Bioinformatics Advance Access published June 6, 2016

R.JIVE for exploration of multi-source molecular data

Michael J. O'Connell [1] and Eric F. Lock [1*]

[1]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

Associate Editor: Dr. Ziv Bar-Joseph

ABSTRACT

Summary: The integrative analysis of multiple high-throughput data sources that are available for a common sample set is an increasingly common goal in biomedical research. JIVE is a tool for exploratory dimension reduction that decomposes a multi-source dataset into three terms: a low-rank approximation capturing joint variation across sources, low-rank approximations for structured variation individual to each source, and residual noise. JIVE has been used to explore multi-source data for a variety of application areas, but its accessibility was previously limited. We introduce **R.JIVE**, an intuitive R package to perform JIVE and visualize the results. We discuss several improvements and extensions of the JIVE methodology that are included. We illustrate the package with an application to multi-source breast tumor data from The Cancer Genome Atlas.

Availability: **R.JIVE** is available via CRAN under the GPLv3 license: https://cran.r-project.org/web/packages/r.jive/.

Contact: oconn725@umn.edu; elock@umn.edu

1 INTRODUCTION

In biomedical research a growing number of platforms and technologies are used to assess diverse but related information. This has motivated a number of methods for the integrative analysis of *multi-source* data, wherein multiple different sources of 'omics' data are available for a common sample set.

A good guiding principle for analyzing multi-source data is to simultaneously model features that are shared across multiple sources and features that are specific to a particular source. A number of recent methods have adopted this strategy, extending well-established techniques such as partial least squares (Löfstedt and Trygg, 2011), canonical correlation analysis (Zhou *et al.*, 2015), non-parametric Bayesian modeling (Ray *et al.*, 2014), non-negative factorization (Yang and Michailidis, 2016), and simultaneous components analysis (Schouteden *et al.*, 2014).

The joint and individual variation explained (JIVE) method (Lock *et al.*, 2013) was developed as a multi-source extension of principal components analysis (PCA). JIVE quantifies the amount of joint (shared) variation between data sources, reduces dimensionality, and allows for visual exploration of joint and individual (source-specific) structure. JIVE can also be used as a processing step prior to the application of other methods, such as clustering techniques (Hellton and Thoresen, 2016). JIVE was designed for the analysis of biomedical data from multiple technologies, but has been used for other diverse applications, such as the analysis of data that were processed using different computational pipelines (Kuligowski *et al.*, 2015) and the analysis of rail commute patterns at different times of day (Jere *et al.*, 2014).

Previously, only spartan Matlab code was available to perform JIVE. In this note we introduce the **R.JIVE** package, for which our intentions are threefold: (1) to improve the accessibility of this method among the bioinformatics community, (2) to implement important extensions and improvements of the JIVE method, and (3) to allow for quick and flexible visualization of JIVE results.

2 FEATURES AND IMPLEMENTATION

2.1 The JIVE method

JIVE decomposes a multi-source dataset into three components: (1) an approximation of rank r capturing joint variation across sources, (2) approximations of rank r_i for structured variation individual to each source i, and (3) residual noise. For dimension reduction and interpretation it is helpful to consider the low-rank approximations in factorized (or "point cloud") form, analogous to PCA. Joint structure corresponds to an r-dimensional sample subspace revealing patterns that explain substantial variability across multiple sources, whereas individual structure corresponds to an r_i-dimensional sample subspace revealing patterns that explain substantial variation in one source but not others.

With given ranks, the JIVE algorithm iteratively estimates joint and individual structure to minimize the overall sum of squared residuals. Simultaneous estimation of individual structure allows the underlying joint structure to be captured more accurately, and vice-versa. A permutation-based approach may be used to specify the ranks, which is important to accurately quantify joint and individual structure. The joint and individual structures are assumed to be orthogonal, which makes the decomposition identifiable.

2.2 Improvements and extensions

In this package, we have made some important additions to the original algorithm:

- Missing data is handled in a straightforward way via SVD using the package **SpatioTemporal** (Lindstrom *et al.*, 2013).
- Ranks may be selected via a BIC-motivated approach inspired by Jere *et al.* (2014), in addition to the permutation approach. Permutation-selected ranks are generally more accurate, but BIC selection can be much faster for certain datasets.
- Orthogonality may be enforced between the estimated individual structures, which improves robustness of the estimates to rank misspecification.

*to whom correspondence should be addressed

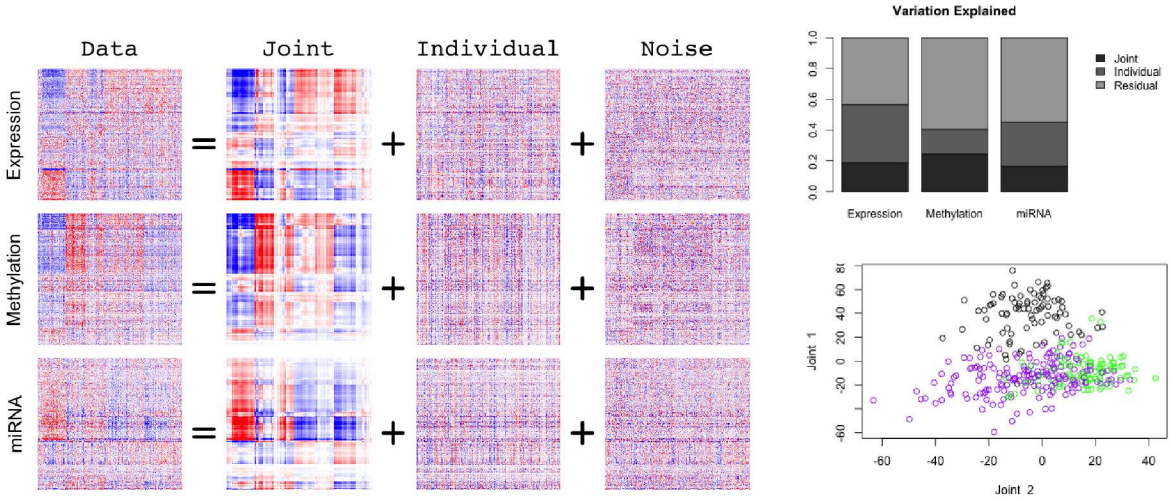

Fig. 1: Application of JIVE to BRCA data. The left panel shows heatmaps of JIVE estimates, with row and column ordering determined by joint structure. The top right panel shows the decomposition of variation for each data source. The bottom right panel shows the two principal components of joint structure, colored by previously determined integrative clusters; cluster 1 is colored black, cluster 2 is purple, and cluster 3 is green.

See the Supplementary Material for further details.

2.3 Package details

The `jive(...)` function takes a multi-source dataset and several options as input, and returns an object of class `jive` with the estimated joint and individual structure. Several functions are provided to summarize and generate quick publication-quality visualizations of the results in the form of a barchart, heatmaps, or PCA plots.

3 EXAMPLE: TCGA BRCA DATA

We illustrate **`R.JIVE`** using publicly available multi-source genomic data for 348 breast cancer (BRCA) tumor samples from The Cancer Genome Atlas data freeze (Cancer Genome Atlas Network, 2012). We apply JIVE to the mRNA expression, DNA methylation, and miRNA data, processed as previously described (Lock and Dunson, 2013). Permutation testing identifies rank 2 joint structure, rank 20 structure individual to expression, rank 12 structure individual to methylation, and rank 18 structure individual to miRNA. Visualizations of the resulting decomposition are shown in Figure 1. Heatmaps of the estimated joint structure show joint patterns shared by the three data sources. The point cloud view of joint structure shows that these patterns largely correspond to three BRCA sample clusters, which were previously identified using integrative clustering as present across these and other genomic sources (Lock and Dunson, 2013). Specifically, one pattern distinguishes Basal-like tumor samples (cluster 1) from other samples; among the remaining samples a subgroup of Luminal A tumors with a low fraction of genomic alteration and improved clinical prognosis (cluster 2) is distinguished.

A vignette to reproduce this analysis, with additional visualizations and interpretation of the results, is available within the R package and as Supplementary Material at *Bioinformatics* online.

Funding: National Institutes of Health (ULI RR033183 & KL2 RR0333182).

Conflict of Interest: none declared.

REFERENCES

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

Hellton, K. and Thoresen, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*. Advance online publication.

Jere, S., Dauwels, J., Asif, M. T., Vie, N. M., Cichocki, A., and Jaillet, P. (2014). Extracting commuting patterns in railway networks through matrix decompositions. In *13th IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 541–546. IEEE.

Kuligowski, J., Pérez-Guaita, D., Sánchez-Illana, Á., León-González, Z., de la Guardia, M., Vento, M., Lock, E. F., and Quintás, G. (2015). Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE). *Analyst*, **140**(13), 4521–4529.

Lindstrom, J., Szpiro, A., Sampson, P. D., Bergen, S., and Oron, A. P. (2013). *SpatioTemporal: Spatio-Temporal Model Estimation*. R package version 1.1.7.

Lock, E., Hoadley, K., Marron, J., and Nobel, A. (2013). Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, **7**(1), 523–542.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, **29**(20), 2610–2616.

Löfstedt, T. and Trygg, J. (2011). OnPLS – a novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, **25**(8), 441–455.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, **30**(10), 1370–1376.

Schouteden, M., Van Deun, K., Wilderjans, T. F., and Van Mechelen, I. (2014). Performing disco-sca to search for distinctive and common information in linked data. *Behavior research methods*, **46**(2), 576–587.

Yang, Z. and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, **32**(1), 1–8.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2015). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Trans Neural Netw Learn Syst*. Advance online publication.