Annals of the Institute of Statistical Mathematics (2021) 73:737–756
https://doi.org/10.1007/s10463-020-00776-x

Determining the number of canonical correlation pairs for high-dimensional vectors

Jiasen Zheng[1] · **Lixing Zhu**[2,3]

Received: 3 February 2020 / Revised: 26 June 2020 / Accepted: 18 August 2020 /
Published online: 12 February 2021
© The Institute of Statistical Mathematics, Tokyo 2021

Abstract
For two random vectors whose dimensions are both proportional to the sample size, we in this paper propose two ridge ratio criteria to determine the number of canonical correlation pairs. The criteria are, respectively, based on eigenvalue difference-based and centered eigenvalue-based ridge ratios. Unlike existing methods, the criteria make the ratio at the index we want to identify stick out to show a visualized "valley-cliff" pattern and thus can adequately avoid the local optimal solutions that often occur in the eigenvalues multiplicity cases. The numerical studies also suggest its advantage over existing scree plot-based method that is not a visualization method and more seriously underestimates the number of pairs than the proposed ones and the AIC and C_p criteria that often extremely over-estimate the number, and the BIC criterion that has very serious underestimation problem. A real data set is analyzed for illustration.

Keywords Canonical correlation matrix · Eigenvalue-based ridge ratios · High dimensionality · The number of canonical correlation pairs

1 Introduction

As the seminal work by Hotelling (1936), canonical correlation analysis (CCA) has been a basic approach in statistics to capture the most of correlation between two multidimensional vectors through much less number of significant canonical variate pairs that are the linear combinations of the original sets of variables. How to

✉ Lixing Zhu
lzhu@hkbu.edu.hk

Jiasen Zheng
2017000815@ruc.edu.cn

[1] School of Statistics, Renmin University of China, Beijing 100872, China

[2] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

[3] School of Statistics, Beijing Normal University, Beijing 100875, China

determine this number to achieve dimension reduction is an important issue. When these vectors are of fixed dimensions, the theory has been very mature. In this paper, we consider this issue when the dimensions are large in the sense that they are proportional to the sample size.

Let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ be two multivariate random vectors with finite second moments and partition the cross-covariance matrix as follows

$$\Sigma_{xy} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{1}$$

where Σ_{11} is the covariance with respect to $\mathbf{x}$ of dimension p, Σ_{22} associates with $\mathbf{y}$ of dimension q, and $\Sigma_{12}^{\top} = \Sigma_{21}$ is the covariance between $\mathbf{x}$ and $\mathbf{y}$. Let $S_{11}, S_{22}, S_{12}^{\top} = S_{21}$ be the sample counterparts to be specified later.

To be precise, as one of the powerful methodologies, CCA realises dimension reduction through exploring the new linear combination of each set to maximize the correlation between the new linear combination pairs. It allows us to summarize the relationship into a much smaller number of new pairs called canonical variates while preserving the main facets of the associations. Specifically, it is to find the linear combinations $(\mathbf{a}^{\top}\mathbf{x}, \mathbf{b}^{\top}\mathbf{y})$ of each set to maximize the canonical correlation

$$\rho = \frac{\mathrm{Cov}(\mathbf{a}^{\top}\mathbf{x}, \mathbf{b}^{\top}\mathbf{y})}{\sqrt{\mathrm{Var}(\mathbf{a}^{\top}\mathbf{x})} \cdot \sqrt{\mathrm{Var}(\mathbf{b}^{\top}\mathbf{y})}} = \frac{\mathbf{a}^{\top}\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}^{\top}\Sigma_{11}\mathbf{a}} \cdot \sqrt{\mathbf{b}^{\top}\Sigma_{22}\mathbf{b}}}, \tag{2}$$

subject to the constraints

$$\mathbf{a}^{\top}\Sigma_{11}\mathbf{a} = \mathbf{b}^{\top}\Sigma_{22}\mathbf{b} = 1,$$

where $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ are called canonical directions. The pairs $(\mathbf{a}^{\top}\mathbf{x}, \mathbf{b}^{\top}\mathbf{y})$ are named canonical variates and the number of nonzero pairs of canonical variates or the number of nonzero canonical correlations is called the dimensionality in CCA.

In this paper, we consider the issue of determining the number of canonical correlation pairs, which can be transferred to the rank of the CCA-based matrix under high-dimensional framework. As mentioned in Hotelling (1936), the optimum solution of (2) can be obtained by directly applying the singular value decomposition (SVD) on the matrix

$$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}.$$

At the sample level, we usually use the corresponding sample covariance matrix instead. Based on the collected sample pairs $\{(\mathbf{x}_j, \mathbf{y}_j) : j = 1, ..., n\}$, note the sample matrix as follows

$$S_{11}^{-1/2}S_{12}S_{22}^{-1/2}. \tag{3}$$

The sample pairs of the canonical variates are

$$U_i = u_i^{\top}S_{11}^{-1/2}\mathbf{x}, \quad V_i = v_i^{\top}S_{22}^{-1/2}\mathbf{y},$$

where u_i, v_i are left and right singular vectors in (3), respectively.

To determine the number of pairs, when p and q are fixed, Akaike's information criterion (AIC) (Akaike 1973) and modified Mallows's C_p (Mallows 1973) which originally from model selection criterion for regression analysis, have been used, see Fujikoshi and Veitch (1979) and Fujikoshi (1985). The criterion based on Bayesian information criterion (BIC) (Schwarz 1978) in canonical correlation analysis was studied by Gunderson and Muirhead (1997). To facilitate the determination in diverging dimension scenarios, Fujikoshi and Sakurai (2009) considered the asymptotic distribution of canonical correlations with $p/n \to c \in [0, 1)$ while q fixed. Fujikoshi (2017a) investigated the asymptotic results under milder conditions. When both p and q go to infinity as n tends to infinity, Bao et al. (2019) gave a more thorough investigation on these topics and used the typical scree plot approach to determine the rank. By using a penalty or a threshold as the tuning parameter, these methods either use the global maximizer/minimizer as the estimators or the minimizer over all indices of the quantities that are larger than a threshold value. However, AIC is not consistent in theory and BIC may seriously rely on the selection of the penalty. When there are equal eigenvalues at the population level, the scree plot-based estimation would underestimate the true rank/number because it has difficulty to well separate those outliers of all eigenvalues. Further, as we know, visualization is a very useful auxiliary tool for practical use. When p and q are large and the number of canonical correlation pairs could also be relatively large, the scree plot is difficult to well present the separation between the outliers and the bulk of other eigenvalues. Several other heuristic approaches for selecting rank, under high-dimensional CCA, have also been studied by Song et al. (2015, 2016). They get the canonical correlation by two steps. Specially, first to do principal component analysis (PCA) for $\mathbf{x}$ and $\mathbf{y}$ to extract the principal components that account for a large fraction of the total variance. Second, to perform CCA in the new low-dimensional sets. But how to use this PCA-CCA technique to decide the number of principal components totally relies on experience. The principal components accounting for each set does not directly explain the correlation between two sets. Further, when the dimensions p and q are proportional to the sample size, the asymptotic behaviors of the PCA and CCA matrix at the sample level is rather different from those of the corresponding matrices at the population level (see Bai et al. (2018) and Bao et al. (2019)), it needs some theoretical investigation on the consistency of the estimator.

In this paper, we attempt to define some criteria that could make a local minimizer as the estimator of k significantly sticking out from all local minimizers. We propose two ridge ratio-based criteria for this purpose. The new criteria have some desirable features. First, thanks to the use of ridge c_n as a tuning parameter that will be specified in Sect. 3 when we construct ratios, the criteria have "valley-cliff" patterns such that the numbers of the corresponding quantities in the criteria can be well isolated and identified at "valley bottom". The identification can be even visualized by plots. This unique nature makes the determination much easier in practice than existing criteria in the literature. Second, the estimation is consistent. Third, they can better handle the multiplicity of nonzero eigenvalues to avoid the underestimation problem some existing criteria encounters in practice. Fourth, again due to the " valley-cliff" pattern, the criteria can also alleviate the multiple local minima problem that could also cause the underestimation problem. To give a better idea about the advantages of the new criteria, we will see from the curves of different criteria presented in Sects. 3 and 5, the centered eigenvalue-based ratio criterion we will propose even has better performance. We also want to point out a limitation of the methods, like any criterion in the literature, we indispensably need to select tuning parameters that could have impact for the performance of the methods. A data-driven approach would be in demand. This issue is beyond the scope of this paper. The investigation is ongoing. It is worth of mentioning that when using the original estimated eigenvalues to construct a criterion to identify the order of matrix, the ridge ratio-based idea has been considered in Zhu et al. (2020). Their method can be feasible when p and q are fixed or would be possibly useful when $p/n \to 0$ and $q/n \to 0$, and p and q are divergent to infinity. However, for the problem studied in this paper with $p/n \to c_1 > 0$ and $q/n \to c_2 > 0$, the completely different asymptotic results of the estimated eigenvalues cause the failure of their method that is not possible to separate the ratio at the order we want to identify and the others. The reasons behind this difficulty will be described in Sect. 3. Thus, we propose eigenvalue difference-based and centered eigenvalue-based objective function to handle the estimation problem in this paper.

The rest of this paper is organized as follows. In Sect. 2 we introduce some necessary preliminaries. The criterion and asymptotic properties are displayed in Sect. 3. Simulation results are conducted in Sect. 4 to compare the proposed estimation with other methods. The analysis of a real data example is presented in Sect. 5. Some technical proofs and numerical studies are included in the supplementary material.

2 Preliminary facts and assumptions

Without loss of generality, in this context we assume $p \geq q$. From Hotelling (1936), the canonical correlation coefficients ρ_i between $\mathbf{x}$ and $\mathbf{y}$ are known as the nonzero singular values of matrix $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$. By Rayleigh quotient, the nonzero singular values of matrix $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2}$ are equivalent to the square roots of the nonzero eigenvalues λ_i of matrix $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, which is called canonical correlation matrix. Thus, the optimum solution of (2) can be obtained by solving the characteristic equation of the latter matrix. To ease the presentation, give the notation as

$$C_{12} := \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{4}$$

Similarly, one can define C_{21} by interchanging the roles of $\mathbf{x}$ and $\mathbf{y}$ in C_{12}. Notice that C_{12} and C_{21} share the same nonzero eigenvalues. Hence, we take matrix C_{12} into consideration throughout the rest of the paper and C_{21} can be similarly handled. Write the eigenvalues of C_{12} in descending order

$$\lambda_1 \geq \lambda_2 \geq \cdots = \lambda_{q_1} > \lambda_{q_1+1} = \cdots = \lambda_q = 0,$$

assuming that at least $\lambda_q = 0$ and the rank of Σ_{12} is equal to q_1.

In estimation, write the sample canonical correlation matrix as

$$S_{11}^{-1}S_{12}S_{22}^{-1}S_{21},$$

and the corresponding eigenvalues as

$$\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_{q_1} \geq \hat{\lambda}_{q_1+1} \geq \hat{\lambda}_{q_1+2} \geq \cdots \geq \hat{\lambda}_q \geq 0.$$

In the settings with fixed p and q, all estimated eigenvalues $\hat{\lambda}_i$ consistently converge to the corresponding ones λ_i. Thus, any criteria that are based on eigenvalues could consistently determine the number of canonical correlation pairs. However, in the regime in which the dimensions p and q are proportional to the sample size, the convergence of the estimated eigenvalues $\hat{\lambda}_i$ becomes very different and thus in general, the number q_1 is not necessarily identifiable. Bao et al. (2019) showed that in theory, only the number of some spiked eigenvalues can be determined. We give the related assumptions below.

Assumption 1 For the dimensionality of vectors, we need following assumptions

$$\begin{aligned} &p/n = a_1(n) \to c_1 \in (0,1),\\ &q/n = a_2(n) \to c_2 \in (0,1),\\ &\text{s.t.} \quad c_1 + c_2 \in (0,1). \end{aligned}$$

When $p \geq q$ and thus $c_1 \geq c_2$.

Note that, under the constraint on c_1 and c_2, we have $0 < c_1 < 1 - c_2$ and $0 < c_2 < 1 - c_1$ and thus, $(c_1c_2)/((1-c_1)(1-c_2)) < 1$. We then consider the following assumption.

Assumption 2 Write $q_1 := \operatorname{rank}(\Sigma_{12})$ for fixed integer q_1. Also write

$$r_c := \sqrt{\frac{c_1c_2}{(1-c_1)(1-c_2)}}. \tag{5}$$

There is a nonnegative integer k such that

$$1 \geq \lambda_1 \geq \cdots \geq \lambda_k > r_c \geq \lambda_{k+1} \geq \cdots \lambda_{q_1} > \lambda_{q_1+1} = \cdots = \lambda_q = 0, \tag{6}$$

and when $\lambda_1 < r_c$, define $k = 0$.

We now present the asymptotic properties of $\hat{\lambda}_i$ for $1 \leq i \leq q$. Their empirical spectral distribution (ESD) is denoted as

$$F_n(x) := \frac{1}{q}\sum_{i=1}^{q} \mathbf{1}_{(-\infty,x]}(\hat{\lambda}_i),$$

where $\mathbf{1}_A(\cdot)$ is the indicator function of A.

Lemma 1 (Wachter 1980) *For independent Gaussian vector* $\mathbf{x} \in \mathbb{R}^p$ *and* $\mathbf{y} \in \mathbb{R}^q$, *when Assumptions* 1 *and* 2 *hold, we have* $F_n(x) \xrightarrow{w} F(x)$, *where* $F(x)$ *gives the density*

$$F'(x) = f(x) = \begin{cases} \frac{1}{2\pi c_2} \frac{\sqrt{(d_+ - x)(x - d_-)}}{x(1-x)}, & \text{if } x \in [d_-, d_+], \\ 0, & \text{otherwise,} \end{cases}$$

where

$$d_\pm = \left(\sqrt{c_1(1 - c_2)} \pm \sqrt{c_2(1 - c_1)} \right)^2. \tag{7}$$

Remark 1 As all estimated eigenvalues follow a continuous distribution with the lower and upper bounds d_- and d_+, we can very easily see that almost all estimated eigenvalues do not converge to the corresponding eigenvalues at the population level. The following lemma from Bao et al. (2019) gives some more detailed results.

Lemma 2 (*Bao et al.* 2019) *Under the conditions in Lemma* 1,

(i): *for* $1 \leq i \leq k$,

$$\hat{\lambda}_i - \gamma_i = O_p(n^{-1/2})$$

where $\gamma_i = \lambda_i(1 - c_1 + c_1\lambda_i^{-1})(1 - c_2 + c_2\lambda_i^{-1})$ *is a function of* λ_i.

(ii): *for i with* $k + 1 \leq i \leq L < q$ *where L is any large, but fixed integer,*

$$\hat{\lambda}_i - d_+ = O_p(n^{-2/3}),$$

where d_+ *is the upper bound mentioned in Lemma* 1.

(iii): $\gamma_i > d_+$ *if* $\lambda_i > r_c$.

These results present the important feature that shows the significant difference of the estimated eigenvalues than that when p and q are fixed in the classical settings. The result (ii) also shows that the weak signals $\lambda_{i+1} \geq \ldots \geq \lambda_{q_1}$ may not be detectable as their estimators converge to the same value as all the estimated eigenvalues $\hat{\lambda}_j$ for $k + 1 \leq j \leq L$ for any large L whose values at the population level are zero at a fast rate $n^{-2/3}$. Clearly, when there are too many weak signals, the estimable order k should be much smaller than q_1, otherwise, k can be close to q_1. These results will be the base for constructing our criteria below.

3 Ridge ratio criteria and properties

We give the two criteria in separate subsections below.

3.1 The eigenvalues difference-based ratios

Define

$$\delta_i = \lambda_i - \lambda_{i+1}, \quad 1 \le i \le q-1. \tag{8}$$

From the description in Sect. 2, we have that $\delta_i \ge 0$ for all i and $\delta_{q_1} > 0$. Further, define the ratios as, if 0/0 is temporarily defined as 1,

$$r_{1,i} = \frac{\delta_{i+1}}{\delta_i} = \frac{\lambda_{i+1} - \lambda_{i+2}}{\lambda_i - \lambda_{i+1}} = \begin{cases} C_i \ge 0, & i < q_1, \\ \dfrac{0}{\delta_{q_1}} = 0, & i = q_1, \\ \dfrac{0}{0} := 1, & q_1 < i \le q-2. \end{cases} \tag{9}$$

It is clear that such a function has a local minimum at $i = q_1$ and all consecutive ratios take the constant value of 1. In other words, the maximum local minimizer is definitely the true value q_1 although there would exist other local minima or not. (This is the case when there are equal positive eigenvalues.) Note that $\frac{0}{0}$ is indeterminant form in general that could cause unstable values when the eigenvalues are estimated. We then modify this criterion by adding a positive ridge value $c_n \to 0$ as $n \to \infty$:

$$r^R_{1,i} = \frac{\delta_{i+1} + c_n}{\delta_i + c_n} = \frac{\lambda_{i+1} - \lambda_{i+2} + c_n}{\lambda_i - \lambda_{i+1} + c_n} = \begin{cases} \ge 0, & i < q_1, \\ \frac{c_n}{\lambda_{q_1} + c_n} \to 0, & i = q_1, \\ \frac{c_n}{c_n} = 1, & q_1 < i \le q-2. \end{cases} \tag{10}$$

The ridge value, as a tuning parameter, has two functions to make the criterion better performed: avoiding the instability of ratios and keeping the property of the ratios at the population level. This is different from the tuning parameters used in the other criteria in the literature. It makes this criterion function has a valley-cliff pattern at the true value of q_1: taking value 0 of the ratio at q_1 and value 1 for all successive ratios. Clearly a criterion can be based on the above to search for q_1 when $\hat{\lambda}_i$ are used instead. However, from Lemma 2 in Sect. 2, we know that q_1 is not identifiable, while k is possible. Recall

$$k := \#\{i : 1 \le i \le q_1, \lambda_i > r_c\}. \tag{11}$$

The sample ratios are as, when c_n is selected properly,

$$\lim_{n\to+\infty} \hat{r}^R_{1,i} = \lim_{n\to+\infty} \frac{\hat{\delta}_{i+1}}{\hat{\delta}_i} = \lim_{n\to+\infty} \frac{\hat{\lambda}_{i+1} - \hat{\lambda}_{i+2} + c_n}{\hat{\lambda}_i - \hat{\lambda}_{i+1} + c_n} = \begin{cases} \ge 0, & i < k, \\ 0, & i = k, \\ 1, & k < i \le L-2. \end{cases} \tag{12}$$

We then construct an estimator as the maximum index of the ratio that is bounded by a constant τ_1 with $0 < \tau_1 < 1$,

$$\hat{k}_1 = \max_{1 \le i \le L-2} \{i : \frac{\hat{\lambda}_{i+1} - \hat{\lambda}_{i+2} + c_n}{\hat{\lambda}_i - \hat{\lambda}_{i+1} + c_n} \le \tau_1\}. \tag{13}$$

The consistency is stated below.

Theorem 1 *Suppose that Assumptions 1 and 2 and the conditions in Lemmas 1 and 2 hold. When* $c_n = \log(\log n)/n^{2/3}$ *and* $n \to \infty$,

$$\mathbb{P}(\hat{k}_1 = k) \to 1. \tag{14}$$

Remark 2 In theory, as long as τ_1 is between 0 and 1, the consistency holds true. Therefore, unlike the thresholding value used in scree-plot methods, the performance of the method is relatively insensitive to its selection. But even though, in practice, when it is too close to 0, underestimation could happen whereas it is too close to 1, the estimator would take large value. Based on our experience in limited numerical studies, we recommend a value of $\tau_1 = 0.5$ as a compromise.

3.2 Centered eigenvalue-based ratios

In the difference-based ratios, δ_i are no longer monotonic with respect to the index i, we now propose another sequence of ratios that are based on the estimated eigenvalues themselves.

At the population level, define the eigenvalues-based ratios as

$$r_{2,i} = \frac{\lambda_{i+1}}{\lambda_i}, \qquad 1 \le i \le q-1. \tag{15}$$

Clearly,

$$r_{2,i} = \frac{\lambda_{i+1}}{\lambda_i} = \begin{cases} \le 1, & \text{for } i < q_1, \\ 0, & \text{for } i = q_1, \\ \frac{0}{0} := 1, & \text{for } q_1 < i \le q-1. \end{cases} \tag{16}$$

Again, to avoid the instability of 0/0, we also add a ridge value $c_n \to 0$ to define ridge ratios:

$$r_{2,i}^R = \frac{\lambda_{i+1} + c_n}{\lambda_i + c_n}, \qquad 1 \le i \le q-1. \tag{17}$$

Then, the redefined ratios show the following property:

$$r_{2,i}^R = \frac{\lambda_{i+1} + c_n}{\lambda_i + c_n} = \begin{cases} \le 1, & \text{for } i < q_1, \\ \frac{c_n}{\lambda_i + c_n} \to 0, & \text{for } i = q_1, \\ \frac{c_n}{c_n} = 1, & q_1 < i \le q - 1. \end{cases} \tag{18}$$

However, when the estimated eigenvalues $\hat{\lambda}_i$ are used, Lemma 2 tells that the above properties cannot continue to hold. To be precise, $\hat{\lambda}_i$ for all i with $1 \le i \le L$ where $L > q_1$ do not converge to zero and then the ratios

$$\lim_{n\to+\infty} \tilde{r}_i^R = \lim_{n\to+\infty} \frac{\hat{\lambda}_{i+1} + c_n}{\hat{\lambda}_i + c_n} = \begin{cases} \frac{\gamma_{i+1}}{\gamma_i} \le 1, & \text{for } i < k, \\ \frac{d_+}{\gamma_k} < 1, & \text{for } i = k, \\ \frac{d_+}{d_+} = 1, & \text{for } k < i \le q_1, \\ \frac{d_+}{d_+} = 1, & \text{for } q_1 < i \le L - 1. \end{cases} \tag{19}$$

In other words, this cannot be a criterion to determine the rank as there is no a clear separation between the ratio at k and all successive ratios. This confirms the comments at the end of Sect. 2. Thus, we consider centered eigenvalues $\hat{\lambda}_i - d_+$ and define modified ridge ratios with the properties:

$$\lim_{n\to+\infty} \hat{r}_i^{MR} = \lim_{n\to+\infty} \left| \frac{\hat{\lambda}_{i+1} - d_+ + c_n}{\hat{\lambda}_i - d_+ + c_n} \right| = \begin{cases} \le 1, & \text{for } i < k, \\ 0, & \text{for } i = k, \\ 1, & \text{for } k < i \le L - 1. \end{cases} \tag{20}$$

This function appears again a “valley-cliff” pattern at the index k. We can then construct an estimator, for a constant τ_2 with $0 < \tau_2 < 1$,

$$\hat{k}_2 = \max_{1\le i\le L-1} \{ i : \hat{r}_i^{MR} \le \tau_2 \}. \tag{21}$$

Theorem 2 *Under the same conditions in Theorem 1, when* $c_n = \log(\log n)/n^{2/3}$ *and* $n \to \infty$,

$$\mathbb{P}(\hat{k}_2 = k) \to 1. \tag{22}$$

Remark 3 Similarly as that for Theorem 1, we also need to select a value of τ_2. We tried several values of τ_2 around 0.8 and found that the simulation results were similar. Thus, $\tau_2 = 0.8$ is recommended.

To better understand these methods, we give the curves of six criteria to visualize their mechanisms. The formal definitions of the criteria will be given in the next section. The data are generated from the model described in the caption of Fig. 1. We see that AIC, BIC and C_p very easily under- or over-determine the rank. The traditional scree plot method, which is not a visualization method, completely relies on the separation between the outliers and others by delicately selecting the thresholding value. However, our criteria could intensify the separation due to the use of ridge ratios. More importantly, the plots in Fig. 1b, c show that, unlike the other criteria, the eigenvalue-based ratio criteria are not continuous functions, while show an useful "valley-cliff" pattern. Particularly for the centered eigenvalue-based ratio criterion, when $i = k + 1$, the ratio takes a very large value. This might be because the kth estimated eigenvalue is closer to the upper bound d_+ than all successive eigenvalues and thus $\hat{r}_{k+1}^{MR}$ could be very large. This "valley-cliff" pattern can make the ratio at the true rank stick out and thus helps a better separation of $\hat{r}_{k}^{MR}$ from all $\hat{r}_{i}^{MR}$ for $i = k + 1, \ldots, L$. Therefore, in practice, this visualization tool can easily determine the number of canonical correlation pairs.

4 Simulation studies

In this section, we conduct some simulations to illustrate the finite sample behaviors of the proposed criteria and to compare with the four competing methods studied in Bao et al. (2019) and Fujikoshi (2017b).

1. The AIC criterion. The formula is:

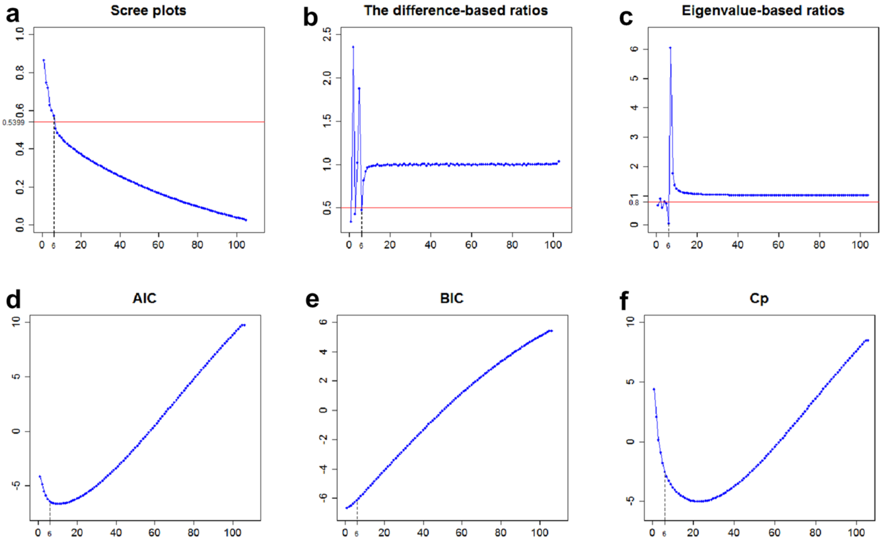

Fig. 1 Model: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7) = (0.8, 0.6, 0.6, 0.4, 0.4, 0.4, 0.2)$, and $\lambda_8 = \cdots = \lambda_q = 0$. Sample size $n = 1000$, and the dimensions $p = 210$, and $q = 105$. The results are based on 1000 replications. The horizontal axis is for the number of eigenvalues.

$$\begin{aligned}\mathrm{AIC}_j =& - n\log\left[\prod_{i=j+1}^{q}(1-\hat{\lambda}_i)\right] + n(p+q) + (p+q+1)\log|S_{xy}| \\ &+ K + 2\{j(p+q-j) + \frac{1}{2}p(p+1) + \frac{1}{2}q(q+1)\},\end{aligned} \tag{23}$$

where $K = 2\log\{\Gamma_{p+q}(\frac{1}{2}n)/(\frac{1}{2}n)^{\frac{1}{2}n(p+q)}\}$, $\Gamma(\cdot)$ is Gamma function. S_{xy} is the sample cross-covariance with respect to Σ_{xy} in (1) and $\prod_{i=q+1}^{q}(\cdot) = 0$.

To bypass the nuisance parameter K and brevity, then use next formula instead of (23) in practice,

$$\begin{aligned}A_j &= \mathrm{AIC}_j - \mathrm{AIC}_q \\ &= -n\log\left[\prod_{i=j+1}^{q}(1-\hat{\lambda}_i)\right] - 2(p-j)(q-j),\end{aligned} \tag{24}$$

Here for $j \in \{0,\dots,q\}$. The estimator is the minimizer of

$$\hat{k}_A = \underset{j\in\{0,1,\dots,q\}}{\arg\min}\ A_j.$$

2. The BIC criterion. Analogous to A_j, the BIC criterion is given as

$$B_j = -n\log\left[\prod_{i=j+1}^{q}(1-\hat{\lambda}_i)\right] - \log(n)(p-j)(q-j).$$

The minimizer of the following is defined as the estimator:

$$\hat{k}_B = \underset{j\in\{0,1,\dots,q\}}{\arg\min}\ B_j.$$

3). The C_p criterion (Fujikoshi and Veitch 1979, (3.11)). The criterion is:

$$C_j = n\sum_{i=j+1}^{q}\frac{\hat{\lambda}_i}{1-\hat{\lambda}_i} - 2(p-j)(q-j).$$

Also the minimizer of the following is defined as the estimator:

$$\hat{k}_C = \underset{j\in\{0,1,\dots,q\}}{\arg\min}\ C_j.$$

4). The scree plot-based criterion (Bao et al. 2019). Denote $\hat{k}_{BM}$ as the maximizer such that

$$\hat{k}_{BM} := \max\{i : \hat{\lambda}_i \geq d_+ + \epsilon_n\}, \tag{25}$$

where ϵ_n is a positive number only depending on n. Bao et al. (2019) selected $\epsilon_n = \log(\log n)/n^{2/3}$.

We now present the simulation results with four different models in Tables 1, 2, 3 and 4. The model settings are described in the captions of the tables.

In the simulations, let the eigenvalues at the population level be $\lambda_1 \geq \lambda_2, \dots, \geq \lambda_{q_1} > \lambda_{q_1+1} = \cdots = \lambda_q = 0$ and $p/q = 2$. The results with the sample of size $n = 1000$ and the data (X_i, Y_i) being generated from the standard normal distribution are reported in Tables 1, 2and 3, but the results in Table 4 are with Student' t distribution with ten degrees of freedom. $\hat{k}_1$ and $\hat{k}_2$ are our estimations. The ridge value in our approach and the tuning parameter in $\hat{k}_{BM}$ are both $\epsilon_n = \log(\log n)/n^{2/3}$ and r_c is a constant in Assumption 1. The thresholding values are $\tau_1 = 0.5$ and $\tau_2 = 0.8$ in the criteria (13) and (21), respectively. The bold lines in all tables are corresponded to the number k, which defined in the criterion (11) in Sect. 3.

Model 1 This model would favor all methods: $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.8, 0.6, 0.4, 0.2)$, $\lambda_5 = \cdots = \lambda_q = 0$ and $p/q = 2$. The results are tubulated in Table 1.

According to Table 1, we can find all the six estimators are performed well in low dimension cases. But when the dimension gets higher, AIC and C_p very seriously overestimate the rank and BIC causes a serious underestimation. Compared

Table 1 The proportions of estimated rank in 1000 replications for Model 1

	$p = 60, r_c = 0.0445, \gamma_4 = 0.2778, d_+ = 0.1674$						$p = 110,$ $r_c = 0.0854, \gamma_4 = 0.3514, d_+ = 0.2956$					
	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$
≤ 2	0	0	0	0	356	0	28	0	0	0	1000	0
3	5	0	0	0	644	0	184	9	99	0	0	0
4	**934**	**992**	**1000**	**816**	**0**	**447**	**684**	**969**	**901**	**487**	**0**	**0**
5	43	8	0	176	0	469	73	22	0	459	0	20
≥ 6	18	0	0	8	0	84	31	0	0	54	0	980
	$p = 160, r_c = 0.1305, \gamma_4 = 0.4330, d_+ = 0.4133$						$p = 210,$ $r_c = 0.1809, \gamma_4 = 0.5226, d_+ = 0.5206$					
≤ 2	70	0	0	0	1000	0	131	0	1	0	1000	0
3	525	203	698	0	0	0	709	494	972	0	0	0
4	**306**	**783**	**302**	**56**	**0**	**0**	**115**	**493**	**27**	**0**	**0**	**0**
5	63	14	0	351	0	0	26	13	0	2	0	0
≥ 6	36	0	0	593	0	1000	19	0	0	998	0	1000
	$p = 260, r_c = 0.2379, \gamma_3 = 0.6644, d_+ = 0.6174$						$p = 310,$ $r_c = 0.2871, \gamma_3 = 0.7222, d_+ = 0.7037$					
≤ 2	347	0	61	0	1000	0	697	7	646	0	1000	0
3	**573**	**654**	**937**	**0**	**0**	**0**	**257**	**670**	**353**	**0**	**0**	**0**
4	46	341	2	0	0	0	29	321	1	0	0	0
5	11	5	0	0	0	0	4	2	0	0	0	0
≥ 6	23	0	0	1000	0	1000	13	0	0	1000	0	1000

Table 2 The proportions of estimated rank in 1000 replications for Model 2

	$p = 60, r_c = 0.0445, \gamma_7 = 0.2778, d_+ = 0.1674$						$p = 110,$ $r_c = 0.0854, \gamma_7 = 0.3514, d_+ = 0.2956$					
	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$
≤ 4	1	0	0	0	117	0	51	0	0	0	1000	0
5	0	0	0	0	510	0	0	0	0	0	0	0
6	7	0	1	0	373	0	122	84	181	1	0	0
7	**926**	**1000**	**999**	**854**	**0**	**584**	**736**	**915**	**819**	**581**	**0**	**33**
≥ 8	66	0	0	146	0	416	91	1	0	418	0	967
	$p = 160, r_c = 0.1305, \gamma_7 = 0.4330, d_+ = 0.4133$						$p = 210,$ $r_c = 0.1809, \gamma_7 = 0.5226, d_+ = 0.5206$					
≤ 4	159	0	0	0	1000	0	270	0	0	0	1000	0
5	0	0	0	0	0	0	1	1	13	0	0	0
6	441	550	855	0	0	0	553	866	982	0	0	0
7	**310**	**449**	**145**	**72**	**0**	**0**	**114**	**133**	**5**	**0**	**0**	**0**
≥ 8	62	1	0	828	0	1000	91	0	0	1000	0	1000
	$p = 260, r_c = 0.2379, \gamma_6 = 0.6644, d_+ = 0.6174$						p=310, $r_c = 0.2871, \gamma_6 = 0.7222, d_+ = 0.7037$					
≤ 4	566	0	50	0	1000	0	854	8	829	0	1000	0
5	17	38	561	0	0	0	43	406	169	0	0	0
6	**351**	**920**	**389**	**0**	**0**	**0**	**70**	**575**	**2**	**0**	**0**	**0**
7	39	42	0	0	0	0	20	11	0	0	0	0
≥ 8	27	0	0	1000	0	1000	13	0	0	1000	0	1000

with AIC, BIC and C_p, our estimators obviously work much better, and $\hat{k}_2$ is uniformly better than $\hat{k}_1$. The classical scree plot-based method $\hat{k}_{BM}$ also works well. Note that when the dimension gets higher, $\hat{k}_1$, $\hat{k}_2$ and $\hat{k}_{BM}$ cannot well estimate the true rank, typically, all these methods gradually have the underestimation issue. When $p = 260$ and 310, γ_3 is just slightly larger than the upper bound d_+ and thus, $k = 3$ although $q_1 = 4$. But overall with high dimension, $\hat{k}_2$ uniformly outperforms the other competitors.

Model 2 Consider the eigenvalue multiplicity on the middle part as $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7) = (0.8, 0.6, 0.6, 0.4, 0.4, 0.4, 0.2)$, $\lambda_8 = \cdots = \lambda_q = 0$. The results are tubulated in Table 2. The results tell us that $\hat{k}_A, \hat{k}_B$ and $\hat{k}_C$ very much either under-estimate or over-estimate the true rank particularly when the dimensions are even moderate. The scree plot method tends to underestimate the rank although it is much better than $\hat{k}_A, \hat{k}_B$ and $\hat{k}_C$. Overall, the original eigenvalue-based ratio criterion $\hat{k}_2$ overwhelms the other competitors. When $p \geq 260$, $k = 6 < q_1 = 7$, the results also show that $q_1 = 7$ is difficult to be determined.

Table 3 The proportions of estimated rank in 1000 replications for Model 3

	$p = 60, Card(q_1) = 5, M_{q_1} = (2,1,1,1)$						$p = 110, Card(q_1) = 7, M_{q_1} = (2,2,2,1)$					
	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$
≤ 4	4	0	0	0	1000	0	15	0	0	0	1000	0
5	**871**	**996**	**1000**	**820**	**0**	**488**	0	0	0	0	0	0
6	74	4	0	173	0	430	98	78	178	0	0	0
7	27	0	0	7	0	80	**684**	**922**	**822**	**581**	**0**	**1**
≥ 8	24	0	0	0	0	2	203	0	0	419	0	999
	$p = 160, Card(q_1) = 8, M_{q_1} = (3,2,2,1)$						$p = 210, Card(q_1) = 10, M_{q_1} = (3,3,2,2)$					
≤ 6	60	0	0	0	1000	0	166	0	0	0	1000	0
7	363	607	845	0	0	0	3	0	3	0	0	0
8	**400**	**393**	**155**	**87**	**0**	**0**	499	691	984	0	0	0
9	68	0	0	458	0	0	109	309	13	0	0	0
10	31	0	0	385	0	0	**85**	**0**	**0**	**1**	**0**	**0**
≥ 11	78	0	0	70	0	1000	128	0	0	999	0	1000
	$p = 260, Card(q_1) = 11, M_{q_1} = (3,3,3,2)$						$p = 310, Card(q_1) = 12, M_{q_1} = (3,4,3,2)$					
≤ 8	462	102	753	0	1000	0	718	47	948	0	1000	0
9	**337**	**873**	**247**	**0**	**0**	**0**	77	739	52	0	0	0
10	84	25	0	0	0	0	**123**	**213**	**0**	**0**	**0**	**0**
11	43	0	0	0	0	0	35	1	0	0	0	0
12	24	0	0	0	0	0	14	0	0	0	0	0
≥ 13	50	0	0	1000	0	1000	33	0	0	1000	0	1000

Model 3 To see whether the method is workable in larger q_1 cases, we consider in this model $Card(q_1) = \lfloor q^{1/2} \rfloor$ that is the largest integer less than or equal to $q^{1/2}$. Recall q is the dimension of vector Y. Use the vector $M_{q_1} = (n_1, n_2, n_3, n_4)$ to present the multiplicity of nonzero eigenvalues of 0.8, 0.6, 0.4, 0.2, respectively. The numbers $n_i, i = 1, 2, 3, 4$ are shown in Table 3. The results indicate that when q_1 is relatively small, all competitors except $\hat{k}_B$ and $\hat{k}_C$ work well. In general, $\hat{k}_B$ does often underestimate while $\hat{k}_C$ overestimates. The methods are gradually and reasonably losing efficiency with increasing q_1. Our criterion $\hat{k}_2$ works the best among all competitors although it also tends to underestimate the true number. We also conduct the simulation with $Card(q_1) = \lfloor q^{2/3} \rfloor$. To save the space, all results are postponed to the supplementary material. The message is similar to that here, showing that when q_1 is large, the estimation works worse. Thus in large q_1 cases, the asymptotic properties would need a further study.

Further, although our criteria are theoretically rooted in Gaussian distribution, it is natural to wonder whether they are in practice feasible to non-Gaussian cases as there are no asymptotic results with non-Gaussian distributions. Consider

Table 4 The proportions of estimated rank in 1000 replications for Model 4 under $t_{df=10}$

	$p = 60, r_c = 0.0445, \gamma_7 = 0.2778, d_+ = 0.1674$						$p = 110,$ $r_c = 0.0854, \gamma_7 = 0.3514, d_+ = 0.2956$					
	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$	$\hat{k}_1$	$\hat{k}_2$	$\hat{k}_{BM}$	$\hat{k}_A$	$\hat{k}_B$	$\hat{k}_C$
≤ 4	0	0	0	0	0	0	0	0	0	0	1000	0
5	0	0	0	0	1000	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	**1000**	**1000**	**1000**	**1000**	**0**	**1000**	**1000**	**1000**	**1000**	**1000**	**0**	**0**
≥ 8	0	0	0	0	0	0	0	0	0	0	0	1000
	$p = 160, r_c = 0.1305, \gamma_7 = 0.4330, d_+ = 0.4133$						$p = 210,$ $r_c = 0.1809, \gamma_7 = 0.5226, d_+ = 0.5206$					
≤ 4	0	0	0	0	1000	0	0	0	0	0	1000	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	**1000**	**1000**	**1000**	**0**	**0**	**0**	**1000**	**1000**	**1000**	**0**	**0**	**0**
7	0	0	0	0	0	0	0	0	0	0	0	0
≥ 8	0	0	0	1000	0	1000	0	0	0	1000	0	1000
	$p = 260, r_c = 0.2379, \gamma_6 = 0.6644, d_+ = 0.6174$						$p = 310,$ $r_c = 0.2871, \gamma_6 = 0.7222, d_+ = 0.7037$					
≤ 4	0	0	0	0	1000	0	1000	0	1000	0	1000	0
5	0	0	1000	0	0	0	0	0	0	0	0	0
6	**1000**	**1000**	**0**	**0**	**0**	**0**	**0**	**1000**	**0**	**0**	**0**	**0**
7	0	0	0	0	0	0	0	0	0	0	0	0
≥ 8	0	0	0	1000	0	1000	0	0	0	1000	0	1000

some examples in simulations. To generate data $(\mathbf{x}, \mathbf{y})$ with non-Gaussian distribution, we first generate W with mean $\mathbb{E}W = 0$ and variance $Var(W) = 1$. Consider the following with the fifth-order polynomial transformation method proposed by Headrick (2002). Specifically, the target variable can be expressed as

$$W = c_0 + c_1 Z + c_2 Z^2 + c_3 Z^3 + c_4 Z^4 + c_5 Z^5, \quad \text{where} \quad Z \sim i.i.d.\ N(0, 1). \tag{26}$$

Recall the central moments of W:

$$\mu_j = \int (w - c)^j dF(w), \quad \text{where} \quad c = \mathbb{E}W,$$

and the definition of cumulants given by Kendall and Stuart (1977)

$$\begin{aligned} &k_1 = \mu_1 = 0, \quad k_2 = \mu_2, \\ &k_3 = \mu_3, \quad k_4 = \mu_4 - 3\mu_2^2, \\ &k_5 = \mu_5 - 10\mu_3\mu_2, \\ &\quad k_6 = \mu_6 - 15\mu_4\mu_2 - 10\mu_3^2 + 30\mu_2^3. \end{aligned}$$

To circumvent the magnitude of k_j toward to arbitrarily large, consider the standardized cumulants

$$\begin{aligned} 0 &= \frac{k_1}{k_2^{1/2}}, & 1 &= \frac{k_2}{k_2}, \\ \gamma_1 &= \frac{k_3}{k_2^{3/2}}, & \gamma_2 &= \frac{k_4}{k_2^2}, \\ \gamma_3 &= \frac{k_5}{k_2^{5/2}}, & \gamma_4 &= \frac{k_6}{k_2^3}. \end{aligned}$$

Note that γ_1 and γ_2 are skewness and kurtosis, respectively. Inasmuch as the odd moments of standard normal distribution are zero, the expectation of W is

$$\mathbb{E}W = c_0 + c_2 + 3c_4 = 0, \tag{27}$$

and the variance of W is derived as $Var(W) = \mathbb{E}W^2 - (\mathbb{E}W)^2$ and can be parameterized as follows

$$Var(W) = c_1^2 + 2c_2^2 + 24c_2c_4 + 6c_1(c_3 + 5c_5) + 3(5c_3^2 + 32c_4^2 + 70c_3c_5 + 315c_5^2) = 1. \tag{28}$$

By the definition above, for any given density function one can calculate the associated $\gamma_i, i = 1, 2, 3, 4$ directly. Therefore, substituting these calculated γ_i into (26) and integrating $\mathbb{E}W^3, \mathbb{E}W^4, \mathbb{E}W^5, \mathbb{E}W^6$ as well as (27) and (28) yield the solutions $c_i, i = 0, 1, 2, 3, 4, 5$.

Further, it is worth noting that taking the block diagonal transformation on $(\mathbf{x}, \mathbf{y})$ to $(\mathbf{Ax}, \mathbf{By})$, the canonical correlation coefficients are invariant as long as matrices $\mathbf{A}_{p\times p}$ and $\mathbf{B}_{q\times q}$ are nonsingular. Hence, to approximate the eigenvalues of matrix defined in (4) we tentatively assume that $\Sigma_{11} = I_p, \Sigma_{22} = I_q$. In other words, we can start with

$$\Sigma_{xy} = \begin{pmatrix} I_p & T \\ T' & I_q \end{pmatrix}, \tag{29}$$

where $T = diag(\sqrt{\lambda_1}, ..., \sqrt{\lambda_{q_1}}) \oplus \mathbf{0}_{(p-q_1)\times(q-q_1)}$. Then, we gather the $(p+q)\times n$ sample matrix of $(\mathbf{x}, \mathbf{y})$ as

$$\begin{pmatrix} \mathscr{X} \\ \mathscr{Y} \end{pmatrix} = \Sigma_{xy}^{1/2} W_1,$$

where the $(p+q)\times n$ matrix W_1 has i.i.d. entries generated by (26), $\mathscr{X}$ and $\mathscr{Y}$ are data matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively.

Model 4 The parameter settings of this model are the same with Model 2 except for the distribution change from Gaussian to Student's t with ten degrees of freedom. The comparison between Tables 2 and 4 provide a very interesting phenomena. That is, the criterion with Student's t-distribution works even better than it with Gaussian distribution. Also, it is somewhat of interest to see that the decision is always 100% in all cases. This suggest that the current theory of CCA would be feasible in some non-Gaussian cases. As this is beyond the score of this paper, we will not give any investigation on relevant theoretical exploration.

To display the performance of the methodology in non-Gaussian distribution cases, we have also taken an asymmetric distribution, chi-square distribution with two degrees of freedom, into consideration. The simulation results resemble to the Gaussian case and are included in the supplementary material. The corresponding constants $c_i, i = 0, 1, 2, 3, 4, 5$ of chi-square distribution can be found in Table 1 of Headrick (2002).

5 A real data example

In this section, we take the breast cancer datasets from The Cancer Genome Atlas (TCGA) project into consideration. This project collected human tumor specimens and conducted molecular studies to reveal higher-order structure of cancer by large-scale genomic data such as mRNA expression arrays and DNA methylation arrays. TCGA contains the data from several diverse genomic platforms on the same cancerous tumor samples. The two datasets used in this paper are available at https://gdc.cancer.gov/about-data/publications, the web portal of The Genomic Data Commons (GDC). The DNA methylation data quantify the methyaltion level using the ratio of intensities methylated and unmethylated alleles, from the first TCGA breast cancer study (Cancer Genome Atlas Network 2012). TCGA only analyzed on 466 breast tumors of the total 940 samples. The mRNA expression data set was collected from Ciriello et al. (2015) which reported the results on nearly twice as many breast tumors from TCGA ($n = 817$) than (Cancer Genome Atlas Network 2012). As Shu et al. (2019), we in this paper focus on 660 samples that are contained in each of the larger group mentioned above. That is, for each selected tumor sample, it records both mRNA expression data and DNA methylation data. Specifically, the assayed 660 clinical samples consist of 112 basal-like, 55 HER2-enriched, 331 luminal A and 162 luminal B tumors. To numerically gauge the extent of suptype distinction, we adopt, as Cabanski et al. (2010) did, the standardization within class sum of squares (SWISS) in practice. That is, noted the matrix $Y = (Y_{ij})_{p\times n}$ as the sample collection of all observations of p-dimensional vector, then the SWISS score that identifies subtype distinctions can be described as follows

$$SWISS(Y) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i,s(j)})}{\sum_{i=1}^{p}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})},$$

where $s(j) = \{k$: samples k and j are signed as the same subtype$\}$, $\bar{Y}_{i,s(j)}$ is the average within samples of the jth subtype on the ith row and $\bar{Y}_{i\cdot}$ is the overall mean of the ith row of the matrix. The SWISS amounts to a proportion of the total variation decided by the variability within the subtypes, hence a lower score does a better subtype distinction. Preprocessing the mRNA expression data, we filter out the subset by appropriately removing the missing data from raw data firstly from the original 20,531 genes. The final subset consists of $p = 265$ variably expressed genes with marginal SWISS ≤ 0.7. In the same way, we select $q = 86$ methylated probes from the original 21,986 probes of DNA methylation data with marginal SWISS ≤ 0.7. The purpose of this part is to determine the number of canonical correlation pairs. We apply our proposed criteria to analyze this genomic datasets.

We summarize the analysis results for this dataset by the used methods in Fig. 2. Note that the tuning parameters c_n and ϵ_n used here are identical to those in the simulations. From Fig. 2, we can see the following. The scree plot in Fig. 2a shows that the canonical correlations are adjacent closely, except the largest one, there is no apparent gap between any two consecutive eigenvalues and then such an auxiliary visualization tool has no way to be useful for the determination in practice. Figure 2b depicts that the values of difference-based ratios are not stable. When $\tau_1 = 0.5$, the estimated rank seems underdeterminated. However, it still offers the information that the number 20 would be a potential value as the value is smaller than $\tau_1 = 0.6 < 1$. From Fig. 2c, we can see that the sequence can obviously have a local minimum at the 16th ratio followed by a very big value of the next ratio. In other words, the 16th eigenvalue is closer to the upper bound d_+ than any successive eigenvalue. Thus, we can consider it as the boundary of the outliers we can separate from others. The visualized curve is very informative. AIC and C_p criteria are clearly to overestimate the rank and BIC to underestimate it. Together with the simulation results showing that the scree plot method is also easy to underestimate

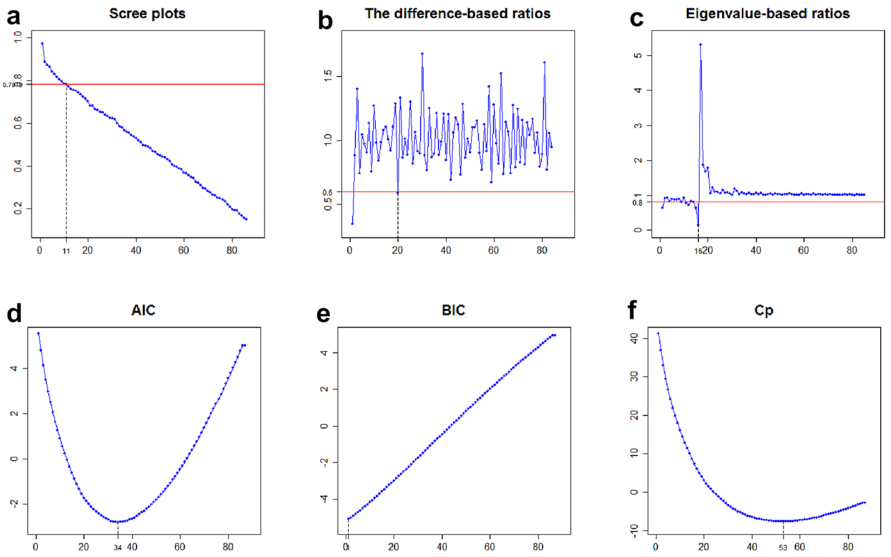

Fig. 2 The 660 tumor samples used in this figure were classified into 4 types by Ciriello et al. We filtered out the mRNA expression genes, $p = 265$, and DNA methylation probes, $q = 86$, from the original data. We figured out the results described by each methods and signed the estimated point by vertical dotted line in sub-figures separately

the rank, we may consider 16 as a reasonable estimation of the potential rank. We also notice that Shu et al. (2019) used a different criterion for the genes and probes, when the SWISS value was set to be 0.9, to determine a much smaller number $k = 2$ of the canonical correlation pairs. When we consider the data with the SWISS value of 0.7, the estimated number should be much larger. Even though, as the dimensionality has been greatly reduced, we can conveniently use these dimension-reduced variable pairs to do further analysis such that possible loss of information could be avoided when the variables pairs are too few.

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/s10463-020-00776-x).

Acknowledgements The authors would like to thank the editor, the associated editor and two anonymous referees for their constructive suggestions and comments that led to the improvement of an early manuscript.

Funding The research described herewith was supported by a Grant (HKBU12303419) from The University Grants Council of Hong Kong, and a grant from The National Natural Science Foundation of China (NSFC11671042).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csáki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281. Budapest: Akadémiai Kaido.

Bai, Z., Choi, K. P., Fujikoshi, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*, *46*(3), 1050–1076.

Bao, Z., Hu, J., Pan, G., Zhou, W. (2019). Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case. *The Annals of Statistics*, *47*(1), 612–640.

Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., Wilkerson, M. D., Marron, J. S., Perou, C. M., Hayes, D. N. (2010). SWISS MADE: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*, *5*(3), e9905.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M., Factor, R. E., Collins, L. C., Allison, K. H., Chen, Y., Jensen, K., Johnson, N. B., Oesterreich, S., Mills, G. B., Cherniack, A. D., Robertson, G., Benz, C., Sander, C., Laird, P. W., Hoadley, K. A., King, T. A., TCGA Research Network, Perou, C. M. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, *163*(2), 506–519.

Fujikoshi, Y. (1985). Two methods for estimation of dimensionality in canonical correlation analysis and the multivariate linear model. In K. Matsushita (Ed.), *Statistical theory and data analysis*, pp. 233–240. Amsterdam: Elsevier Science.

Fujikoshi, Y. (2017a). High-dimensional asymptotic distributions of characteristic roots in multivariate linear models and canonical correlation analysis. *Hiroshima Mathematical Journal*, *47*(3), 249–271.

Fujikoshi, Y. (2017b). High-dimensional properties of AIC, BIC and C_p for estimation of dimensionality in canonical correlation analysis. *SUT Journal of Mathematics*, *53*(1), 59–72.

Fujikoshi, Y., Sakurai, T. (2009). High-dimensional asymptotic expansions for the distributions of canonical correlations. *Journal of Multivariate Analysis*, *100*(1), 231–242.

Fujikoshi, Y., Veitch, L. (1979). Estimation of dimensionality in canonical correlation analysis. *Biometrika*, *66*(2), 345–351.

Gunderson, B., Muirhead, R. (1997). On estimating the dimensionality in canonical correlation analysis. *Journal of Multivariate Analysis*, *62*(1), 121–136.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis*, *40*(4), 685–711.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3–4), 321–377.

Kendall, M., Stuart, A. (1977). *The advanced theory of statistics* 4th ed. New York: Macmillan.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*(4), 661–675.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shu, H., Wang, X., Zhu, H. (2019). D-CCA: A decomposition-based canonical correlation analysis for high-dimensional datasets. *Journal of the American Statistical Association*, *115*, 292–306. https://doi.org/10.1080/01621459.2018.1543599.

Song, Y., Schreier, P. J., Roseveare, N. J. (2015). Determining the number of correlated signals between two data sets using PCA-CCA when sample support is extremely small. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3452–3456. South Brisbane, QLD: IEEE.

Song, Y., Schreier, P. J., Ramírez, D., Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing*, *128*, 449–458.

Wachter, K. W. (1980). The limiting empirical measure of multiple discriminant ratios. *The Annals of Statistics*, *8*(5), 937–957.

Zhu, X., Guo, X., Wang, T., Zhu, L. (2020). Dimensionality determination: A thresholding double ridge ratio approach. *Computational Statistics and Data Analysis*, *146*, 106910.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.