AN ISA ALGORITHM WITH UNKNOWN GROUP SIZES IDENTIFIES MEANINGFUL CLUSTERS IN METABOLOMICS DATA

Harold W. Gutch[1,2], *Jan Krumsiek*[2,3], *Fabian J. Theis*[2,3]

[1] Max-Planck-Institute for Dynamics and Self-Organization
Department of Nonlinear Dynamics, 37073 Göttingen, Germany
[2] Technical University Munich, Department of Mathematics
85748 Garching, Germany
[3] Institute of Bioinformatics and Systems Biology
Helmholtz Zentrum München, 85764 Neuherberg, Germany

ABSTRACT

Independent Subspace Analysis (ISA) denotes the task of linearly separating multivariate observations into statistically independent multi-dimensional sources, where dependencies only exist within these subspaces but not between them. So far ISA algorithms have mostly been described in the context of known group sizes. Here, we extend a previously proposed ISA algorithm based on joint block diagonalization of 4-th order cumulant matrices to separate subspaces of unknown sizes. Further automated interpretation of the demixed sources then requires a means of recovering the subspace structure within them, and we propose two distinct methods for this. We then apply the method to a novel application field, namely clustering of metabolites, which seems to be well-fit to the ISA model. We are able to successfully identify dependencies between metabolites that could not be recovered using conventional methods.

1. INTRODUCTION

If $\mathbf{S}$ is an independent, real d-dimensional random vector, Independent Component Analysis (ICA) [2] denotes the task of recovering $\mathbf{S}$ given only mixtures $\mathbf{X} := \mathbf{AS}$ (where $\mathbf{A}$ is in $\mathrm{Gl}(d)$, the invertible $d \times d$ matrices with entries in $\mathbb{R}$) and efficient algorithms are available solving this task under slight approximations [1, 9] (up to the obvious indeterminacies of permutation and scaling). However, an arbitrary random vector $\mathbf{X}$ in general does not need to admit such a decomposition. Independent Subspace Analysis (ISA) generalizes this model by assuming no longer full independence of $\mathbf{S}$ (i.e. independence between all components of $\mathbf{S}$), but instead allowing some dependencies within $\mathbf{S}$. The components of $\mathbf{S}$ that share some dependencies are called the *subspaces* of $\mathbf{S}$, the indeterminacies generalize to permutation of whole subspaces and arbitrary choice of basis within the single subspaces, and the task then is recovery of $\mathbf{S}$ up to these indeterminacies. Such a decomposition always exists, and if the subspaces are *irreducible* (i.e. they cannot be further decomposed) it is unique [7].

Recently, an algorithmic approach for ISA based on the ideas of the well known JADE approach for ICA was discussed [18]. There the idea in JADE of minimizing the cross-dependencies in the 4-th order cumulant tensor and then performing joint diagonalization of a set of matrices reflecting the dependency structure was generalized to Joint Block Diagonalization (JBD), which was estimated via an ad-hoc algorithm. This empirical approach performed an approximation of JBD via Joint Diagonalization of the symmetrized cumulant matrices, which was observed to give correct recoveries apart from a final permutation stage.

We present a refined algorithmic approach to this task: Given a random vector $\mathbf{X}$, we will exploit the 4-th order information contained in it and algorithmically solve the ISA task by JBD of a set of so called (4-th order) cumulant matrices of $\mathbf{X}$ using a probabilistic JBD-algorithm [12] that has been used successfully in the context of ICA [6, 13] After this, we will estimate the subspace structure of the recovered estimation of $\mathbf{S}$. Due to the inherent noise in the data, the algorithms solving these tasks have to allow a certain fuzziness of the input data, and both steps, JBD and subspace recovery, allow adjusting of this fuzziness via threshold parameters.

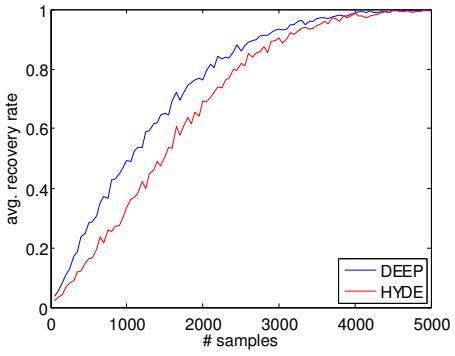

Figure 1: Evaluation of the two block recovery algorithms HYDE and DEEP, performed on a $(2+2+2)$-dimensional data set consisting of the letters A, B and C, similar to [18]. Plotted is number of samples vs. recovery rate.

2. INDEPENDENT SUBSPACE ANALYSIS VIA JOINT BLOCK DIAGONALIZATION

We briefly repeat the formal definition of the ISA Model and explain why it can be solved via Joint Block Diagonalization.

2.1 The ISA model

Let $\mathbf{S}$ be a d-dimensional real-valued random vector admitting a decomposition $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$ into independent groups, i.e. for every $1 \leq i \leq N$, the group $\mathbf{S}_i$ is independent from the rest, where we make no assumptions on the subspace sizes $\dim(\mathbf{S}_i)$. Assume now an $\mathbf{A} \in \mathrm{Gl}(d)$ and let $\mathbf{X} := \mathbf{AS}$. In this situation, ISA denotes the task of recovering the decomposition $(\mathbf{S}_1, \ldots, \mathbf{S}_N)$ given only $\mathbf{X}$. Obviously recovery of the components $\mathbf{S}_i$ is at most possible up to permutation of whole blocks and basis changes within the subspaces. If the subspaces $\mathbf{S}_i$ additionally are irreducible, this decomposition is unique up to these two indeterminacies [7]. It is easy to see that such a decomposition always exists (if we allow the trivial decomposition into a single block, if $\mathbf{S}$ itself already is irreducible).

2.2 Joint Block Diagonalization

A concise definition of joint block diagonality can be given as follows: Assume an integer d and an ordered set of positive integers $\mathbf{k} := \{1 = k_1 < \cdots < k_{N+1} = d+1\}$. Then the $\mathbf{k}$-block mask is the $d \times d$ matrix $\mathbf{B}$ defined entry-wise such that $b_{ij} = 1$ if $k_l \leq i, j < k_{l+1}$ for some integer $1 \leq l \leq N$ and $b_{ij} = 0$ otherwise. A $d \times d$ matrix

$\mathbf{M}$ is $\mathbf{B}$-*diagonal*, if $b_{ij}m_{ij} = m_{ij}$ for every $1 \leq i,j \leq d$. A set of matrices $\mathfrak{M} := \{\mathbf{M}_1, \ldots, \mathbf{M}_n\}$ is $\mathbf{B}$-diagonal, if every $\mathbf{M}_k$ $(1 \leq k \leq n)$ is, and it is jointly $\mathbf{B}$-*diagonalizable* if there is some orthogonal $\mathbf{A}$ such that $\mathbf{A}\mathfrak{M}\mathbf{A}^\top$ is jointly $\mathbf{B}$-diagonal. If $\mathbf{B} \neq \mathbf{B}'$ both are block masks and $b_{ij} \geq b'_{ij}$ for every $1 \leq i,j \leq d$, then the block mask $\mathbf{B}'$ is *finer* than $\mathbf{B}$. A set of $d \times d$ matrices $\mathfrak{M}$ is minimally $\mathbf{B}$-diagonalizable if it is $\mathbf{B}$-diagonalizable and if there is no block mask $\mathbf{B}'$ finer than $\mathbf{B}$ such that $\mathfrak{M}$ is $\mathbf{B}'$-diagonalizable. If $\mathfrak{M}$ is jointly $\mathbf{B}$-diagonal, we say that the i-th and the j-th block (where $1 \leq i < j \leq N$) of $\mathfrak{M}$ are equivalent if there is some orthogonal $\mathbf{T}$ such that $\mathbf{T}\mathbf{M}^{(i)}\mathbf{T}^\top = \mathbf{M}^{(j)}$ for every $\mathbf{M} \in \mathfrak{M}$, where $\mathbf{M}^{(l)}$ is the l-th block of $\mathbf{M}$. It should be noted that this definition coincides exactly with the intuition of joint block diagonality.

Given a set of $d \times d$ matrices $\mathfrak{M} := \{\mathbf{M}_1, \ldots, \mathbf{M}_n\}$, Joint Block Diagonalization (JBD) is the task of finding an orthogonal matrix $\mathbf{A}$ such that $\mathbf{A}\mathfrak{M}\mathbf{A}^\top =: \mathfrak{M}' =: \{\mathbf{M}'_1, \ldots, \mathbf{M}'_n\}$ is minimally $\mathbf{B}$-block diagonal for some block mask $\mathbf{B}$. It it known that a set $\mathfrak{M}$ yields a unique joint block diagonalizer $\mathbf{A}$, as long as in the minimally block diagonal set $\mathfrak{M}'$ no two blocks are equivalent [19].

2.3 Quadricovariances

Generalizing the JADE approach to solve not the ICA problem but ISA, we restrict our analysis to only 4-th order information and w.l.o.g., we assume $\mathbf{S}$ to be centered. In this case the entries of the so-called 4-th order cumulant tensor can be written as

$$\begin{aligned}\operatorname{cum}(S_i,S_j,S_k,S_l) = \mathbf{E}\{S_iS_jS_kS_l\} - \mathbf{E}\{S_iS_j\}\mathbf{E}\{S_kS_l\} - \\ \mathbf{E}\{S_iS_k\}\mathbf{E}\{S_jS_l\} - \mathbf{E}\{S_iS_l\}\mathbf{E}\{S_jS_k\}\end{aligned}$$

(for $1 \leq i,j,k,l \leq d$), and this expression is 0 whenever at least two of the arguments are statistically independent (within the subspace spanned by S_i, S_j, S_k and S_j). Based on this, we look at the so-called $d \times d$ quadricovariances (or cumulant matrices) $\mathbf{Q}_\mathbf{S}(\mathbf{M})$ of $\mathbf{S}$. Here the parameter $\mathbf{M}$ also is a $d \times d$ matrix and for a given $\mathbf{M}$, the according cumulant matrix is defined entrywise as $\mathbf{Q}_\mathbf{S}(\mathbf{M})_{ij} = \sum_{k,l}^{d} \operatorname{cum}(S_i,S_j,S_k,S_l)m_{kl}$. For white $\mathbf{S}$, the quadricovariance of $\mathbf{S}$ at $\mathbf{M}$ can be written as

$$\begin{aligned}\mathbf{Q}_\mathbf{S}(\mathbf{M}) = \mathbf{E}\{\mathbf{S}\mathbf{S}^\top(\mathbf{S}^\top\mathbf{M}\mathbf{S})\} - \\ \mathbf{E}\{\mathbf{S}^\top\mathbf{M}\mathbf{S}\} - (\mathbf{M}+\mathbf{M}^\top)\end{aligned}$$

and if $\mathbf{A}$ is an $m \times d$ matrix with orthogonal rows of unit norm, the cumulant matrices of $\mathbf{AS}$ are $\mathbf{Q}_\mathbf{AS}(\mathbf{M}) = \mathbf{A}\mathbf{Q}_\mathbf{S}(\mathbf{A}^\top\mathbf{M}\mathbf{A})\mathbf{A}^\top$. If S_i and S_j are statistically independent, every cumulant in the entrywise definition of $\mathbf{Q}_\mathbf{S}$ is 0, so in this case the (i,j)-th entry of $\mathbf{Q}_\mathbf{S}(.)$ always is 0. By assumption, $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$ where dependencies only exist within subspaces $\mathbf{S}_i$, so there is some block mask $\mathbf{B}$ defined by the sizes $\dim(\mathbf{S}_i)$ of the subspaces such that every matrix $\mathbf{Q}_\mathbf{S}(.)$ is $\mathbf{B}$-diagonal. We here additionally assume that the set $\{\mathbf{Q}_\mathbf{S}(\mathbf{M}) | \mathbf{M} \in \operatorname{Mat}(d \times d)\}$, where $\operatorname{Mat}(d \times d)$ are the $d \times d$ matrices with entries in $\mathbb{R}$ admits no finer block diagonalization.

Assume now that we are only given $\mathbf{X} = \mathbf{AS}$. Performing the usual whitening step of the observed sources and assuming decorrelation of the single source subspaces, we may assume that $\mathbf{A}$ is orthogonal. Due to the orthogonality of $\mathbf{A}$,

$$\mathbf{A}^\top\mathbf{Q}_\mathbf{X}(\mathbf{M})\mathbf{A} = \mathbf{Q}_\mathbf{S}(\mathbf{A}^\top\mathbf{M}\mathbf{A})$$

for any $\mathbf{M} \in \operatorname{Mat}(d \times d)$, where on the left hand side, the matrices $\mathbf{Q}_\mathbf{X}(\mathbf{M})$ are known, and the right hand side is $\mathbf{B}$-diagonal. As this block structure is minimal for the set $\mathbf{Q}_\mathbf{S}(.)$, our task reduces to JBD of the set $\mathbf{Q}_\mathbf{X}(.)$. As for the set of of matrices $\mathbf{M}$ where we will evaluate $\mathbf{Q}_\mathbf{X}(.)$, we here follow the choice from JADE letting this set be $\{\mathbf{M}_{ij} | 1 \leq i \leq j \leq d\}$ where

$$\begin{aligned}\mathbf{M}_{ii} &= \mathbf{e}_i\mathbf{e}_i^\top && (1 \leq i \leq d) \\ \mathbf{M}_{ij} &= (\mathbf{e}_i\mathbf{e}_j^\top)/\sqrt{2} && (1 \leq i < j \leq d)\,.\end{aligned} \tag{1}$$

3. BLOCK RECOVERY

The algorithm from [13] performs the JBD task of a set $\mathfrak{M}$ by construction of a single symmetric matrix $\mathbf{M}_0$ which has the property that a matrix diagonalizing $\mathbf{M}_0$ already performs JBD of $\mathfrak{M}$. Due to this approach, the algorithm performs JBD without actually having to estimate any subspace sizes. However, for further automated processing of the subspaces, recovery of the block structure is essential, which we now estimate from the recovered data $\mathbf{S}$. We again restrict ourselves to the use of of 4-th order information. As before we assume that not all cumulants $\operatorname{cum}(S_i,S_j,.,.)$ vanish if S_i and S_j lie in the same subspace. In the theoretical limit case of perfect information and in the absence of noise, we could simply recover the dependency graph of $\mathbf{S}$ by gathering every pair of nodes i and j into a common subspace if $\operatorname{cum}(S_i,S_j,.,.) \not\equiv 0$, but under the assumption of noise (both due to the availability of only a finite number of samples, and noise inherent in the system) this would result in just one large d dimensional dependency block, which clearly is not desired.

A simple approach to recover the block structure of a matrix $\mathbf{M}$ is to read it as the adjacency matrix of a graph G, then recovery of the blocks of $\mathbf{M}$ is equivalent to recovery of the connected components of G and efficient algorithms are available achieving this task [16]. We present two approaches that try to capture the dependencies within within $\mathbf{S}$ as a function of the 4-th order cumulant tensor of $\mathbf{S}$ and some real-valued threshold ε and then collect all dependent components together, marking these as a recovered subspace. Note that this problem can be seen as a community detection problem, which is a hot topic in the field of graph theory.

3.1 HYDE: Hypergraph dependency estimation

For the first approach, we view the components of $\mathbf{S}$ as the nodes of a hypergraph H, where every hyperedge connects at most 4 nodes. The weight of the hyperedge connecting the i-th, j-th, k-th and the l-th node is estimated as $|\operatorname{cum}(S_i,S_j,S_k,S_l)|$, and after choosing a threshold ε, we remove all hyperedges with weight lower than ε, resulting in a in general no longer fully connected hypergraph H_{prune}. Collapsing H_{prune} then into a symmetric, unweighted (standard) graph G with d nodes by defining that nodes i and j of G are connected iff nodes i and j of H_{prune} are via at least one hyperedge, we now simply collect the components of G, telling us which components S_i of the recoveries $\mathbf{S}$ are connected and thus belong to a single subspace. We call this approach "HYDE" (*HY*pergraph based *D*ependency *E*stimation), and accordingly denote the threshold ε used in the pruning step ε_{HYDE} from now on.

3.2 DEEP: Pairwise dependency estimation

For the second approach, we want to employ a purely pairwise measure of dependence on the components, i.e. we seek a dependency matrix Δ where the (i,j)-th entry reveals something about the dependency between S_i and S_j, larger entries corresponding to larger dependencies. For this we exploit all pairwise information in the cumulant tensor by setting

$$\begin{aligned}\Delta_{ij} = |\operatorname{cum}(S_i,S_i,S_i,S_j)| + |\operatorname{cum}(S_i,S_i,S_j,S_j)| \\ + |\operatorname{cum}(S_i,S_j,S_j,S_j)|\,.\end{aligned}$$

Reading a thresholded (by some ε) version of Δ as the adjacency matrix of a graph, we again recover the block structure of $\mathbf{S}$ by extraction of the connected components of this graph. We call this approach "DEEP" (*DE*pendency *E*stimation in *P*airs), and denote the threshold ε_{DEEP} from now on.

3.3 Block parameter scanning

For a fixed $\mathbf{S}$, the set of all possible block structures returned by both HYDE and DEEP for all possible values of the parameters ε_{HYDE} and ε_{DEEP} is not ordered. While both HYDE and DEEP estimate the finest block structure for small enough values of ε_{HYDE} and ε_{DEEP}, and both estimate the coarsest block structure for large enough values of the thresholds, the two approaches can return

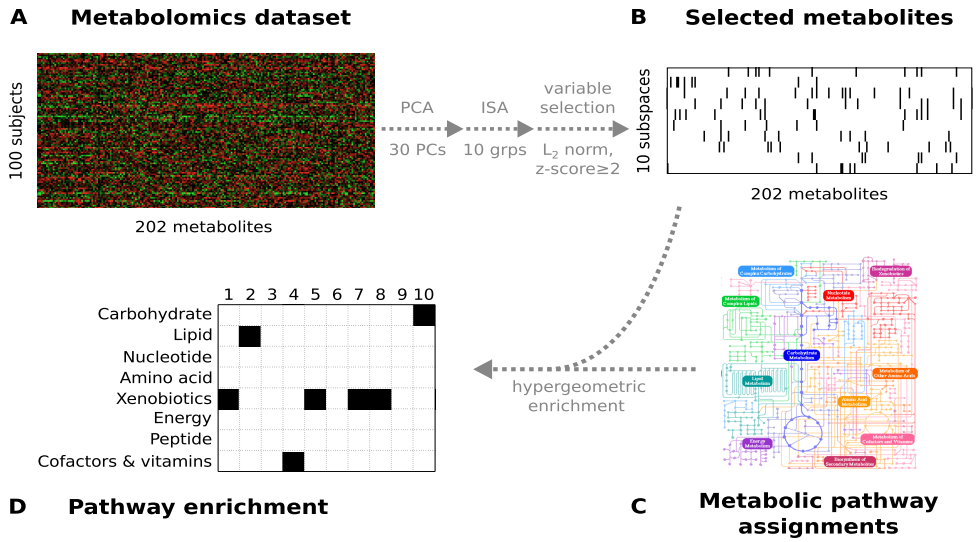

Figure 2: Workflow for the application our subspace JADE variant to metabolomics data. **A.** We used data from a German population cohort [15], containing 202 measured metabolites from 100 study participants. **B.** After application of the algorithm and subsequent metabolite selection, we obtained 10 groups containing subsets of the overall metabolite panel. **C.** For each metabolite, a single major metabolic pathway is annotated, describing its overall position on the metabolic map (network graphic from KEGG PATHWAY website [8]). **D.** Using hypergeometric enrichment analysis we assessed whether our recovered groups specifically enrich metabolites from a given class.

distinct block structures for intermediate values. For example, if $d = 3$, HYDE might group the first two components into a single subspace for intermediate threshold values, while DEEP might group the last two components into a single subspace. However, both of the two approaches themselves return a totally ordered set of block structures for the possible set of allowed thresholds $\mathbb{R}^+$: For any two values $\varepsilon_1 < \varepsilon_2$, the structure returned by HYDE (resp. DEEP) for a threshold of $\varepsilon_{HYDE} = \varepsilon_2$ (resp. ε_{DEEP}) will be at most as fine as the structure returned by it when choosing the threshold $\varepsilon_{HYDE} = \varepsilon_1$ (resp. ε_{DEEP}). Continuously changing the respective threshold from an arbitrarily small value to a large enough value is guaranteed to return all possible block structure in order. This allows the use of either method to find a block structure with a given fixed number of subspaces.

3.4 Simulations

In order to evaluate the validity of the approaches to block recovery, we generated 3 independent 2-dimensional signals, corresponding two the three letters A, B and C respectively, along the lines of [18], giving us in total a 6-dimensional (unmixed) signal. We picked $N = 50t$ samples from this signal for every $t \in \{1, \ldots, 100\}$, calculated the sample cumulants from these samples and used HYDE and DEEP to estimate the subspaces in the signal, not selecting specific thresholds ε_{HYDE} and ε_{DEEP}, but instead fixing the number of subspaces to be found to 3. We repeated this for $k = 1000$ runs for every choice of N. Figure 1 shows the results of these batch runs, where the x-axis depicts the number of samples selected, and the y-axis depicts the average percentage of runs where the three subspaces selected correspond to the correct components $\{(1,2),(3,4),(5,6)\}$. For $N = 50$, HYDE (resp. DEEP) estimates the correct subspaces only in 1.5% (resp. 3.3%) of the trials, only slightly better than pure guessing (which would select the correct subspaces with $p = 1/90$), but rising numbers of samples lead to higher recovery rates, with both algorithms recovering the subspaces correctly in at least 90% of the trials for $N \geq 3000$.

4. APPLICATION

In the following we will apply our subspace variant of JADE to a large-scale metabolomics data set. The approach follows the ideas of [17] where ICA was successfully used for a similar task, however due to lack of space we cannot go into the same level of detail and have to restrict out analysis to comparison with a single method, for which we chose k-means clustering. Metabolomics is a newly arising field aiming at the measurement of all metabolites, that is small metabolic compounds like sugars, fatty acids and amino acids, in a given biosample [5]. Cellular metabolism is driven by a set of strongly interconnected and overlapping metabolic pathways [3, 11]. Our approach now attempts to recover independent metabolite profiles, each of which stands for a separate "direction" or cellular process in metabolic space, and whose mixture finally gives rise to the measured metabolite concentration data. We first describe the steps taken and the results gained, before discussion the choice of parameters.

4.1 Description and analysis of metabolite data set

We used fasting blood serum data for 100 participants from the KORA F3 population cohort with 202 metabolites measured by ESI-MS/MS [15] (Figure 2A). The metabolite panel contains substances from various parts of the metabolism, including central energy metabolism, lipid synthesis and degradation, amino acid metabolism, xenobiotics pathways (e.g. for caffeine metabolism) and so on. For each metabolite in the dataset, one of the following 8 major metabolic pathways is annotated: *Carbohydrate, Lipid, Nucleotide, Amino acid, Xenobiotics, Energy, Peptide* and *Cofactors and vitamins*. The dataset was logarithmized in order to obtain roughly normally distributed metabolite concentrations (cf. [10]).

Every test subject corresponds to an input dimension of the data, while every metabolite corresponds to a single sample. In order to reduce the work load, we first performed PCA of the data set to the first 30 dimensions (accounting for 98.7% of the observed variance). We then calculated the sample cumulant matrices for the data set, evaluated for the matrices as in (1). Joint Block Diagonalization of this set then was performed using the JBD algorithm introduced in [13], where the non-trivial choice of the error controlling parameter ε_{JBD} was performed heuristically by setting it within the first larger gap in the eigenvalue spectrum (Figure 3A). Manual inspection (Figure 3B) suggested we fix the number of groups to at most 15. This coincided with [17], where the number of sources was fixed to 10, which we used for the number of groups. We now have a combination of 10 groups $\mathbf{S}_1, \ldots, \mathbf{S}_{10}$ of mixtures of study

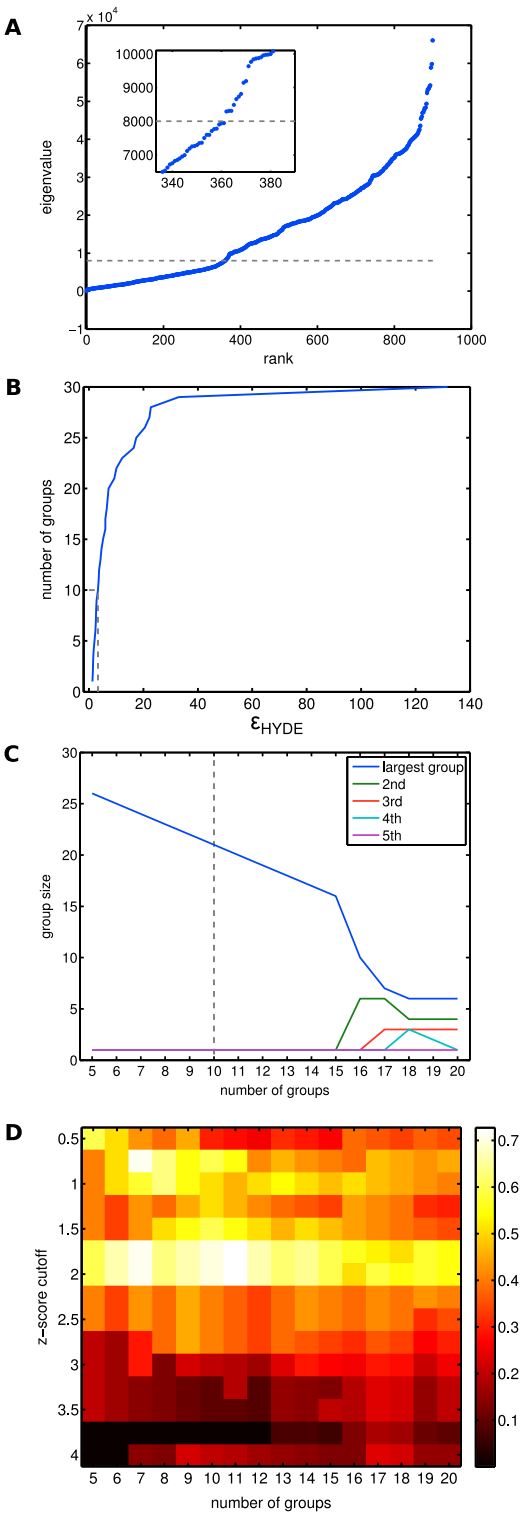

Figure 3: **A.** Eigenvalues of JBD matrix, inset showing first gap around $\varepsilon_{JBD} = 8000$. **B.** Development of the number of groups with respect to the HYDE epsilon parameter. **C.** Number of components in the five largest groups as a function of the number of groups. **D.** Heatmap displaying the fraction of enriched groups (between 0=no enriched groups, and 1=all groups are enriched), for ranging values of the z-score cutoff and the number of groups chosen.

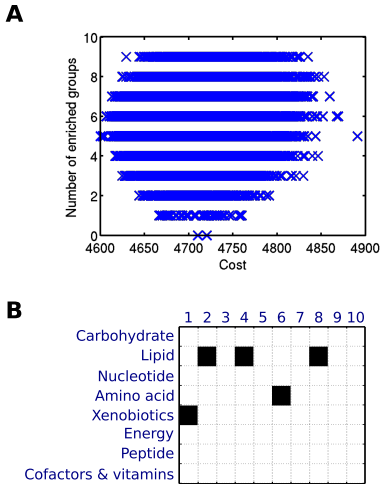

Figure 4: **A.** Cost value for 10^6 runs of the k-means algorithm, plotted against the number of enriched groups in each clustering. The best solution was obtained for a clustering with five enriched groups. **B.** Enrichment matrix for the best solution. Three groups enrich the lipid class, and two groups enrich amino acids and xenobiotics, respectively.

participants of arbitrary dimension (30 in total) and 202 metabolites, each showing some level of expression in all of the subspaces. To obtain a binary assignment of whether a metabolite "belongs" to a subspace or not, we calculated the L_2 norm for each column (i.e. each metabolite) within each subspace $\mathbf{S}_k$, and applied an absolute cutoff of 2 (Figure 2B). Next, we assessed whether the recovered metabolite groups are reasonable from a biological point-of-view, by performing a hypergeometric enrichment test [4] with $\alpha = 0.05$ for the 8 annotated pathways in each group. High enrichment values indicate good overlaps of our subspaces and known biological pathways, see also [17].

4.2 Results

In total, 7 of the 10 recovered groups display significant enrichments for one of the metabolite classes. We found that three metabolic pathways *Lipid*, *Cofactors and vitamins* and *Carbohydrate* are specifically enriched in groups 2, 4 and 10, respectively (Figure 2C+D). This indicates for distinct intracellular regulations of these major parts of the metabolism. For instance, one independent subspace contains a significantly high number of metabolites belonging to the lipid class, indicating that a large number of lipid metabolites are coregulated with respect to metabolic pathway activity. Interestingly, four groups enrich xenobiotic metabolism, indicating that this represents a separated part of metabolism with varying overlapping effects. We detected no enrichment for the classes *Nucleotide*, *Amino acid*, *Energy* and *Peptide*, which is most likely an effect of the ubiquitous roles of the metabolic pathways. For example, since energy metabolism takes place in many reactions throughout the metabolic map, it cannot be expected to show a specific enrichment in one of the recovered groups in our data.

4.3 Choice of parameters

For our primary analysis we chose a set of *ad-hoc* parameters for the different parts of the method. We first evaluate how changes in these parameters affect the results, with respect to both the overall results structure and the biologically-driven quality assessment. The employed JBD algorithm requires an input parameter ε_{JBD} declaring a threshold up to which eigenvalues are to be seen numerically identical to 0. The first 350 eigenvalues change very gradually, making it hard to choose a cut-off within these, but there is a more abrupt increase after the 360-th eigenvalue, so we select ε_{JBD} such

that it includes only the points up to here (Figure 3A). Although DEEP slightly outperformed HYDE on our toy data set (see Fig. 1), comparison of the final output with known pathways (the enrichment step) revealed that HYDE performed better on the real data set, and we therefore restrict presentation of the results to this analysis. By ranging the value of ε_{HYDE}, we are able to obtain any desired granularity in the grouping, from one large group containing all sources up to singleton groups (Figure 3B), where the latter case corresponds to regular ICA-alike grouping. Inspecting how the size of each recovered group changes with the number of groups chosen allows us to estimate the general subspace structure in the data. In our dataset, multiple groups containing more than one component emerge if we choose 16 or more groups to be recovered (Figure 3C). Finally, we investigated how the z-score cutoff (used to obtain the binary metabolite assignment from the original source matrix) and the number of groups affect metabolic class enrichment (Figure 3D). The somewhat arbitrary choice of 10 groups turned out to be very close to the optimum in the investigated value range (which lies at a z-score cutoff of 2 with 11 groups).

4.4 Comparison with conventional clustering

In order to further assess the quality of our approach and underscore its novelty, we compared our results to the well-established k-means clustering method with k=10. Since k-means inherently runs into local minima during cluster centroid optimization, we repeated the algorithm 10^6 times. The best solution, i.e. the lowest cost, was achieved for a clustering that contained a total of 5 groups enriched for one metabolic pathway class (Figure 4A). Inspecting this best solution, we found that the lipid class is enriched by three groups, and the amino acid and xenobiotics classes are enriched by the other two groups (Figure 4B). We point out that this differs from the results we obtained from the independent subspace analysis, where only two rather localized parts of metabolism (lipid and xenobiotics metabolism) pop up and, furthermore, the metabolically very heterogeneous amino acid pathway appears enriched. This finding stresses that our ISA approach detects biologically reasonable signals that could not be uncovered by simple clustering approaches.

5. CONCLUSION

We have introduced two approaches to recover subspace sizes within a multivariate data set, demonstrating the validity of the approach on toy data. In an ISA approach, the algorithm was successfully applied to a real world data set that was previously separated via a subspace variant of JADE. We have demonstrated that our combination of algorithms recovers generally reasonable sets of metabolites from a large-scale metabolomics dataset.

On the theoretical side, more sophisticated cluster detection algorithms may result in further improvements of these results. From a data interpretation point of view, the next steps in this research include a closer inspection of the actual groups beyond simple enrichment analysis, and biological interpretation of why certain metabolites are grouped in the independent mixture model. Comparisons with not only k-means, but other approaches from [17] and other ISA algorithms (e.g. [14]) are a step in yet a further direction.

Acknowledgements

This research was partially supported by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (project CoReNe), by the European Union within the ERC grant LatentCauses (grant agreement number 259294), by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center Diabetes Research (DZD e.V.), and by the BMBF-funded "Medizinische Systembiologie - MedSys" initiative (subproject SysMBo, project label 0315494A). Jan Krumsiek is supported by a PhD student fellowship from the "Studienstiftung des Deutschen Volkes".

References

[1] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *Radar and Signal Processing, IEE Proceedings F*, 140(6):362–370, 1993.

[2] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, Jan 1994.

[3] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci USA*, 104(6):1777–1782, Feb 2007.

[4] R. Gentleman and S. Falcon. Hypergeometric testing used for gene set enrichment analysis. In *Bioconductor Case Studies*, Use R, pages 207–220. Springer New York, 2008.

[5] J. L. Griffin. The cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos Trans R Soc Lond B Biol Sci*, 361(1465):147–161, Jan 2006.

[6] H. W. Gutch, T. Maehara, and F. J. Theis. Second order subspace analysis and simple decompositions. *Proc. LVA and Signal Separation 2010*, pages 370–377, 2010.

[7] H. W. Gutch and F. J. Theis. Independent subspace analysis is unique, given irreducibility. *Proc. ICA, Lecture Notes in Computer Science*, 4666:49–56, 2007.

[8] http://www.genome.jp/kegg/pathway.html. KEGG PATHWAY Database. 2011.

[9] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[10] J. Krumsiek, K. Suhre, T. Illig, J. Adamski, and F. J. Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst Biol*, 5(1):21, Jan 2011.

[11] H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin. The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135, 2007.

[12] T. Maehara and K. Murota. Algorithm for error-controlled simultaneous block-diagonalization of matrices. Dec 2009.

[13] T. Maehara and K. Murota. Error-controlling algorithm for simultaneous block-diagonalization and its application to independent component analysis. *JSIAM Letters*, 2:131–134, Dec 2010.

[14] B. Póczos, Z. Szabó, M. Kiszlinger, and A. Lőrincz. Independent process analysis without a priori dimensional information. *Proc. ICA, Lecture Notes in Computer Science*, 4666:252–259, Jan 2007.

[15] K. Suhre, C. Meisinger, A. Döring, E. Altmaier, P. Belcredi, C. Gieger, D. Chang, M. V. Milburn, W. E. Gall, K. M. Weinberger, H.-W. Mewes, M. H. de Angelis, H.-E. Wichmann, F. Kronenberg, J. Adamski, and T. Illig. Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. *PLoS One*, 5(11):e13953, 2010.

[16] R. Tarjan. Depth-first search and linear graph algorithms. *Switching and Automata Theory, 1971, 12th Annual Symposium on*, pages 114–121, 1971.

[17] A. E. Teschendorff, M. Journée, P. A. Absil, R. Sepulchre, and C. Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput Biol*, 3(8):e161, 2007.

[18] F. J. Theis. Towards a general independent subspace analysis. *Proc. NIPS*, pages 1361–1368, Jan 2006.

[19] J. H. M. Wedderburn. *Lectures on Matrices*, volume 17 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, New York, 1934.