National Physical Laboratory

NPL REPORT MS 18

Traceability for computationally – intensive metrology: Benchmarking reference pairs for Principal Components Analysis

P M Harris

MAY 2015

National Measurement System

NPL Report MS 18

Traceability for computationally-intensive metrology: Benchmarking reference pairs for Principal Components Analysis

P M Harris
Mathematics and Modelling Group

May 2015

ABSTRACT

Ten benchmarking reference pairs for the computational aim of Principal Components Analysis are described, and instructions for how the benchmarking reference pairs can be used to test a candidate software implementation of the computational aim are given. The report is a deliverable of the EMRP joint research project NEW06 "Traceability for computationally-intensive metrology" (TraCIM).

ISSN 1754–2960

National Physical Laboratory,
Hampton Road, Teddington, Middlesex, United Kingdom TW11 0LW

We gratefully acknowledge the financial support of the UK Department for Business, Innovation and Skills (National Measurement Office)

Approved on behalf of NPLML by Prof Alistair Forbes,
Science Area Leader for the Mathematics & Modelling Group

Contents

1 Introduction

The report describes ten benchmarking reference pairs for the computational aim of Principal Components Analysis (PCA) [1]. It also describes how the benchmarking reference pairs can be used to test a candidate software implementation of the computational aim.

The report is organised as follows. The computational aim addressed in this work is recorded in section 2, and the form of the reference pairs is described in section 3. Considerations about the numerical uncertainty associated with the reference pairs are given in section 4, and performance metrics for comparing reference and test results are defined in section 5. Details of the particular reference pairs that are provided are given in section 6 together with information about the formats of the files used to store the reference pairs. Finally, a procedure for using the benchmarking reference pairs to assess the numerical performance of test software for the computational aim is presented in section 7.

1.1 Disclaimer

The reference pairs are made available from the TraCIM Computational Aims Database and Website. NPL does not offer any validation or certification against the results obtained by using the reference pairs. Use of the website and the reference pairs is subject to the terms laid out in the following Disclaimer.[1]

1. Although care has been taken in preparing the information supplied on the website, to the fullest extent permitted at law, NPL acting on behalf of EURAMET is providing this website and its contents "as is" and makes no (and expressly disclaims all) representations or warranties of any kind, express or implied, with respect to the website and the information, content, materials or products included therein, including, without limitation, warranties of merchantability and fitness for a particular purpose. In addition, NPL does not represent or warrant that the information accessible via the website is accurate, complete or current.

2. To the fullest extent permitted at law, neither NPL or any of its affiliates, directors, employees or other representatives will be liable for damages arising out of or in connection with the use of this website or the information, content, materials or products included on this site. This is a comprehensive limitation of liability that applies to all damages, loss of data, income or profit, loss or damage to property and claims of third parties. For avoidance of doubt, NPL does not limit its liability for death or personal injury to the extent that it arises as a result of the negligence of NPL, its affiliates, directors, employees, or other representatives.

   In addition:

[1] See also www.tracim-cadb.npl.co.uk/conditions.php.

(a) NPL does not and cannot guarantee the accuracy, reliability, or completeness; nor can NPL be held responsible for any errors or omissions; nor can NPL be held liable for any infringement of the rights of any third party by information posted to the website.

(b) Your use of the information held on this site is at your sole risk.

(c) NPL shall not be liable for any direct, indirect, incidental, special, consequential or exemplary damages, including but not limited to damages for loss of profits, goodwill, use, data or other intangible losses, resulting from:

   i. the use of or inability to use the information held on this website;
   ii. the cost of procurement of goods and services resulting from any data, information, or services obtained through or from the website;
   iii. unauthorised access or alternation to your contributions;
   iv. statements or conduct of any third party on the website;
   v. any other matter relating to the information held on the website.

2 Computational aim

2.1 Problem definition

Given the mean-centred matrix $\boldsymbol{X}$ of dimension $n \times p$, i.e.,

$$\sum_{i=1}^{n} x_{ij} = 0, \quad j = 1, \ldots, p,$$

and the vector $\boldsymbol{v} = (v_1, \ldots, v_p)^\top$, define

$$\mathrm{var}(\boldsymbol{X}\boldsymbol{v}) = \frac{1}{n} ||\boldsymbol{X}\boldsymbol{v}||_2^2,$$

where $||\boldsymbol{z}||_2$ is the Euclidean norm of vector $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$:

$$||\boldsymbol{z}||_2 = \sqrt{\sum_{i=1}^{n} z_i^2}.$$

Let $\boldsymbol{v}_1$ be a solution to

$$\max_{||\boldsymbol{v}||=1} \mathrm{var}(\boldsymbol{X}\boldsymbol{v}),$$

and $\boldsymbol{v}_k,\ k = 2, \ldots, \ell = \min(n, p)$, solutions to

$$\max_{||\boldsymbol{v}||=1} \mathrm{var}(\boldsymbol{X}\boldsymbol{v}) \quad \text{subject to} \quad \boldsymbol{v}_j^\top \boldsymbol{v}_k = 0,\ j = 1, \ldots, k-1.$$

Furthermore, for $k = 1, \ldots, \ell$, define

$$\boldsymbol{y}_k = \boldsymbol{X}\boldsymbol{v}_k \quad \text{and} \quad \lambda_k = n\text{var}(\boldsymbol{y}_k).$$

The values $\lambda_k,\ k = 1, \ldots, \ell$, are the *principal components* for the matrix $\boldsymbol{X}$, the vectors $\boldsymbol{v}_k,\ k = 1, \ldots, \ell$, are the *principal component loadings*, and the vectors $\boldsymbol{y}_k,\ k = 1, \ldots, \ell$, are the *principal component scores*.

2.2 Equivalence of solutions

Let $m \le \ell$ be the number of distinct principal components, and denote those distinct values by $\mu_j,\ j = 1, \ldots, m$. Furthermore, let ℓ_j be the number of principal components equal to μ_j, and denote the set of indices of those principal components by $K_j \subseteq \{1, \ldots, \ell\}$. The principal component loadings $\boldsymbol{v}_k,\ k \in K_j$, corresponding to the principal component μ_j are not uniquely defined: any vectors that span the subspace defined by $\boldsymbol{v}_k,\ k \in K_j$, constitute a solution to the problem defined in section 2.1. For example, if $\ell_j = 1$, then the vector $\boldsymbol{v}_k, k \in K_j$, is determined only up to its sign.

Let $\boldsymbol{V}_j$ be the matrix of dimension $p \times \ell_j$ containing in its columns the solution $\boldsymbol{v}_k,\ k \in K_j$. Then, the principal component loadings for $k \in K_j$ are defined by $\boldsymbol{V}_j$ and its equivalence class

$$[\boldsymbol{V}_j] = \left\{ \boldsymbol{W}_j : \exists\, \boldsymbol{C} \text{ of dimension } \ell_j \times \ell_j \text{ such that } \boldsymbol{C}^\top \boldsymbol{C} = \boldsymbol{I}_{\ell_j} \text{ and } \boldsymbol{W}_j = \boldsymbol{V}_j \boldsymbol{C} \right\},$$

i.e., the principal component loadings contained in $\boldsymbol{W}_j$ are orthonormal and define the same sub-space as those contained in $\boldsymbol{V}_j$. For example, if $\ell_j = 1$, then $\boldsymbol{C} = 1$ or $\boldsymbol{C} = -1$, and

$$[\boldsymbol{V}_j] = \{\boldsymbol{V}_j, -\boldsymbol{V}_j\}.$$

Two implementations of the computational aim are considered equivalent if their solutions comprise

1. the same principal components, and
2. for each distinct principal component, the principal component loadings are members of the same equivalence class.

3 Reference pairs

Suppose $\boldsymbol{X}$ of dimension $n \times p$ is a mean-centred matrix. Then, a reference pair consists of:

- The reference data $\boldsymbol{X}$;

- The corresponding reference results $\{\lambda_k,\ I_k,\ \boldsymbol{v}_k,\ k = 1, \ldots, \ell\}$, where $\ell = \min(n, p)$ and $I_k = j \Leftrightarrow \lambda_k = \mu_j$.[2]

4 Numerical uncertainty

Let $\widehat{\lambda}_k = \lambda_k^{\mathrm{f}}(\boldsymbol{X})$ be a value for the kth principal component[3], for example, as might be returned by a reference or test software implementation of the computational aim. The closeness of agreement of the value with the corresponding reference value λ_k is given by

$$d(\lambda_k, \widehat{\lambda}_k) = \lambda_k - \widehat{\lambda}_k, \quad k = 1, \ldots, \ell.$$

Let $\widehat{\boldsymbol{v}}_k = \boldsymbol{v}_k^{\mathrm{f}}(\boldsymbol{X}),\ k \in K_j$, be principal component loadings corresponding to the jth distinct principal component μ_j, for example, as might be returned by a reference or test software implementation of the computational aim. The closeness of agreement between these loadings and the reference loadings $\boldsymbol{v}_k,\ k \in K_j$, is measured by the angles between each loading and the subspace defined by the reference loadings, and is given by

$$d(\widehat{\boldsymbol{v}}_k, \boldsymbol{V}_j) = 2 \sin^{-1} \left(\frac{1}{2} \left\| \widehat{\boldsymbol{v}}_k - \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|} \right\| \right), \quad k \in K_j, \quad j = 1, \ldots, m,$$

where

$$\boldsymbol{w}_k = \boldsymbol{V}_j \boldsymbol{V}_j^{\top} \widehat{\boldsymbol{v}}_k$$

is the projection of $\widehat{\boldsymbol{v}}_k$ onto the subspace and $\boldsymbol{V}_j$ is the matrix of dimension $p \times \ell_j$ containing in its columns the reference loadings $\boldsymbol{v}_k,\ k \in K_j$.

The numerical uncertainty associated with a reference result is expressed in two ways:

- By the *accuracy* $\{A(\lambda_k),\ A(\boldsymbol{v}_k),\ k = 1, \ldots, \ell\}$ of the reference results;
- By the *sensitivity* $\{S(\lambda_k),\ S(\boldsymbol{v}_k),\ k = 1, \ldots, \ell\}$ of the reference results to perturbations in the reference data.

The accuracy of each reference result is given by its closeness of agreement with a result obtained using an extended-precision software implementation of the computational aim, which is considered to behave as a reference software implementation of the computational aim. The sensitivity of each reference result is expressed as a sensitivity coefficient. Random perturbations are applied to the reference data, and the sensitivity coefficient is determined as the quotient of a measure for the variation of the reference result divided by a measure for the variation of the reference data. The measure for the variation of the reference

[2]The indices $I_k,\ k = 1, \ldots, \ell$, are related to the index sets $K_j,\ j = 1, \ldots, m$, as follows: $I_{K_{jk}} = j$, $k = 1, \ldots, \ell_j$, $j = 1, \ldots, m$.

[3]Here, the notation λ_k^{f} is used to distinguish the *function* that maps the matrix $\boldsymbol{X}$ onto the kth principal component from the *value* λ_k of the principal component.

result is the standard deviation of the calculated accuracies of the reference result compared to the results obtained for the perturbed data sets using an extended-precision software implementation of the computational aim. The extended-precision software implementation of the computational aim is again considered to behave as a reference software implementation of the computational aim. The measure for the variation of the reference data is given by the relative standard deviation of the perturbations to the individual components of the input data.

5 Performance metrics

For completeness, a check on the number of principal components and principal component loadings returned by test software should be made before tests on the values of the principal components and principal component loadings are undertaken. Such a check can be implemented as a (simple) comparison of two integers, viz., the number of principal components returned by the test software and the reference value $\ell = \min\{n, p\}$. Considerations about the numerical uncertainty associated with the reference value ℓ are not required.

Let $\{\lambda_k^{\mathrm{t}},\ \boldsymbol{v}_k^{\mathrm{t}},\ k = 1, \ldots, \ell\}$ be the principal components and principal component loadings for the reference data $\boldsymbol{X}$ provided by a test software implementation of the computational aim operating with floating-point relative accuracy ϵ_{w}.[4] The test results are compared with reference results in the following way:

1. Calculate
$$D(\lambda_k^{\mathrm{t}}) = |d(\lambda_k, \lambda_k^{\mathrm{t}})|, \quad k = 1, \ldots, \ell, \tag{1}$$
and
$$D(\boldsymbol{v}_k^{\mathrm{t}}) = |d(\boldsymbol{v}_k^{\mathrm{t}}, \boldsymbol{V}_j)|, \quad k \in K_j, \quad j = 1, \ldots, m. \tag{2}$$

2. Calculate
$$D_A(\lambda_k^{\mathrm{t}}) = \max\left\{D(\lambda_k^{\mathrm{t}}), A(\lambda_k)\right\}, \quad k = 1, \ldots, \ell,$$
and
$$D_A(\boldsymbol{v}_k^{\mathrm{t}}) = \max\left\{D(\boldsymbol{v}_k^{\mathrm{t}}), A(\boldsymbol{v}_k)\right\}, \quad k = 1, \ldots, \ell,$$
which accounts for the numerical accuracy associated with the reference results.

3. Calculate
$$P(\lambda_k^{\mathrm{t}}) = \log_{10}\left(1 + \frac{D(\lambda_k^{\mathrm{t}})}{S(\lambda_k)\epsilon_{\mathrm{w}}}\right), \quad k = 1, \ldots, \ell, \tag{3}$$
and
$$P(\boldsymbol{v}_k^{\mathrm{t}}) = \log_{10}\left(1 + \frac{D(\boldsymbol{v}_k^{\mathrm{t}})}{S(\boldsymbol{v}_k)\epsilon_{\mathrm{w}}}\right), \quad k = 1, \ldots, \ell, \tag{4}$$
which accounts for the sensitivity of the reference results.

[4]The distance from 1 to the next largest floating-point number.

For a generic test result q^{t} for which the reference result is q, the performance metric $P(q^{\mathrm{t}})$ measures the number of decimal digits of accuracy lost by the test software *in addition to* the number that would be expected to be lost by reference software for the computational aim operating with the same floating-point relative accuracy. Here, the number of decimal digits of accuracy that would be expected to be lost by reference software is measured by the sensitivity of the reference result.

For example, suppose

$$A(q) = 10^a \epsilon_{\mathrm{w}}, \quad S(q) = 10^b, \quad D(q^{\mathrm{t}}) = 10^c \epsilon_{\mathrm{w}}.$$

Consider first the case $c \geq a$, i.e., the reference result has more significant digits of accuracy than the test result. Then,

$$P(q^{\mathrm{t}}) = \log_{10}\left(1 + \frac{10^c \epsilon_{\mathrm{w}}}{10^b \epsilon_{\mathrm{w}}}\right) = \log_{10}\left(1 + 10^{c-b}\right),$$

and

- $P(q^{\mathrm{t}}) \approx 0$ when $c < b$, or
- $P(q^{\mathrm{t}}) \approx c - b$ when $c \geq b$.

In the alternative case when $c < a$, i.e., the numerical accuracy of the reference result is not adequate to distinguish numerically between the test and reference results, then

$$P(q^{\mathrm{t}}) = \log_{10}\left(1 + \frac{10^a \epsilon_{\mathrm{w}}}{10^b \epsilon_{\mathrm{w}}}\right) = \log_{10}\left(1 + 10^{a-b}\right),$$

and the performance of the test software is measured by the numerical accuracy of the reference result.

6 Benchmarking reference pairs

Benchmarking reference pairs are provided with the properties described in sections 6.1 to 6.10. Specifications of the formats of the ASCII data and results files for each benchmarking reference pair are given in section 6.11.

6.1 PCA_test01

$n = 20$, $p = 10$, with $m = 10$ distinct principal components between 10^0 and 10^{-1} each with multiplicity one.

6.2 PCA_test02

$n = 20$, $p = 10$, with $m = 10$ distinct principal components between 10^0 and 10^{-2} each with multiplicity one.

6.3 PCA_test03

$n = 20$, $p = 10$, with $m = 10$ distinct principal components between 10^0 and 10^{-3} each with multiplicity one.

6.4 PCA_test04

$n = 20$, $p = 10$, with $m = 10$ distinct principal components between 10^0 and 10^{-4} each with multiplicity one.

6.5 PCA_test05

$n = 20$, $p = 10$, with $m = 10$ distinct principal components between 10^0 and 10^{-5} each with multiplicity one.

6.6 PCA_test06

$n = 100$, $p = 10$, with $m = 10$ distinct principal components each with multiplicity one.

6.7 PCA_test07

$n = 100$, $p = 15$, with $m = 15$ distinct principal components each with multiplicity one.

6.8 PCA_test08

$n = 5$, $p = 10$, with $m = 5$ distinct principal components each with multiplicity one.

6.9 PCA_test09

$n = 15$, $p = 10$, with $m = 5$ distinct principal components each with multiplicity two.

6.10 PCA_test10

$n = 25$, $p = 10$, with $m = 8$ distinct principal components, the largest and smallest having multiplicity two and the others multiplicity one.

6.11 File formats

The file 'fname_X.txt' containing reference data has the following format:

$$\begin{array}{ll} n & p \\ X_{11} & \\ \vdots & \\ X_{n1} & \\ X_{12} & \\ \vdots & \\ X_{n2} & \\ X_{1p} & \\ \vdots & \\ X_{np} & \end{array}$$

The file 'fname_L.txt' containing reference results for the principal components has the following format:

$$\begin{array}{ll} \ell & \\ \lambda_1 & I_1 \\ & \vdots \\ \lambda_\ell & I_\ell \end{array}$$

The file 'fname_Lunc.txt' containing the accuracies and sensitivities of the reference results for the principal components has the following format:

$$\begin{array}{ll} \ell & \\ A(\lambda_1) & S(\lambda_1) \\ & \vdots \\ A(\lambda_\ell) & S(\lambda_\ell) \end{array}$$

The file 'fname_V.txt' containing reference results for the principal component loadings has the following format:

$p \quad \ell$
V_{11}
$\vdots$
V_{p1}
V_{12}
$\vdots$
V_{p2}
$V_{1\ell}$
$\vdots$
$V_{p\ell}$

The file 'fname_Vunc.txt' containing the accuracies and sensitivities of the reference results for the principal component loadings has the following format:

ℓ
$A(\boldsymbol{v}_1) \quad S(\boldsymbol{v}_1)$
$\vdots$
$A(\boldsymbol{v}_\ell) \quad S(\boldsymbol{v}_\ell)$

7 Procedure

A procedure for using the benchmarking reference pairs described in section 6 to assess the numerical performance of test software for PCA is as follows:

1. Read reference data $\boldsymbol{X}$ from a file 'fname_X.txt' (see section 6.11);
2. Read corresponding reference results $\{\lambda_k,\ I_k,\ \boldsymbol{v}_k,\ k = 1, \ldots, \ell\}$ from files 'fname_L.txt' and 'fname_V.txt' (see section 6.11);
3. Read numerical accuracies $\{A(\lambda_k),\ A(\boldsymbol{v}_k),\ k = 1, \ldots, \ell\}$ and numerical sensitivities $\{S(\lambda_k),\ S(\boldsymbol{v}_k),\ k = 1, \ldots, \ell\}$ of the reference results from files 'fname_Lunc.txt' and 'fname_Vunc.txt' (see section 6.11);
4. Apply test software to obtain test results $\{\lambda_k^{\mathrm{t}},\ \boldsymbol{v}_k^{\mathrm{t}},\ k = 1, \ldots, \ell\}$;
5. Set the working precision ϵ_{w} used to obtain the test results;
6. Evaluate the absolute differences (1) and (2) betweeen the test and reference results;
7. Evaluate the performance metrics (3) and (4) for the test results;

8. Compare the values of the absolute differences and performance metrics calculated above with requirements on, respectively, the absolute and relative numerical performance of the test software.

Acknowledgements

The EMRP is jointly funded by the EMRP participating countries within EURAMET and the European Union.

References

[1] Harris, P. M. 2013 Computational Aim 'en/D/0/000016' for Principal Components Analysis (PCA) *TraCIM Computational Aims Database*