Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **72**, 1023–1040
https://doi.org/10.1093/jrsssc/qlad039
Advance access publication 8 June 2023
Original Article

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Association Plots: visualizing cluster-specific associations in high-dimensional correspondence analysis biplots

Elzbieta Gralinska and Martin Vingron

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany

Address for correspondence: Martin Vingron, Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany. Email: vingron@molgen.mpg.de

Abstract

In molecular biology, just as in many other fields of science, data often come in the form of matrices or contingency tables with many observations (rows) for a set of variables (columns). While projection methods like principal component analysis or correspondence analysis (CA) can be applied for obtaining an overview of such data, in cases where the matrix is very large the associated loss of information upon projection into two or three dimensions may be dramatic. However, when the set of variables can be grouped into clusters, this opens up a new angle on the data. We focus on the question of which observations are associated to a cluster and distinguish it from other clusters. CA employs a geometry geared towards answering this question. We exploit this feature in order to introduce *Association Plots* for visualizing cluster-specific observations in complex data. Regardless of the data matrix dimensionality Association Plots are two-dimensional and depict the observations associated to a cluster of variables. We demonstrate our method on two small data sets and then use it to study a challenging genomic data set comprising >10,000 samples. We show that Association Plots can clearly highlight those observations which characterise a cluster of variables.

Keywords: association, correspondence analysis, gene expression, marker genes

1 Introduction

A fundamental question in genomic data analysis is the following: Given a cluster of samples (variables), which genes (observations) are characteristically highly expressed in these samples, i.e. are associated to these samples? Approaches to this question occur in many forms, be it as biclustering (Pontes et al., 2015; Tanay et al., 2002) or in the search for marker genes. While for small data sets answering this question is fairly easy, for data sets with a higher number of samples the situation is more complex and poses a significant challenge to analysis and visualisation methods currently available.

Although methods such as principal component analysis (PCA) have been successfully employed for many years, they have serious limitations when applied to large and complex data sets. For such data, the first two or three principal components may explain only a small fraction of the total variance, often well below 5% of the total variance. This renders the low-dimensional representation, obtained after projection, effectively pointless due to the massive loss of information. We are faced with the challenge of visualising information from higher dimensions.

In this work, we focus on the analysis and visualisation of a large data set in the presence of a known clustering of variables. This is a realistic setting since clustering has become a routine step in data analysis, or, in many cases, a clustering is imposed naturally by the structure of the data.

Received: October 29, 2020. **Revised:** July 28, 2022. **Accepted:** April 14, 2023

Given a cluster, we will define an Association Plot, which can visualise the cluster-specific gene sets in a manner independent of the size of the data set. Association Plots are derived from correspondence analysis (CA) (Benzécri, 1973; Greenacre, 2017, 1984), itself a data projection technique closely related to PCA. In CA, the joint embedding of observations (rows) and variables (columns) from a data matrix in one space conveys information as to the association between variables and observations. We take advantage of this CA property for our definition of Association Plots. The Association Plot is planar, although it is not a projection, and it stays two-dimensional even when representing very high-dimensional data.

This article is organised as follows. First, we will introduce the formalism behind CA and Association Plots. Two example data sets, one simulated and one from economics, will serve to illustrate the new visualisation. Subsequently, we will introduce denoising based on singular value decomposition (SVD) and address statistical and scoring issues about the visual patterns one observes. Lastly, we will use the new method to study a very large and difficult to interpret gene expression data set.

2 Contingency tables and association

We assume that data are given in the form of a matrix with non-negative entries. The methods we are going to use have traditionally been applied to contingency tables, which contain count data and are therefore also integer valued. For our purposes, it suffices that the entries of the matrix are non-negative but not necessarily integer. In terms of application, we think, e.g. of gene-expression values determined using the RNA-seq technology (Wang et al., 2009). A row corresponds to a gene's expression values, while a column corresponds to the biological sample in which the measurement was performed. Like with any contingency table, we will look at row frequencies and column frequencies, and compare observed frequencies in a matrix cell with the cell's expected frequency.

Our particular question pertains to the associations between rows and columns, or between rows and a cluster of columns. For a mathematical definition of association, we follow the logic of contingency tables, where the likelihood ratio for the entry in cell (i, j) (the entry in the ith row and jth column) to be a product of chance would be $\frac{p_{ij}}{e_{ij}}$, the observed frequency in the cell divided by the expected frequency. When this ratio is close to 1, we have little reason to believe that there is an association, whereas a large ratio hints at an association between the row and the column. Subtracting 1 from the ratio we can write this as $\frac{p_{ij}-e_{ij}}{e_{ij}}$, which will be near 0 in the absence of an association. We call this quantity the 'association ratio' and abbreviate it as $a(i, j)$.

Since we are also interested in the association between an observation and a cluster of variables, we proceed to add up the association ratio for the respective cells of the matrix. Let $\mathcal{C}$, $|\mathcal{C}| = K$ be a cluster of K variables $j_1, \ldots, j_K$ which behave similarly. We extend the notation to clusters by defining

$$a(i, \mathcal{C}) := \frac{1}{K} \sum_{l=1}^{K} a(i, j_l)$$

and call this the association ratio of the observation with the cluster $\mathcal{C}$.

3 A rehash of CA

CA is a projection method for visually representing a data matrix in a high-dimensional space. While it was originally intended for analysis of contingency tables, Gower et al., 2011, p. 290, state that 'the elements of **X** [the data matrix] may contain any non-negative values'. Unlike PCA, CA does not submit the data matrix itself to a singular value decomposition, but the 'object of interest' is the matrix of Pearson residuals derived from the data matrix. CA is computed in the following steps. Let M be a matrix with G rows and C columns. By m_{gc}, we denote a value in the g^{th} row and c^{th} column, and by m_{++} the grand total of M. One calculates an observed proportion matrix P with elements $p_{gc} = m_{gc}/m_{++}$ and uses this for calculating row and column masses. The g^{th} row mass, p_{g+}, is defined as the sum across the g^{th} row, and the c^{th} column mass,

p_{+c} , as the sum of all values in the c^{th} column of P. With the expected proportions $e_{gc} = p_{g+} * p_{+c}$ Pearson residuals $f_{gc} = (p_{gc} - e_{gc})/\sqrt{e_{gc}}$ are computed. For comparison to contingency tables, note that the sum of the squares of all Pearson residuals, multiplied by the grand total m_{++} (Greenacre & Blasius, 1994, (3.2.12)) would form the χ^2 statistic.

In analogy to PCA, it is now this matrix $F = (f_{gc})$ that gets submitted to singular value decomposition, factoring it into the product of three matrices: $F = USV^T$. S is a diagonal matrix and its diagonal elements, s_{cc}, are known as singular values of F. Furthermore, the matrices U (with elements u_{gc}) and V (with elements v_{gc}) contain left- and right-singular vectors, respectively, which are represented by the columns. From the left and right singular vectors coordinates of points for rows and columns in an high-dimensional space are computed. The coordinates of $v^{(g)}$ depicting the g^{th} row are defined as $v_n^{(g)} = u_{gn}/\sqrt{p}_{g+} * s_{nn}$, for $n = 1, \ldots, m$, where $m = \min(G, C) - 1$ (Greenacre, 2017, (A.8)). The coordinates of $\omega^{(c)}$ representing the c^{th} column are given by $\omega_n^{(c)} = v_{cn}/\sqrt{p}_{+c}$ (Greenacre, 2017, (A.7)). In the literature (Greenacre, 2017), this choice of scaling is called 'asymmetric', with the rows represented in 'principal coordinates' and the columns in 'standard coordinates'.

It is a key feature of CA that one can interpret the joint map of points for rows, the v's, and columns, the ω's, in the same space. The (full) dimension of this space will be $\min(G, C) - 1$, which above we called m, (Greenacre, 2017, p. 203) and both sets of points can be thought of as elements of $\mathbb{R}^m$. We will refer to this space as the CA space. Traditionally, one visualises only the first 2 or 3 dimensions and calls this a biplot. In the examples below we will, where meaningful, also depict the two-dimensional biplot, although only for illustration purposes. The focus of our exposition, however, is on large data sets where too much information would be lost upon projection into two dimensions. Instead of 'explained variance' that is used in PCA, CA speaks of 'inertia'. Inertia is the sum of the squares of the elements of the matrix F, or in other words, the χ^2 statistic divided by the grand total m_{++} (Greenacre, 2017, p. 28). Just like the explained variance in PCA, inertia gets approximated increasingly better with increasing number of dimensions used.

The geometry of the Association Plots to be defined below rests on two key features of CA:

- A column point can be expressed as a weighted sum of row points, where the weights are related to the contribution of the rows to the particular column in P. This is Greenacre's transition equation (Greenacre, 2017, (A.16)). It is the reason for the clustering of similar rows (observations) and similar columns (variables), respectively, in CA space.
- What is even more important for our application is that the inner product of i^{th} row- and j^{th} column-vector approximates the respective association ratio, i.e.

$$a(i, j) = \langle v^{(i)}, \omega^{(j)} \rangle + \epsilon$$

where in the m-dimensional CA space the error term $\epsilon = 0$. This is Greenacre's reconstitution formula (Greenacre, 2017, formula (13.5) or (A.14)), and it is a consequence of the SVD which was used to compute the coordinates. It pertains directly to our goal of describing association in geometrical terms: Due to the inner product, the row-vectors (observations) that are associated to a column-vector (variables) lie in the direction of that column; the more aligned the two are, and the longer the vectors, the higher the association. When a low-dimensional projection is permissible, this feature is usually clearly visible.

Note that the reconstitution formula also allows for generalisation to clusters: row-points that are associated to a cluster of columns (variables) lie in the direction of these (clustered) variables. Since the inner product is bilinear, it also means that the association ratio sums nicely over groups of observations or clusters of variables. This will be utilised below.

4 Association Plots

A cluster of variables $\mathcal{C} = j_1, \ldots, j_K$ can be represented by the centroid of its variable vectors $\omega^{(j_l)}$, $l = 1 \ldots K$ in CA space. We call this centroid $\vec{X}$:

$$\vec{X} := \frac{1}{K} \sum_{l=1}^{K} \omega^{(j_l)}$$

Due to linearity, we can see the average association ratio also in the inner product between an observation and the vector from the origin to this centroid. Let $\vec{r}$ be an observation-vector in CA space representing row r of the data. Then, we can express the association between r and the cluster $\mathcal{C}$ as:

$$a(r, \mathcal{C}) = \frac{1}{K} \sum_{l=1}^{K} \langle r, \omega^{(j_l)} \rangle + \epsilon = \langle \vec{r}, \vec{X} \rangle + \epsilon$$

This is a trivial consequence of the reconstitution formula and the definitions from above.

This observation forms the basis for a simple two-dimensional visualisation of the high-dimensional information. The inner product is determined by the length of $\vec{r}$, $|\vec{r}|$, the length of $\vec{X}$, $|\vec{X}|$, and the angle between the two vectors. We call the angle $\phi(\vec{r})$, or just ϕ where the context is clear. In this notation, the inner product from above can be written as

$$\langle \vec{r}, \vec{X} \rangle = |\vec{r}|\, |\vec{X}|\, \cos(\phi(\vec{r}))$$

Therefore, it makes sense to introduce a two-dimensional representation where the x-axis corresponds to the direction of the centroid vector, and we represent $\vec{r}$ by the following x- and y-coordinates:

$$x(\vec{r}) := |\vec{r}| \cos(\phi(\vec{r}))$$
$$y(\vec{r}) := |\vec{r}| \sin(\phi(\vec{r}))$$

Clearly, $|\vec{r}| \cos(\phi(\vec{r}))$ is the length of the projection of $\vec{r}$ onto $\vec{X}$, or $\langle \vec{r}, \frac{\vec{X}}{|\vec{X}|} \rangle$, and $|\vec{r}| \sin(\phi(\vec{r}))$ is the length of the orthogonal distance of $\vec{r}$ to $\vec{X}$, or $\left| \vec{r} - \frac{x(\vec{r})}{|\vec{X}|} \vec{X} \right|$. Also, $|\vec{r}|$ is equal to the length of the vector $(x(\vec{r}), y(\vec{r}))^T$. We define the Association Plot for cluster $\mathcal{C}$ as the two-dimensional plot where each observation-vector $\vec{r}$ in CA space is represented by these two-dimensional points $(x(\vec{r}), y(\vec{r}))$.

Introducing $\tilde{X}$ as the two-dimensional vector

$$\begin{pmatrix} |\vec{X}| \\ 0 \end{pmatrix} =: \tilde{X}$$

we can ascertain the conservation of the inner products, and with it the association ratio, between CA space and Association Plot:

$$a(r, \mathcal{C}) = \langle \vec{r}, \vec{X} \rangle + \epsilon = |\vec{r}| \cos(\phi) |\vec{X}| + \epsilon = \langle \begin{pmatrix} |\vec{r}| \cos(\phi) \\ y(\vec{r}) \end{pmatrix}, \begin{pmatrix} |\vec{X}| \\ 0 \end{pmatrix} \rangle + \epsilon = \langle \begin{pmatrix} x(\vec{r}) \\ y(\vec{r}) \end{pmatrix}, \tilde{X} \rangle + \epsilon.$$

This demonstrates that the association ratio can be seen both as an inner product in the m-dimensional CA space as well as an inner product in the two-dimensional Association Plot. When the full m dimensions are used for the vectors $\vec{r}$ and $\vec{X}$, the error term ϵ will be 0. In Section 7, we will propose to use fewer dimensions than m, actually relying on the approximation.

From this simple line of algebra above, one also notes that $y(\vec{r})$ gets multiplied with 0. This means that the inner product, and with it the association ratio between observation and cluster, is constant along vertical lines in the Association Plot. Intuitively, this is due to the fact that by definition the association ratio $a(r, \mathcal{C})$ calculated for a given cluster of variables is not influenced by other variables or clusters in the data, which might also attract an observation. The angle ϕ contributes information because the less competition there is for an observation from other clusters, the smaller will be ϕ, whereas when an observation is shared also by other clusters, this will be reflected in a larger ϕ. This will be visible in the example below, and will be studied further in the section on significance of the visual patterns.

To aid interpretation, one can also embed samples into the Association Plot using the same coordinate system of projection length onto the centroid vector vs. orthogonal distance to it. The examples to follow will illustrate that closely clustered sample points in an Association Plot indicate a coherent cluster, while widely spread out points indicate a heterogeneous cluster. When one suspects that two clusters are close to each other, one can also display their respective positions in the Association Plot to check on the proximity between clusters. This will be done for the Genotype-Tissue Expression (GTEx) data in Section 9.

5 Demonstration of Association Plots on two simple data sets

While the ideas laid out here are really meant for the analysis of very large data sets, we want to first show in two simple examples how Association Plots work. We will start with a simulated data set and then proceed to employment data from the International Labor Organization.

We generated a simulated data set of 100 rows (observations) and 15 columns (variables). The 15 variables fall into 5 clusters of 3 variables each. This was generated using the function *make_blobs* from the Python library *scikit-learn* (Pedregosa et al., 2011) which can be used for generating isotropic Gaussian point clouds for clustering (see Section 10). The clusters are clearly visible upon projection into a two-dimensional correspondence analysis biplot (Figure 1), in which the observations (small circles) span the space between the five clusters of variables (crosses).

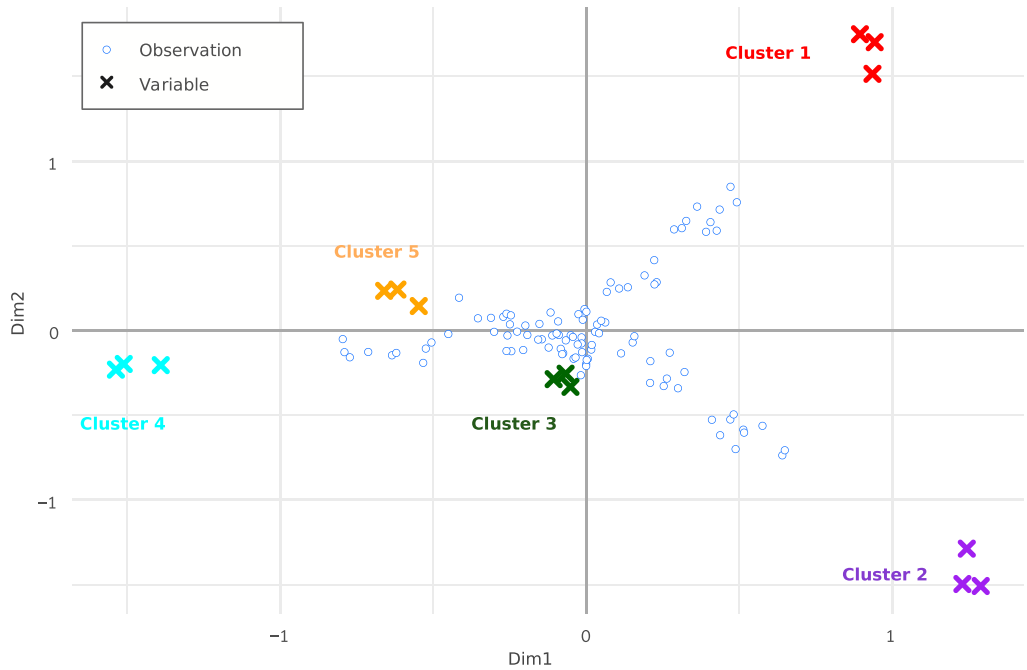

Figure 1. Two-dimensional correspondence analysis projection of the simulated data. The projection arranges the cluster members in proximity to each other. Observations pointing toward Clusters 1, 2, and 4 are recognisable, while a direction toward Cluster 3 or 5 is not visible in this two-dimensional projection.

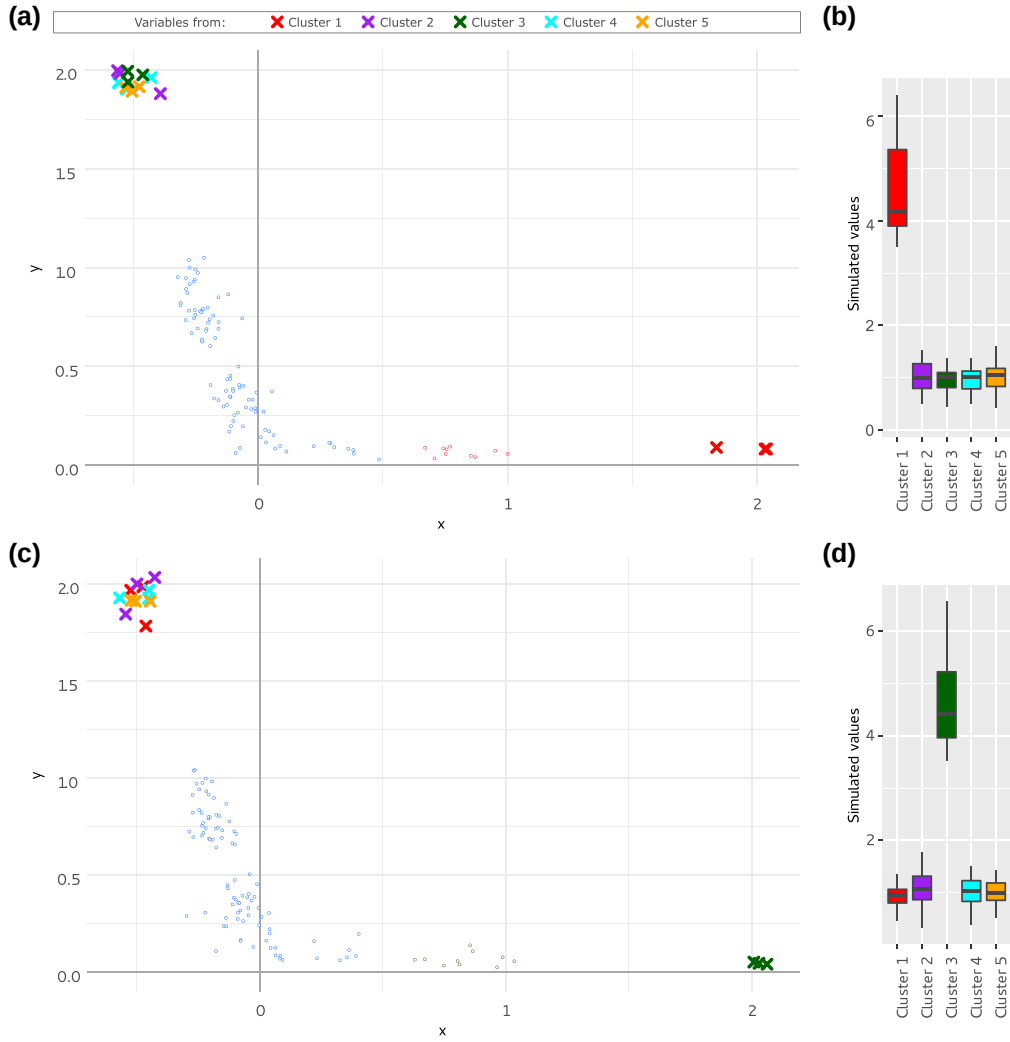

Figure 2. Identification of observations associated to a given cluster from the simulated data using Association Plots. (a–c) Association Plots were generated for (a) Cluster 1 and (c) Cluster 3 of variables using four correspondence analysis dimensions. 10 observations with the highest association for a given cluster (according the generated Association Plot) are located furthest to the right and highlighted in color ((a) red and (c) green). (b, d) Over-representation of 10 detected observations associated to (b) Cluster 1 or (d) Cluster 3.

We aim at delineating the observations associated to a given cluster. As described in Section 3 (reconstitution formula), in CA, space the observations characteristic for a given cluster of variables will be located in the direction of this cluster. Therefore, in the two-dimensional CA projection of the simulated data the observations which are located in the direction of Cluster 1 (red crosses) are expected to be associated to this cluster. In the CA biplot the observations located in the direction towards Cluster 1 are easy to see because the planar projection happens to nicely resolve this.

Alternatively, the observations associated to Cluster 1 can be visualised in the Association Plot for this cluster (Figure 2a). In this plot, one can observe that all three variables belonging to Cluster 1 are clearly separated from the other variables and the observations associated to Cluster 1 are located in the right part of the plot. This selection of observations is confirmed in Figure 2b where one can see that the 10 observations located most to the right (red circles in Figure 2b) are indeed highly over-represented in Cluster 1 in comparison to the remaining clusters.

The challenge, however, lies in delineating the observations associated to clusters which are not as nicely visible in the low-dimensional biplot as Cluster 1. For example, identification of observations associated to Cluster 3 (dark green crosses) from the simulated data is not possible based on the two-dimensional data projection. The Association Plot generated for Cluster 3 (Figure 2c) reveals a set of observations associated to this cluster, which can again be confirmed in a box plot of the observations (Figure 2d).

Our second illustrative example is a real data set and therefore not as 'clean' as the simulated data. It comes from the International Labour Organization and describes people's sector of employment in 233 countries (International Labour Organization, 2020). Employment sectors are agricultural sector, industry sector, services sector, or unemployed. Data are further divided according to gender (m/f) and year (2000, 2005, 2010, and 2015). This results in a matrix with 233 rows for the countries and 32 columns for employment category, gender, and year. For example, the column 'industry, female, 2005' would represent the percentage of female labour force of a given country that was employed in industry in 2005. Details on the data set can be found in Section 10.

Data categories are 'agriculture', 'industry', 'services', and 'unemployed', further refined by gender. These form homogeneous groups in the data which is easily confirmed by visual inspection of the two-dimensional CA projection (Figure 3). The year from which the data comes seems to have a lesser influence - at least in the 2D projection.

The blue dots in the biplot represent the countries and their location within the plot provides a clue on the employment profile of that country in 2000–2015. For example, countries in which a high percentage of the labour force was employed in the agricultural sector are located towards the agricultural sector cluster. Based on this, it is possible to identify countries which were the employment leaders in the agricultural sector. This is a simple application of the interpretation of the directions as discussed with the reconstitution formula. However, the clusters for industry and services lie very close to each other in the 2D projection, and it is hard to discern whether there are countries specifically associated to either of the two categories.

Figure 3. Two-dimensional correspondence analysis projection of the employment data. The projection allows discerning four main clusters: 'agriculture sector', 'services sector', 'industry sector', and 'unemployed'. Each of these clusters contains data points from 4 years (2000, 2005, 2010, and 2015). The location of the countries, represented in the biplot by points, implies their employment profiles in the years 2000–2015.

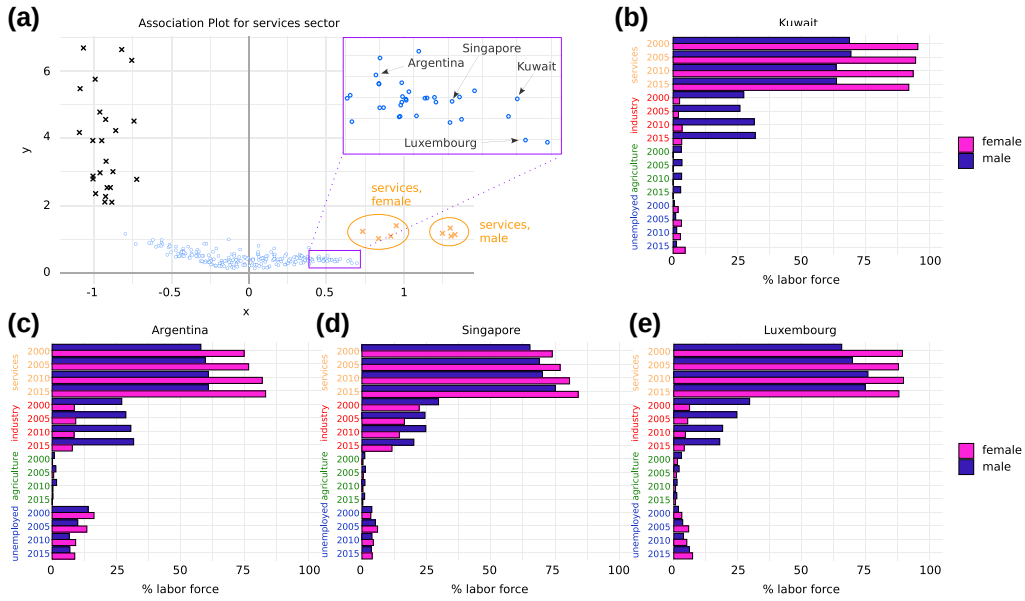

Figure 4. Association Plot for services sector. (a) The Association Plot was generated from the employment data using eight correspondence analysis dimensions. The countries with the highest percentage of population employed in the services sector (yellow crosses, on the right side) are located towards the right part of the plot. The names of four example leader countries are shown. The countries with the lowest proportion of population employed in services are located towards the left part of the plot, in the direction of other employment sectors (black crosses, on the left side). (b–e) Employment profiles of the example leaders in services sector: (b) Kuwait, (c) Argentina, (d) Singapore, and (e) Luxembourg. The presented barplots illustrate the percentages of female and male population in given country employed in different sectors in the years 2000–2015.

Finding out which countries are the leaders in, say, the services sector is made possible by the respective Association Plot (Figure 4a). In the figure, one can clearly see that the service category (represented by yellow crosses) is separated from the other employment clusters (black crosses). The yellow crosses are in turn divided into two groups, which correspond to the male and female data. The countries with the highest percentage of labour force employed in the services are located towards the right part of the plot. Starting from right, the leaders are: Hong Kong, Luxembourg, Guam, and Kuwait followed by Macao, Saudi Arabia, Brunei Darussalam, Singapore, Netherlands, and the UK. Representative barplots for some of these countries are given in Figure 4b–e. The countries with the lowest percentage of the labour force employed in services are on the opposite side of the Association Plot (Burundi, Ruanda, Central African Republic, Niger, and Rwanda). The Association Plot visualises this information while it would be invisible in the low-dimensional projection. The Association Plot for the industry cluster (Online Supplementary Material, Figure S1) actually indicates a lack of countries exclusively associated to industry. This is represented by a clear distance between industry categories and countries, as well as by the lack of a tail of points extending toward the cluster members in the Association Plot generated for the services sector.

6 Dimensions: noise reduction rather than projection into two or three dimensions

Traditionally, both PCA and CA are performed with the goal of depicting the information either in the plane or, possibly, in three dimensions. Where this implies a large loss of information, the SVD can still provide for noise reduction by cancelling the dimensions that belong to the small singular values. This is common practice in the analysis of large data sets. Employing this mechanism implies a projection into a still high-dimensional subspace, which cannot be visualised but at least maintains the relevant information in the data.

For our purposes, we adopt this procedure for noise reduction in the CA space, i.e. before computing Association Plots. The estimation of the number of dimensions that should be retained, can be based on e.g. the number of clusters in the investigated data, or can be done computationally be analysing the scree plot, i.e. the plot showing the sorted singular values from largest to smallest. We draw on the 'elbow rule' (Ciampi et al., 2005) which is based on a scree plot for original and randomised data. A randomised data matrix is generated by permuting the values for each variable separately. Subsequently, CA is applied to the resulting matrix and a vector of sorted randomised singular values is calculated. After repeating these steps multiple times, we compute an average of the first singular values, the second singular values, etc. The number of dimensions to retain is then chosen as the intersection point between the actual singular values and the average of the random singular values. See Online Supplementary Material, Figure S4 for an example.

When data happen to fall nicely into clusters, the number of dimensions to be retained should roughly reflect the number of clusters in the data. For example, the Association Plot in the employment example above was computed for eight dimensions. However, many papers have been written on the problem of selection of the right number of singular values (see the literature on spectral clustering, e.g. Von Luxburg, 2007; Zelnik-Manor & Perona, 2005). Luckily, visual inspection shows that the Association Plot is fairly robust with respect to the precise choice of dimension for the computation. An example is shown in Online Supplementary Material, Figure S2, where Association Plots were calculated for the GTEx example (see below) based on 37, 96, or 225 dimensions kept from the CA. These numbers were obtained using three different approaches for selecting optimal number of dimensions (Ciampi et al., 2005; Greenacre & Blasius, 1994) and which are described in Section 10. Inspection of the cluster-specific gene sets shows large overlaps across these choices.

7 Statistical significance and scoring of visual patterns

In the examples above, we have interpreted the association of an observation r to a cluster of variables by how large its x-value $x(r)$ is in the Association Plot for this cluster. The reasoning is that $x(r)$ is proportional to the association ratio with the length of the cluster centroid vector as proportionality factor: $a(r, \mathcal{C}) = |\vec{r}| \cos(\phi)|\vec{X}| = x(\vec{r})|\vec{X}|$. Thus, we first focus on estimating the statistical significance for $x(r)$, i.e. how far to the right a point lies in the Association Plot. Furthermore, we are typically looking at large numbers of observations such that a multiple test correction is necessary. To this end, we will also estimate FDR q-values (Storey & Tibshirani, 2003).

We estimate an empirical null distribution for $x(r)$ from permuted data. Following the SAM (Significance Analysis of Microarrays) method (Tusher et al., 2001), we randomly permute observations within each row of the data matrix. This can be repeated to generate larger samples for the null distribution. Applying CA alone to such randomised data would lead to a dense, roughly ellipsoidal cloud of observation points around the origin. For a random Association Plot, however, it does not suffice to randomise the data, but one also needs to define a random cluster of variables in order to orient the Association Plot in the direction of the cluster centroid. Additionally, we have observed that the cardinality of the cluster also influences the appearance of a random Association Plot. Therefore, in the generation of a null distribution for $x(r)$ in an Association Plot we also emulate the cluster size of the real Association Plot.

Using the randomly generated data, we can read off the empirical p-values for the $x(r)$-values of a real data set. The false discovery rate (FDR, Benjamini & Hochberg, 1995) is defined as the expectation of the ratio of false positives (F_D) over the rejected hypotheses (S_D, like in 'Significant'). Storey and Tibshirani (Storey & Tibshirani, 2003) introduce the ratio itself as the q-value assigned to the level of a particular observation.

In order to estimate F_D, we need to estimate what part of the data points comes from the null. Storey and Tibshirani (2003) proceed via the uniformity of p-values under a distributional law in order to estimate the proportion of true positives. In the absence of a distributional law, one uses empirical estimates stemming from the comparison of a (random) control experiment with the real experiment (see e.g. Zhang et al., 2008). In our problem, we have the Association Plot to give us a very clear picture of the true positive points. When plotting the randomised points, one observes a

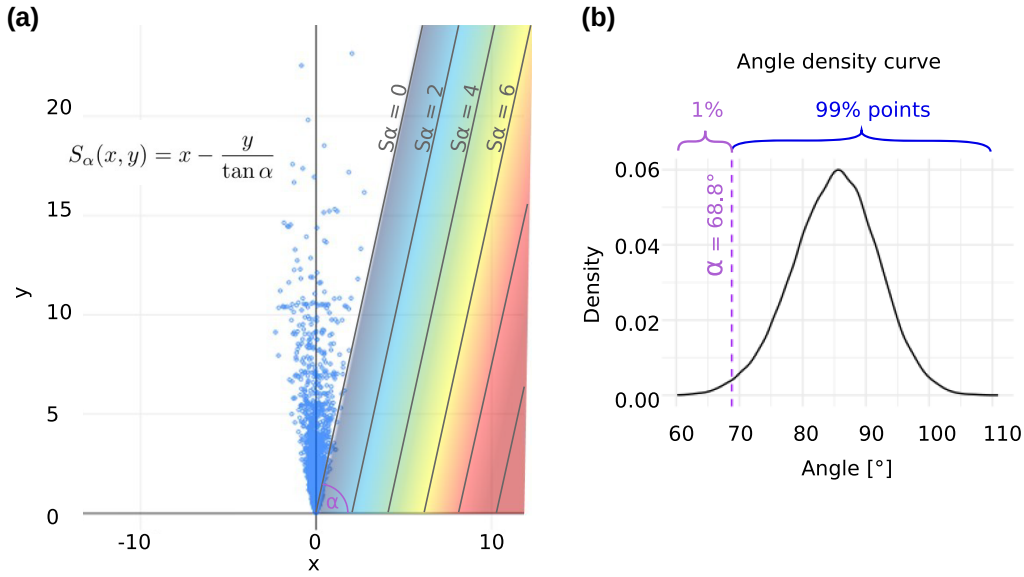

Figure 5. Scoring system of candidate genes. (a) Association Plot generated from randomised Genotype-Tissue Expression data for a 'pseudo'-cluster comprising 600 samples. For each point the angle between a given point and the *x*-axis was calculated. 1% of points with the lowest angle determines the α threshold, which will be further used for calculating the gene scores S_α in the original data. More information on the Genotype-Tissue Expression data is presented in Section 8. We use this example as a base for computing the S_α scores in the Association Plot for the heart tissue from Figure 7a. (b) Estimated density of the points upon varying the angle from 0 to 180°. In this example, the threshold of 1% resulted in $\alpha = 68.8°$.

V-shaped cloud of points, which is the Association Plot's view on the dense cloud of points around the origin in CA space. See Figure 5a for an example. The width of the 'V' provides information as to the area occupied by chance, to the right of which true positives are found in the real data. In order to count the number of true positives, we determine the angle of the ray delineating the 'V' towards the right. Therefore, for each point from Figure 5a we first compute the angle between a given point and the *x*-axis. The smoothed histogram of the computed angles, presented in Figure 5b and computed using the R function *density*, allows then for the choice of a threshold on the angle α. One can, e.g. choose the line delineating 1% of the points to the right of the V which in the example of the figure would imply a slope of 68.8°. Points to the right of this line we count as true positives. Of course, this is a slight overestimate but in practice we see that the number of true positives by far outweighs possible null points from this rule.

Based on the above reasoning, we can estimate the number of true positives and thus also the number of points stemming from the null distribution. Following Storey and Tibshirani (2003), we call this latter number m_0 and estimate the expected number of false positives above t, $E(F_D(t))$, as $m_0 * pval(t)$. t takes the x-values of the data points to the right side of the Association Plot and $pval(t)$ denotes the empirical p-value. $E(S_D(t))$ is just the number of points with $x(r)$ above the particular level t. The q-values $E(F_D(t))/E(S_D(t))$ may occasionally increase as t increases. As a remedy, (Storey & Tibshirani, 2003), use the quantity $\hat{q}$ where such a larger q-value is substituted by the prior (small) one. This $\hat{q}$ is what we report together with the p-values. See Table 1 for an example output showing for the most significant measurements $x(r)$, p-value, and $\hat{q}$.

When focusing on the association ratio and $x(r)$, we do not account for the (valuable) information conveyed by the y-axis $y(r)$ of a point in the Association Plot. Points near the x-axis, i.e. with $y(r)$ small, are more specific to the chosen cluster, while a large $y(r)$ indicates that other clusters also compete for a point. Thus, besides the association ratio we also propose a second, heuristic statistic denoted S_α. Among points of equal $x(r)$, this statistic is intended to give preference to the points with smaller $y(r)$. For the definition, we again draw on the shape of the random point cloud as described above. With the delineating line of angle α defined by the same

Table 1. Statistical significance of genes from the Association Plot generated for the heart cluster from the GTEx data.

Gene name	$x(r)$	p-value	$\hat{q}$
BMP10	4.43	8E−05	0.038
MYL7	4.38	8E−05	0.038
NPPB	4.30	8E−05	0.038
NPPA	4.27	8E−05	0.038
MYL4	4.20	8E−05	0.038
AC093642.6	4.15	8E−05	0.038
MYBPC3	4.08	8E−05	0.038
TNNI3	4.04	8E−05	0.038
AC018464.3	4.03	8E−05	0.038
MYH6	3.99	0.0001	0.038
TNNT2	3.98	0.0001	0.038
RP3-527G5.1	3.94	0.0001	0.038
RP11-532N4.2	3.92	0.0001	0.038
RP11-264A11.1	3.90	0.0001	0.038
SBK2	3.78	0.00012	0.039
TECRL	3.72	0.00016	0.039
ACTC1	3.67	0.00016	0.039
NKX2-5	3.62	0.00016	0.039
ANKRD1	3.46	0.00016	0.039
LINC00881	3.43	0.00016	0.039
MYBPHL	3.43	0.00016	0.039
NMRK2	3.07	0.00018	0.040
NPY6R	2.96	0.00018	0.040
MIR208B	2.90	0.00022	0.042
RP11-480G7.1	2.89	0.00022	0.042
RP11-432J24.5	2.83	0.00022	0.042
RP11-245G13.2	2.71	0.00022	0.042
CSRP3	2.68	0.00024	0.042
MYOZ2	2.68	0.00024	0.042
MYL3	2.44	0.00028	0.046
XIRP1	2.43	0.00028	0.046
FABP3	2.30	0.00034	0.054
HSPB3	2.19	0.00038	0.058
	…		
	…		
NEBL	1.47	0.00102	0.093
	…		
	…		
ATP5B	0.19	0.02526	0.725
	…		
	…		

criteria, the heuristic scoring function $S_\alpha(x, y)$ for an individual point (x, y) in the Association Plot is defined as:

$$S_\alpha(x, y) = x - \frac{y}{\tan \alpha}.$$

S_α is 0 along the delineating line of degree α which the simulation yielded and which in Figure 5a is annotated with '$S_\alpha = 0$'. The scoring function S_α is designed such that parallels to this line constitute level lines of increasing S_α as they shift to the right (see Figure 5a). This serves the purpose of giving higher scores to points further towards the right, while at the same time decreasing the scores as one moves upward. With this choice of S_α one can distinguish among observations which otherwise would have the same association ratio with respect to a cluster. The data analysis below will illustrate and support this choice of the scoring function S_α.

8 Analysing the GTEx data

The 'GTEx' data (Melé et al., 2015) is a very large data set comprising 11,688 tissue samples generated by RNA-seq from 30 human tissues collected from different donors (see Methods for GTEx data processing and Online Supplementary Material, Table S1 for number of samples per tissue). Each column contains one of many replicates for each tissue. The tissue information provides a natural clustering of the columns, and we interpret each tissue as one cluster. The rows of the matrix correspond to the genes whose expression levels were measured, of which we use the 5,000 genes with the largest variance across columns of the chi-square component matrix. Altogether the matrix has a size of $5{,}000 \times 11{,}688$. Several aspects of the data have been analysed in the publications of the GTEx consortium (see e.g. Melé et al., 2015) or in conjunction with specialised methods development (see, e.g. Dey et al., 2017). The question we address is 'Which genes are associated to a certain tissue?'. In biology, this corresponds to the search for so-called marker genes for a tissue.

We first conducted CA and projected the data into a three-dimensional subspace. The three dimensions of this projection together explain only ca. 24% of the inertia in this data set. The plot is clearly organised around the three directions for the tissues: pancreas, blood, and pituitary gland (Figure 6a). Genes that are located in one of these three directions are known to be specifically expressed in the tissue represented by the given direction. Alternatively, to delineate all genes specific for the given tissue one can generate an Association Plot. In Figure 6b, we present the Association Plot computed for the cluster of pancreas samples. Pancreas is the tissue which is clearly separated from other tissues in the three-dimensional projection of the data, and thus, the Association Plot for pancreas (Figure 6b) contains no surprise: Many genes point to the right, in the direction of the pancreas centroid. We provide a zoom into the right tail of the plot. This clearly shows a set of pancreas-specific genes, of which we coloured in red the known pancreas marker genes as determined by the Human Protein Atlas (Uhlén et al., 2015).

The real challenge, however, lies in the invisibility of the remaining 27 tissues in the 3D-projection. These tissues cannot be distinguished from each other since they form a dense cloud around the origin of the coordinate system. Consequently, it is also impossible in the 3D-projection to identify marker genes for these tissues. Yet, any of these 'invisible' tissues can be analysed using an Association Plot to extract further tissue-specific information. Figure 6c depicts the Association Plot generated for the liver samples. Again, a zoom into the right tail of the plot shows numerous genes, a subset of which are known as liver-specific marker genes as given by the Human Protein Atlas (Uhlén et al., 2015). One can in principle go through the tissues in this manner and for each of them generate one Association Plot which will highlight genes located in the direction of the given tissue cluster in the CA space, and thus, whose expression is characteristic for this tissue. As another representative example heart-specific genes are shown in the Association Plot generated for the heart cluster (Figure 6d).

Until this point, we have only focused on the associations between genes and a cluster of samples revealed by Association Plots. However, in addition to this, Association Plots can also provide information on the similarities between clusters. Although in the GTEx data, we can distinguish

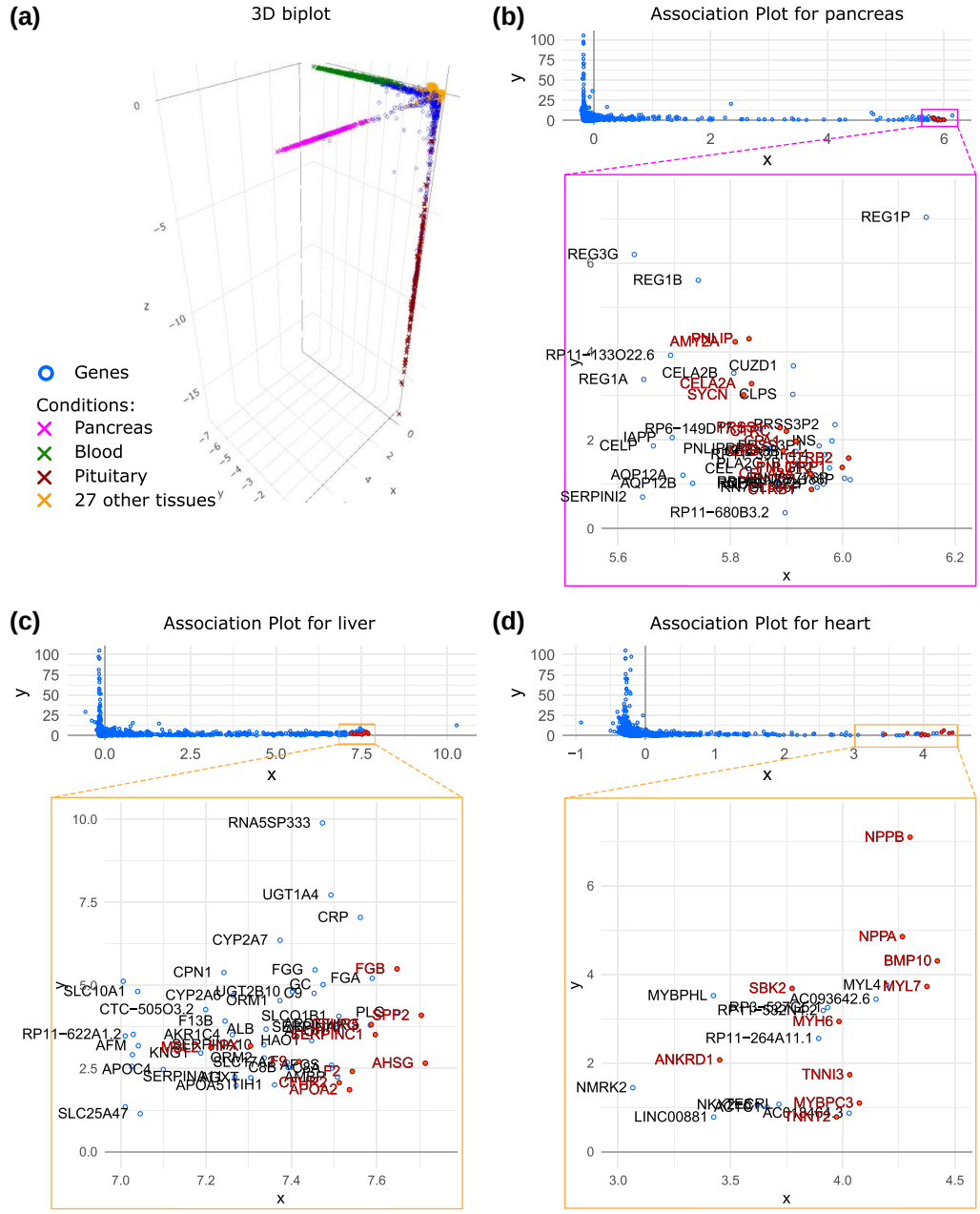

Figure 6. Applying correspondence analysis to Genotype-Tissue Expression data. (a) Classical correspondence analysis projection into 3D subspace, which allows discerning three different tissues (pancreas, blood, and pituitary). Other tissues remain lumped in the centre of the observed structure. Association Plots enable obtaining further tissue-specific information. (b–d) The Association Plots were generated based on the first 96 correspondence analysis dimensions and can be used for delineating pancreas- (b), liver- (c), and heart-specific (d) genes. Such genes are located in the right bottom part of the plot. Red colour indicates marker genes from Human Protein Atlas for a given tissue. The presented Association Plots serve as examples. Each tissue can be inspected separately and respective marker genes can be visualised.

30 clusters representing 30 distinct tissues, between some of these clusters we expect to find gene expression similarities. For instance, due to their biological identities, tissue pairs such as nerve and brain, or muscle and heart are expected to share the characteristic expression of a higher amount of genes than with the remaining tissues.

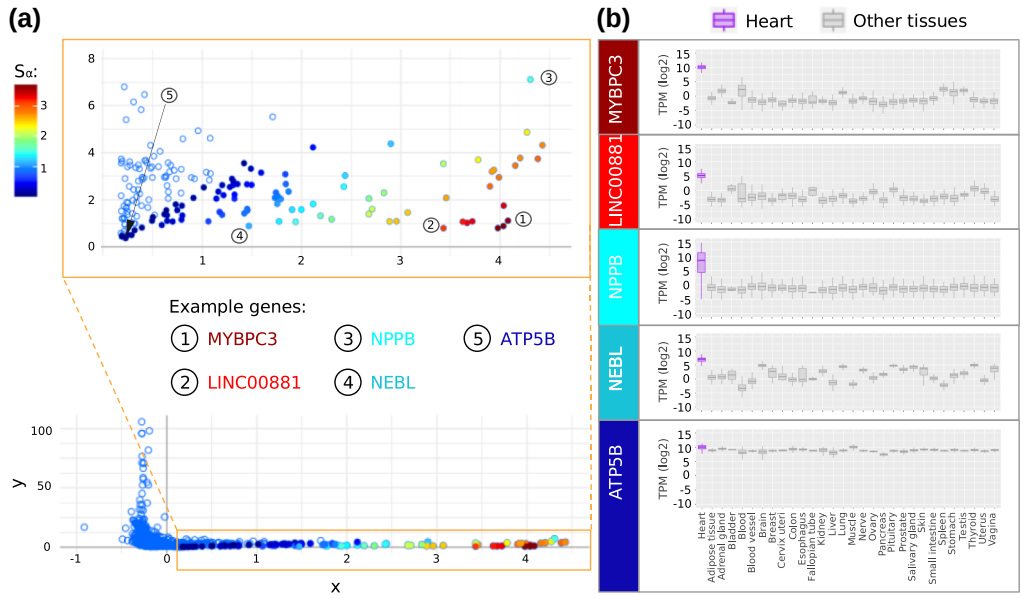

Figure 7. Heart-specific genes coloured by their S_α values. (a) Association Plot was generated for heart samples from Genotype-Tissue Expression data using 96 correspondence analysis dimensions. The colour of a gene (circle) refers to its S_α value. The higher the S_α value, the higher the specificity of the given gene for heart samples. (b) Boxplots illustrating the expression of five example genes (MYBPC3, LINC00881, NPPB, NEBL, and ATP5B) across 30 tissues. The boxplots generated for genes with high S_α values clearly show their upregulation in the heart tissue in comparison to other tissues. When comparing the expression profiles of the five highlighted genes, MYBPC3 shows the highest upregulation in the heart tissue, whereas ATP5B the lowest.

The similarity between muscle and heart is visible in the Association Plot for muscle (Online Supplementary Material, Figure S3a). There, many heart samples (red crosses) point towards the muscle samples (orange crosses), which indicates an overlap between the gene expression profiles of these two tissues. The remaining 28 tissues are located in the left part of the plot and are well separated. We also highlighted the lung samples (green crosses) which are not expected to show any similarity to muscle. Indeed, lung samples are isolated from the muscle samples and are located in the left part of the plot. Similarly, nerve and brain share many genes, and in the Association Plot for nerve samples (Online Supplementary Material, Figure S3b) the brain samples (dark pink crosses) also point towards the right. This visualises that in CA space there is a small angle between the brain samples and the nerve samples, and thus they are closer to each other than to the remaining samples.

When studying any Association Plot, one also faces the question of which points to call associated to the cluster. Here, we can apply the significance considerations from Section 7. Table 1 lists the p-values and $\hat{q}$-values for points in the Association Plot for heart. It is apparent that the points from the zoomed-in insert in Figure 6d are significant even after multiple test correction. A $\hat{q}$-value cutoff of 0.1 accepts points far further to the left than the ones shown in the insert as exemplified by the gene NEBL. This gene is also highlighted in Figure 7a.

The score S_α allows to further distinguish among genes of equal $x(r)$ but at different distance to the x-axis. Figure 7a shows the heart-specific genes coloured by the S_α value. The angle α is taken from the randomised GTEx data, which is the data set displayed in Figure 5. For more details on data randomisation and the S_α score calculation, see Section 7. The ranking of genes provided by the colour code is a rather intuitive ranking. Some examples of the actual distribution of gene expression values over the tissues are shown in Figure 7b and provide an intuition for the associations depicted in these ranked Association Plots.

The Association Plots for the GTEx data were computed based on the first 96 dimensions of the CA space, as determined by the elbow rule (see Online Supplementary Material, Figure S4 for the elbow plot). Online Supplementary Material, Figure S2 shows the very similar images for 37, 96, and 225 dimensions.

9 Discussion

Motivated mostly by biological questions, we developed and presented here a novel visualisation method, Association Plots, to study cluster-specific genes in a contingency table. As a generic data visualisation method, Association Plots do not make particular distributional assumptions as would be commonly made in the field of gene expression. For example, the widely used DESeq2 (Love et al., 2014) program for determining differentially expressed genes assumes that replicate data follow a negative binomial distribution. In this context, we rather see Association Plots as a visualisation tool and do not claim the same statistical rigour as the specialised methods.

One problem in the statistical treatment of Association Plots is the difficulty in defining an appropriate null distribution. We have approached the question by row-wise randomisation of the data matrix combined with assuming an arbitrarily defined cluster of the same size as the cluster under study. We think that this is a reasonable approach to the problem, but further research on null models that involve clusters may be needed in the future.

Generally, data sets organised in the form of a matrix or contingency table are ubiquitous, and today their size and complexity can be intimidating. Cluster information is also frequently available, be it a natural clustering stemming from the nature of the data, or a computed one. With large data, a traditional exploratory visualisation method like, e.g. PCA, may provide little help, whereas Association Plots offer the capability to visualise and interact with the sets of rows (or observations) that are associated to a cluster of columns (or variables). Association Plots can provide intuitive visualisation in a manner only dependent on the clusters to be studied but independent of the size of the data set. Dimension reduction is applied merely for the purpose of noise reduction and not to a degree where the visualisation would cancel significant amounts of information from the data.

At the heart of our method lies the geometry of the CA biplot, where the direction from the origin towards a column/variable allows for an interpretation in terms of row–column association. This gives rise to the representation of rows by orthogonal distance to this direction vector. What is visualised is essentially a cone from the origin centred on a column or a cluster centroid. This makes the geometry of CA a necessary prerequisite for Association Plots.

An aspect often overlooked in data analysis methods is whether a method will indicate that its assumptions are not met. Association Plots can also reveal that a given cluster of variables is in fact not a coherent cluster. In this case, the typical structure of observations pointing in the direction of the cluster centroid in CA space will be dispersed, and, as a consequence, the right tail in the Association Plot generated for this cluster will be short and not clearly visible anymore. Flagging the violation of a clustering assumptions is a desirable feature of our method.

We believe that the viewpoint on data analysis described opens up many further interesting questions. Most prominently, we plan to study the connection between Association Plots and biclustering (Pontes et al., 2015; Tanay et al., 2002). Also the connection to spectral clustering (Von Luxburg, 2007; Zelnik-Manor & Perona, 2005) mentioned above seems worth pursuing.

10 Data and methods

10.1 Employment data

The employment data set was downloaded on 06 November 2020 from Ilostat database (International Labour Organization, 2020). The data describes sectors of employment in 233 countries and groups of countries. In the example presented in this study, we focused on the data from years 2000, 2005, 2010, and 2015, from the eight following category types: 'employment in agriculture, female', 'employment in agriculture, male', 'employment in services, female', 'employment in services, male', 'employment in industry, female', 'employment in industry, male', 'unemployment, female', and 'unemployment, male'. Additionally, to obtain the percentages of total female or male labour force employed in three different sectors the data from the employment categories were multiplied by the factor $1 - x/100$, where x describes the percentage of the labour force for given year and gender, which was unemployed (data read from categories 'unemployment, female' or 'unemployment, male'. At the end, the resulting matrix of 233 rows and 32 columns was submitted to CA and the analysis was done using eight CA dimensions.

10.2 Simulated data

The simulated data from Section 5 was generated using the function *make_blobs* from the Python library *scikit-learn* (Pedregosa et al., 2011). We chose five 100-dimensional Gaussians, each specified through its centre and standard deviation. The large number of 100 features was chosen for illustrating the arrangement of observations in the CA plot. From those Gaussians we had the function generate three samples from each cluster. The centres were chosen far enough in the all-positive quadrant of 100-dimensional space such that with a selected standard deviation of 0.3 all generated samples contained only positive values. The following parameters were used: *n_samples*=15, *n_features*=100, *cluster_std*=.3, *centers*=[(1,5,4,3,5,5,6,6,4,4,4,4,4,4,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),

(3,10,4,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,5,5,6,6,4,4,4,4,4,4,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),

(2,5,10,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,5,5,6,6,4,4,4,4,4,4,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),

(3,1,4,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,5,5,6,6,4,4,4,4,4,4,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),

(2,8,4,3,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,5,5,6,6,4,4,4,4,4,4,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,1,1,1,1)]

This simple setup already serves well to illustrate the importance of directions and associations between variables and observations in interpretation of CA plots. Finally, CA was applied to the transpose of the resulting matrix.

10.3 GTEx data

RNA-seq data of postmortem non-disease human tissues was retrieved from GTEx Portal (Carithers et al., 2015). The files 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz' containing gene TPM values and 'GTEx_v7_Annotations_SampleAttributesDS.txt' containing sample annotations (both files available from https://gtexportal.org/home/datasets) were downloaded on 03 July 2019. A detailed description of data processing procedures is available from https://gtexportal.org/home/documentationPage.

CA was computed using a subset of 5,000 genes with the highest variance across 11,688 columns of the chi-square component matrix, which was calculated based on the input gene expression matrix. For generating the Association Plots in Figure 6, the first 96 CA dimensions (number obtained using the elbow rule) were considered. S_α scores and p-values shown in Table 1 were computed based on the randomised data obtained from 10 permutations.

10.4 Approaches for selecting optimal number of CA dimensions

The step of generating an Association Plot for a chosen set of samples is preceded by dimension reduction of the data for the purpose of noise reduction. To this end, one needs to set a number of smaller singular values to 0 and reconstruct the original data. We have implemented three approaches from the literature for determining the number of dimensions that should be retained:

The implemented methods are:

1. The 'average rule' retains those dimensions that explain more inertia than one dimension on average (Greenacre & Blasius, 1994).
2. The '80% rule' retains the minimum number of dimensions, which in total account for more than 80% of total inertia (Greenacre & Blasius, 1994).
3. The 'elbow rule' is based on scree plots of randomised data. The dimension is read off from the point where the actual scree plot of the data enters the band of randomised singular values (Ciampi et al., 2005).

In the GTEx data example, these approaches resulted in 225, 37, and 96 dimensions, respectively, and the Association Plots generated using these three numbers of dimensions are shown in Online Supplementary Material, Figure S2.

10.5 Human tissue marker genes

For validation of the Association Plots generated for GTEx data, the lists of genes with the highest levels of enriched expression in liver, heart, and pancreas were obtained on 18 March 2019 from the Human Protein Atlas (Uhlén et al., 2015), available from http://v18.proteinatlas.org.

10.6 Software package 'APL' and code availability

The code to produce Association Plots is written (mostly) in R using shiny library (Chang et al., 2019) and is available from the GitHub repository https://github.com/elagralinska/APL. Only the SVD routine (torch.svd) is taken from the Python3 torch package and gets called from the R code. The software not only displays Association Plots but also allows for interactive exploration of the plots and supports enrichment analysis of genes.

10.7 Data availability

GTEx data are available under https://gtexportal.org/home/datasets). The files 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz' and 'GTEx_v7_Annotations_SampleAttributesDS.txt' were downloaded on 03 July 2019. The employment data are available in Ilostat database (International Labour Organization, 2020) and was downloaded on 06 November 2020.

Acknowledgments

We thank our colleagues Robert Schöpflin, Peter Holderrieth, and Verena Laupert (née Heinrich) for providing feedback and criticism on the developing manuscript.

Conflict of interest: The authors declare no conflict of interest.

Funding

E.G. was funded by the International Max Planck Research School for Computational Biology and Scientific Computing (IMPRS-CBSC).

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series C* online.

References

Benjamini Y., & Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benzécri J.-P. (1973). *L'analyse des correspondances*. Dunod.

Carithers L. J., Ardlie K., Barcus M., Branton P. A., Britton A., Buia S. A., Compton C. C., DeLuca D. S., Peter-Demchok J., Gelfand E. T., & Guan P. (2015). A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreservation and Biobanking*, *13*(5), 311–319. https://doi.org/10.1089/bio.2015.0032

Chang W., Cheng J., Allaire J., Xie Y., & McPherson J. (2019). *Shiny: Web application framework for R*. R package version 1.3.2.

Ciampi A., González Marcos A., & Castejón Limas M. (2005). Correspondence analysis and 2-way clustering. *SORT*, *29*(1).

Dey K. K., Hsiao C. J., & Stephens M. (2017). Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics*, *13*(3), e1006599. https://doi.org/10.1371/journal.pgen.1006599

Gower J. C., Lubbe S. G., & Le Roux N. J. (2011). *Understanding biplots*. Wiley.

Greenacre M. (2017). *Correspondence analysis in practice*. Chapman and Hall/CRC.

Greenacre M., & Blasius J. (1994). *Correspondence analysis in the social sciences: Recent developments and applications*. Academic Press.

Greenacre M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press.

International Labour Organization (2020). Ilostat database [database]. Available from World Development Indicators: http://data.worldbank.org/indicator/SL.SRV.EMPL.MA.ZS, http://data.worldbank.org/indicator/SL.SRV.EMPL.FE.ZS, http://data.worldbank.org/indicator/SL.IND.EMPL.MA.ZS, http://data.worldbank.org/indicator/SL.IND.EMPL.FE.ZS, http://data.worldbank.org/indicator/SL.AGR.EMPL.MA.ZS, http://data.worldbank.org/indicator/SL.AGR.EMPL.FE.ZS, http://data.worldbank.org/indicator/SL.UEM.TOTL.MA.ZS, http://data.worldbank.org/indicator/SL.UEM.TOTL.FE.ZS.

Love M. I, Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Melé M., Ferreira P. G., Reverter F., DeLuca D. S., Monlong J., Sammeth M., Young T. R., Goldmann J. M., Pervouchine D. D., Sullivan T. J., & Johnson R. (2015). The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., & Vanderplas J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pontes B., Giráldez R., & Aguilar-Ruiz J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, *57*, 163–180. https://doi.org/10.1016/j.jbi.2015.06.028

Storey J. D., & Tibshirani R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Tanay A., Sharan R., & Shamir R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, *18*(suppl_1), S136–S144. https://doi.org/10.1093/bioinformatics/18.suppl_1.S136

Tusher V. G., Tibshirani R., & Chu G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(9), 5116–5121. https://doi.org/10.1073/pnas.091062498

Uhlén M., Fagerberg L., Hallström B. M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., Kampf C., Sjöstedt E., Asplund A., & Olsson I. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Von Luxburg U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Wang Z., Gerstein M., & Snyder M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Zelnik-Manor L., & Perona P. (2004). Advances in Neural Information Processing Systems 17 (NIPS 2004).

Zhang Y., Liu T., Meyer C. A., Eeckhoute J., Johnson D. S., Bernstein B. E., Nusbaum C., Myers R. M., Brown M., Li W., & Liu X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), 1–9. https://doi.org/10.1186/gb-2008-9-9-r137