Linear Algebra and its Applications 565 (2019) 208–224

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Regression-aware decompositions

Mark Tygert

ARTICLE INFO

Article history:
Received 3 March 2018
Accepted 11 December 2018
Available online 17 December 2018
Submitted by D. Needell

MSC:
65F30
62J05
62H25

Keywords:
Least squares
Regression
Canonical correlation analysis
Principal component analysis
Singular value decomposition
Interpolative decomposition

ABSTRACT

Linear least-squares regression with a "design" matrix A approximates a given matrix B via minimization of the spectral- or Frobenius-norm discrepancy $\|AX - B\|$ over every conformingly sized matrix X. Also popular is low-rank approximation to B through the "interpolative decomposition," which traditionally has no supervision from any auxiliary matrix A. The traditional interpolative decomposition selects certain columns of B and constructs numerically stable (multi)linear interpolation from those columns to all columns of B, thus approximating all of B via the chosen columns. Accounting for regression with an auxiliary matrix A leads to a "regression-aware interpolative decomposition," which selects certain columns of B and constructs numerically stable (multi)linear interpolation from the corresponding least-squares solutions to the least-squares solutions X minimizing $\|AX - B\|$ for all columns of B. The regression-aware decompositions reveal the structure inherent in B that is relevant to regression against A; they effectively enable supervision to inform classical dimensionality reduction, which classically has been restricted to strictly unsupervised learning.

E-mail address: tygert@aya.yale.edu.

https://doi.org/10.1016/j.laa.2018.12.015
0024-3795/

1. Introduction

A common theme in multivariate statistics and data analysis is detecting and exposing low-dimensional latent structure governing two sets of vectors (each set could consist of realizations of a vector-valued random variable, for example). Widely used methodologies for this include linear least-squares regression and the canonical correlation analysis (CCA) of [9] ([9] discusses all previously developed methods mentioned below). Subsection 3.4 below combines the advantages of both principal component analysis (PCA) and linear least-squares regression, leveraging a single data set's intrinsic low-rank structure as well as low-rank structure in the data's interaction with another data set; this combination, "regression-aware principal component analysis," amounts to a variant of CCA informed by the regression.

The key to this combination is the construction in Subsection 3.3 of an analogous regression-aware interpolative decomposition, which provides an efficient means of performing subset selection for general linear models, especially in the simplified (while less canonical) formulation of Subsection 3.5. Concretely, the regression-aware interpolative decomposition selects certain columns of a given matrix B, then constructs numerically stable (multi)linear interpolation from corresponding least-squares solutions to the least-squares solutions $X = A^{\dagger}B$ minimizing $\|AX - B\|$ for all columns of B (here, A is the design matrix in the regression, $A^{\dagger}$ is the pseudoinverse of A, and $\|AX - B\|$ is the spectral or Frobenius norm). The presentation below first focuses on approximating X directly via the interpolation, in Subsection 3.2, then focuses on approximating AX (as AX approximates B), in the rest of the present paper.

Section 4 illustrates these methods via several numerical experiments. The other sections set the stage: Section 2 reviews pertinent prior mathematics, and is essential in the sequel. Section 3 introduces the regression-aware decompositions. More specifically, Subsection 2.1 specifies notational conventions. Subsection 2.2 defines and summarizes facts about interpolative decompositions. Subsection 3.1 formulates a general construction. Subsection 3.2 specializes the general formulation of Subsection 3.1 to the case of linear least-squares regression, albeit simplistically. Subsection 3.3 then provides the most useful formulation. Subsection 3.4 leverages Subsection 3.3 to interpret a kind of CCA as a regression-aware decomposition. Subsections 3.5 and 3.6 provide computationally simpler alternatives to Subsections 3.3 and 3.4, respectively. Subsections 4.1–4.5 present five illustrative numerical examples.

2. Preliminaries

This section sets notation (in Subsection 2.1) and reviews the interpolative decomposition (in Subsection 2.2), both of which are used throughout the remainder of the paper.

2.1. Notation

This subsection sets notational conventions used throughout the present paper.

All discussion pertains to matrices whose entries are real- or complex-valued. For any matrix A, we denote the adjoint (conjugate transpose) by A^* and the Moore–Penrose pseudoinverse by $A^\dagger$; we use $\|A\|$ to denote the same norm throughout the paper, either the spectral norm or the Frobenius norm (unitary invariance of the norm will be important in Section 3), and we denote by $(A^*A)^{-1/2}$ the pseudoinverse of the self-adjoint square root of A^*A. Detailed definitions of all these are available in the exposition of [5]. A proof that $X = A^\dagger B$ minimizes $\|AX - B\|$ for any conformingly sized matrices A and B — for both the spectral norm and the Frobenius norm — is available, for example, in Appendix B of [12]; accordingly, we refer to $X = A^\dagger B$ as "the" minimizer of $\|AX - B\|$. All decompositions will be accurate to a user-specified precision $\epsilon > 0$.

2.2. Interpolative decomposition

This subsection reviews the interpolative decomposition (ID).

The ID dates at least to [3]; however, modern applications owe much to [13] and [6], among others (this is also related to the CX decomposition of [2] and others, though technically the CX decomposition omits the ID's requirement for numerical stability). The software and documentation of [11] describe some common algorithms for computing IDs, based on the contributions of [1] and [8], which prove the following.

Theorem 1. *Suppose that m and n are positive integers, and B is an $m \times n$ matrix.*

Then, for any real number $\alpha \geq 1$ and for any positive integer k with $k \leq m$ and $k \leq n$, there exist a $k \times n$ matrix P and an $m \times k$ matrix C whose columns constitute a subset of the columns of B, such that

1. *some subset of the columns of P makes up the $k \times k$ identity matrix,*
2. *no entry of P has an absolute value greater than α,*
3. *the spectral norm of P is at most $\sqrt{\alpha^2 k(n-k)+1}$,*
4. *the least (that is, the kth greatest) singular value of P is at least 1,*
5. *$B = CP$ when $k = m$ or $k = n$, and*
6. *when $k < m$ and $k < n$,*

$$\|B - CP\|_2 \leq \sqrt{\alpha^2 k(n-k)+1}\,\sigma_{k+1}, \tag{1}$$

where $\|B - CP\|_2$ is the spectral norm of the difference $B - CP$, and σ_{k+1} is the $(k+1)$th greatest singular value of B (also, σ_{k+1} is the spectral norm $\|B - \tilde{B}\|_2$ minimized over every $\tilde{B}$ whose rank is at most k).

Existing algorithms for computing C and P in Theorem 1 are computationally expensive when $\alpha = 1$, so normally we instead use $\alpha = 2$. We often select k in the theorem so

that $\|B - CP\|$ is at most some specified precision ϵ. We say that C collects together a subset of the columns of B and that P is an interpolation matrix, expressing to precision ϵ each column of B as a linear combination of the subset collected together into C. The factorization into the product of C and P is known as an interpolative decomposition (ID). Properties 1–4 of Theorem 1 ensure that the ID is numerically stable.

3. Mathematical constructions

This section develops mathematical theory for regression-aware decompositions, starting with the interpolative decomposition (ID) in Subsections 3.1–3.3, progressing to the singular value decomposition (SVD) in Subsection 3.4, and then simplifying the requisite numerical computations in Subsections 3.5 and 3.6. The procedures suggested in Subsections 3.5 and 3.6 are numerically stable, whereas the procedures in earlier subsections of the present section can be numerically unstable in naive implementations; Subsections 3.5 and 3.6 may seem more abstruse, yet are preferable for computations in finite-precision arithmetic.

3.1. An ID with an auxiliary matrix

This subsection provides a general formulation, of which the following two subsections are special cases.

Given matrices A and B of sizes conforming for the product AB, we can form an ID of AB, collecting together a subset of the columns of AB into a matrix AC, where C collects together a subset of the columns of B, together with an interpolation matrix P, such that

$$\|AB - ACP\| \leq \epsilon. \tag{2}$$

Expressing (2) as

$$\|A(B - CP)\| \leq \epsilon, \tag{3}$$

we may view this as interpolating stably and accurately to all columns of B from the subset collected together in C, provided that the accuracy of the interpolation is measured via the "norm" in (3) involving A,

$$\|D\|_A = \|AD\| \tag{4}$$

for any matrix D of size conforming for the product AD, including $D = B - CP$.

3.2. An ID for regression

This subsection constructs an ID that is informed by linear least-squares regression, attaining high accuracy when measuring errors directly on the least-squares solutions

(which is a terrible idea in the typical, numerically rank-deficient case of interest for dimensionality reduction). The following subsection alters the simplistic yet instructive formulation of the present subsection, instead measuring errors via the residuals of the least-squares fits.

Substituting the pseudoinverse $A^\dagger$ for A in Subsection 3.1, we obtain the following: Given matrices A and B of sizes conforming for the product $A^\dagger B$, we can form an ID of $A^\dagger B$ such that

$$\|A^\dagger B - A^\dagger C P\| \leq \epsilon, \tag{5}$$

where P is an interpolation matrix, and C collects together a subset of the columns of B. Denoting by X the minimizer of $\|AX - B\|$ given by $X = A^\dagger B$ and by Y the minimizer of $\|AY - C\|$ given by $Y = A^\dagger C$, we may express (5) as

$$\|X - YP\| \leq \epsilon. \tag{6}$$

Thus, the selected columns of B collected together into C enable accurate interpolation from the corresponding least-squares solutions to the least-squares solutions for all columns of B.

3.3. A regression-aware ID

This subsection constructs a decomposition which answers the question of how a matrix B looks under the general linear model with a given design matrix A, that is, how B looks under the regression which minimizes $\|AX - B\|$. "Looks" means that the decomposition provides a subset of the columns of B such that the least-squares solutions for the subset can stably and to high precision be (multi)linearly interpolated to the least-squares solutions (X) for all columns of B, at least when measuring accuracy via the residuals $\|AX - B\|$. [Statisticians, beware: X denotes the solution $X = A^\dagger B$ to the linear least-squares regression minimizing $\|AX - B\|$, not the design matrix. The design matrix is A.]

Here, given a matrix A, we define

$$S = (A^* A)^{-1/2} A^*; \tag{7}$$

notice that

$$AA^\dagger = S^* S. \tag{8}$$

Substituting S for A in Subsection 3.1, we obtain the following: Given matrices A and B of sizes conforming for the product SB, we can form an ID of SB having an interpolation matrix P such that

$$\|SB - SCP\| \leq \epsilon, \tag{9}$$

where C collects together a subset of the columns of B (or, equivalently, SC collects together a subset of the columns of SB). Denoting by X the minimizer of $\|AX-B\|$ given by $X = A^\dagger B$ and by Y the minimizer of $\|AY - C\|$ given by $Y = A^\dagger C$, combining (8), (9), the unitary invariance of the norm, and the fact that each singular value of S defined in (7) is either 1 or 0 yields that

$$\|AX - AYP\| = \|AA^\dagger[B - CP]\| = \|S^*S[B - CP]\| = \|SB - SCP\| \leq \epsilon. \tag{10}$$

As (10) shows, the selected columns of B collected together into C enable numerically stable interpolation from the corresponding least-squares solutions to the least-squares solutions for all columns of B, to high precision, when measuring accuracy via the residuals. Indeed, (10) yields that

$$\big|\|AX - B\| - \|AYP - B\|\big| \leq \|(AX - B) - (AYP - B)\| \leq \epsilon. \tag{11}$$

Remark 2. The nonzero singular values and corresponding right singular vectors of $(A^*A)^{-1/2}A^*$ and of the self-adjoint square root of $AA^\dagger$ are the same; note also that $AA^\dagger$ is the self-adjoint square root of itself — $AA^\dagger$ is an orthogonal projector. So, constructing an ID of $AA^\dagger B$ could select the same columns of B and produce the same interpolation matrix P as the above procedure, which constructs an ID of $(A^*A)^{-1/2}A^*B$. However, using $(A^*A)^{-1/2}A^*$ is more efficient when A is tall and skinny.

3.4. Regression-aware principal component analysis

The singular value decomposition (SVD) provides an alternative to using IDs. Given matrices A and B of sizes conforming for the product A^*B, the SVD of $(A^*A)^{-1/2}A^*B$ provides a kind of regression-aware principal component analysis (PCA), as PCA and SVD are more or less the same. This is basically the canonical correlation analysis (CCA) of [9], though CCA usually involves whitening B to $B(B^*B)^{-1/2}$ prior to taking the SVD: the most popular formulation of CCA forms the SVD of $(A^*A)^{-1/2}A^*B(B^*B)^{-1/2}$. That said, the SVD of $(A^*A)^{-1/2}A^*B$ has the interpretation developed in the previous subsection as a regression-aware PCA, even without whitening.

In detail, defining S via (7), we can form a low-rank approximation of SB with matrices U, Σ, and V such that

$$\|SB - U\Sigma V^*\| \leq \epsilon, \tag{12}$$

where the columns of U are orthonormal, as are the columns of V, the entries of Σ are all nonnegative and are zero off the main diagonal, and the column span of U lies in the column span of S. Denoting by X the minimizer of $\|AX - B\|$ given by $X = A^\dagger B$,

combining (8), (12), the unitary invariance of the norm, and the fact that each singular value of S defined in (7) is either 1 or 0 yields that

$$\|AX - S^*U\Sigma V^*\| = \|AA^\dagger B - S^*U\Sigma V^*\| = \|S^*SB - S^*U\Sigma V^*\| = \|SB - U\Sigma V^*\| \leq \epsilon. \tag{13}$$

In particular, combining (7) and (13) yields that

$$\begin{aligned} \big|\|AX - B\| - \|AT\Sigma V^* - B\|\big| &= \big|\|AX - B\| - \|S^*U\Sigma V^* - B\|\big| \\ &\leq \|(AX - B) - (S^*U\Sigma V^* - B)\| \leq \epsilon, \end{aligned} \tag{14}$$

where

$$T = (A^*A)^{-1/2}U. \tag{15}$$

This shows that the reduced-rank representation $T\Sigma V^*$ permits reconstruction of B effectively as accurately as $X = A^\dagger B$ permits reconstruction (via AX) of B, with X being the best possible minimizer of $\|AX - B\|$. Admittedly, the interpretation here is not as satisfying as that in Subsection 3.3, but is clearly strongly related just the same. Singular vectors are linear combinations of the original vectors, whereas the columns selected in Subsection 3.3 are simply a subset of the original vectors.

3.5. Simpler computations

This subsection provides a computationally simpler (albeit less natural) version of Subsection 3.3.

Here, given a matrix A, we form a pivoted QR decomposition

$$A = QR\Pi, \tag{16}$$

where the columns of Q are orthonormal, Π is a permutation matrix, and R is an upper-triangular (or upper-trapezoidal) matrix whose entries on the main diagonal are all nonzero; notice that

$$AA^\dagger = QQ^*. \tag{17}$$

Substituting Q^* for A in Subsection 3.1, we obtain the following: Given matrices A and B of sizes conforming for the product Q^*B, we can form an ID of Q^*B, collecting together a subset of the columns of Q^*B into a matrix Q^*C, such that

$$\|Q^*B - Q^*CP\| \leq \epsilon, \tag{18}$$

where P is an interpolation matrix, and C collects together a subset of the columns of B. Denoting by X the minimizer of $\|AX - B\|$ given by $X = A^\dagger B$ and by Y the minimizer

of $\|AY - C\|$ given by $Y = A^\dagger C$, combining (17), (18), the unitary invariance of the norm, and the fact that the columns of Q from (16) are orthonormal yields that

$$\|AX - AYP\| = \|AA^\dagger[B - CP]\| = \|QQ^*[B - CP]\| = \|Q^*B - Q^*CP\| \leq \epsilon. \quad (19)$$

Thus, the selected columns of B collected together into C enable numerically stable interpolation from the corresponding least-squares solutions to the least-squares solutions for all columns of B, to high precision, when measuring accuracy via the residuals. In fact, (19) yields that

$$\big|\|AX - B\| - \|AYP - B\|\big| \leq \|(AX - B) - (AYP - B)\| \leq \epsilon. \quad (20)$$

3.6. Another way to regression-aware PCA

This subsection provides a computationally simpler (albeit less natural) version of Subsection 3.4.

In the present subsection, given a matrix A, we form a pivoted QR decomposition

$$A = QR\Pi, \quad (21)$$

where the columns of Q are orthonormal, Π is a permutation matrix, and R is an upper-triangular (or upper-trapezoidal) matrix whose entries on the main diagonal are all nonzero. We can form a low-rank approximation of Q^*B with matrices U, Σ, and V such that

$$\|Q^*B - U\Sigma V^*\| \leq \epsilon, \quad (22)$$

where the columns of U are orthonormal, as are the columns of V, the entries of Σ are all nonnegative and are zero off the main diagonal, and the column span of U lies in the column span of Q^*. Denoting by X the minimizer of $\|AX - B\|$ given by $X = A^\dagger B$, combining (17), (22), the unitary invariance of the norm, and the fact that the columns of Q from (21) are orthonormal yields that

$$\|AX - QU\Sigma V^*\| = \|AA^\dagger B - QU\Sigma V^*\| = \|QQ^*B - QU\Sigma V^*\| = \|Q^*B - U\Sigma V^*\| \leq \epsilon. \quad (23)$$

In particular, combining (21) and (23) yields that

$$\begin{aligned}\big|\|AX - B\| - \|AT\Sigma V^* - B\|\big| &= \big|\|AX - B\| - \|QU\Sigma V^* - B\|\big| \\ &\leq \|(AX - B) - (QU\Sigma V^* - B)\| \leq \epsilon, \end{aligned} \quad (24)$$

where

$$T = \Pi^{-1} R^{\dagger} U. \tag{25}$$

So, the reduced-rank representation $T\Sigma V^*$ permits reconstruction of B effectively as accurately as $X = A^{\dagger}B$ permits reconstruction (via AX) of B, with X minimizing $\|AX-B\|$. Since the columns of QU are orthonormal ($(QU)^*(QU)$ is the identity matrix), the singular values of $AT\Sigma V^* = QU\Sigma V^*$ are the diagonal entries of Σ.

4. Numerical examples

This section discusses several illustrative examples. The section first presents two examples with synthetic data, in Subsections 4.1 and 4.2, then considers real data, in Subsections 4.3–4.5. Software for running the examples is available at http://tygert.com/rad.tar.gz

Several of the examples display biplots, in the rightmost halves of Figs. 2–9; regarding biplots, please consult [4] or [7]. The horizontal coordinates of the black circular dots in the biplots are the leading "scores" — the greatest singular value times the entries of the corresponding left singular vector; the vertical coordinates of the black circular dots are the next leading scores — the second greatest singular value times the entries of the corresponding left singular vector. The horizontal coordinates of the tips of the gray lines in the biplots are the entries of the right singular vector corresponding to the greatest singular value; the vertical coordinates of the tips of the gray lines are the entries of the right singular vector corresponding to the second greatest singular value.

In Figs. 2–9, Q_A refers to a matrix whose columns form an orthonormal basis for the column span of A, and Q_B refers to a matrix whose columns form an orthonormal basis for the column span of B; the singular values σ_k of $(Q_A)^*Q_B$ are those arising in the canonical correlation analysis (CCA) between A and B, whereas the singular values σ_k of $(Q_A)^*B$ are those arising in the regression-aware principal component analysis (RAPCA) of B for A. The singular values for the RAPCA also determine the spectral-norm accuracy of the regression-aware interpolative decomposition (RAID), commensurate with formula (1).

4.1. Potential theory

This subsection considers points on concentric circles of radii 0.9, 1, and 1.1, as illustrated in Fig. 1. The interaction between a unit charge at point p and a unit charge at point q is $\ln(\|p-q\|)$, the potential energy for the Laplace equation in two dimensions, where $\|p-q\|$ denotes the Euclidean distance between p and q. We are interested in the interactions of the test charges in Fig. 1 with the original charges, but only those components of the interactions that are representable by the interactions of the test charges with the supervisory charges. There are 80 test charges spread evenly around the circle of radius 1. There are 20 original charges spread evenly around the top-left quadrant of the circle of radius 0.9. There are 20 supervisory charges spread evenly around the

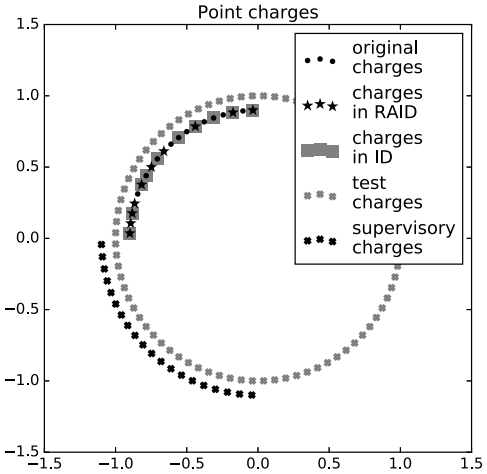

Fig. 1. Example from Subsection 4.1.

bottom-left quadrant of the circle of radius 1.1. The entries of an 80×20 matrix B are the natural logarithms of the distances between the test charges and the original charges, normalized such that the spectral norm $\|B\|_2$ becomes 1. The entries of an 80×20 matrix A are the natural logarithms of the distances between the test charges and the supervisory charges, normalized by the same factor as B. The ID of B considered here selects 10 representative charges from the original charges; the RAID of B for A selects a different set of 10. The spectral norm of the difference between B and its reconstruction from the ID is 0.016. The spectral-norm error of the RAID is 0.25E–10; the spectral-norm error is $\|AX - AYP\|_2$ from the left-hand side of (19). (For reference, $\min_X \|AX - B\|_2 = 0.67$, where $\|AX - B\|_2$ is the spectral norm of $AX - B$.) Thus, the interactions between the test charges and the 10 charges selected by the RAID are sufficient to capture to very high accuracy the interactions between the test charges and all the original charges, at least those components that are representable by the interactions between the test charges and the supervisory charges.

4.2. Synthetic time–series

This subsection analyzes a synthetic multivariate time–series, specifically the matrix C with $m = 10{,}000{,}000$ rows and $n = 10$ columns constructed as follows: We start with all entries being i.i.d. standard normal variates. Then, we multiply the first 5 columns by 1,000,000, and in each of the last 5 columns set all entries equal to the entry in the last row (so that the last 5 columns are just multiples of each other). Finally, we add to every entry 0.01 times the product of the row and column indices (this is a rank-1 perturbation). We let A be the first block of $m - 1$ rows of C and let B be the last block of $m - 1$ rows (so A includes the first row of C but not the last, whereas B includes the last row of C but not the first), dividing each entry of A and B by the same factor such that the spectral norm $\|B\|_2$ becomes 1.

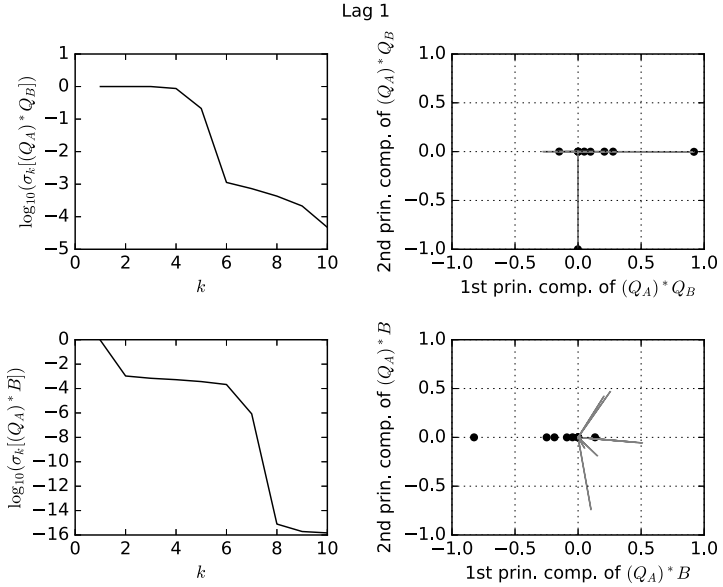

Fig. 2. Example from Subsection 4.2.

The ID of B considered here selects 4 representative columns from the originals; the RAID of B for A selects a different set of 4. Specifically, the ID ends up selecting columns 2–5, whereas the RAID ends up selecting columns 1, 2, 5, and 10 — the ID entirely misses the last 5 columns (which were multiples of each other prior to adding the rank-1 perturbation), whereas the RAID includes one of the second 5 columns (namely, the last). The spectral norm of the difference between B and its reconstruction from the ID is 0.80. The spectral-norm error of the RAID is 0.00039. (For reference, $\min_X \|AX-B\|_2 = 0.79$, where $\|AX - B\|_2$ is the spectral norm of $AX - B$.) Thus, the 4 columns selected by the RAID are sufficient to capture to high accuracy the entire multivariate time–series in B, at least its components that are linearly predictable with the previous lag of the time series from C (this lag is the time series in A).

Fig. 2 displays the singular values both for the matrix in the CCA between A and B and for the RAPCA of B for A (the former are in the top-left plot; the latter are in the bottom-left plot). Regarding the biplots in the rightmost half of Fig. 2, please consult [4] or [7]. Fig. 2 shows that the spectral-norm accuracy of the rank-4 RAPCA is similar to the excellent accuracy of the corresponding RAID, whereas the spectral-norm accuracy of the rank-4 CCA is very poor.

4.3. Electricity loads

This subsection considers electricity meter readings for 370 clients of a utility company from Portugal, with 140,256 readings per customer in total; this data from [10] is available at http://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014

Table 1
Example from Subsection 4.3.

l	$\min_X \|AX - B\|_2$	ID error	RAID error
100	.075	.020	.0037
200	.094	.020	.0030
300	.098	.020	.0029

together with its complete detailed specifications. We collect together the data into a $140{,}256 \times 370$ matrix C. In this subsection, we rescale each column of C so that its Euclidean norm becomes 1, thus "equalizing" clients. For varying values of a lag l (namely $l = 100$, 200, 300), we let A be the block of all rows of C except the last l, and let B be the block of all rows of C except the first l. We then divide each entry of both A and B by the same value, such that the spectral norm $\|B\|_2$ becomes 1.

The ID of B considered here selects 200 representative columns from the originals; the RAID of B for A selects a different set of 200. Table 1 reports the spectral-norm accuracies attained. Figs. 3–5 display the singular values both for the matrix in the CCA between A and B and for the RAPCA of B for A (the former are in the top-left plot of each figure; the latter are in the bottom-left plot of each figure). Regarding the biplots in the rightmost halves of Figs. 3–5, please consult [4] or [7]. Figs. 3–5 show that the spectral-norm accuracy of the rank-200 RAPCA is similar to the high accuracy of the corresponding RAID, whereas the spectral-norm accuracy of the rank-200 CCA is two orders of magnitude worse.

4.4. Electricity loads transposed

The present subsection considers the same data as in the previous subsection, Subsection 4.3, but now we let B be the transpose of the block of the last 100,000 rows of C, and let A be the transpose of the block of the 300 rows just before B. We then divide each entry of both A and B by the same number, such that the spectral norm $\|B\|_2$ becomes 1. The aim here is to select a small number, say 3, of the columns of B that represent all 100,000 columns of B, or rather represent those components which are linearly predictable with columns from A. Thus, whereas the previous subsection selected representative clients, with the clients' histories regressed against the lagged histories, the present subsection selects representative times that the electricity meters were read, with each time-slice of meter readings predicted from the early readings collected together in A. Transposing makes the columns refer to time-slices rather than clients (rows then refer to clients).

As just mentioned, the ID of B considered here selects 3 representative columns from the originals; the RAID of B for A selects a different set of 3. The spectral norm of the difference between B and its reconstruction from the ID is 0.12. The spectral-norm error of the RAID is 0.044. (For reference, $\min_X \|AX - B\|_2 = 0.22$, where $\|AX - B\|_2$ is the spectral norm of $AX - B$.)

Table 2
Example from Subsection 4.5.

l	$\min_X \|AX - B\|_2$	ID error	RAID error
20	.41	.81	.16
40	.41	.78	.15
60	.42	.78	.13

Fig. 6 displays the singular values both for the matrix in the CCA between A and B and for the RAPCA of B for A (the former are in the top-left plot; the latter are in the bottom-left plot). Fig. 6 shows that the spectral-norm accuracy of the rank-3 RAPCA is similar to the accuracy of the corresponding RAID, whereas the CCA is vacuous (the logarithms of all singular values in the top-left plot of Fig. 6 are equal to 0 to nearly the machine precision of 0.22E–15).

4.5. Motion capture

This subsection considers real-valued features derived from motion-capture data of a person gesticulating; for each of 1,743 successive instants, there are 50 real numbers characterizing the gesticulator's motions and positions. This data of [14] and [10] is available at http://archive.ics.uci.edu/ml/datasets/Gesture+Phase+Segmentation together with its complete detailed specifications. We collect together the data into a $1{,}743 \times 50$ matrix C. For varying values of a lag l (namely $l = 20, 40, 60$), we let A be the block of all rows of C except the last l, and let B be the block of all rows of C except the first l. We rescale each column of A and each column of B so that their Euclidean norms become 1. We then divide each entry of both A and B by the same factor, such that the spectral norm $\|B\|_2$ becomes 1.

Here, we consider an ID which selects 2 representative columns of B and a RAID which selects a different 2. Table 2 reports the spectral-norm accuracies attained. Figs. 7–9 display the singular values both for the matrix in the CCA between A and B and for the RAPCA of B for A (the former are in the top-left plot of each figure; the latter are in the bottom-left plot of each figure). Regarding the biplots in the rightmost halves of Figs. 7–9, please consult [4] or [7]. Figs. 7–9 show that the spectral-norm accuracy of the rank-2 RAPCA is similar to the accuracy of the corresponding RAID, whereas the spectral-norm accuracy of the rank-2 CCA is nearly the worst possible.

Acknowledgement

We would like to thank Facebook Artificial Intelligence Research for supporting this work.

Appendix A. Additional figures

This appendix displays Figs. 3–9; see Section 4 for descriptions of these figures.

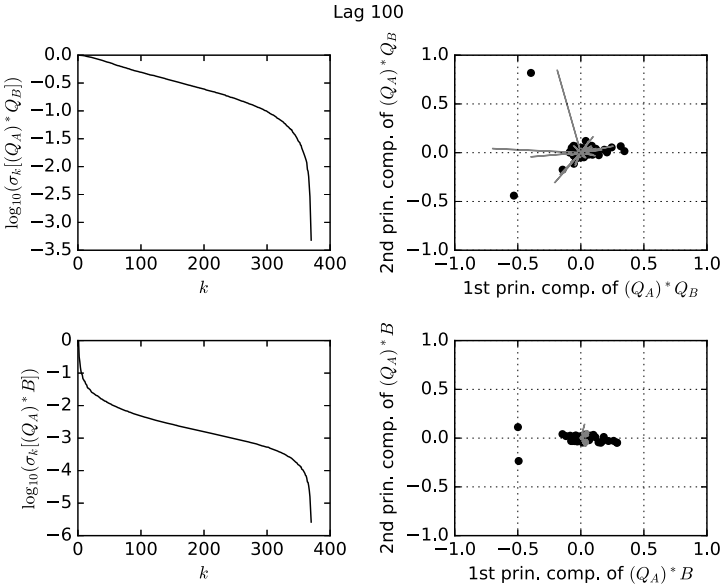

Fig. 3. Example from Subsection 4.3 with $l = 100$.

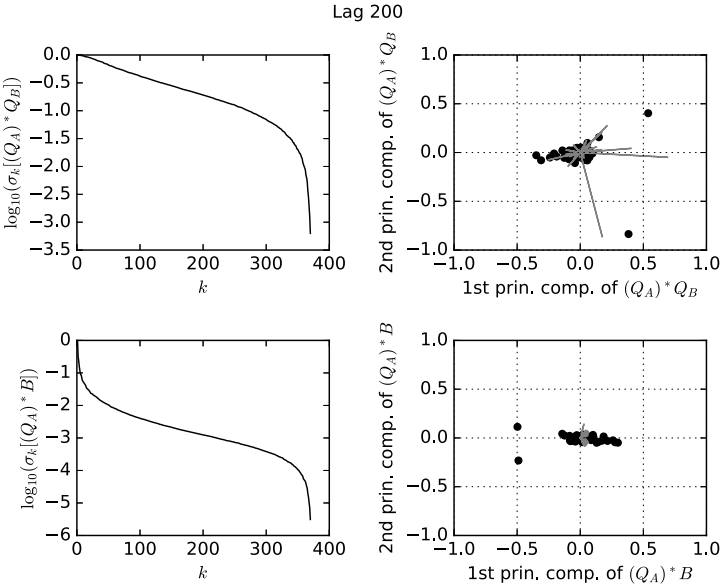

Fig. 4. Example from Subsection 4.3 with $l = 200$.

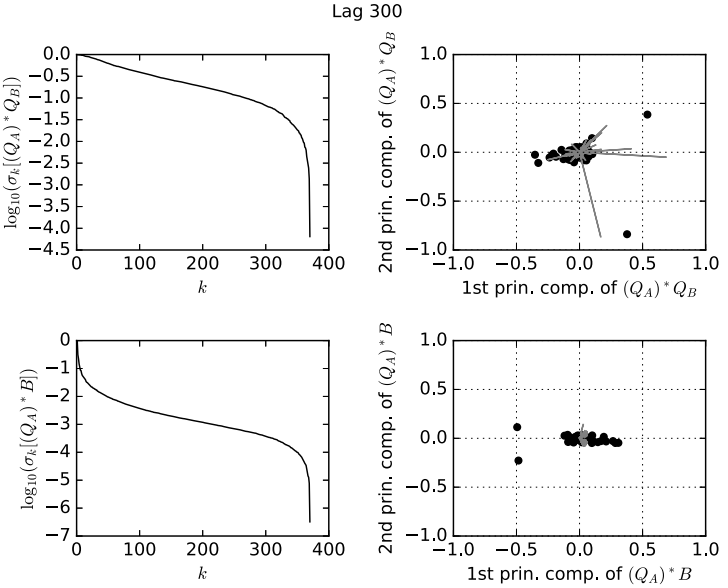

Fig. 5. Example from Subsection 4.3 with $l = 300$.

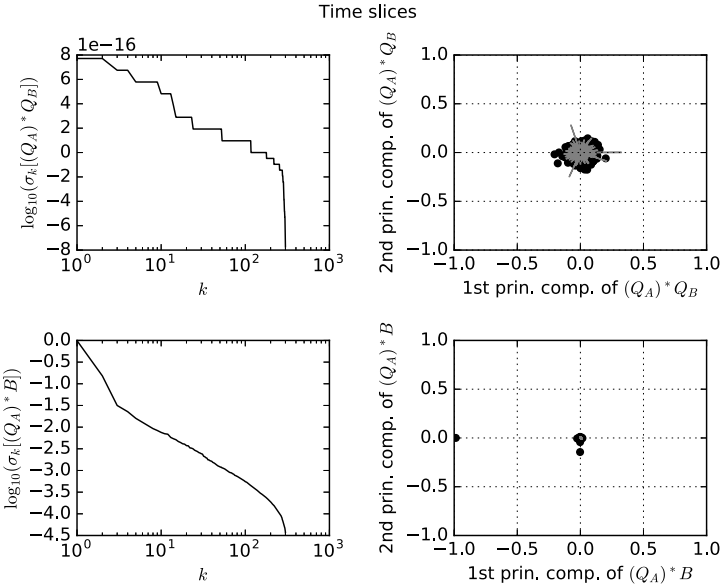

Fig. 6. Example from Subsection 4.4.

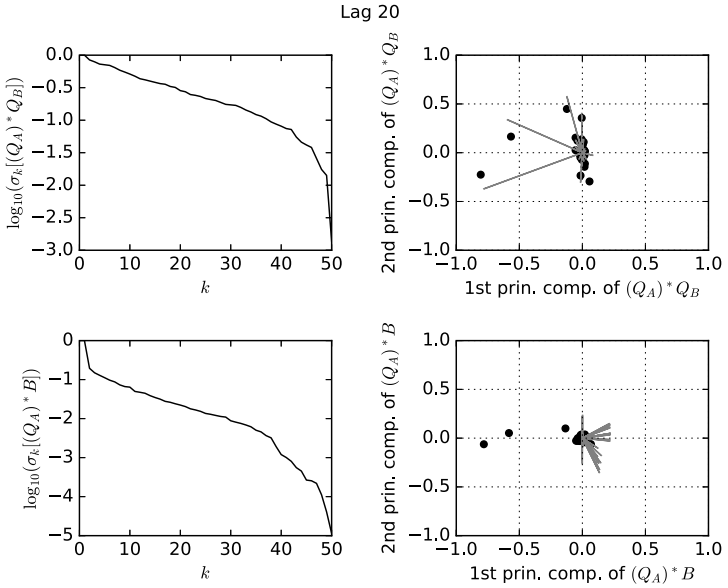

Fig. 7. Example from Subsection 4.5 with $l = 20$.

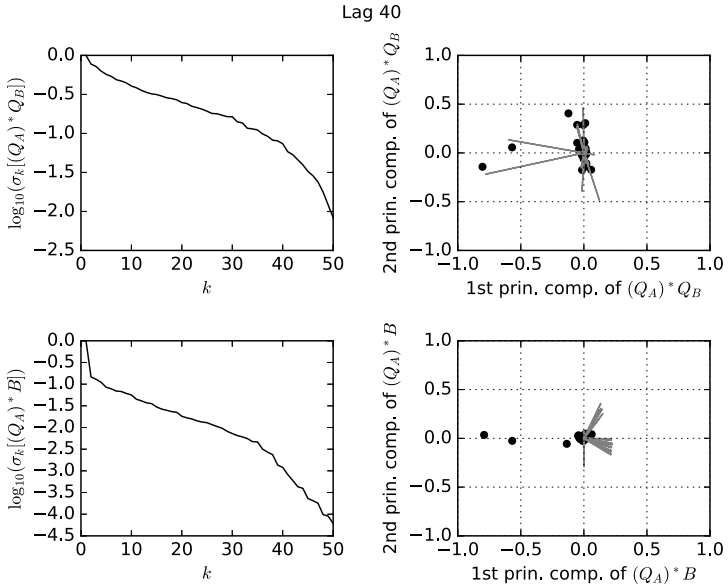

Fig. 8. Example from Subsection 4.5 with $l = 40$.

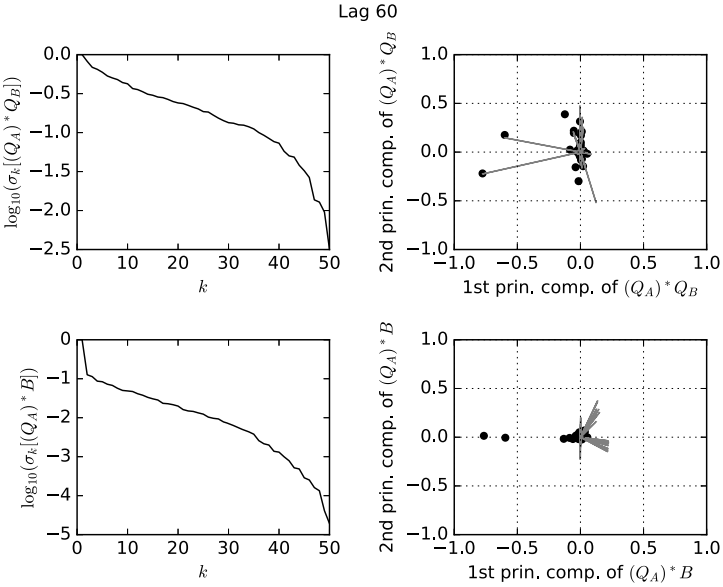

Fig. 9. Example from Subsection 4.5 with $l = 60$.

References

[1] H. Cheng, Z. Gimbutas, P. Martinsson, V. Rokhlin, On the compression of low-rank matrices, SIAM J. Sci. Comput. 26 (2006) 1389–1404.
[2] P. Drineas, M.W. Mahoney, S. Muthukrishnan, Relative-error CUR matrix decompositions, SIAM J. Matrix Anal. Appl. 30 (2008) 844–881.
[3] M. Fekete, Über die verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten, Math. Z. 17 (1923) 228–249.
[4] K.R. Gabriel, The biplot graphic display of matrices with application to principal component analysis, Biometrika 58 (1971) 453–467.
[5] G. Golub, C. Van Loan, Matrix Computations, 4th ed., Johns Hopkins University Press, 2012.
[6] S.A. Goreinov, E.E. Tyrtyshnikov, The maximal-volume concept in approximation by low-rank matrices, in: V. Olshevsky (Ed.), Structured Matrices in Mathematics, Computer Science, and Engineering I, vol. 280, American Mathematical Society, Providence, RI, 2001, pp. 47–52.
[7] J. Gower, S. Lubbe, N. le Roux, Understanding Biplots, John Wiley & Sons, 2011.
[8] M. Gu, S.C. Eisenstat, Efficient algorithms for computing a strong rank-revealing QR factorization, SIAM J. Sci. Comput. 17 (1996) 848–869.
[9] H. Hotelling, Relations between two sets of variables, Biometrika 28 (1936) 321–377.
[10] M. Lichman, UCI machine learning repository, http://archive.ics.uci.edu/ml, 2018.
[11] P. Martinsson, V. Rokhlin, Y. Shkolnisky, M. Tygert, ID: a software package for low-rank approximation of matrices via interpolative decompositions, Version 0.4, http://tygert.com/software.html, March 2014.
[12] A. Szlam, A. Tulloch, M. Tygert, Accurate low-rank approximations via a few iterations of alternating least squares, SIAM J. Matrix Anal. Appl. 38 (2017) 425–433.
[13] E.E. Tyrtyshnikov, Incomplete cross approximation in the mosaic-skeleton method, Computing 64 (2000) 848–869.
[14] P.K. Wagner, S.M. Peres, C.A.M. Lima, F.A. Freitas, R.C.B. Madeo, Gesture unit segmentation using spatial-temporal information and machine learning, in: Proc. 27th Internat. Florida Artificial Intelligence Research Soc. Conf., AAAI Press, 2014, pp. 101–106.