously by our earlier method. As it turns out, this is a more sensitive criterion in the sense it may also identify many more mildly outlying observations whose presence was perhaps obscured in our earlier approach when two antieigenvalues were compared for similar magnitudes. The results are given in Tables 7-10. Observations have been rearranged by the decreasing magnitudes of the $\eta_{i,1}$ values.

To be specific, Kendall's data shows observations 4 and 14 with significantly smaller values of $\delta_{i,p}$ and $\eta_{i,1}$ compared to other values. This is consistent with our previous evaluation. For Daniel-Wood data, observation numbers 3, 10 and 1 are found to be outlying (due to relatively larger x_5, x_2 and x_1 values). In case of Cork data of Rao, although none of the observations was determined to be severely outlying in our earlier analysis, a few more mildly outlying observations are now found by this approach (Observation numbers 12, 15, 16, 18, 19). This identification is consistent with what was observed by Khattree and Naik (1999) using various graphical methods such as biplots. Also, in case of Chatterjee, Hadi and Price's data, the observation number 3 (due to x_3 being relatively larger) and observation number 15 (due to x_1 and x_2 being relatively smaller) are also identified although their departures still appear to be subdued. These facts are graphically and more effectively illustrated using the *scree plots* given in Figure 1, where a sudden vertical drop, or lack of it, indicates the presence or absence of outlyingness.

One can also arrive at $\delta_{i,p}$ or $\eta_{i,1}$ as the yardsticks for outlyingness through certain other criteria. Specifically, to measure the distance between $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, one may consider as measures, the eigen or antieigenvalues of $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$, with respect to $\mathbf{X}'\mathbf{X}$ (see Rao, 2005) or alternatively, use the determinant of matrix $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}(\mathbf{X}'\mathbf{X})^{-1}$. In either case, in view of Theorem 1, the final criterion rests on $\delta_{i,p}$ and $\eta_{i,1}$.

Computationally, evaluation of eigenvalues $\delta_{i,p}$ and hence of antieigenvalues $\eta_{i,1}$ is rather straight forward and in fact, does not even require the explicit evaluation of eigenvalues or of the square root matrices – issues which can be a severe computational burden if the data set was large. To be specific, in view of Theorem 1, at most one of the eigenvalues of $\mathbf{G}_i(= \mathbf{U}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_i'$, where $\mathbf{U}_i$ is an upper triangular matrix so that $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)} = \mathbf{U}_i'\mathbf{U}_i)$ is not equal to 1 (in fact, less than 1). In view of Theorem 2, $\delta_{i,p} = tr(\mathbf{G}_i) - p + 1 = 1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. which is simply a quadratic form in the i^{th} data-row.

When using the $\mathbf{X}$ matrix, unlike the leverage value calculations, our approach to outlying observation detection as outlined here is *not* model based. Polynomial or cross-product terms which may be important in the model have not been considered and columns corresponding to these in the $\mathbf{X}^*$ matrix do play an important role in the computation of leverage values. Thus, $\delta_{i,p} = 1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ is in general *not* the model based leverage corresponding to the i^{th} observation (and this may be true in addition to the fact that in any model based approach, the $\mathbf{X}$ matrix contains a constant column corresponding to the intercept of the model). In view of this subtle difference, Khattree (2019) chooses to call $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ as *Emphasis* of the observation $\mathbf{x}_i$ rather than leverage. It will be same as leverage only if the assumed model had no intercept and no polynomial or cross product terms. In fact, that is exactly the point being made here. There may be observations which are simply different from the rest of the data without any consideration of assumed model whatsoever and they should be detected and examined at the very early stages of data cleaning prior to defining any specific model. The clean data may then be intended to be used as a reference dataset for a number of future studies. Consideration of antieigenvalues and *emphasis* measures help us do just that.

Multicollinearity and outlyingness can occasionally go hand in hand. Outlying observations can sometimes introduce or mask the multicollinearity. Fortunately, antieigenvalues can be utilized to assess both and hence provide a useful approach to identify "*collinearity - outlying*" points. We may measure the *collinearity - outlyingness* as the relative change in the antieigenvalues of $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$ compared to those of $\mathbf{X}'\mathbf{X}$. For $j = 1, 2, \cdots, r$,

$$\gamma_{i,j} = \frac{\eta_{i,j} - \eta_j}{\eta_j} \tag{3}$$

measures this change for $i = 1, 2, \cdots, n$. A similar measure can be defined in terms of the generalized antieigenvalue,

$$\Gamma_i = \frac{\Delta_i - \Delta}{\Delta}. \tag{4}$$

The measures in Equations (3) and (4) are computed for all four data sets. Instead of giving their values in several tables, it may be easier to plot and interpret $\gamma_{i,j}$ and Γ_i graphically. Specifically, we graphically identify the collinearity - outlyingness of individual data points by plotting $\gamma_{i,j}$ (or Γ_i) against theoretical (normal) quantiles in what is equivalent to a Q-Q plot. It is so, since the values of the nonoutlying points are likely to be more or less random (possibly approximately normally distributed). One set of such plots are given in Figure 2 for the four data sets. The statistics used is $\Delta^{1/r}$.

Table 7: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Kendall's Data]

Serial No.	Deleted	x_1	x_2	x_3	x_4	$\delta_{j,p}$	$\eta_{j,1}$
No.	Obs. (j)						
1	18	13.2	6.6	2.0	5.8	0.93	99.93
2	3	20.6	12.5	2.3	7.0	0.90	99.85
3	20	20.7	9.6	3.1	5.9	0.88	99.80
4	10	25.5	12.9	1.9	7.3	0.87	99.77
5	15	31.2	11.6	2.4	6.5	0.87	99.76
6	16	22.7	10.1	3.3	6.2	0.87	99.76
7	11	26.5	14.9	2.4	6.7	0.86	99.72
8	1	13.0	9.7	1.5	6.4	0.85	99.67
9	7	12.7	5.7	2.9	6.7	0.85	99.66
10	5	20.5	14.2	1.9	6.9	0.82	99.49
11	6	10.0	6.7	2.2	7.0	0.82	99.49
12	2	10.0	7.5	1.5	6.5	0.81	99.47
13	8	36.5	15.7	2.3	7.2	0.81	99.44
14	19	11.1	6.7	2.2	7.2	0.81	99.44
15	17	31.2	9.6	2.4	6.0	0.80	99.39
16	12	22.3	8.4	4.0	7.0	0.78	99.26
17	9	37.1	14.3	2.1	7.2	0.74	98.83
18	13	30.8	7.4	2.7	6.4	0.67	98.10
19	**14**	**25.3**	**7.0**	**4.8**	**7.3**	**0.58**	**96.34**
20	**4**	**33.8**	**19.0**	**2.8**	**5.8**	**0.48**	**93.68**

Remark: Most-outlying observations are highlighted in **bold**.

It is important to interpret these graphs in Figure 2 carefully. Larger positive values not falling on the straight line pattern indicate an improvement in terms of multicollinearity when the particular observation is deleted. In other words, these observations when present, tend to introduce multicollinearity. Similarly, observations with values which are more negative and away from the overall straight line pattern will tend to mask the multicollinearity. Thus from the graphs in Figures 2, it is easy to conclude that for Kendall's data, inclusion of observations numbered 4 and 14 tend to mask the problem, if any, of multicollinearity in the data. A more drastic instance of masking of multicollinearity is vividly seen in Daniel-Wood data. Observation number 3 clearly stands out at the lower left end of that figure. The 10^{th} observation also does so and masks some multicollinearity although it is not as excessive. For Hadi, Chatterjee and Price data, the inclusion of 10^{th} observation somewhat but not excessively increases the multicollinearity. As observed previously, this data set was already found to be highly ill-conditioned. Lastly, in case of Cork data, 18^{th} and 12^{th} observations are towards the higher end. Their inclusion possibly increases the multicollinearity slightly. However, as we have noted earlier, this data set is relatively well behaved in terms of multicollinearity.

It must be emphasized that this analysis of multicollinearity-outlyingness is irrelevant when with or without the particular observation, various measures of multicollinearity indicate a lack of it and is of interest only when the data exhibit a situation when the inclusion/exclusion of an observation drastically alters that situation and makes the data look much better or much worse than otherwise. This scenario is well illustrated by observation number 3 in case of Daniel-Wood data.

4. An evaluation of a large data set: red wine data

So far, we have purposely chosen data sets which could be presented in their entirety and allowed us to see the differences made by certain observations in the eigen-structure of the data. Would the large size of data obscure these changes since a single observation in a large dataset will supposedly have very small fraction of contribution to the overall structure? While this query is difficult to answer theoretically, we apply our approach on a relatively large data set available from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/`) consisting of 1599 observations on the quality of red wine. The four variables considered here are measurements on, x_1 = fixed acidity, x_2 = chlorides, x_3 = free sulfur dioxide and x_4 = pH for various wine samples.

As a process of data cleaning for such a large data, we will attempt to identify the observations which cause drastic changes in the eigen-structure using the techniques described earlier. We also evaluate the multicollinearity-outlyingness. As a representative subsample, the first ten, middle ten and last ten observations of the dataset are given in Table 11, along with a few cases that we have identified as outlying. Various statistics considered here have been calculated. Specifically, upon deleting one observation at a time for all 1599 observations, we compute the first and second antieigenvalues, and the generalized antieigenvalue. As a reference set to compare, we also compute these values for the entire data without deleting any observation. In view of small magnitudes of certain quantities, whenever needed, we have reported these quantities upto five decimal places.

Analysis results in four observations which stand out. For the observations numbered 82, 107, 152 and 259, the changes in the first antieigenvalue and the generalized antieigenvalue are substantial relative to entire data and compared to other cases of deletion. As can be seen in Table 11 (Columns 6 and 7) for the subsample listed above, compared to entire data, these quantities hardly change in the cases of one at a time deletions of the other 1595 observations.

Columns 8, 9 and 10 of the same table also present the values of $\gamma_{i,1}$, $\gamma_{i,2}$ and Γ_i which measure the collinearity- outlyingness. Clearly for the observations 82, 107, 152 and 259 these values by far stand out. The corresponding negative values indicate that these observations tend to mask the multicollinearity present in the data. For all other observations, the corresponding values are minuscule and practically insignificant.

Why do observations 152 and 259 stand out? It is clear that the value of x_2 is substantially larger while the value x_4 is substantially smaller compared to other observations. For observations 82 and 107, it is their large (but not as large as that for observations 152 and 259) x_2 values which make them outlying.

One may ask, “Why not just use the leverage function as specified in standard textbooks instead of the approach that we suggest?” If the usual leverage measure is used to identify outlying observations and if we use the recommended rule to identify outlying observations as those which have their leverage values higher than twice the mean leverage $(= \frac{2p}{n} = \frac{2\times 4}{1599} = 0.005)$, then that procedure ends up identifying a total of 189 observations as outlying! As an alternative recommendation, if we just choose those observations whose leverage values clearly stand out (with an existence of gap between them and rest of the leverage values) then it will identify observations 152 and 259 (leverages = 0.0806 and 0.0801 respectively) but not the observations 82 and 107 whose leverage values (= 0.0420 and 0.0416 respectively) are not as prominent compared to others. To keep these tables manageable, we have not printed all leverage values in Tables 11 and 12. Our approach, although more computer intensive, appears to be much more effective. Automating the above calculations can make such identifications quick and efficient, especially when the data may also contain other nonrandom noises such as freak values due to transcription errors.

Table 12 presents the analysis of the same data minus the observations numbered 82, 107 152 and 259. A comparison of corresponding entries in row 0 (*i.e.*, when “Deleted Obs.=none”) in Tables 11 and 12 shows that the removal of above four observations changes the statistics in Columns 6 and 7 of Table 11 substantially. Substantially smaller values in Table 12 suggest that this data set has much more multicolinearity than it originally showed, which was earlier masked by these four observations. Further, deletion of any of the remaining 1595 observations causes little change in the values of $\eta_{i,1}, \Delta_i, \gamma_{i,1}, \gamma_{i,2}$ and Γ_i, leading to the conclusion that there are no more outlying observations in the data set (However, an approach based on leverage values still declares 189 outlying observations!). Also the data set now has no outlying observation induced multicollinearity.

5. Concluding remarks

The work presented here introduces the use of eigen-structure and antieigenvalues for data cleaning early on after data collection but prior to modeling. This helps us identify and evaluate the quality of data and identify the possible anomalies within the data. Our work also supplements existing useful diagnostics techniques and are beneficial in providing the valuable insights into the data. One possibility to calibrate our proposed metrics may be via the probability distribution of the antieigenvalues. Some related work by Martin Singull can be found at `https://users.mai.liu.se/maroh70/pres/iclaa2017.pdf` .

One important recurring question is whether or not to center and/or standardize the data before subjecting them to these tools of analysis of antieigen-structure. It is obvious that

Table 8: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Daniel and Wood's Data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	$\delta_{j,p}$	$\eta_{j,1}$
1	6	22.32	6.17	2.85	66.47	2.43	0.86	99.72
2	9	21.96	4.65	6.06	64.07	2.32	0.81	99.47
3	12	21.34	6.07	2.93	67.03	2.56	0.81	99.45
4	4	24.60	5.85	2.80	64.18	2.40	0.79	99.30
5	2	25.96	3.48	5.06	63.15	2.32	0.78	99.23
6	13	21.94	5.57	2.68	67.71	2.44	0.77	99.12
7	8	23.54	4.83	7.21	62.03	2.24	0.71	98.50
8	11	22.48	5.00	7.46	62.72	2.24	0.70	98.40
9	14	25.72	4.12	6.06	61.05	2.08	0.68	98.23
10	5	25.04	3.86	2.11	66.57	2.36	0.64	97.55
11	7	20.93	4.64	5.74	66.26	2.08	0.60	96.80
12	**1**	**27.68**	**3.76**	**1.98**	**64.97**	**2.48**	**0.54**	**95.55**
13	**10**	**21.44**	**8.81**	**1.19**	**66.64**	**2.48**	**0.30**	**83.89**
14	**3**	**21.86**	**5.75**	**2.77**	**65.02**	**5.04**	**0.01**	**23.22**

Remark: Most-outlying observations are highlighted in **bold**.

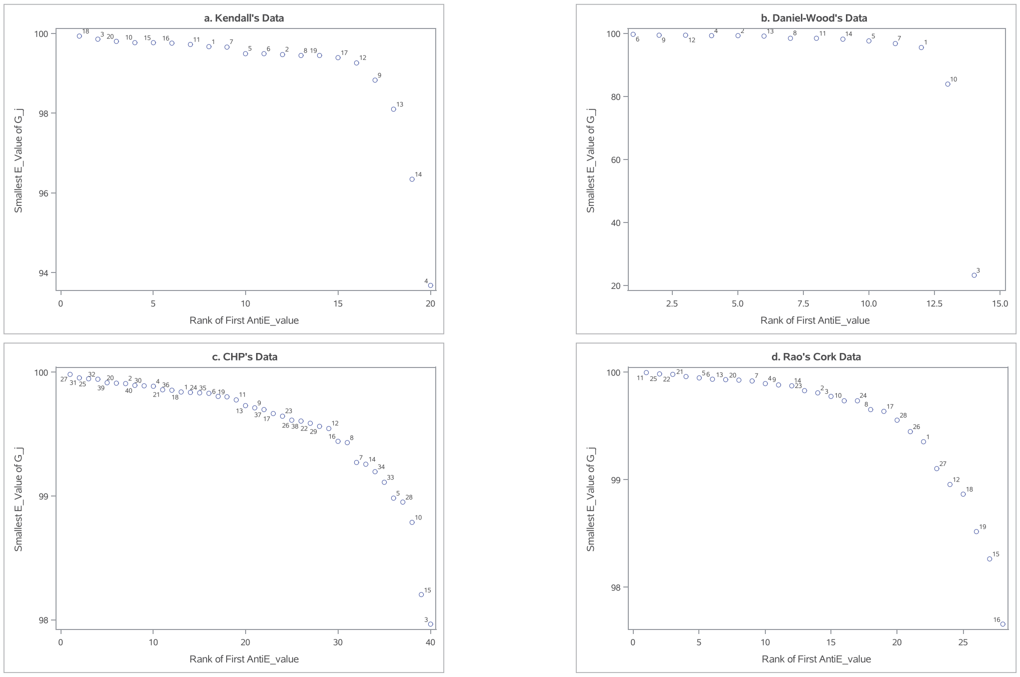

Figure 1: Scree plots for smallest antieigenvalues of $\mathbf{G}_i$ upon deleting an observation

Table 9: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [Chatterjee, Hadi and Price's data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	x_5	x_6	$\delta_{j,p}$	$\eta_{j,1}$
1	27	83	78	71	11	8	233	0.96	99.98
2	31	34	44	65	7	9	143	0.94	99.95
3	25	61	53	79	6	5	193	0.94	99.95
4	32	71	34	56	8	9	162	0.93	99.94
5	39	80	55	61	11	1	197	0.92	99.91
6	20	67	65	62	13	12	196	0.92	99.91
7	2	27	70	31	6	6	129	0.92	99.91
8	40	82	88	54	14	7	225	0.91	99.89
9	30	64	30	81	10	10	176	0.91	99.89
10	4	92	62	92	5	8	247	0.91	99.89
11	21	38	26	44	10	8	110	0.90	99.86
12	36	61	35	55	13	0	152	0.90	99.85
13	18	106	87	82	18	7	276	0.89	99.84
14	1	49	79	76	8	15	205	0.89	99.84
15	24	53	55	60	8	0	170	0.89	99.83
16	35	57	69	72	5	4	200	0.89	99.83
17	6	31	54	34	14	11	119	0.88	99.80
18	19	97	98	71	12	8	266	0.88	99.80
19	11	62	62	81	9	1	207	0.87	99.78
20	13	45	65	84	19	13	195	0.86	99.73
21	9	98	72	71	12	-1	242	0.86	99.71
22	37	29	45	47	13	13	123	0.86	99.70
23	17	78	102	84	5	7	266	0.85	99.66
24	23	54	100	50	11	15	205	0.84	99.64
25	26	60	108	104	17	8	273	0.84	99.61
26	38	82	105	81	20	9	268	0.84	99.60
27	22	56	32	99	16	8	188	0.83	99.59
28	29	89	121	71	8	8	283	0.83	99.56
29	12	25	11	7	9	9	45	0.83	99.55
30	16	111	52	93	11	13	256	0.81	99.44
31	8	114	85	84	17	20	285	0.81	99.43
32	7	105	60	47	5	10	212	0.78	99.27
33	14	92	75	63	9	20	232	0.78	99.26
34	34	112	105	123	5	12	340	0.78	99.19
35	33	88	30	87	13	0	207	0.77	99.11
36	5	67	42	94	16	3	202	0.75	98.98
37	28	74	125	66	16	4	265	0.75	98.95
38	10	15	59	99	15	11	174	0.73	98.79
39	**15**	**27**	**26**	**82**	**4**	**17**	**134**	**0.68**	**98.20**
40	**3**	**115**	**92**	**130**	**0**	**9**	**339**	**0.67**	**97.96**

Remark: Most-outlying observations are highlighted in **bold**.

Table 10: Detecting outlyingness, raw data and smallest eigenvalue δ_p values, smallest antieigenvalue $\eta_{j,1}(\times 100)$ for $\mathbf{G}_j$ Matrix [C. R. Rao's Cork data]

Serial No.	Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\delta_{j,p}$	$\eta_{j,1}$
1	11	32	30	34	28	0.98	99.99
2	25	39	36	39	31	0.97	99.98
3	22	32	30	30	32	0.96	99.98
4	21	39	35	34	37	0.94	99.96
5	5	32	32	35	36	0.94	99.94
6	6	30	35	34	26	0.93	99.93
7	13	54	46	60	52	0.93	99.93
8	20	46	38	37	38	0.93	99.93
9	7	39	39	31	27	0.92	99.92
10	4	41	29	36	38	0.91	99.90
11	9	37	40	31	25	0.91	99.88
12	14	47	51	52	43	0.90	99.87
13	23	60	50	67	54	0.89	99.83
14	2	60	53	66	63	0.88	99.81
15	3	56	57	64	58	0.87	99.77
16	10	33	29	27	36	0.86	99.74
17	24	35	37	48	39	0.86	99.74
18	8	42	43	31	25	0.85	99.65
19	17	79	65	70	61	0.84	99.63
20	28	48	54	57	43	0.83	99.55
21	26	50	34	37	40	0.81	99.45
22	1	72	66	76	77	0.80	99.35
23	27	43	37	39	50	0.76	99.10
24	**12**	**63**	**45**	**74**	**63**	**0.75**	**98.96**
25	**18**	**81**	**80**	**68**	**58**	**0.74**	**98.86**
26	**19**	**78**	**55**	**67**	**60**	**0.71**	**98.51**
27	**15**	**91**	**79**	**100**	**75**	**0.69**	**98.26**
28	**16**	**56**	**68**	**47**	**50**	**0.65**	**97.65**

Remark: Most-outlying observations are highlighted in **bold**.

Table 11: Analysis of red wine data: complete data (n = 1599)

Deleted Obs. (j)	x_1	x_2	x_3	x_4	$\eta_{j,1}$ ($\times$100)	Δ_j ($\times$100)	$\gamma_{j,1}$	$\gamma_{j,2}$	Γ_j
none	**.**	**.**	**.**	**.**	**0.462**	**0.124**	**0**	**0**	**0**
1	7.4	0.076	11	3.51	0.462	0.124	0.01259	-0.00722	-0.00360
2	7.8	0.098	25	3.20	0.463	0.124	0.04972	0.05319	0.02659
3	7.8	0.092	15	3.26	0.462	0.124	0.02127	0.02221	0.01111
4	11.2	0.075	17	3.16	0.462	0.124	0.02282	-0.01682	-0.00841
5	7.4	0.076	11	3.51	0.462	0.124	0.01259	-0.00722	-0.00360
6	7.4	0.075	13	3.51	0.462	0.124	0.01615	-0.00458	-0.00229
7	7.9	0.069	15	3.30	0.462	0.124	0.01886	0.02010	0.01005
8	7.3	0.065	15	3.39	0.462	0.124	0.01720	0.00219	0.00110
9	7.8	0.073	9	3.36	0.462	0.124	0.00872	0.01594	0.00797
10	7.5	0.071	17	3.35	0.462	0.124	0.02458	0.01650	0.00825
82	**7.8**	**0.464**	**22**	**3.13**	**0.453**	**0.122**	**-2.00423**	**-2.00403**	**-1.00708**
107	**7.8**	**0.467**	**18**	**3.08**	**0.453**	**0.122**	**-2.06405**	**-2.06336**	**-1.03705**
152	**9.2**	**0.610**	**32**	**2.74**	**0.445**	**0.119**	**-3.83596**	**-3.89969**	**-1.96923**
259	**7.7**	**0.611**	**8**	**3.06**	**0.444**	**0.119**	**-4.02277**	**-4.00309**	**-2.02198**
797	8.7	0.126	24	3.10	0.462	0.124	0.02903	0.01822	0.00911
798	9.3	0.038	21	3.24	0.462	0.124	0.00023	-0.00877	-0.00438
799	9.4	0.082	5	3.29	0.462	0.124	0.00489	0.06907	0.03453
800	9.4	0.082	5	3.29	0.462	0.124	0.00489	0.06907	0.03453
801	7.2	0.082	26	3.25	0.463	0.124	0.05499	0.06207	0.03103
802	8.6	0.068	8	3.23	0.462	0.124	0.00490	0.03901	0.01951
803	5.1	0.044	14	3.56	0.462	0.124	0.00432	-0.12472	-0.06238
804	7.7	0.114	14	3.24	0.462	0.124	0.00626	0.00741	0.00371
805	8.4	0.084	4	3.26	0.462	0.124	0.00423	0.05255	0.02628
806	8.2	0.052	4	3.33	0.462	0.124	-0.01058	0.02851	0.01426
1590	6.6	0.073	29	3.29	0.463	0.124	0.06490	0.08170	0.04084
1591	6.3	0.077	26	3.32	0.463	0.124	0.05366	0.05161	0.02581
1592	5.4	0.089	16	3.67	0.462	0.124	0.02095	-0.10075	-0.05038
1593	6.3	0.076	29	3.42	0.463	0.124	0.06502	0.07259	0.03630
1594	6.8	0.068	28	3.42	0.463	0.124	0.05875	0.06587	0.03293
1595	6.2	0.090	32	3.45	0.463	0.124	0.07723	0.10128	0.05063
1596	5.9	0.062	39	3.52	0.463	0.124	0.10568	0.18234	0.09113
1597	6.3	0.076	29	3.42	0.463	0.124	0.06502	0.07259	0.03630
1598	5.9	0.075	32	3.57	0.463	0.124	0.07692	0.08773	0.04386
1599	6.0	0.067	18	3.39	0.462	0.124	0.02646	-0.02163	-0.01081

Remark: Most-outlying observations are highlighted in **bold**. *Top row corresponds to entire data with no deletion.*

Table 12: Analysis of red wine data: after deleting obs. No. 82, 107, 152 and 259 ($n = 1595$)

Deleted	x_1	x_2	x_3	x_4	$\eta_{j,1}$	Δ_j	$\gamma_{j,1}$	$\gamma_{j,2}$	Γ_j
Obs. (j)					(×100)	(×100)			
none	**.**	**.**	**.**	**.**	**0.132**	**0.0793**	**0**	**0**	**0**
1	7.4	0.076	11	3.51	0.132	0.0792	-0.10861	-0.12180	-0.060853
2	7.8	0.098	25	3.20	0.132	0.0793	-0.02016	0.07747	0.038792
3	7.8	0.092	15	3.26	0.132	0.0792	-0.04344	-0.03128	-0.015577
4	11.2	0.075	17	3.16	0.132	0.0792	-0.01573	-0.03490	-0.017385
5	7.4	0.076	11	3.51	0.132	0.0792	-0.10814	-0.12132	-0.060614
6	7.4	0.075	13	3.51	0.132	0.0792	-0.10118	-0.10024	-0.050070
7	7.9	0.069	15	3.30	0.132	0.0792	-0.04437	-0.03356	-0.016718
8	7.3	0.065	15	3.39	0.132	0.0792	-0.07958	-0.06261	-0.031244
9	7.8	0.073	9	3.36	0.132	0.0792	-0.07065	-0.10248	-0.051188
10	7.5	0.071	17	3.35	0.132	0.0792	-0.05981	-0.02949	-0.014680
797	8.7	0.126	24	3.10	0.132	0.0793	0.00877	0.03356	0.016842
798	9.3	0.038	21	3.24	0.132	0.0793	0.01014	0.02349	0.011809
799	9.4	0.082	5	3.29	0.132	0.0793	0.02025	-0.00452	-0.002193
800	9.4	0.082	5	3.29	0.132	0.0793	0.02030	-0.00437	-0.002119
801	7.2	0.082	26	3.25	0.132	0.0793	0.01850	0.04528	0.022701
802	8.6	0.068	8	3.23	0.132	0.0793	0.02001	0.00788	0.004003
803	5.1	0.044	14	3.56	0.132	0.0792	-0.11208	-0.11019	-0.055043
804	7.7	0.114	14	3.24	0.132	0.0793	0.01638	0.01733	0.008728
805	8.4	0.084	4	3.26	0.132	0.0793	0.01609	0.00342	0.001776
806	8.2	0.052	4	3.33	0.132	0.0793	0.01011	-0.00048	-0.000178
1590	6.6	0.073	29	3.29	0.132	0.0793	0.08935	0.04278	0.021454
1591	6.3	0.077	26	3.32	0.132	0.0793	0.09028	0.02157	0.010850
1592	5.4	0.089	16	3.67	0.132	0.0793	0.04197	0.02997	0.015049
1593	6.3	0.076	29	3.42	0.132	0.0793	0.09004	0.03221	0.016167
1594	6.8	0.068	28	3.42	0.132	0.0793	0.08935	0.04278	0.021454
1595	6.2	0.090	32	3.45	0.132	0.0793	0.09028	0.02157	0.010850
1596	5.9	0.062	39	3.52	0.132	0.0793	0.09030	-0.00125	-0.000560
1597	6.3	0.076	29	3.42	0.132	0.0793	0.09051	0.03229	0.016210
1598	5.9	0.075	32	3.57	0.132	0.0793	0.08664	0.01229	0.006207
1599	6.0	0.067	18	3.39	0.132	0.0793	0.08434	0.06733	0.033726

Remark: Upon deletion of observation numbers 82, 107, 152 and 259, there are no more outlying observations. Top row corresponds to entire data (1599 - 4 = 1595 obs.) with no further deletions.

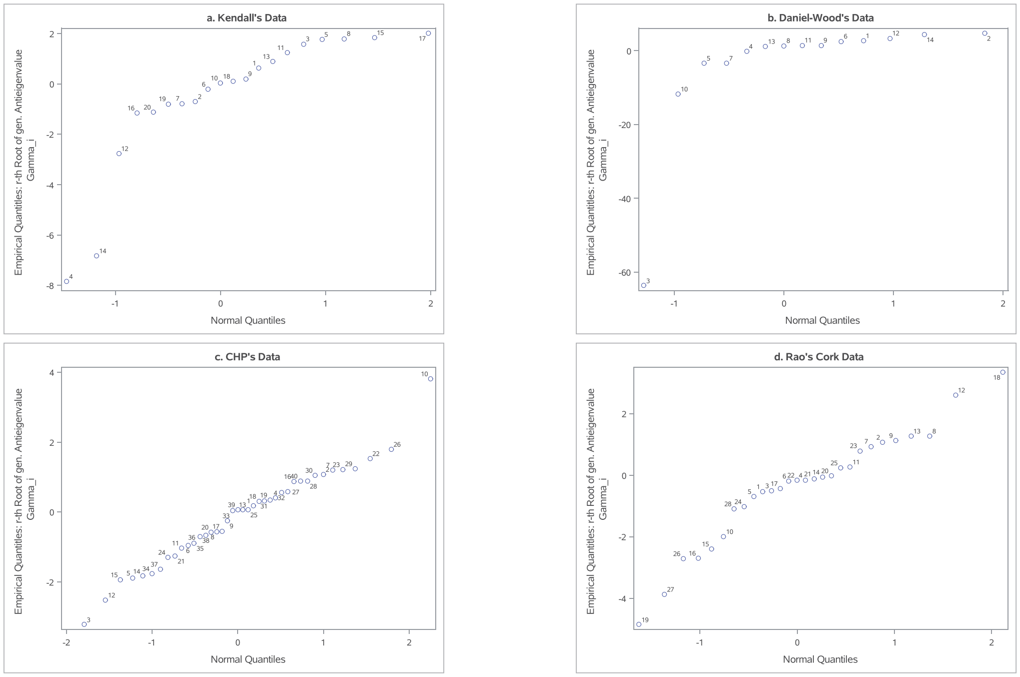

Figure 2: Scree plots for smallest antieigenvalues of $\mathbf{G}_i$ upon deleting an observation

the two sets of antieigenvalues will differ for standardized and unstandardized data. Shifting and scaling have substantial effects on eccentricities of the corresponding $p-$ dimensional ellipsoids. This is especially so since both the means as well as standard deviations can be very sensitive to outliers. Naik and Khattree (1996) provide a detailed discussion of pitfalls of blindly standardizing the data. Our (obvious) suggestion, thus, is to look into both possibilities since the use of unstandardized and standardized data for the subsequent modeling purposes are both acceptable practices.

Another natural question one may pose is, how about identification of a subset of s most outlying observations? How can we assess whether or not a particular subset of observations as a whole, is outlying? Khattree (2019) has extensively dealt with this issue with *Emphasis* as a measure. As it turns out, under that criterion, this problem is equivalent to sequential identification and deletion of observations, one by one, based on the maximum outlyingness using the matrices $\mathbf{G}_i$ and the method described in the present work. Thus, our approach here, at least partially eliminates the problem of first determining an appropriate choice of s and then looking at the computationally intensive task of evaluating the outlyingness of all possible ${}^nC_s$ subsets of s observations. When antieigenvalues are used as the criteria, whether or not such a sequential deletion is possible, is an issue that is still needed to be explored.

The numbers of observations as well as number of variables play important roles and an observation may have substantially less outlying effect on eigen-structure if it was only one of the several thousand observations. We see this in case of red wine data when out of a total of 1599 observations only four stand out. Thus it is difficult to give a firm and universal threshold value for the determination of an observation's outlyingness. For large data sets, graphical methods and plots such as those given in this work provide a useful approach to identify such patterns.

The approach is admittedly computer intensive and thus there is a genuine need for developing an efficient and effective algorithm based on the methodology presented here. This is an important direction for future work as contemporary data sets are often very large in terms of number of variables as well as number of data points. We have not addressed these algorithmic-efficiency issues in this paper. Further research is needed in this direction.

We must also realize that the context here is that of data cleaning for a large dataset without any assumed model. Such data will often have missing values. With a model-free approach, it will be difficult to incorporate various types of missingness in our approach. However, assuming that the missingness is *completely at random*, our suggestion is to first impute the missing values appropriately and then perform the data cleaning. Since we intend to not assume any model on data, an approach to imputation based on empirical copula has been suggested by Lun and Khattree (2019, 2020, 2024). To what extent the performance may be affected by imputation is yet another aspect which can be explored via simulation studies.

What happens if the number of variables is greater than the number of observations? The genomics data, which very often are inevitably quite noisy, typically have this situation. How to clean data in such a case is admittedly a difficult problem. However, this situation allows us to address a very different problem still relevant in the context. Since our approach is based entirely on eigen-structure and since the nonzero eigenvalues of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are same, for "$p > n$" situation, our approach can perhaps be adopted to evaluate the quality of variables. Since the interpretations entirely change, this problem needs a further careful consideration.

Acknowledgment and competing interests

The author wishes to thank two referees for their thoughtful comments. This work was supported, in part, by funding from the NSF Award FAIN: 2244091. The author declares no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. Finally, this article is dedicated to the memory of Professor C. R. Rao.

Supplementary Material

The original Wine data are available from the website

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality .

ORCID Information

Ravindra Khattree: 0000-0002-9305-2365

References

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* John Wiley and Sons: New York.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons: New York.

Chatterjee, S. and Hadi, A. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley and Sons: New York.

Chatterjee, S., Hadi, A., and Price, B. (2006). *Regression Analysis by Example.* John Wiley and Sons: New York.

Chu, X. and Ilyas, I. F. (2016). Qualitative data cleaning. *Proceedings of the VLDB Endowment*, **9**, 1605–1608.

Chu, X., Ilyas, I. F., Krishnan, S., and Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2201–2206.

Cuntoor, N. P. and Chellappa, R. (2006). Key frame-based activity representation using antieigenvalues. In *Asian Conference on Computer Vision*, pages 499–508. Springer.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data.* John Wiley and Sons: New York.

Guo, C., Jin, M., Guo, Q., and Li, Y. (2018). Antieigenvalue-based spectrum sensing for cognitive radio. *IEEE Wireless Communications Letters*, **8**, 544–547.

Ilyas, I. F. and Chu, X. (2019). *Data Cleaning.* Association for Computing Machinery.

Johnson, R. A. and Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis.* Pearson: London, UK.

Kendall, M. G. (1975). *Multivariate Analysis.* Griffin: London.

Khattree, R. (2001). On the calculation of antieigenvalues and antieigenvectors. *Journal of Interdisciplinary Mathematics*, **4**, 195–199.

Khattree, R. (2002). On generalized antieigenvalue and antieigenmatrix of order r. *American Journal of Mathematical and Management Sciences*, **22**, 89–98.

Khattree, R. (2003). Antieigenvalues and antieigenvectors in statistics. *Journal of Statistical Planning and Inference*, **114**, 131–144.

Khattree, R. (2010). Antieigenvalues provide a bound on realized signal to noise ratio. *Journal of Statistical Planning and Inference*, **140**, 2846–2848.

Khattree, R. (2014). Antieigenvalues and antieigenvectors. *Wiley Stats Ref: Statistics Reference Online*, **1**, 131–134.

Khattree, R. (2019). A note on effects on the eigenstructure of a data matrix when deleting a subset of observations. *Journal of the Indian Society of Agricultural Statistics*, **73**, 11–17.

Khattree, R. and Bahuguna, M. (2019). An alternative data analytic approach to measure the univariate and multivariate skewness. *International Journal of Data Science and Analytics*, **7**, 1–16.

Khattree, R. and Naik, D. N. (1999). *Applied Multivariate Statistics with SAS Software, Second Edition.* SAS Publishing: Cary NC/John Wiley and Sons: New York.

Lun, Z. and Khattree, R. (2019). Multiple imputation for skewed multivariate data: A marriage of the MI and COPULA procedures. In *Proceedings of the SAS Global Forum, Paper 3605-2019*.

Lun, Z. and Khattree, R. (2020). Imputation for non-normal multivariate continuous data using copula transformation. *Proceedings of Joint Statistical Meeting, 2020 - Survey Research Methods Section*, 1922–1930.

Lun, Z. and Khattree, R. (2024). A general approach for imputation of non-normal continuous data based on copula transformation. *Communications in Statistics-Simulation and Computation*, **53**, 567–594.

Mason, R. L. and Gunst, R. F. (1985). Outlier-induced collinearities. *Technometrics*, **27**, 401–407.

Naik, D. N. and Khattree, R. (1996). Revisiting Olympic track records: Some practical considerations in the principal component analysis. *The American Statistician*, **50**, 140–144.

Rao, C. R. (1948). Tests of significance in multivariate analysis. *Biometrika*, **35**, 58–79.

Rao, C. R. (2005). Antieigenvalues and antisingularvalues of a matrix and applications to problems in statistics. *Research Letters in the Information and Mathematical Sciences*, **8**, 53–76.

Timm, N. H. (2002). *Applied Multivariate Analysis*. Springer: Switzerland.

Tran, N. Q. H. and Khattree, R. (2024). Supervised learning via eigen-structures. *Preprint, Under preparation*, .

Wang, S.-G. and Nyquist, H. (1991). Effects on the eigenstructure of a data matrix when deleting an observation. *Computational Statistics and Data Analysis*, **11**, 179–188.

APPENDIX

Appendix 1

Two theorems referred in main text are stated here. These are the special cases of results given in Khattree (2019) Proofs have been omitted.

Theorem 1. *Let* $\mathbf{X}$ *be an* $n \times p$ *matrix with* $n > p$ *and rank* p*. Define* $\mathbf{A} = \mathbf{X}'\mathbf{X}$ *and* $\mathbf{B}_i = \mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$ *and suppose* $\mathbf{B}_i = \mathbf{U}_i'\mathbf{U}_i$ *where* $\mathbf{U}_i$ *is upper triangular. Then, for the ordered eigenvalues* $\delta_1 \geq \delta_2 \geq ... \geq \delta_p$ *of* $\mathbf{G}_i = \mathbf{U}_i\mathbf{A}^{-1}\mathbf{U}_i'$*,* $\delta_j = 1$ *for* $j = 1, 2, ..., (p-1)$*.*

Theorem 2. *The smallest eigenvalue of* $\mathbf{G}_i = \mathbf{U}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{U}_i'$ *where* $\mathbf{U}_i$ *is the upper triangular square root matrix of* $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}$*, defined by* $\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)} = \mathbf{U}_i'\mathbf{U}_i$ *is* $\delta_{i,p} = 1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$*.*

Appendix 2

SAS code which generated Table 11

```
/*
data on red wine from :
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Attribute information:

   For more information, read [Cortez et al., 2009].

   Input variables (based on physicochemical tests):
   1 - fixed acidity
   2 - volatile acidity
   3 - citric acid
   4 - residual sugar
   5 - chlorides
   6 - free sulfur dioxide
   7 - total sulfur dioxide
   8 - density
   9 - pH
   10 - sulphates
   11 - alcohol
   Output variable (based on sensory data):
   12 - quality (score between 0 and 10)

*/

data wine; *Red wine data;
infile "C:\Users\Desktop\winequality.txt" ;
input y1-y12;
run;
```

```
data wine; set wine;
x1 = y1; *x1 = fixed acidity;
x2 = y5; *x2 = chlorides;
x3 = y6; *x3 = free sulfur dioxide;
x4 = y9; *x4 = pH;
keep x1-x4;
run;
proc standard data=wine mean=0 std=1
              out=stndtest;
     var x1-x4;
run;
%let mydataset = wine; ***use this to analyze original raw data;
*%let mydataset = stndtest; ***use this to analyze standardized data;


*options nolog;         ***suppresses log file**;


%macro multicol(count = );

%do del_row = 0 %to &count;
data uci; set &mydataset;run;
proc iml; use uci; read all into x;
xpx = x‘*x;
if &del_row = 0 then do; smallxpx = xpx; end;
if &del_row > 0 then do;
t&del_row =x[ {&del_row}, ];

smallxpx = xpx - t&del_row‘*t&del_row;end;
call eigen(lambda, p, smallxpx);
create evalues from lambda;
append from lambda;close evalues;
quit;
proc transpose data = evalues out = evaluesvar; run;
data evaluesvar&del_row; set evaluesvar;
anti1 =100*2*sqrt(col1*col4)/(col1+col4);
anti2 = 100*2*sqrt(col2*col3)/(col2+col3);
gen_anti = anti1*anti2/100;
rt_gen_anti = (anti1*anti2)**(1/2) ;
obser = &del_row;
run;
proc datasets library=work nolist;
   append base=work.antieig data=work.evaluesvar&del_row force;
run;
proc delete library = work data = evaluesvar&del_row;run;
%end;
```

```
%mend multicol ;
%multicol(count = 1599); ***no. of complete obs = count = 1599;
title;
footnote "Actual values are multiplied by 100";
data wine; set wine; obser = _n_;run;
proc sort data = wine; by obser;run;
proc sort data = antieig; by obser;run;
data combine; merge wine antieig; by obser;run;
*****************Calculation of gamma values of  Section 3*********;
data gamma; set antieig;

***The numbers below are obtained from the output
when no observations were deleted.;

anti1all = .46228;;
gamma1 = 100*(anti1-anti1all)/anti1all;
gen_antiall = .12419;
gamma2 = 100*(gen_anti -gen_antiall)/gen_antiall;
rt_gen_antiall = 3.52406;
gamma3 = 100*(rt_gen_anti -rt_gen_antiall)/rt_gen_antiall;
run;
proc sort data = wine; by obser;run;
proc sort data = gamma; by obser;run;
data combine2; merge wine gamma;
by obser;run;
data antieigsmall2; set combine2;
if (obser in (0 82 107 152 259) or obser < 11 or obser gt 1589
or (obser > 796 and obser < 807));
run;
data Table11; set antieigsmall2;
keep obser x1 x2 x3 x4 anti1   gen_anti gamma1 gamma2 gamma3 ;
run;
proc export  data = table11
outfile =
'C:\Users\khattree\Desktop\DataCleaningTable11OfPaper.txt'
replace; ***Output is stored in the .txt file;
run;
```