An ExPosition of Multivariate Analysis with the Singular Value Decomposition in R

Derek Beaton[a,*], Cherise R. Chin Fatt[a], Hervé Abdi[a,*]

[a]*School of Behavioral and Brain Sciences, University of Texas at Dallas, MS: GR4.1, 800 West Campbell Road, Richardson, TX 75080–3021, USA*

Abstract

ExPosition is a new comprehensive `R` package providing crisp graphics and implementing multivariate analysis methods based on the singular value decomposition (SVD). The core techniques implemented in *ExPosition* are: principal components analysis, (metric) multidimensional scaling, correspondence analysis, and several of their recent extensions such as barycentric discriminant analyses (*e.g.,* discriminant correspondence analysis), multitable analyses (*e.g.,*multiple factor analysis, STATIS, and DISTATIS), and non-parametric resampling techniques (*e.g.,* permutation and bootstrap). Several examples highlight the major differences between *ExPosition* and similar packages. Finally, the future directions of *ExPosition* are discussed.

Keywords: Singular value decomposition, `R`, principal components analysis, correspondence analysis, bootstrap, partial least squares

`R` code for examples are found in Appendix A and Appendix B. Release packages can be found on CRAN at `http://cran.r-project.org/web/packages/ExPosition/`. Code from this article as well as release and development versions of the packages can be found at the authors' code repository: `http://code.google.com/p/exposition-family/`

1. An ExPosition

The singular value decomposition (SVD; Yanai et al., 2011) is an indispensable statistical technique used in many domains, such as neuroimaging (McIntosh and Mišić, 2013), complex systems (Tuncer et al., 2008), text reconstruction (Gomez and Moens, 2012), sensory analyses (Husson et al., 2007), and genetics (Liang, 2007). The SVD is so broadly used because it is the core of many multivariate statistical techniques (Lebart et al., 1984), including principal components analysis (PCA; Jolliffe, 2002; Abdi and Williams, 2010a), correspondence analysis (CA; Benzécri, 1973; Hill, 1974; Greenacre, 1984), (metric) multidimensional scaling (MDS; Torgerson, 1958; Borg, 2005), and partial least squares (PLS; Wold et al., 1984; Bookstein, 1994). In turn, these methods have many extensions such as multi-table analyses—*e.g.,* multiple factor analysis, or STATIS (Lavit et al., 1994; Abdi et al., 2012c, 2013b; Bécue-Bertaut and Pagès, 2008)—three-way distance analysis—*e.g.,* DISTATIS (Abdi et al., 2005)—and numerous variants of PLS (Esposito Vinzi and Russolillo, 2013; Abdi et al., 2013a) that span regression (Tenenhaus, 1998; Abdi, 2010), correlation (McIntosh and Lobaugh, 2004; Krishnan et al., 2011), and path-modeling (Tenenhaus et al., 2005). Finally, more recent extensions include generalized (Takane et al., 2006) and regularized methods (Le Floch et al., 2012).

*Corresponding authors
Email addresses: derekbeaton@utdallas.edu (Derek Beaton), herve@utdallas.edu (Hervé Abdi)

R (R Development Core Team, 2010) provides several interfaces to the SVD and its derivatives, but many of these tend to have diverse, and at times idiosyncratic, inputs and outputs and so a more unified package dedicated to the SVD could be useful to the R community. *ExPosition*—a portmanteau for *Ex*ploratory Analysis with the Singular Value Decom*Position*—provides for R a comprehensive set of SVD-based methods integrated into a common framework by sharing input and output structures. This suite of packages comprises: `ExPosition` for one table analyses (*e.g.,* PCA, CA, MDS), `TExPosition`, for two-table analyses (*e.g.,* barycentric discriminant analyses and PLS), and `MExPosition` for multi-table analyses (*e.g.,* MFA, STATIS, and DISTATIS). Also included in this this suite are `InPosition` and `TInPosition` that implement permutation, bootstrap, and cross-validation procedures.

This paper is outlined as follows: Section 2 presents the singular value decomposition and notations, Section 3 describes the differences between *ExPosition* and other packages, Section 4 show several examples that illustrate features not readily available in other packages, and finally, Section 5 elaborates on future directions. In addition, Appendix A and Appendix B include the code referenced in this paper. Throughout the paper the suite of packages is referred to as *ExPosition* or "the *ExPosition* family" and `ExPosition` as the package specific for one table analyses.

2. The Singular Value Decomposition

Matrices are in upper case bold (*i.e.,* $\mathbf{X}$), vectors in lowercase bold (*i.e.,* $\mathbf{x}$, and variables in lowercase italics (*i.e.,* x). The identity matrix is denoted $\mathbf{I}$. Matrices, vectors, or items labeled with $_{\mathrm{I}}$ are associated to rows and matrices, vectors, or items labeled with $_{\mathrm{J}}$ are associated to columns.

The SVD generalizes the eigenvalue decomposition (EVD) to rectangular tables (Abdi, 2007a; Greenacre, 1984; Lebart et al., 1984; Yanai et al., 2011; Jolliffe, 2002; Williams et al., 2010). Specifically, the SVD decomposes an I by J matrix, $\mathbf{X}$, into three matrices:

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ with } \mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I} \tag{1}$$

where $\mathbf{\Delta}$ is the L by L diagonal matrix of the singular values, (where L is the rank of $\mathbf{X}$), and $\mathbf{P}$ and $\mathbf{Q}$ are (respectively) the I by L and J by L orthonormal matrices of the left and right singular vectors. In the PCA tradition, $\mathbf{Q}$ is also called a *loading* matrix and a singular value with its corresponding pair of left and right singular vectors define a *component*.

Squared singular values, denoted $\lambda_\ell = \delta_\ell^2$, are the eigenvalues of both $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. Each eigenvalue expresses the variance of $\mathbf{X}$ extracted by the corresponding pair of left and right singular vectors. An eigenvalue divided by the sum of the eigenvalues gives the proportion of the total variance—denoted τ_ℓ for the ℓ-th component—explained by this eigenvalue,

it is computed as:

$$\tau_\ell = \frac{\lambda_\ell}{\sum \lambda_\ell}. \tag{2}$$

The sets of factor scores for rows (I items) and columns (J items) are computed as (see Eq. 1):

$$\mathbf{F}_\mathrm{I} = \mathbf{P}\boldsymbol{\Delta} \text{ and } \mathbf{F}_\mathrm{J} = \mathbf{Q}\boldsymbol{\Delta} \tag{3}$$

for the rows and columns of $\mathbf{X}$, respectively. Rewriting Eqs. 1 and 3 shows that factor scores can also be computed as a projection of the data matrix on the singular vectors:

$$\mathbf{F}_\mathrm{I} = \mathbf{P}\boldsymbol{\Delta} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{X}\mathbf{Q} \text{ and } \mathbf{F}_\mathrm{J} = \mathbf{Q}\boldsymbol{\Delta} = \mathbf{Q}\boldsymbol{\Delta}\mathbf{P}^\mathsf{T}\mathbf{P} = \mathbf{X}^\mathsf{T}\mathbf{P}. \tag{4}$$

Eq. 4 also indicates also how to compute factor scores (and loadings) for supplementary elements (a.k.a., "out of sample"; Gower, 1968, see also Section 4.2). There are also additional indices, derived from factor scores, whose function is to guide interpretation. These include *contributions*, *squared distances to the origin*, and *squared cosines*.

The contribution of an element to a component quantifies the importance of the element to the component. Contributions are computed as the ratio of an element's squared factor score by the component eigenvalues:

$$c_{\mathrm{I}i,\ell} = \frac{f^2_{\mathrm{I}\,i,\ell}}{\lambda_\ell} \text{ and } c_{\mathrm{J}j,\ell} = \frac{f^2_{\mathrm{J}\,j,\ell}}{\lambda_\ell}. \tag{5}$$

Next the squared distances to the the origin are computed as the sum of the squared distances of each element:

$$d_{\mathrm{I}i,\ell}^2 = \sum_{\ell} f_{\mathrm{I}i,\ell}^2 \text{ and } d_{\mathrm{J}j,\ell}^2 = \sum_{\ell} f_{\mathrm{J}j,\ell}^2. \tag{6}$$

Finally the squared cosines are the angles of elements from the origin, and indicate the quality of representation of a component to an element:

$$r_{\mathrm{I}i,\ell} = \frac{f_{\mathrm{I}i,\ell}^2}{d_{\mathrm{I}i,\ell}^2} \text{ and } r_{\mathrm{J}j,\ell} = \frac{f_{\mathrm{J}j,\ell}^2}{d_{\mathrm{J}j,\ell}^2}. \tag{7}$$

The *generalized* SVD (GSVD) provides a weighted least squares decomposition of $\mathbf{X}$ by incorporating constraints, on the rows and the columns (GSVD; Greenacre, 1984; Abdi and Williams, 2010a,b; Abdi, 2007a). These constraints—expressed by positive definite matrices—are, here, called *masses* for the rows and *weights* for the columns. Masses are denoted by an I by I (almost always) diagonal matrix denoted $\mathbf{M}$ and weights are denoted by a J by J (often diagonal) matrix denoted $\mathbf{W}$. The GSVD decomposes the matrix $\mathbf{X}$ into three matrices (compare with Eq. 1):

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ where } \mathbf{P}^{\mathsf{T}}\mathbf{M}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q} = \mathbf{I}. \tag{8}$$

The GSVD generalizes many linear multivariate techniques (given appropriate masses, weights and preprocessing of $\mathbf{X}$) such as CA and discriminant analysis. Note that with the "triplet notation"—which is a general frame-

work to formalize multivariate techniques (see, *e.g.,* Caillez and Pagès, 1976; Thioulouse, 2011; Escoufier, 2007)—the GSVD of $\mathbf{X}$ under the constraints of $\mathbf{M}$ and $\mathbf{W}$ is equivalent to the statistical analysis of the triplet $(\mathbf{X}, \mathbf{W}, \mathbf{M})$.

3. ExPosition: Rationale and Features

`R` has many native functions—*e.g.,* `svd()`, `princomp()`, and `cmdscale()`—and add-on packages —*e.g.,* `vegan` (Oksanen et al., 2013), `ca` (Nenadic and Greenacre, 2007), `FactoMineR` (Lê et al., 2008) and `ade4` (Dray and Dufour, 2007)—to perform the SVD and the related statistical techniques. The *ExPosition* family has a number of features not available (or not easily available) in current `R` packages: (1) a battery of inference tests (via permutation, bootstrap, and cross-validation) and (2) several specific SVD-based methods. Furthermore, *ExPosition* provides a unified framework for SVD-based techniques and therefore was designed around three main tenets: common notation, core analyses, and modularity. The following sections compare other `R` packages to the *ExPosition* family in order to illustrate *ExPosition*'s specific features.

3.1. Rationale and Design principles

There are three fundamental SVD methods: PCA for quantitative data, CA for contingency and categorical data, and MDS for dissimilarity and distance data. Each method has many extensions which typically rely on the same preprocessing pipelines as their respective core methods. Therefore, `ExPosition` contains three "core" functions: `corePCA`, `coreCA`, `coreMDS`

that (respectively) perform the baseline aspects (*e.g.,* preprocessing, masses, weights) of each "core" technique. Each core function is an interface to `pickSVD` and returns a comprehensive output (see Eqs. 3, 5, 6, and 7). While `corePCA` and `coreMDS` are fairly straightforward implementations of PCA and MDS, `coreCA` provides some important features not easily found in other packages (*e.g.,* Hellinger, see 3.2.1 for details).

Because all techniques pass through a generalized SVD function in `ExPosition`—*i.e.,* `pickSVD`—the output from *ExPosition* contains a common structure. The returned output is listed in Table 1. When the size of the data is very large (*i.e.,* when the analysis can be computationally expensive), `pickSVD` uses the EVD (see Abdi, 2007a, for SVD and EVD equivalence). `pickSVD` decomposes a matrix *after* it has passed through one of the `core*` methods. The `core*` methods in *ExPosition* provide more detailed output for the I row items and the J column items (in Table 2).

3.2. Modularity and Feature set

The *ExPosition* family is partitioned into multiple packages. These partitions serve two purposes: to identify the packages suitable for a given analysis and to afford development independence. Each partition serves a specific analytical concept: `ExPosition` for one-table analyses, `TExPosition` for two-table analyses, and `MExPosition` for multi-table methods. The inference packages (which include, *e.g.,* permutation and bootstrap) follow the same naming convention: `InPosition` and `TInPosition`.

Table 1: ExPosition output variables, associated to the SVD, common to all techniques. This table uses R's list notation, which includes a $ preceding a variable name.

SVD matrix or vector	Variable name	Description
$\mathbf{P}$	`$pdq$p`	Left singular vectors
$\mathbf{Q}$	`$pdq$q`	Right singular vectors
$\mathbf{\Delta}$	`$pdq$Dd`	Diagonal matrix of singular values
$\text{diag}\{\mathbf{\Delta}\}$	`$pdq$Dv`	Vector of singular values
$\text{diag}\{\mathbf{\Lambda}\}$	`$eigs`	Vector of eigen values
τ	`$t`	Vector of explained variances
$\mathbf{m}$	`$M`	Vector of masses (most techniques)
$\mathbf{w}$	`$W`	Vector of weights (most techniques)

3.2.1. Fixed-effects features

The function `coreCA` from `ExPosition` includes several distinct features such as symmetric vs. asymmetric plots (available in `ade4` and `ca`), eigenvalue corrections for MCA (available in `ca`), and Hellinger analysis (only available through MDS in `vegan` and `ape`).

`TExPosition` includes (barycentric) discriminant analyses and partial least squares methods. The partial least squares methods are derivatives of Tucker's inter-battery analysis (Tucker, 1958; Tenenhaus, 1998), also called Bookstein PLS, PLS-SVD or PLS correlation (Krishnan et al., 2011). There are two forms of PLS in `TExPosition`: (1) an approach for quantitative data (Bookstein, 1994), frequently used in neuroimaging (McIntosh et al., 1996; McIntosh and Lobaugh, 2004) and (2) a more recently developed approach

Table 2: ExPosition output variables common to all the `core*` methods.

I rows	Item	J columns
`$fi`	Factor Scores	`$fj`
`$di`	Squared Distances	`$dj`
`$ri`	Cosines	`$rj`
`$ci`	Contributions	`$cj`

for categorical data (Beaton et al., 2013). The discriminant methods in `TExPosition` are special cases of PLS correlation: barycentric discriminant analysis (BADA; Abdi et al., 2012a,b; St-Laurent et al., 2011; Buchsbaum et al., 2012) for quantitative data and discriminant correspondence analysis (DICA; Williams et al., 2010; Pinkham et al., 2012; Williams et al., 2012) for categorical or contingency data.

`MExPosition` is designed around the STATIS method. However, there are numerous implementations and extensions of STATIS, such as MFA, ANISO-STATIS, COVSTATIS, CANOSTATIS, and DISTATIS. As of now, `MExPosition` is the only package to provide an easy interface to all of the STATIS derivatives (see Abdi et al., 2012c).

3.2.2. prettyGraphs

The `prettyGraphs` package was designed especially to create "publication-ready" graphics for SVD-based techniques. All *ExPosition* packages depend on `prettyGraphs`. `prettyGraphs` includes standard visualizers (*e.g.,* com-

ponent maps, correlation plots) as well as additional visualizers not available in other packages (*e.g.,* contributions to the variance, bootstrap ratios). Further, `prettyGraphs` handles aspect ratio problems found in some multivariate analyses (as noted in Meyners et al., 2013). *ExPosition* provides interfaces to `prettyGraphs` (*e.g.,* `epGraphs`, `tepGraphs`) to allow users more control over visual output, without creating each graphic individually. Finally, `prettyGraphs` can visualize results from other packages (see Appendix A).

3.2.3. Permutation

Permutation tests in *ExPosition* are implemented *via* the "random-lambda" approach (see *Rnd-Lambda* in Peres-Neto et al., 2005) because it typically performs well, is conservative, and is computationally inexpensive. All these features are critical when analyzing "big data" sets such as those found, for example, in neuroimaging or genomics.

For all `*InPosition` methods, permutation tests evaluate the "significance" of components. However, it should be noted that other permutation methods (Dray, 2008; Josse and Husson, 2011) may provide better estimates for components selection. For all CA-based and discriminant methods, *ExPosition* tests overall (omnibus) inertia (sum of the eigenvalues). Finally, an R^2 test is performed for the discriminant techniques (BADA, DICA; Williams et al., 2010). Permutation tests similar to these are available in some SVD-based analysis packages, such as `ade4`, `FactoMineR`, and `permute` which can be used with `vegan`.

3.2.4. Bootstrap

The bootstrap method of resampling (Efron and Tibshirani, 1993; Chernick, 2008) is used for two inferential statistics: confidence intervals and bootstrap ratio statistics (a Student's t-like statistic; McIntosh and Lobaugh, 2004; Hesterberg, 2011). Bootstrap distributions are created by treating each bootstrap sample as supplementary data to the fixed-effects space.

Bootstrap ratios are performed for all methods to identify the variables that significantly contribute to the variance of a component. Under standard assumptions, these ratios are distributed as a Student's t and therefore a "significant" bootstrap ratio will need to have a magnitude larger than a critical value (*e.g.,* 1.96 for a large N corresponds to $\alpha = .05$). Additionally, for discriminant techniques, confidence (from bootstrap) and tolerance (fixed-effects) intervals are computed for the groups and displayed with peeled convex hulls (Greenacre, 2007). When two confidence intervals do not overlap, the corresponding groups are considered significantly different (Abdi et al., 2009). While some bootstrap methods are available in similar packages, these particular tests are only available in the *ExPosition* packages.

3.3. Leave one out

The *ExPosition* family includes leave-one-out (LOO) cross-validation for classification purposes (Williams et al., 2010). Each observation is, in turn, (1) left out, (2) predicted from out of sample, and then, (3) assigned to a group. While leave-one-out is available from `MADE4` and `FactoMineR`,

ExPosition uses LOO for classification estimates (*i.e.,* BADA, DICA).

4. Examples of ExPosition

Several brief examples of `ExPosition` are presented. Each example highlights (1) the specific features of *ExPosition* and, (2) how to interpret the results. Basic set up and code for each analysis are in Appendix B. All examples use an illustrative data set built into `ExPosition` called `beer.tasting.notes` which is an example of one person's personal tasting notes. `beer.tasting.notes` also includes supplementary data (*e.g.,* additional measures, design matrices). `R` code and *ExPosition* parameters are presented in `monotype font`.

First are illustrations of PCA and BADA (sometimes called between class analysis or mean centered PLSC; Baty et al., 2006; Krishnan et al., 2011). However, PCA and BADA are presented via `InPosition` and `TInPosition`, as they provide an extensive set of inferential tests unavailable elsewhere. Next, is an illustration of MCA with χ^2 vs. Hellinger analysis. Hellinger is an appropriate choice when χ^2 is too sensitive to population size (Rao, 1995b; Escofier, 1978). Finally, the `MExPosition` package—which provides an interface to many STATIS derivatives (Abdi et al., 2012c)—is illustrated `MExPosition` with DISTATIS: a STATIS generalization of MDS.

4.1. PCA *Inference Battery*

PCA is available in `ExPosition`, like in many other packages. However, `InPosition` provides two types of inferential analyses. The first are permu-

tation tests (see Section 3.2.3) to determine which, if any, components are significant. The second are bootstrap ratio tests of the measures. The data to illustrate PCA consist of a matrix of tasting notes of 16 flavors (columns) for 29 craft beers (rows) from the United States. Additionally, there is a design matrix (a.k.a. group coded, disjunctive coding) to indicate to which style each beer belongs (according to Alström and Alström, 2012).

In all `ExPosition` methods, data matrices are passed as `DATA`. Further, a design matrix (either a single vector or a dummy-coded matrix with the same rows as `DATA`) is passed as `DESIGN` and determines the specific colors assigned to each observation from `DATA` (*i.e.,* observations from the same group will have the same color when plotted). In this analysis, the data are centered (`center` = `TRUE`) but not scaled (`scale` = `FALSE`). Bootstrap ratios whose magnitude is larger than `crit.val` are considered significant. The default `crit.val` is equal to 2 (Abdi, 2007b, analogous to a t- or Z-score with an associated p value approximately equal to .05). `test.iters` permutation and bootstrap samples are computed (in the same loop for efficiency). See Appendix B for code and additional data details.

4.1.1. Interpretation

Many SVD-based techniques are visualized with component maps in which row or column factors scores are used as coordinates to plot the corresponding items. On these maps, distance between data points reflects their similarity. The dimensions can also be interpreted by looking at the items with large positive or negative loadings. In addition, permutation

tests provide p values that can used to identify the reliable dimensions.

Figure 1a. shows a component map of the row items (beers) colored by their style (automatically selected via `prettyGraphs`). The component labels display the percentage of explained variance and p-values per component. Components 1 and 2 are significant (from the permutation test) and explain 28.587% ($p < .001$) and 19.845% ($p < .001$) of the total variance, respectively. Figure 1a. suggests that beers with similar brewing styles cluster together. For example, all of the "saison-farmhouse" are on the right side of Component 1 (in orange). Note that in Figure 1a, beers are plotted with circles whose size reflect the beer contribution to the variance (*i.e.,* `$ci`) of the components used to draw the map. In PCA, column items (flavors) are in general plotted separately (by default). Figure 1b. indicates what flavors (1) are alike and (2) make these beers alike. For example, all the beers at the top of Component 2 (*e.g.,* Consecration, La Folie, and Marrón Acidifié) are sour beers (through barrel aging, wild yeast strains, and/or additional bacteria such as lactobacillus) and this is confirmed by the position of the column "sour" at the top of Component 2 (cf Figure 1b). By default, two plots for the variables are included for a PCA: (1) the plot in which the loadings serve as coordinates (Figure 1b) and the size of the dots reflect the contributions (*e.g.,*) importance of the variables for the dimensions used, and (2) a plot—called the circle of correlation plot—in which the correlation between the the factor scores and the variables are used as coordinates (Figure 1c). The last plot includes a unit circle because the sum of these

Table 3: Bootstrap ratios for the first two components of the PCA. Bold values indicate bootstrap ratios whose magnitude exceed 2 (*i.e.,* "significant").

	Component 1	Component 2
Alcoholic	−0.506	−0.148
Dark Fruit	**−3.902**	1.632
Citrus Fruit	**3.082**	1.252
Hoppy	**3.357**	**−2.238**
Floral	**3.035**	**−2.282**
Spicy	**2.345**	**3.403**
Herbal	**2.033**	−0.541
Malty	**−3.529**	**−2.05**
Toffee	**−2.764**	**−2.379**
Burnt	**−2.255**	0.383
Sweet	**−3.505**	−0.614
Sour	−0.022	**4.241**
Bitter	1.495	1.013
Astringent	**2.009**	**2.797**
Body	−0.496	−1.187
Linger	0.390	−0.060

squared correlations cannot exceed 1. The closer a variable is to the circle, the more "explained" by the dimensions the variable is.

Plotting items as a function of their contributed variance (`$ci` or bootstrap ratios) provide immediate visual information about the importance of items. This feature is available through the `prettyPlot` function in `prettyGraphs` package. Other visualizations for SVD-based analyses do not typically provide this feature. In Figures 1b. and c., the flavors (variables) are colored using their bootstrap ratios. Variables colored in

grey do not significantly contribute to either visualized component [*i.e.,* `abs`(bootstrap ratio) < `crit.val`]. Variables colored with purple significantly contribute to the horizontal axis (here: Component 1) and variables in green significantly contribute to the vertical axis (here: Component 2). Variables colored in red significantly contribute to both plotted components. In sum, Component 1 is defined as acidic vs. sweet (*e.g.,* "citrus fruit" vs. "dark fruit") whereas Component 2 is defined largely by "sour". Some items, such as "hoppy," contribute significantly to both components. The graphs suggest that beers in the lower right quadrant are characterized by "hoppy" and "floral" characteristics.

4.2. BADA *Inference Battery*

BADA is illustrated with the same data set as in Section 4.1 because there exists data and design matrices. Because BADA is a discriminant technique, there are more inference tests available than for plain PCA. The additional tests include: (1) classification accuracy, (2) omnibus effect (sum of eigenvalues), (3) bootstrap ratios and confidence intervals for groups, and finally, (4) a squared coefficient statistic (R^2), computed as the $\frac{\text{between-groups variance}}{\text{total variance}}$. This coefficient quantifies the quality of the assignments of the beers to their categories (Williams et al., 2010).

`TInPosition` uses permutation to generate distributions for (1) components (just as with PCA in `InPosition`), (2) omnibus inertia (sum of the eigenvalues), and 3) R^2. Bootstrap resampling generates distributions to create (1) bootstrap ratios for the measures (just as with PCA in `InPosition`)

and for the groups, and (2) to create confidence intervals around the groups. Finally, classification accuracies are computed for fixed-effects and for random effects (via leave-one-out).

4.2.1. Interpretation

Because BADA is a PCA-based technique, the graphical and numerical outputs are essentially the same as those of PCA with, however, a few important differences. First, BADA plots have both active and supplementary elements: the group averages are active rows (from the decomposed matrix) and the original observations (*e.g.,* the beers) are supplemental rows which are projected onto the component space.

The graphical output for BADA provides tolerance peeled hulls that envelope all or a given proportion the observations that belong to a group (Figure 2a.). Mean confidence intervals for the groups are also plotted with peeled hulls (see Figure 2b.). When group confidence intervals, on any (significant) components, do not overlap, groups can be considered significantly different. For example, Figure 2 shows that "Sour" and "Misc" are significantly different from each group. In contrast "Pale" and "Saison" do not differ from each other. In Figure 2a. groups and items are colored based on bootstrap ratios (just as in PCA): grey items do not contribute to either component, purple items contribute to Component 1, green items contribute to Component 2, and red items contribute to both components (See Table 4 for the bootstrap ratio values).

Furthermore, `TInPosition` performs three separate tests based on per-

Table 4: Bootstrap ratios for the first two components of the BADA. Bold values indicate bootstrap ratios whose magnitude exceed 2 (*i.e.,* "significant").

(a) Flavors

	Component 1	Component 2
Alcoholic	−0.189	0.847
Dark Fruit	**6.789**	0.007
Citrus Fruit	**−2.093**	**−2.374**
Hoppy	**−5.172**	0.15
Floral	**−3.323**	−0.055
Spicy	0.382	**−2.999**
Herbal	−1.391	−0.506
Malty	**2.944**	**5.103**
Toffee	**2.617**	**2.423**
Burnt	**2.122**	0.22
Sweet	1.818	1.836
Sour	**5.019**	**−7.641**
Bitter	−0.621	−1.131
Astringent	−0.963	**−2.203**
Body	−0.57	1.7
Linger	−0.173	−0.386

(b) Groups

	Component 1	Component 2
PALE	**−6.968**	0.565
SOUR	**4.734**	**−5.199**
SAISON	**−8.905**	**−3.138**
MISC	1.498	**4.152**

mutation resampling. After $1,000$ permutations, R^2 (reliability of assignment to groups) and omnibus inertia are significant ($R^2 = .610, p < .001$, $\sum \lambda_\ell = 0.390, p < .001$). These tests indicate that the assignment of individuals to groups (R^2) and the overall structure of the data ($\sum \lambda_\ell$) are not due to chance (*i.e.,* "are significant"). Additionally, Components 1 and 2 are significant, (56.457%, $p < .001$; 36.391%, $p < .001$, respectively) whereas Component 3 does not reach significance (7.151%, $p = .073$). Inference results are found in the `$Inference.Data` list in output from `TInPosition`. Finally, `TInPosition` provides output for leave one out estimates of classification. Classification accuracy for the fixed effects model is 82%, whereas the random effect model (assessed from the leave-one-out procedure) accuracy is 62% (Table 5).

4.3. Hellinger vs. χ^2

There are three substantial differences between the CA and MCA implementations of `ExPosition` versus those in other packages as `ExPosition` is currently the only package to offer together: (1) symmetric vs. asymmetric plots (available in `ade4` and `ca`), (2) eigenvalue corrections and adjustments (for MCA only; available in `ca`), and (3) χ^2 vs. Hellinger distance (only available through MDS in `vegan` and `ape`).

Because asymmetric factor scores (Abdi and Williams, 2010b; Greenacre, 2007; Escofier, 1978) and eigenvalue corrections (Benzécri, 1979; Greenacre, 2007) are well known amongst CA users, MCA is illustrated with the lesser known feature: χ^2 distance (the standard) vs. Hellinger distance (Rao,

Table 5: Classification and classification accuracy with (a) fixed and (b) random effects.

(a) Fixed (82%)

	PALE	SOUR	SAISON	MISC
PALE	**8**	0	0	1
SOUR	0	**5**	0	0
SAISON	1	1	**5**	2
MISC	0	0	0	**6**

(b) LOO-CV (62%)

	PALE	SOUR	SAISON	MISC
PALE	**4**	0	2	1
SOUR	0	**5**	0	0
SAISON	4	1	**3**	2
MISC	1	0	0	**6**

1995a,b; Escofier, 1978; Cuadras et al., 2006). The Hellinger distance was developed as an alternative for the standard χ^2 distance for CA-based methods to palliate CA's insensitivity to small marginal frequencies (Escofier, 1978; Rao, 1995b). MCA (χ^2 vs. Hellinger) is illustrated with the data used in the PCA and BADA examples. Data were recoded to be categorical ('LOW", "MidLOW", "MidHIGH", or "HIGH") within each column. See Appendix B for details.

4.3.1. Interpretation

Figures 3a. and b. show the χ^2 MCA analysis. Components 1 and 2 are largely driven by Astringent.LOW and Toffee.HIGH which occur only once, and 2 twice, respectively. The data illustrate the relevance of the choice of the Hellinger distance rather than the standard χ^2: MCA based on the χ^2 distance is very sensitive to outliers (Figure 3b.) whereas the analysis with the analysis with the Hellinger distance is not (Figure 3d.). With the Hellinger distance analysis, Chocolate.Bock and Chocolate.Stout are no longer outliers (Figures 3c) and share qualities that make them similar to other beers (Three.Philosophers). In both analyses, beers are grouped together in a meaningful fashion. For example, the Saisons are found in the lower right quadrants; malty and sweet beers are on the left side of the component map (Figures 3a. vs. c.).

4.4. DiSTATIS

`MExPosition` is a package designed for multi-table analyses based on multiple factor analysis and STATIS. `MExPosition` uniquely provides direct

interfaces (*i.e.,* functions) to many related techniques and specific derivatives of STATIS (*e.g.,* MFA, COVSTATIS, ANISOSTATIS, and DISTATIS). While some packages may include STATIS (*e.g.,* `ade4`) or MFA (*e.g.,* `FactoMineR`), DISTATIS (*i.e.,* `DistatisR`), no other package offers as many derivatives as `MExPosition`.

Prior analyses (particularly, PCA and MCA) indicate that, sometimes, beers of different styles cluster together. For example: Pliny the Elder (Imperial IPA) and Trade Winds (Tripel) or Endeavour (Imperial IPA) and Sisyphus (Barleywine). These relationships bring up a question: are there aspects of flavor that are not based entirely on style (*e.g.,* particular malts and hops), such as (1) in-house yeast strains and (2) water source? In this analysis, physical distances (in meters) between breweries are used as proxies of water source, yeast strains, and other geographically-sensitive factors. The `rjson` package (Couture-Beil, 2013) was used to retrieve distances between cities via Google Maps API (Google, Inc, 2013). A distance matrix was derived from `beer.tasting.notes` with the `dist` function. There are now two distance matrices that can be analyzed in two different ways: 1) separately with MDS or 2) together with DISTATIS. Figure 4a. shows the MDS analysis of flavors. This map is interpreted with the same rules as PCA (Figure 1). Figure 4b shows the MDS analysis of the physical distance between breweries. Either MDS alone provides partial information with respect to beer style or flavor perception.

DISTATIS can analyze both distance matrices simultaneously. Figures 5a.

shows that DISTATIS reveals some very interesting characteristics of the beers. First, saisons and sours, by comparison to the original analyses, are largely unaffected by physical distance. These styles appear to maintain their flavor properties regardless of location. Second, the remaining beers, across styles, are not as separable as saisons or sours. This suggests that some (standard) beer styles in fact are sensitive to regional factors (*e.g.,* water source).

5. Conclusions

This paper introduced a suite of SVD-based analysis packages for `R`, called *ExPosition*, that offers a simple and unified approach to SVD analyses through a set of core functions. While *ExPosition* offers a number of features unavailable elsewhere, there are still several future directions for the *ExPosition* family. First, because very large data sets are now more routine, an obvious step forward is to include faster decompositions. For example, a faster analysis could be achieved via an `R` interface to more efficient `C` libraries (Eddelbuettel and Sanderson, 2013). Next, `MExPosition` will include decompositions of each table based on "mixed-"data types (as in Lê et al., 2008; Bécue-Bertaut and Pagès, 2008). That is, if a user provides several contingency tables (CA), a nominal table (MCA), and several scaled tables (PCA), `MExPosition` will correctly normalize and decompose each table. Massive studies, such as ADNI (`http://www.adni-info.org`), collect a wide array of mixed data, and as such, methods like mixed data STATIS will become critically important. Additionally, `TExPosition` will

include all partial least squares correlation (PLSC) techniques (see, *e.g.,* the PLSC software [for neuroimaging] available *only* for Matlab[1]). Further, all available *ExPosition* methods will include multi-block projection (Williams et al., 2010; Abdi et al., 2012a,b). Finally, `InPosition` will (1) extend to `MExPosition` (*i.e.,* `MInPosition`), (2) include more inferential methods, such as split-half resampling (which provides estimates for prediction and reliability; Strother et al., 2002) and, (3) various permutation approaches (Peres-Neto et al., 2005). To note, there exist recent approaches that are more accurate for SVD-based techniques (Dray, 2008; Josse and Husson, 2011).

To conclude, *ExPosition* offers a very wide array of features for analyses: it is easily extendable through the core functions (see Appendix A) and implements many descriptive methods (*e.g.,*PCA, CA, MDS), their derivatives (*e.g.,*BADA, STATIS, and DISTATIS), extensive visualizers, and inferential tests (via permutation, bootstrap, and cross-validation). Currently, no other package for `R` offers such a comprehensive approach for SVD-based techniques.

6. Acknowledgments

Many thanks are due to the editor and associate editor of this journal, one anonymous reviewer, and to Stéphane Dray for their help and constructive comments on previous versions of this paper. Many people have

[1]by McIntosh, Chau, Lobaugh, & Chen available at http://www.rotman-baycrest.on.ca/index.php?section=84

been instrumental in the development and testing of *ExPosition* and for feedback on this paper. See the complete author list (`?ExPosition`) and acknowledgments [`acknowledgements()`] in `ExPosition`.

This appendix includes code either (1) to illustrate a feature or (2) required to run the examples.

Appendix A. Illustrations

This section provides illustrations of code to exhibit particular features of *ExPosition*.

Appendix A.1. PLS *Correlation Methods*

To illustrate the usefulness of modularity, an analysis core, and common notation, we present the code required to perform a PLSC (Tucker, 1958; McIntosh et al., 1996; Krishnan et al., 2011). In this example, we center and scale (sum of squares equal to 1) two data sets `X` and `Y`.

```
X <- expo.scale(beer.tasting.notes$sup.data[,1:2],scale="SS1",center=TRUE)
Y <- expo.scale(beer.tasting.notes$data,scale="SS1",center=TRUE)
```

Next, we call `corePCA()` instead of a plain SVD. We do this because `corePCA` provides a comprehensive set out of output that we would otherwise need to compute if we called just `svd()`.

```
pls.out <- corePCA(t(X) %*% Y)
```

Finally, we compute the latent variables (*i.e.,* the rows of the data are projected as supplementary elements):

```
Lx <- supplementalProjection(X,pls.out$fi,Dv=pls.out$pdq$Dv)
Ly <- supplementalProjection(Y,pls.out$fj,Dv=pls.out$pdq$Dv)
```

Appendix A.2. prettyGraphs beyond ExPosition

`prettyGraphs` is a package designed to create high-quality graphics for the *ExPosition* family. However, `prettyGraphs` can be used to visualize other data or results from other packages. The following code illustrates how to use `prettyPlot` from the `prettyGraphs` package to plots results obtained from analyses performed with the `ade4` and `FactoMineR` packages:

```
#for ade4
data(deug)
deug.dudi <- dudi.pca(deug$tab, center = deug$cent,
scale = FALSE, scan = FALSE)
inertia <- inertia.dudi(deug.dudi,row.inertia = T)$row.abs
prettyPlot(deug.dudi$li,
contributionCircles=TRUE,
contributions=inertia)

# for FactoMineR
data(decathlon)
res.pca <- PCA(decathlon, quanti.sup = 11:12, quali.sup=13,graph=FALSE)
prettyPlot(res.pca$ind$coord,
contributionCircles=TRUE,
contributions=res.pca$ind$contrib)
```

Appendix B. Required code

Here, we illustrate how to use a number of features across *ExPosition*. We use the same data set—built into `ExPosition`—across all examples. The data consist of 16 flavor notes (columns) collected on 29 craft beers (rows) brewed in the United States. Included is a design matrix (same constraints as the data), which is group coded (a.k.a. disjunctive coding). The design matrix reflects a particular style per beer (styles according to Alström and Alström, 2012).

Appendix B.1. PCA *Inference Battery*

The following code runs the example described in Section 4.1. `InPosition` is introduced with a simple and familiar example: PCA. In order to perform PCA with `InPosition`, we use the function `epPCA.inference.battery()`, which calls `epPCA()` in `ExPosition`. For this example, we will use the parameters `DATA`, `DESIGN`, `scale`, `make_design_nominal` and `test.iters`. Data are initialized as such:

```
these.rows <- which(rowSums(beer.tasting.notes$region.design[,-5])==1)
BEER <- beer.tasting.notes$data[these.rows,]
STYLES<-beer.tasting.notes$style.design[these.rows,]
```

PCA with inference test battery:

```
beer.taste.res.style <-
epPCA.inference.battery(DATA = BEER,
scale = FALSE,
```

```
DESIGN = STYLES,
make_design_nominal = FALSE,
test.iters = 1000)
```

Fixed effects and plotting data are found in `beer.taste.res.style$Fixed.Data`, and inference results are found in `beer.taste.res.style$Inference.Data`.

Appendix B.2. BADA *Inference Battery*

The following code runs the example in Section 4.2. With BADA, we aimed to investigate the properties of beers classified as "'pale," "saison,", "sour," and "miscellaneous." We use BADA to reveal differences (and similarities) between these beer categories. Data are initialized as:

```
these.rows <- which(rowSums(beer.tasting.notes$region.design[,-5])==1)
BEER <- beer.tasting.notes$data[these.rows,]
DESIGN <- beer.tasting.notes$pale.sour.style[these.rows,]
```

and analysis is performed as:

```
beer.bada <- tepBADA.inference.battery(DATA = BEER,
scale = FALSE,
DESIGN = DESIGN,
make_design_nominal = FALSE,
test.iters = 1000)
```

Appendix B.3. Hellinger vs. χ^2

The following code runs the example in Section 4.3. In this example, we still use the the same beer data as in the PCA and BADA examples, but

we have transformed the data into categorical data. In fact, the data are inherently ordinal and data may be better analyzed with MCA. For this example, we recoded each column into 4 bins and perform χ^2 MCA and MCA with Hellinger:

```
 these.rows <- which(rowSums(beer.tasting.notes$region.design[,-5])==1)
BEER <- beer.tasting.notes$data[these.rows,]
STYLES<-beer.tasting.notes$style.design[these.rows,]
BEER.recode <-
apply(BEER,2,cut,breaks=4,labels=c(‘‘LOW’’,‘‘MidLOW’’,‘‘MidHIGH’’,‘‘HIGH’’))
rownames(BEER.recode) <- rownames(BEER)
```

Then perform χ^2 MCA:

```
mca.res <- epMCA(DATA = BEER.recode,
make_data_nominal = TRUE,
DESIGN = STYLES,
make_design_nominal = FALSE,
correction = NULL)
```

And finally perform Hellinger MCA:

```
hellinger.res <- epMCA(DATA = BEER.recode,
make_data_nominal = TRUE,
DESIGN = STYLES,
make_design_nominal = FALSE,
hellinger = TRUE,
```

```
symmetric = FALSE,
correction = NULL)
```

Appendix B.4. DiSTATIS

The following code runs the example in Section 4.4. DISTATIS is a generalization of MDS to multiple distance tables. The aim of this analysis is to find if flavor perception is driven by factors beyond style, such as yeast, water source, or "le terroir" (geophysical factors). Data are set up as:

```
 these.rows <- which(rowSums(beer.tasting.notes$region.design[,-5])==1)
BEER <- beer.tasting.notes$data[these.rows,]
STYLES<-beer.tasting.notes$style.design[these.rows,]
BEER.DIST <- dist(BEER,upper=TRUE,diag=TRUE)
phys.dist <- beer.tasting.notes$physical.distances
```

Then we compute two separate MDS analyses. One for perceived flavors:

```
flav<-epMDS(DATA=BEER.DIST,
DESIGN=STYLES,
make_design_nominal =FALSE)
```

And the next based on physical distance between breweries:

```
phys.dist <- beer.tasting.notes$physical.distances
phys<-epMDS(DATA=phys.dist,
DESIGN=STYLES,
make_design_nominal =FALSE)
```

To combine the two matrices in a single analysis, we use DISTATIS

```
table <- c(rep("flavors",ncol(BEER.DIST)),rep("meters",ncol(phys.dist)))
flavor.phys.dist <- cbind(BEER.DIST,phys.dist)
demo.distatis <- mpDISTATIS(flavor.phys.dist,
DESIGN=STYLES,
make_design_nominal =FALSE,
sorting='No',
normalization='MFA',
table=table)
```

DISTATIS produces a compromise between perceived taste and physical distance between each beer.

Abdi, H., 2007a. Singular value decomposition (svd) and generalized singular value decomposition (gsvd). In: Salkind, N. J. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks CA, pp. 907–912.

Abdi, H., 2007b. Z-scores. In: Salkind, N. J. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks CA, pp. 1057–1058.

Abdi, H., 2010. Partial least squares regression and projection on latent structure regression (pls regression). Wiley Interdisciplinary Reviews: Computational Statistics 2 (1), 97–106.

Abdi, H., Chin, W., Esposito Vinzi, V., Russolillo, G., Trinchera, L., 2013a. New Perspectives in Partial Least Squares and Related Methods. Springer Verlag, New-York.

Abdi, H., Dunlop, J. P., Williams, L. J., 2009. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage 45, 89–95.

Abdi, H., Valentin, D., O'Toole, A., Edelman, B., 2005. Distatis: The analysis of multiple distance matrices. In: Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA, pp. 42–47.

Abdi, H., Williams, L., 2010a. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459.

Abdi, H., Williams, L., Beaton, D., Posamentier, M., Harris, T., Krishnan, A., Devous Sr, M., 2012a. Analysis of regional cerebral blood flow data to discriminate among alzheimer's disease, frontotemporal dementia, and elderly controls: A multiblock barycentric discriminant analysis (mubada) methodology. Journal of Alzheimer's Disease 31, s189–s201.

Abdi, H., Williams, L., Connolly, A., Gobbini, M., Dunlop, J., Haxby, J., 2012b. Multiple subject barycentric discriminant analysis (musubada): How to assign scans to categories without using spatial normalization. Computational and Mathematical Methods in Medicine 2012, 1–15.

Abdi, H., Williams, L., Valentin, D., 2013b. Multiple factor analysis: Principal component analysis for multi-table and multi-block data sets. Wiley Interdisciplinary Reviews: Computational Statistics 5, 149–179.

Abdi, H., Williams, L., Valentin, D., Bennani-Dosse, M., 2012c. Statis and distatis: optimum multitable principal component analysis and three way metric multidimensional scaling. Wiley Interdisciplinary Reviews: Computational Statistics 4, 124–167.

Abdi, H., Williams, L. J., 2010b. Correspondence analysis. In: Salkind, N. J., Dougherty, D. M., Frey, B. (Eds.), Encyclopedia of Research Design. Sage, Thousand Oaks, CA, pp. 267–278.

Alström, J., Alström, T., Jun. 2012. Beeradvocate.com.
URL `http://beeradvocate.com/`

Baty, F., Facompré, M., Wiegand, J., Schwager, J., Brutsche, M. H., 2006. Analysis with respect to instrumental variables for the exploration of microarray data structures. BMC bioinformatics 7 (1), 422.

Beaton, D., Filbey, F. M., Abdi, H., 2013. Integrating partial least squares and correspondence analysis for nominal data. In: Proceedings in Mathematics and Statistics: New perspectives in Partial Least Squares and Related Methods. Springer-Verlag, pp. 81–94.

Bécue-Bertaut, M., Pagès, J., 2008. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. Computational Statistics & Data Analysis 52 (6), 3255 – 3268.

Benzécri, J., 1973. L'analyse des données. Vol. 2. Paris: Dunod.

Benzécri, J., 1979. Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. Cahiers de l'Analyse des Données 4, 377–378.

Bookstein, F., 1994. Partial least squares: a dose–response model for measurement in the behavioral and brain sciences. Psycoloquy 5 (23).

Borg, I., 2005. Modern multidimensional scaling: Theory and applications. Springer.

Buchsbaum, B., Lemire-Rodger, S., Fang, C., Abdi, H., 2012. The neural basis of vivid

memory is patterned on perception. Journal of Cognitive Neuroscience 24, 1–17.

Caillez, F., Pagès, J., 1976. Introduction à l'Analyse des Données. SMASH, Paris.

Chernick, M., 2008. Bootstrap methods: A guide for practitioners and researchers. Vol. 619. Wiley-Interscience.

Couture-Beil, A., 2013. rjson: JSON for R. R package version 0.2.12.
URL http://CRAN.R-project.org/package=rjson

Cuadras, C. M., Cuadras, D., Greenacre, M. J., 2006. A comparison of different methods for representing categorical data. Communications in Statistics–Simulation and Computation® 35 (2), 447–459.

Dray, S., 2008. On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. Computational Statistics & Data Analysis 52 (4), 2228–2237.

Dray, S., Dufour, A., 2007. The ade4 package: implementing the duality diagram for ecologists. Journal of statistical software 22 (4), 1–20.

Eddelbuettel, D., Sanderson, C., 2013. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. Computational Statistics & Data Analysis in press.
URL http://dx.doi.org/10.1016/j.csda.2013.02.005

Efron, B., Tibshirani, R., 1993. An introduction to the bootstrap. Vol. 57. Chapman & Hall/CRC.

Escofier, B., 1978. Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée 26 (4), 29–37.

Escoufier, Y., 2007. Operators related to a data matrix: A survey. In: COMPSTAT: Proceedings in Computational Statistics; 17th Symposium Held in Rome, Italy, 2006. Physica Verlag;, New York, pp. 285–287.

Esposito Vinzi, V., Russolillo, G., 2013. Partial least squares algorithms and methods. Wiley Interdisciplinary Reviews: Computational Statistics 5 (1), 1–19.

Gomez, J. C., Moens, M.-F., 2012. Pca document reconstruction for email classification. Computational Statistics & Data Analysis 56 (3), 741 – 751.

Google, Inc, 2013. Google Maps.

Gower, J., 1968. Adding a point to vector diagrams in multivariate analysis. Psychometrika 55, 582–585.

Greenacre, M., 1984. Theory and applications of correspondence analysis. Academic Press.

Greenacre, M. J., 2007. Correspondence analysis in practice. CRC Press.

Hesterberg, T., 2011. Bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics 3, 497–526.

Hill, M. O., 1974. Correspondence analysis: A neglected multivariate method. Journal of the Royal Statistical Society. Series C (Applied Statistics) 23 (3), 340–354.

Husson, F., e, S. L., Pagés, J., 2007. Variability of the representation of the variables resulting from pca in the case of a conventional sensory profile. Food Quality and Preference 18 (7), 933 – 937.

Jolliffe, I., 2002. Principal Component Analysis. Springer Series in Statistics. Springer-Verlag, New-York.

Josse, J., Husson, F., 2011. Selecting the number of components in principal component analysis using cross-validation approximations. Computational Statistics & Data Analysis.

Krishnan, A., Williams, L. J., McIntosh, A. R., Abdi, H., 2011. Partial least squares (PLS) methods for neuroimaging: A tutorial and review. NeuroImage 56 (2), 455 – 475.

Lavit, C., Escoufier, Y., Sabatier, R., Traissac, P., 1994. The act (statis method). Computational Statistics & Data Analysis 18 (1), 97–119.

Lê, S., Josse, J., Husson, F., et al., 2008. Factominer: An r package for multivariate analysis. Journal of statistical software 25 (1), 1–18.

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J., Duchesnay, E., 2012. Significant correlation between a set of genetic polymorphisms

and a functional brain network revealed by feature selection and sparse partial least squares. NeuroImage 63 (1), 11–24.

Lebart, L., Morineau, A., Warwick, K. M., 1984. Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Liang, F., 2007. Use of svd-based probit transformation in clustering gene expression profiles. Computational Statistics & Data Analysis 51 (12), 6355 – 6366.

McIntosh, A., Bookstein, F., Haxby, J., Grady, C., 1996. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage 3 (3), 143–157.

McIntosh, A., Lobaugh, N., 2004. Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23, S250–S263.

McIntosh, A. R., Mišić, B., 2013. Multivariate statistical analyses for neuroimaging data. Annual Review of Psychology 64 (1), 499–525.

Meyners, M., Castura, J. C., Thomas Carr, B., 2013. Existing and new approaches for the analysis of CATA data. Food Quality and Preference 30 (2), 309–319.

Nenadic, O., Greenacre, M., 2007. Correspondence analysis in r, with two- and three-dimensional graphics: The ca package. Journal of Statistical Software 20 (3), 1–13.
URL `http://www.jstatsoft.org`

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., 2013. vegan: Community Ecology Package. R package version 2.0-6.
URL `http://CRAN.R-project.org/package=vegan`

Peres-Neto, P. R., Jackson, D. A., Somers, K. M., 2005. How many principal components? stopping rules for determining the number of non-trivial axes revisited. Computational Statistics & Data Analysis 49 (4), 974–997.

Pinkham, A. E., Sasson, N. J., Beaton, D., Abdi, H., Kohler, C. G., Penn, D. L., 2012. Qualitatively distinct factors contribute to elevated rates of paranoia in autism and

schizophrenia. Journal of Abnormal Psychology 121.

R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org

Rao, C., 1995a. A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa 19 (1), 23–63.

Rao, C., 1995b. The use of hellinger distance in graphical displays of contingency table data. Multivariate Statistics 3, 143–161.

St-Laurent, M., Abdi, H., Burianová, H., Grady, C., 2011. Influence of aging on the neural correlates of autobiographical, episodic, and semantic memory retrieval. Journal of Cognitive Neuroscience 23 (12), 4150–4163.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D., 2002. The quantitative evaluation of functional neuroimaging experiments: The npairs data analysis framework. NeuroImage 15 (4), 747–771.

Takane, Y., Yanai, H., Hwang, H., 2006. An improved method for generalized constrained canonical correlation analysis. Computational Statistics & Data Analysis 50 (1), 221 – 241.

Tenenhaus, M., 1998. La régression PLS: théorie et pratique. Paris: Technip.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y., Lauro, C., 2005. Pls path modeling. Computational Statistics & Data Analysis 48 (1), 159 – 205.

Thioulouse, J., 2011. Simultaneous analysis of a sequence of paired ecological tables: a comparison of several method. Annals of Applied Statisticse 5, 2300–2325.

Torgerson, W., 1958. Theory and Methods of Scaling. Wiley, New-York.

Tucker, L. R., Jun. 1958. An inter-battery method of factor analysis. Psychometrika 23 (2), 111–136.

Tuncer, Y., Tanik, M. M., Allison, D. B., 2008. An overview of statistical decomposition

techniques applied to complex systems. Computational Statistics & Data Analysis 52 (5), 2292 – 2310.

Williams, L., Abdi, H., French, R., Orange, J., 2010. A tutorial on Multi-Block discriminant correspondence analysis (MUDICA): a new method for analyzing discourse data from clinical populations. Journal of Speech, Language and Hearing Research 53, 1372–1393.

Williams, L., Dunlop, J., Abdi, H., 2012. Effect of age on variability in the production of text-based global inferences. PloS one 7 (5).

Wold, S., Ruhe, A., Wold, H., Dunn, III, W., 1984. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing 5 (3), 735–743.

Yanai, H., Takeuchi, K., Takane, Y., 2011. Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition. Springer-Verlag, New-York.

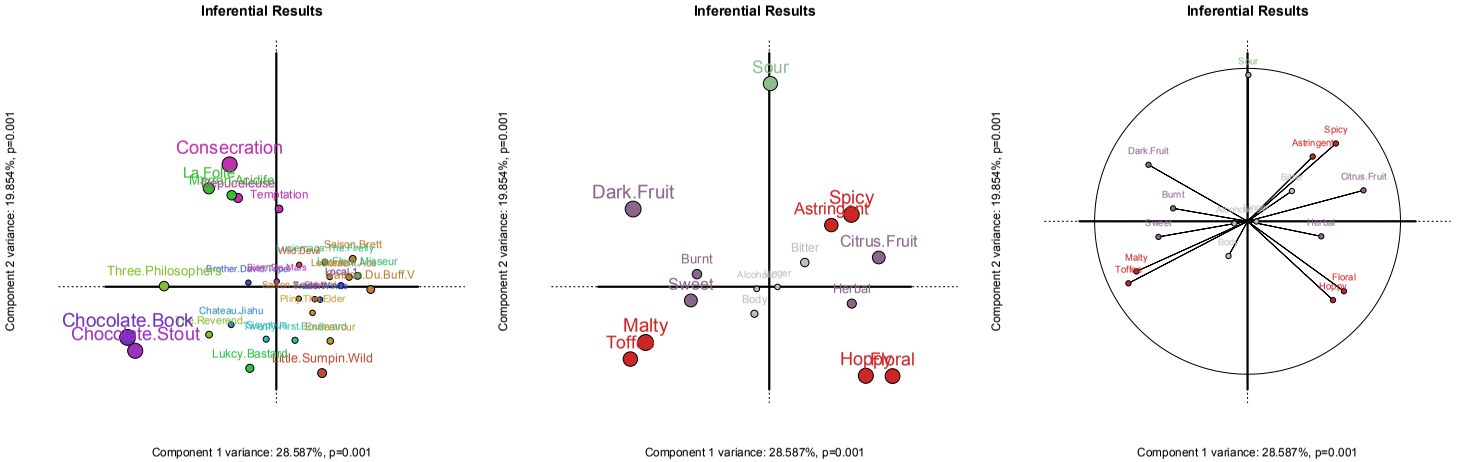

Figure 1: a. Component map with factor scores of beer (rows). b. Component map with factor scores of flavors (columns). c. Correlation between flavors (columns) and components (axes). A principal component analysis component map of the observations (rows) on Components 1 and 2. This map features 20 craft beers across 16 styles. Beers are colored by their respective style. Certain styles—such as saisons, sours, and wilds—have unique and consistent flavor profiles within their type. Furthermore, particular beer styles are strongly associated to particular flavors. For example, "Sour" is strongly associated to "Consecration" and "La Folie."

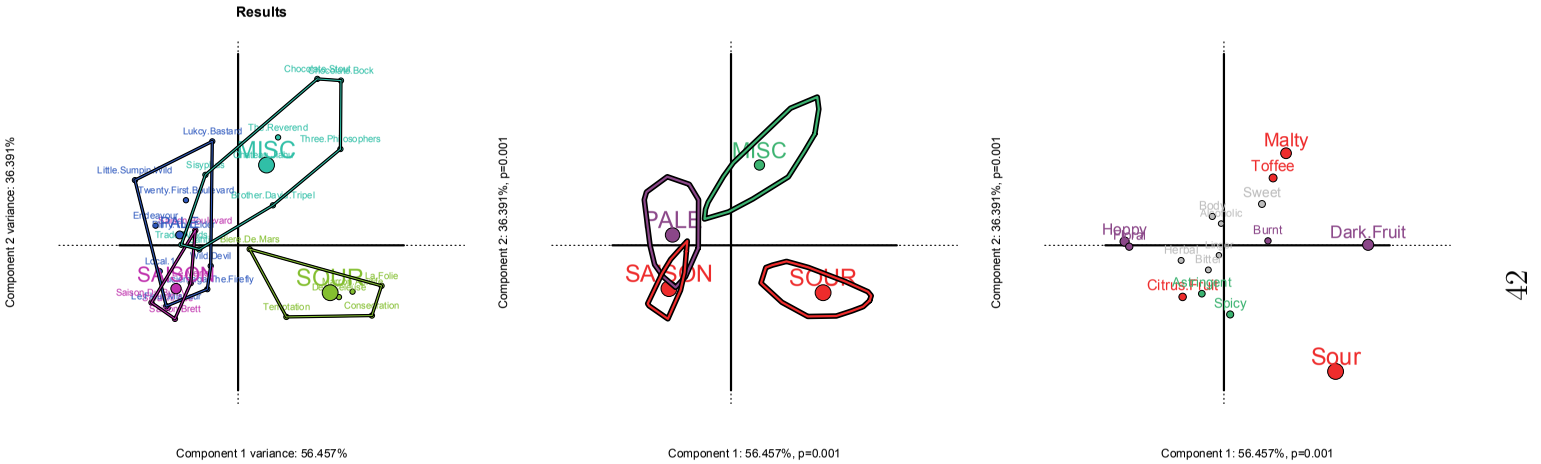

Figure 2: a). Component map with factor scores of groups and beers. b). Component map with confidence intervals around the groups. c). Component scores of measures colored by bootstrap ratio tests. A BADA illustrates which groups are significantly different and which measures help separate groups.

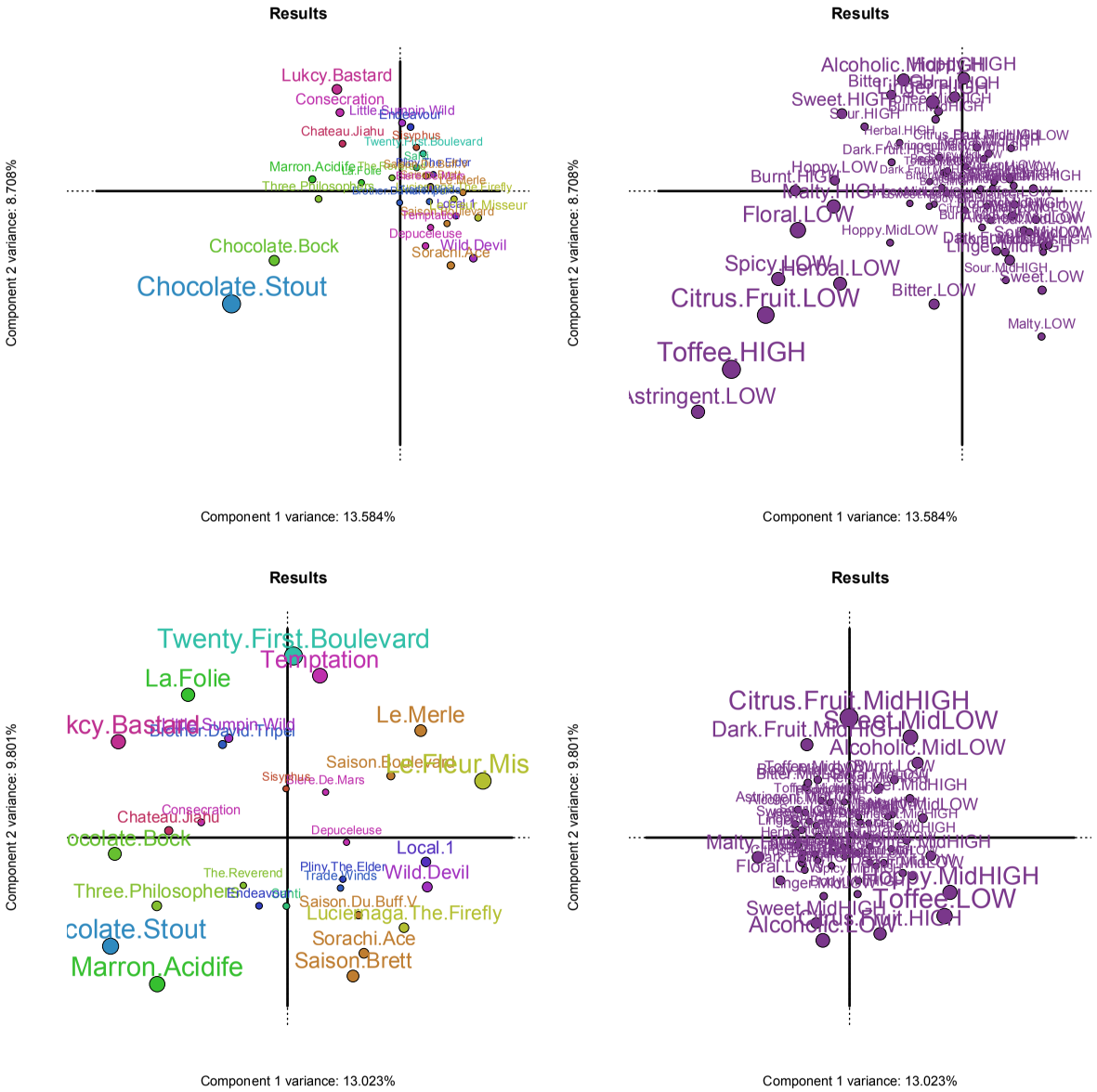

Figure 3: a. (top left) and b. (top right) illustrate factor maps in MCA with χ^2 distance factor scores. c. (bottom left) and d. (bottom right) illustrate factor maps in MCA with Hellinger distance factor scores.

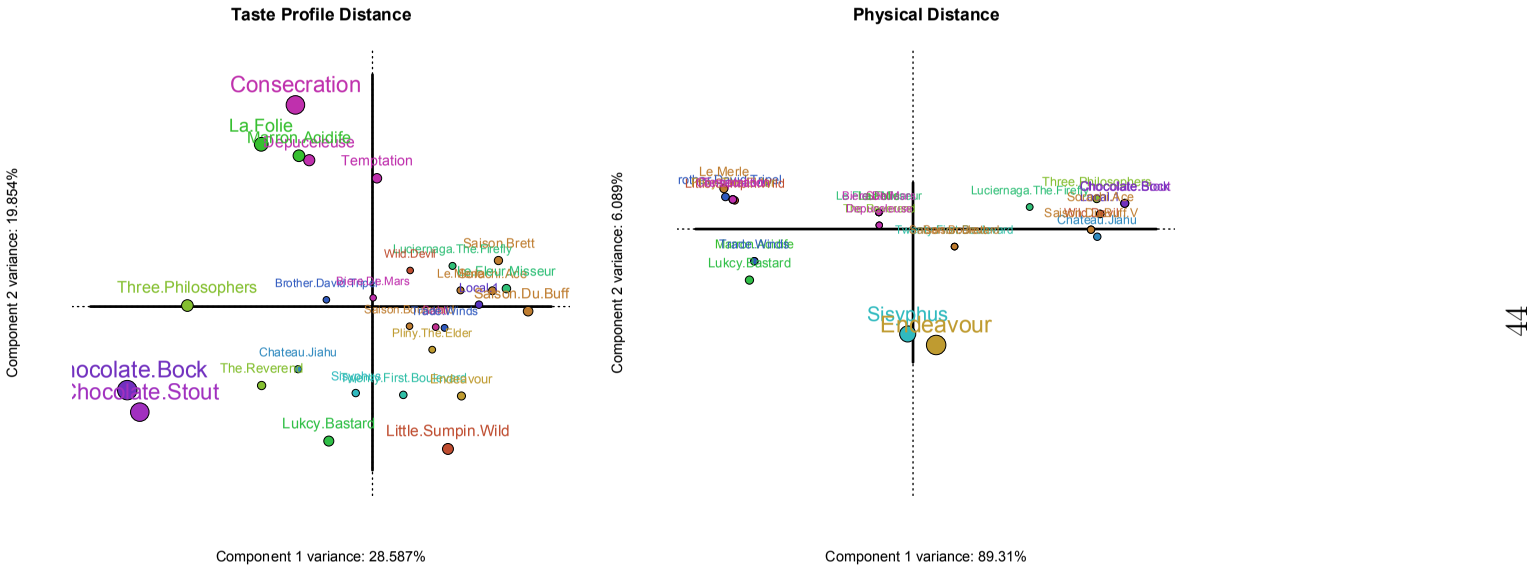

Figure 4: a. (left) shows a MDS analysis of the ratings for each beer. b. (right) shows a MDS analysis of the physical distances between the brewery.

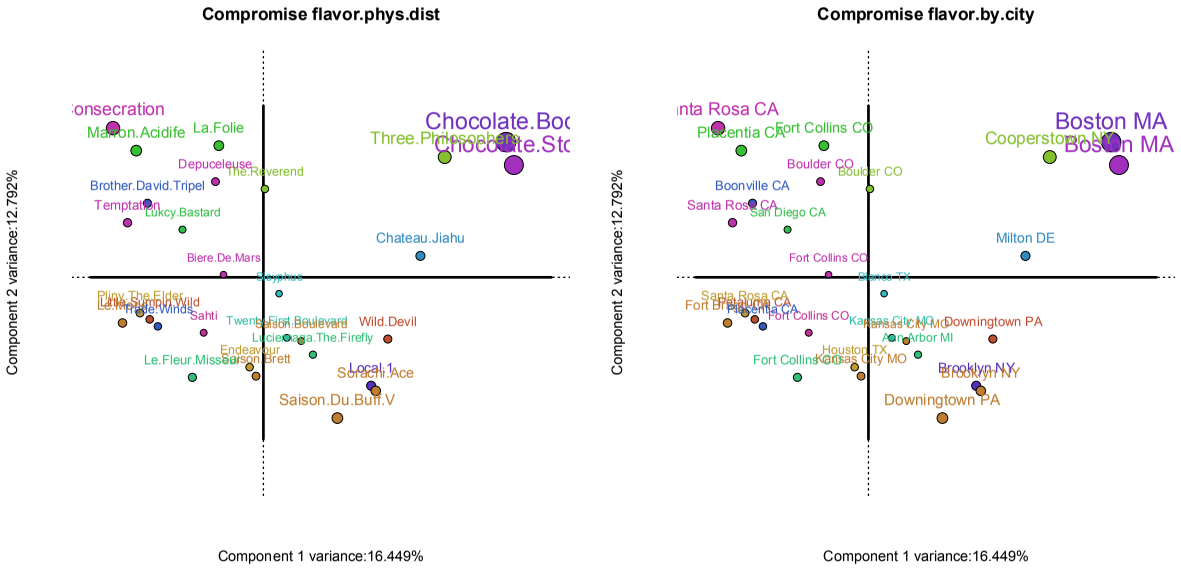

Figure 5: a. (left) shows the compromise analysis between "flavor" and physical distances for our beer data set. b. (right) is the same data, but each beer is labeled by their city of origin.