Memory Efficient Principal Component Analysis for the Dimensionality Reduction of Large Mass Spectrometry Imaging Datasets

Race, Alan; Steven, Rory; Palmer, Andrew; Styles, Iain; Bunch, Josephine

DOI:
10.1021/ac302528v

Document Version
Peer reviewed version

Citation for published version (Harvard):
Race, A, Steven, R, Palmer, A, Styles, I & Bunch, J 2013, 'Memory Efficient Principal Component Analysis for the Dimensionality Reduction of Large Mass Spectrometry Imaging Datasets', *Analytical Chemistry*, vol. 85, no. 6, pp. 3071-3078. https://doi.org/10.1021/ac302528v

Supporting information for: Memory Efficient Principal Component Analysis for the Dimensionality Reduction of Large Mass Spectrometry Imaging Datasets

Alan M. Race,[†,‡,¶] Rory T. Steven,[†,¶] Andrew D. Palmer,[†,‡,¶] Iain B. Styles,[*,‡] and Josephine Bunch[*,†,¶]

Centre for Physical Sciences of Imaging in the Biomedical Sciences, School of Chemistry, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom., School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom., and School of Chemistry, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.

E-mail: i.b.styles@cs.bham.ac.uk; j.bunch@bham.ac.uk

Abstract

Here we provide all mass spectrometry imaging experimental details, a detailed description of the peak detection algorithm, a figure demonstrating a comparison between the proportion of an entire organ that can be processed with *princomp* and the proposed method and a figure to show the clustering results for k=2...10. We also provide a MATLAB implementation of the algorithm.

[*]To whom correspondence should be addressed
[†]PSIBS, University of Birmingham
[‡]School of Computer Science, University of Birmingham
[¶]School of Chemistry, University of Birmingham

Materials

Methanol (HPLC grade) used in preparation of matrix and analyte solutions was purchased from Fisher Scientific (Leicestershire, UK). Trifluoroacetic acid (TFA, 99% purity) and MALDI matrices α-cyano-4-hydroxycinnamic acid (CHCA) and 4-nitroaniline (PNA) were purchased from Sigma Aldrich (Dorset, UK).

Single rat brain image

MALDI MS imaging of formalin fixed rat brain was recently described.[1] In summary, a formalin fixed rat brain was frozen and bisected along the midline. 12 μm sagital sections were thaw mounted onto stainless steel MALDI target plates and coated in 5 mgmL^{-1} CHCA prepared in 80% methanol (0.1% trifluoracetic acid) using an automated matrix deposition system (TM sprayer from HTX Technologies, NC, U.S.A.). Sections were coated using 8 deposition cycles at 150 °C, 10 psi, a flow rate of 0.25 mL/min with a stage velocity of 500 mm/min. Positive ion MALDI MS imaging was carried out on a QqTOF (Qstar Elite) mass spectrometer (Applied Biosystems, Foster City, U.S.A.), using an Nd:YAG (355 nm) laser operated at 20% available power (2.1 μJ) and a repetition rate of 500 Hz. Sequential spectra were acquired at a resolution of 100 μm^2 using the "dynamic pixel" setting (oMALDI, 5.1). External calibration was performed using Analyst QS 1.1.

3-D mouse brain image

Mouse brain samples were stored at 193 K until sectioned. Beginning within 2 mm parallel to the sagittal midline and working toward the midline, 14 μm serial sections were collected (Leica CM1810) and thaw mounted onto MALDI imaging plates. The serial sections were thaw mounted onto a single standard stainless steel imaging plate (Applied Biosystems). Samples were coated in matrix (PNA) solution (20 mgmL^{-1} in 80% CH_3OH, 0.1 % TFA) using an artist airbrush (Draper

Air Tools Airbrush Kit (Hampshire, UK)) propelled by dry N_2. MALDI TOF MSI analysis was carried out on a QSTAR XL QqTOF instrument using Analyst QS 1.1 with oMALDI server 5.1 (Applied Biosystems). An Nd:YVO_4 (Elforlight: SPOT-10-100-355) DPSS laser (355 nm), was triggered by a Thurlby Thandar Instruments (Huntingdon, Cambridgeshire) TGP110 10MHz Pulse Generator. All imaging data were acquired at a pulse energy of 12 mJ and rep rate of 5kHz. All MALDI MS images collected in positive ion mode using a acquisition at a speed of 1 mms^{-1}, with pixel dimensions of 100 μm. External calibration was performed using Analyst QS 1.1. Two sections were excluded from further analysis due to a tissue tear which affected the registration significantly.

Second Derivative Gradient Peak Detection

Where the first derivative of a spectrum crosses the y-axis ($y = 0$), the spectrum trace has altered from increasing to decreasing, or vice-versa indicating the presence of a maxima (peak) or minima (trough). To distinguish between a peak and a trough, the second derivative of the spectrum is calculated and if the value of the second derivative at the peak/trough *m/z* bin is negative then it is a peak, however if it is positive then it is a trough. Only peaks were retained; troughs introduced through the Savitzky-Golay smoothing were discarded.

References

(1) Carter, C.; McLeod, C.; Bunch, J. *Journal of the American Society for Mass Spectrometry* **2011**, *22*, 1991–1998.

(2) Fonville, J. M.; Carter, C.; Cloarec, O.; Nicholson, J. K.; Lindon, J. C.; Bunch, J.; Holmes, E. *Analytical Chemistry* **2012**, *84*, 1310–1319.

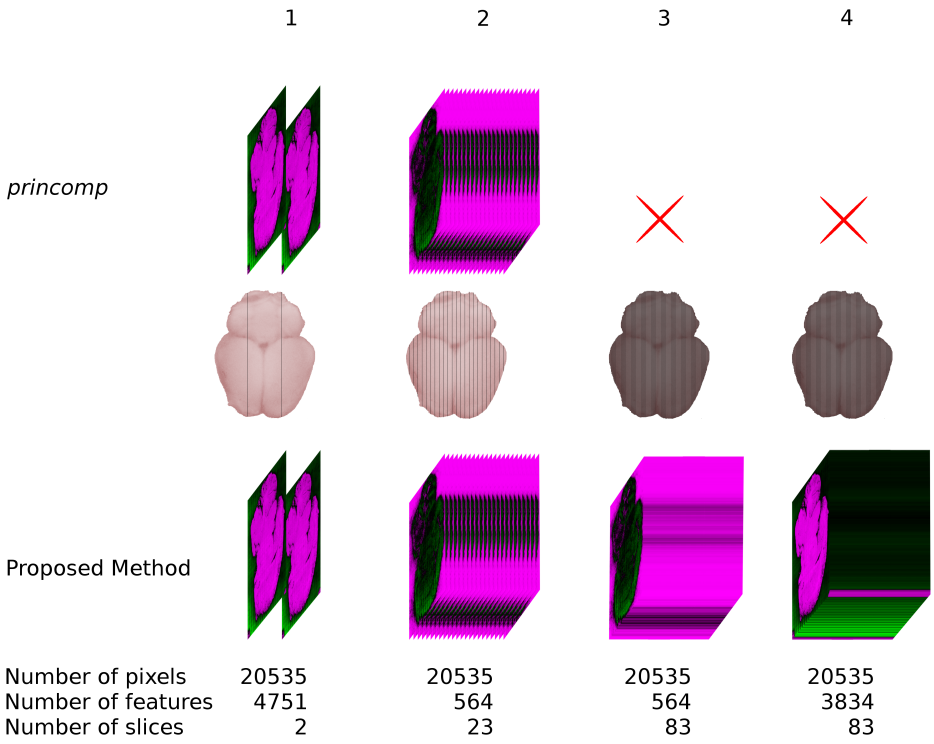

Figure 1: Simulated 3D MALDI MSI data of a 2 cm x 1 cm rat brain through repetition of a single 12 μm section image and the corresponding first principal component score image when using *princomp* (top) and the proposed method (bottom). Column 1 shows the maximum number of sections (2 sections) that could be retained if the data was binned at 0.2 *m/z* (4751 bins), the standard bin width used in BIOMAP (Novartis), and then analysed with *princomp* with 8 GB RAM. Distance between sections would be 325 μm. Column 2 shows the maximum number of sections (23 sections) that could be retained if informative peaks were extracted[2] (564 extracted *m/z* bins) prior to PCA using *princomp*. Distance between sections would be 30 μm. Column 3 shows the PCA results if the entire brain was sectioned and analysed (83 sections) and informative peaks were extracted.[2] The red cross indicates *princomp* failed due to memory limitations. Column 4 shows the PCA results if the entire brain was sectioned and analysed (83 sections) and all detected peaks (3834 detected peaks) were retained when determined from the entire data set.

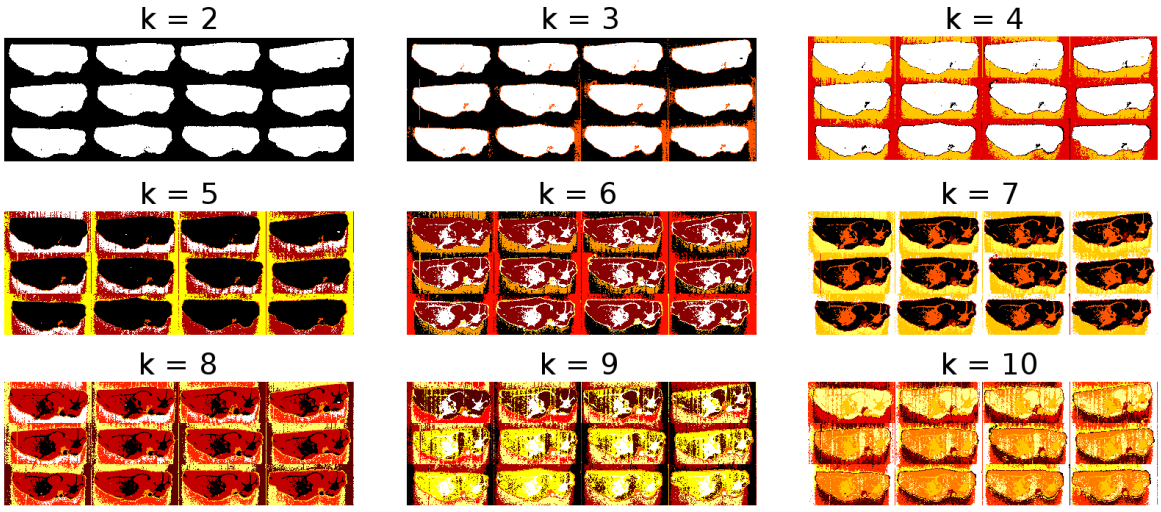

Figure 2: *k*-means applied to PC 1-40 scores of serial mouse brain sections with varying values for $k = 2...10$.