Benchmarks

Sequential projection pursuit principal component analysis – dealing with missing data associated with new -omics technologies

Bobbie-Jo M. Webb-Robertson^{1*}, Melissa M. Matzke¹, Thomas O. Metz², Jason E. McDermott¹, Hyunjoo Walker³, Karin D. Rodland⁴, Joel G. Pounds⁵, Katrina M. Waters¹

¹Computational Biology & Bioinformatics, Pacific Northwest National Laboratory, Richland, WA, USA, ²Omic Biological Applications, Pacific Northwest National Laboratory, Richland, WA, USA, ³Software Systems & Architecture, Pacific Northwest National Laboratory, Richland, WA, USA, ⁴Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA, and ⁵Systems Toxicology, Pacific Northwest National Laboratory, Richland, WA, USA

BioTechniques 54:165-168 (March 2013) doi 10.2144/000113978

Keywords: Principal component analysis, sequential projection pursuit, missing data, censored data, optimization

Supplementary material for this article is available at www.BioTechniques.com/article/113978.

Principal Component Analysis (PCA) is a common exploratory tool used to evaluate large complex data sets. The resulting lower-dimensional representations are often valuable for pattern visualization, clustering, or classification of the data. However, PCA cannot be applied directly to many -omics data sets generated by newer technologies such as label-free mass spectrometry due to large numbers of non-random missing values. Here we present a sequential projection pursuit PCA (sppPCA) method for defining principal components in the presence of missing data. Our results demonstrate that this approach generates robust and informative low-dimensional data representations compared to commonly used imputation approaches.

Principal Component Analysis (PCA) is a straight-forward mathematical method for reducing the dimensionality of large multivariate data sets to facilitate visualization and exploration of the data. Users can often identify patterns quickly in biological data related to specific phenotypes or other phenomena of interest, such as outlier values (1). A primary caveat, however, is that PCA requires a complete data matrix, forcing users to either filter their data to completeness or to impute missing values. Either approach changes the structure of the data set and, conse-

quently, the resulting inferences from the PCA.

Recently developed label-free mass spectrometry (MS) analysis of biological samples is a great example of a new data type for which existing PCA methodologies cannot be applied directly. Liquid chromatography (LC)-MS-based proteomics has a limited range of detection based on the mass spectrometer model and the biological material being analyzed. This means that although some of the data may be missing at random (MAR), many data points (for example, peptide abundances) are missing

due to non-random effects such as qualitative changes between biological groups or feature identification issues (2–7). Thus, these missing values are not missing at random (NMAR). There are practical PCA approaches that rely on methods to impute or model the missing values (8,9).

We introduce a modified form of sequential projection pursuit (SPP) PCA for evaluation of data sets containing both MAR and NMAR values, commonly found in left-censored data sets, without the need for data reduction or imputation (Online Methods). Projection pursuit is a generic term for methods that reveal the clustering structure of data in low-dimensional space through the optimization of a projection index (for example, entropy) (10). SPP is an algorithmic approach to perform the optimization task of project pursuit in a sequential manner to reduce computational complexity (11,12). Furthermore, SPP can be combined with PCA by simply defining the projection index as “variance”. Thus, our sppPCA approach uses SPP to optimize variance by ignoring missing values in the variance computation when estimating the first n principal components. Supplemental Figure 1 shows the overall algorithmic approach for this optimization. The sppPCA algorithm is available as a free executable program at www.biopilot.org/docs/Software/sppPCA.php; it was written in MatLab version R2011a and implemented through a Java graphical user interface (GUI). A protocol for running the program is available in the Supplementary Material.

An existing LC-MS mouse proteomics data set from an experiment evaluating the effect of diet induced obesity [Regular weight (RW) or obese (OB)] and inhaled endotoxin [controls (SC) or lipopolysaccharide (LPS)] was used to evaluate sppPCA in comparison to common imputation approaches (13). This data set contains information on 4,803 peptides across 32 mice (8 for each of the 4 factor combinations) of which 1,745 peptides were observed in all 32 mice. We first evaluate the full data set using PCA linear discriminant analysis (PCA-DA) on the first two principal components with (i) sppPCA, (ii) limit of detection imputation (1/2minPCA), defined as 1/2 of the minimum observed peptide value and (iii) regularized expectation maximization (14,15) imputation (remPCA). These three methods are evaluated in terms of classification accuracy using repeated 10-fold cross-validation and compared with PCA of the data subset of complete peptides. We then utilize a subset of the complete data (926 peptides that have a p-value greater than 0.095 by Kruskal-Wallis univariate statistics) as a starting point for a simulation-

Scientific Imaging Redefined.

xSCCELL Digital Scientific Camera

Features:

1000 fps at Full 1024x1024 Resolution
15µm² pixel size for Low Light Imaging
~65% Peak Quantum Efficiency
Read Noise <2 e⁻ at Full Resolution
Cooling to -30° for Low Dark Current

Powered by:

InXite
CMOS Redefined.

InXite™ is a proprietary CMOS-based technology which brings an entirely new level of speed and clarity to scientific imaging. InXite™ is available exclusively from PHOTONIS.

Applications:

- > Super Resolution Imaging
- > Spinning Disk Confocal Microscopy
- > Fluorescence Imaging
- > High Throughput Screening
- > Optical Computing Tomography

For those who ^{need} ~~want~~ clear, visible results.

The xSCCELL Digital Scientific Camera by PHOTONIS is now available in three unique models to suit your specific application.

Call or visit our web site to learn more details about the xSCCELL Digital Scientific Camera or to place your order today.

xSCCELL
Digital Scientific Camera

For more details:

visit www.xsccellcamera.com
email imaging@photonis.com
or call +1 215 428 0501

PHOTONIS

Figure 1. Comparison of the PCA-DA results of the real LC-MS experimental data. Differences in classification accuracy between PCA methods can be observed in the boxplot of classification accuracies obtained from 100 repetitions of 10-fold cross-validation on the score vectors associated with the top two principal components. Inset into the boxplot are comparative 95% confidence intervals used for a Tukey multiple test comparison. There is significant evidence (p-value < 0.05) that the average classification accuracy of sppPCA > remPCA > Complete > 1/2minPCA.

Figure 2. Trellis graph of PCA-DA classification accuracies in the context of left-censored and missing at random data. Each imputation approach, remPCA and 1/2minPCA, were compared with sppPCA using a paired *t*-test at each left censored and MAR level. A (+) and (*) indicated that sppPCA yields significantly larger accuracies than remPCA and 1/2minPCA, respectively, at a p-value threshold of 0.05. The (++) and (**) are significant after a Bonferroni multiple testing correction. Lastly the (-) indicates the few scenarios where remPCA was greater than sppPCA at a p-value threshold of 0.05. There were no circumstances where remPCA was greater than sppPCA after a Bonferroni correction.

based study evaluating the effects of left-censored and MAR values on the PCA-DA results. Data sets were constructed from the complete LC-MS data that contained a combination of left-censored and MAR data with percentages ranging from 0% to 25% in increments of 5%.

Figure 1 shows the classification accuracy in the four groups based on 100 repetitions of 10-fold cross-validation. The low-dimensional scores based on both remPCA and sppPCA perform better than classification using only the subset containing complete values, but 1/2minPCA results in poorer accuracy on

Table 1. Significance of factors based on ANOVA.

Factor Name	Factor Levels	P-value
PCA Method (PM)	1/2minPCA remPCA sppPCA	1.7e-49
Left Censored (LC)	0, 5, 10, 15, 20, 25 (%)	2.4e-54
Missing at Random (MAR)	0, 5, 10, 15, 20, 25 (%)	2.6e-3
PM x LC x MAR		3.3e-7
Replication	Ten repeated simulations	0.39

average. This is due to 1/2minPCA producing components that separate based on missing data rather than biological effects due to changes in variance structure (Supplemental Figure 2). Analysis of Variance (ANOVA) demonstrates that there is a difference in accuracy based on the first two components (p -value < $8e-70$). A multiple comparison procedure (Tukey's test) demonstrates that in fact all four methods are significantly different at a p -value less than 0.05 (insert within Figure 1) and that sppPCA results in significantly larger classification accuracy.

Figure 2 is a trellis graph of the simulation study results for various levels of left-censored and MAR data (points represent the 10 repetitions of the full simulation). Symbols on the bottom right corner (+, ++, -, *, **) represent a significant difference based on a paired t -test between each imputation approach and sppPCA. In general, sppPCA shows a trend of significant improvement over the other two approaches as the percent of left-censored data increases. Comparing classification accuracy over all levels of missing data, we found that sppPCA was significantly larger than both 1/2minPCA and remPCA with p -values of $3.9e-8$ and $4.1e-4$, respectively. Table 1 gives a global

comparison using ANOVA to compare classification accuracy values based on the factors (i) PCA method (PM), (ii) percent left-censored, (iii) percent MAR, (iv) the interaction of the first three factors, and (v) simulation replication. All factors are significant except for simulation replication. The highly significant results due to PCA methodology and the percentage of left-censored data are visually supported by Figure 2. As the percentage of left-censored data increases, the impact of the imputation approach (1/2min or rem) on the PCA results increases. The effect of factor MAR is significant, but not to the degree of the PCA method or left-censored data.

In conclusion, the sppPCA method presented here allows researchers to perform PCA on new -omics data sets containing NMAR data. The results of the low-dimensional projections of the data are not skewed by inaccurate estimates of variance, which are often introduced by imputation. In these examples, the limit of detection imputation approach (1/2minPCA) reduces the accuracy of classification by PCA-DA because the estimates of variance are skewed by this low value insertion. Better results were produced by remPCA compared to 1/2minPCA and

limit of detection; however, sppPCA performs best in respect to classification accuracy on both the real data and the simulation study. The largest gains in accuracy are associated with larger amounts left-censored data, which commonly occur in MS-based proteomics and metabolomics studies. The primary disadvantage of sppPCA is computation time, which depends on the amount of missing data and the number of samples and variables; thus, sppPCA will take considerably longer to obtain the principal component estimates than remPCA. Analysis of the data set used in this study with our algorithm will require several hours to complete on a typical desktop computer. However, since the effects of imputation will generally be unknown (9) and PCA typically does not require replicate computations, we advocate that researchers evaluating -omics data sets with large amounts of left-censored data utilize data analysis methods that do not require imputation whenever possible.

Acknowledgments

Computational work was supported by the National Institutes of Health (NIH) through grant 1R0111GM084892 (B.J.W) and the Clinical Proteomics Tumor Analysis Consortium (CA160019) (K.D.R.). The metabolomics example data in the software were generated under NIH grant DK070146 (T.O.M) and the proteomics data were generated under NIH grant U54-016015 (J.G.P.). Metabolomics and proteomics data were collected and processed in the Environmental Molecular Sciences Laboratory (EMSL). EMSL is a national scientific user facility supported by the Department of Energy. All work was performed at Pacific Northwest National Laboratory (PNNL), which is a multiprogram national laboratory

X-MAN™ Isogenic Cell Lines

Genetically Defined | Quality Controlled | Confidence Inspiring

Horizon Discovery's X-MAN™ Isogenic Cell Lines are genetically matched and functionally validated cell line pairs that differ only by a precise genomic modification. These predictive patient models allow for the definitive comparison of mutant vs. normal biology for highly relevant experimental results.

horizon
precision genome editing
www.horizondiscovery.com

operated by the Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06-76RL01830.

Competing interests

The authors declare no competing interests.

References

1. Ringnér, M. 2008. What is principal component analysis? *Nat. Biotechnol.* 26:303-304.
2. Wang, H., Y. Fu, R. Sun, S. He, R. Zeng, and W. Gao. 2006. An SVM scorer for more sensitive and reliable peptide identification via tandem mass spectrometry. *Pac. Symp. Biocomput.* 303-314.
3. Schlatter, D.M., J.E. Dazard, M. Dharsee, R.M. Ewing, S. Ilchenko, I. Steward, G. Christ, and M.R. Chance. 2009. Urinary protein profiles in a rat model for diabetic complications. *Mol. Cell. Proteomics* 8:2145-2158.
4. Dakna, M., K. Harris, A. Kalousis, S. Carpenter, W. Kolch, J.P. Schanstra, M. Haubitz, A. Vlahou, et al. 2010. Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinformatics* 11:594.
5. Tuli, L., T.H. Tsai, R.S. Varghese, J.F. Xiao, A. Cheema, and H.W. Ransom. 2012. Using a spike-in experiment to evaluate analysis of LC-MS data. *Proteome Sci.* 10:13.
6. Webb-Robertson, B.J., L.A. McCue, K.M. Waters, M.M. Matzke, J.M. Jacobs, T.O. Metz, S.M. Varnum, and J.G. Pounds. 2010. Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. *J. Proteome Res.* 9:5748-5756.
7. Webb-Robertson, B.J., W.R. Cannon, C.S. Oehmen, A.R. Shah, V. Gurumoorathi, M.S. Lipton, and K.M. Waters. 2010. A support vector machine model for the prediction of proteotypic peptides for accurate mass and time proteomics. *Bioinformatics* 26:1677-1683.
8. Ilin, A. and T. Raiko. 2010. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.* 11:1957-2000.
9. García-Laencia, P.J., J. Sancho-Gómez, and A.R. Figueiras-Vidal. 2009. Pattern classification with missing data: a review. *Neural Comput. Appl.* 19:263-282.
10. Friedman, J. and J.W. Tukey. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Tran Comput C-23*: 881-890.
11. Guo, Q., F. Questier, D.L. Massart, C. Boucon, and S. de Jong. 2000. Sequential projection pursuit using genetic algorithms for data mining of analytical data. *Anal. Chem.* 72:2846-2855.
12. Webb-Robertson, B.J., K.H. Jarman, S.D. Harvey, C. Posse, and B.W. Wright. 2005. An improved optimization algorithm and a Bayes factor termination criterion for sequential projection pursuit. *Chemom. Intell. Lab. Syst.* 77:149-160.
13. Tilton, S.C., K.M. Waters, N.J. Karin, B.J. Webb-Robertson, R.C. Zangar, K.M. Lee, D.J. Bigelow, J.G. Pounds, and R.A. Corley. Diet-induced obesity reprograms the inflammatory response of the marine lung to inhaled endotoxin. *Toxicol. Appl. Pharmacol.* 267: 137-148.
14. Little, R.J. and D.B. Rubin. 2002. *Statistical analysis with missing data*. Wiley-Interscience, Hoboken, NJ.
15. Schneider, T. 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* 14:853-871.

Received 19 September 2012; accepted 28 January 2013.

Address correspondence to Bobbie-Jo Webb-Robertson, Computational Biology & Bioinformatics, Pacific Northwest National Laboratory, Richland, WA, USA. E-mail: bj@pnnl.gov.

To purchase reprints of this article, contact: biotechniques@fosterprinting.com

Unlock your assay with custom aptamers.

For more information and demonstrated successes with our aptamers visit us at www.basepairbio.com

Anxious for antibodies?

- Aptamers are **more stable** and **more cost effective** than antibodies
- Comparable **affinity** and **specificity**
- Proprietary methods can generate custom **validated** nucleic acid ligands in as few as **three weeks**
- **Experience** with a variety of difficult targets, **success** where antibodies have failed
- Complete **intellectual property** transfer possible
- **Free** aptamer development available through NCI grant, contact us for more details

Base Pair Biotechnologies, Inc.
8058 El Rio St. • Houston, TX 77054
(978) 609-4199 • info@basepairbio.com