Coupled Matrix Factorization with Sparse Factors to Identify Potential Biomarkers in Metabolomics

Evrim Acar*, Gözde Gürdeniz†, Morten A. Rasmussen*, Daniela Rago†, Lars O. Dragsted† and Rasmus Bro*

*Department of Food Science, Faculty of Science, University of Copenhagen
Email: {evrim, mortenr, rb}@life.ku.dk
†Department of Human Nutrition, Faculty of Science, University of Copenhagen
Email: {gozg, dara, ldra}@life.ku.dk

Abstract—Metabolomics focuses on the detection of chemical substances in biological fluids such as urine and blood using a number of analytical techniques including Nuclear Magnetic Resonance (NMR) spectroscopy and Liquid Chromatography-Mass Spectroscopy (LC-MS). Among the major challenges in analysis of metabolomics data are (i) joint analysis of data from multiple platforms and (ii) capturing easily interpretable underlying patterns, which could be further utilized for biomarker discovery. In order to address these challenges, we formulate joint analysis of data from multiple platforms as a coupled matrix factorization problem with sparsity constraints on the factor matrices. We develop an all-at-once optimization algorithm, called CMF-SPOPT (Coupled Matrix Factorization with SParse OPTimization), which is a gradient-based optimization approach solving for all factor matrices simultaneously. Using numerical experiments on simulated data, we demonstrate that CMF-SPOPT can capture the underlying sparse patterns in data. Furthermore, on a real data set of blood samples collected from a group of rats, we use the proposed approach to jointly analyze metabolomic data sets and identify potential biomarkers for apple intake.

Keywords—Coupled matrix factorization; sparsity; gradient-based optimization; missing data; metabolomics

I. INTRODUCTION

With the ability to collect massive amounts of data as a result of technological advances, we are commonly faced with data sets from multiple sources. For instance, metabolomics studies focus on detection of a wide range of chemical substances in biological fluids such as urine and plasma using a number of analytical techniques including Liquid Chromatography-Mass Spectroscopy (LC-MS) and Nuclear Magnetic Resonance (NMR) Spectroscopy. NMR, for example, is a highly reproducible technique and powerful in terms of quantification. LC-MS, on the other hand, allows the detection of many more chemical substances in biological fluids but only with lower reproducibility. These techniques often generate data sets that are complementary to each other [1]. Data from these complementary methods, when analyzed together, may enable us to capture a larger proportion of the complete metabolome belonging to a specific biological system. However, currently, there is a significant gap between data collection and knowledge extraction: being able to collect a vast amount of relational data from multiple sources, we cannot still analyze these data sets in a way that shows the overall picture of a specific problem of interest, e.g., exposure to a specific diet.

To address this challenge, data fusion methods have been developed in various fields focusing on specific problems of interest, e.g., missing link prediction in recommender systems [2], and clustering/community detection in social network analysis [3], [4]. Data fusion has also been studied in metabolomics mostly with a goal of capturing the underlying patterns in data [5] and using the extracted patterns for prediction of a specific condition [6] (see [1] for a comprehensive review on data fusion in omics).

Matrix factorizations are the common tools in data fusion studies in different fields. An effective way of jointly analyzing data from multiple sources is to represent data from different sources as a collection of matrices. Subsequently, this collection of matrices can be jointly analyzed using collective matrix factorization methods [7], [8].

Nevertheless, applicability of available data fusion techniques is limited when the goal is to identify a limited number of variables, e.g., a few metabolites as potential biomarkers. Matrix factorization methods, without specific constraints on the factors, would reveal dense patterns, which are difficult to interpret. Therefore, motivated by the applications in metabolomics, in this paper, we formulate data fusion as a coupled matrix factorization model with penalties to enforce sparsity on the factors in order to capture sparse patterns. Our contributions in this paper can be summarized as follows:

- Formulating a coupled matrix factorization model with penalties to impose sparsity on factor matrices,
- Developing a gradient-based optimization algorithm for solving the smooth approximation of the coupled matrix factorization problem with sparsity penalties,
- Demonstrating the effectiveness of the proposed model/algorithm in terms of capturing the underlying sparse patterns in data using simulations,
- Identifying potential apple biomarkers based on joint analysis of metabolomics data sets collected on blood samples of a group of rats.

The rest of the paper is organized as follows. In Section II,
we introduce our coupled matrix factorization model with penalties to impose sparsity and a gradient-based optimization algorithm for fitting the model. Section III demonstrates the performance of the proposed approach on both simulated and real data. In Section IV, we survey the related work, and, finally, conclude in Section V.

II. CMF-SPOPT

In this section, we first introduce our model for coupled matrix factorization (CMF) with penalty terms to enforce sparsity on the factor matrices and discuss the extension of the model to coupled analysis of incomplete data. We then present our algorithmic framework called CMF-SPOPT (Coupled Matrix Factorization with SParse OPTimization), which fits the proposed model using a gradient-based optimization method.

A. Model

We consider joint analysis of multiple matrices with one mode in common using coupled matrix factorization to capture the underlying sparse factors. We first discuss the formulation of coupled matrix factorization, which has previously been studied in various data fusion studies [2], [8], [9]. Without loss of generality, suppose matrices $X \in \mathbb{R}^{I \times J}$ and $Y \in \mathbb{R}^{I \times K}$ have the first mode in common. The objective function for their joint factorization can be formulated as:

$$f(A, B, C) = \left\| X - AB^T \right\|^2 + \left\| Y - AC^T \right\|^2$$  \hspace{1cm} (1)

where $\left\| \cdot \right\|$ denotes the Frobenius norm for matrices and the 2-norm for vectors. The goal is to find the matrices $A \in \mathbb{R}^{I \times R}$, $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$ that minimize (1). Note that $A$, i.e., the factor matrix extracted from the shared mode, is common in factorization of both $X$ and $Y$.

In this paper, we extend the formulation in (1) by adding penalty terms in order to impose sparsity on factor matrices $B$ and $C$, and reformulate the objective function as:

$$f(A, B, C) = \left\| X - AB^T \right\|^2 + \left\| Y - AC^T \right\|^2 + \lambda \sum_{r=1}^{R} \| b_r \|_1 + \lambda \sum_{r=1}^{R} \| c_r \|_1 + \alpha \sum_{r=1}^{R} \| a_r \|^2$$  \hspace{1cm} (2)

where $b_r$ and $c_r$ correspond to the $r$th column of $B$ and $C$, respectively. $\left\| \cdot \right\|_1$ denotes 1-norm of a vector and is defined as $\sum |x_i|$. $\lambda$ and $\alpha$ are penalty parameters with $\lambda, \alpha \geq 0$.

This formulation is motivated by metabolomics applications, where we often have different types of measurements on the same samples. For instance, $X$ may correspond to a samples by features matrix constructed using LC-MS measurements while $Y$ may be a matrix in the form of samples by chemical shifts constructed using NMR measurements. In most metabolomics applications, we need the underlying sparse patterns in variables dimensions, e.g., metabolites, in order to relate diseases or dietary interventions with a small set of variables. Therefore, we impose sparsity only in the variables modes by adding the 1-norm penalty, which has shown to be an effective way of enforcing sparsity [10]. The 2-norm penalty on the factors in the samples mode, i.e., the last term in (2), is added to handle the scaling ambiguity. Since there is a scaling ambiguity in the matrix factorization given above, i.e., $\hat{X} = \left( \eta A \right) \left( \frac{1}{\eta} B \right) = AB$, without penalizing the norm of the factors in the samples mode, the sparsity penalty would not have the desired effect.

1) Smooth Approximation: In order to minimize the objective function (2), we need to deal with a non-differentiable optimization problem due to the 1-norm terms. However, by replacing the 1-norm terms with differentiable approximations, it can be converted into a differentiable problem. Here, we approximate the terms with 1-norm using the “epsL1” function [11] and rewrite (2) as:

$$f(A, B, C) = \left\| X - AB^T \right\|^2 + \left\| Y - AC^T \right\|^2 + \lambda \sum_{r=1}^{R} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon}$$  \hspace{1cm} (3)

where $b_{jr}$ denotes the entry in the $j$th row, $r$th column of $B$. Note that, for sufficiently small $\epsilon > 0$, $\sqrt{x^2 + \epsilon} = |x|$.

2) Missing Data: In the presence of missing data, we can still jointly factorize matrices and extract sparse patterns by fitting the coupled model only to the known data entries. Suppose $X$ has missing entries and let $W \in \mathbb{R}^{I \times J}$ indicate the missing entries of $X$ such that $w_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is known}, \\ 0 & \text{if } x_{ij} \text{ is missing}, \end{cases}$ for all $i \in \{1, \ldots, I\}$ and $j \in \{1, \ldots, J\}$. To jointly analyze matrix $Y$ and the incomplete matrix $X$, we can then modify the objective function (3) as

$$f_W(A, B, C) = \left\| W \ast (X - AB^T) \right\|^2 + \left\| Y - AC^T \right\|^2 + \lambda \sum_{r=1}^{R} \sum_{j=1}^{J} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon}$$  \hspace{1cm} (4)

where $\ast$ denotes the Hadamard (element-wise) product.

The formulations in (3) and (4) easily generalize to joint factorization of more than two matrices, each with
underlying sparse factors in the variables mode. In our objectives, we give equal weights to the factorization of each data matrix, and in the experiments, we divide each data set by its Frobenius norm so that the model does not favor one part of the objective. However, determining the right weighting scheme remains to be an open research question.

B. Algorithm

With the smooth approximation, we have obtained differentiable objective functions in (3) and (4), which can be solved using any first-order optimization algorithm [12]. In order to use a first-order optimization method, we only need to derive the gradient. The gradient of $f_W$ in (4), which is a vector of size $P = R(I + J + K)$, can be formed by vectorizing the partial derivatives with respect to each factor matrix and concatenating them all, i.e.,

$$
\nabla f_W = \begin{bmatrix}
\text{vec} \frac{\partial f_W}{\partial A} \\
\text{vec} \frac{\partial f_W}{\partial B} \\
\text{vec} \frac{\partial f_W}{\partial C}
\end{bmatrix}
$$

Let $Z = AB^T$. Assuming each term of $f_W$ in (4) is multiplied by $\frac{1}{2}$ for the ease of computation, the partial derivatives of $f_W$ with respect to factor matrices, $A$, $B$ and $C$, can be computed as:

$$
\frac{\partial f_W}{\partial A} = (W * Z - W * X)B - YC + AC^T C + \alpha A
$$

$$
\frac{\partial f_W}{\partial B} = (W * Z - W * X)^T A + \frac{\lambda}{2} B / (B * B + \epsilon) \frac{1}{2}
$$

$$
\frac{\partial f_W}{\partial C} = -Y^T A + CA^T A + \frac{\lambda}{2} C / (C * C + \epsilon) \frac{1}{2}
$$

where the operator $/ \text{ denotes element-wise division.}$

Traditional approaches for coupled matrix factorizations are based on alternating algorithms [8], [9], where the optimization problem is solved for one factor matrix at a time by fixing the other factor matrices. While alternating algorithms are widely-used, direct nonlinear optimization methods solving for all factor matrices simultaneously have better convergence properties within the context of matrix factorizations with missing entries [13] and shown to be more accurate in the case of tensor factorizations [14]. Therefore, we use a gradient-based optimization algorithm to solve the non-convex optimization problem in (4). Neither alternating nor all-at-once approaches can guarantee to reach the global optimum. The computational cost per iteration is the same for both alternating and gradient-based approaches (See [13], [14] for in-depth comparison of alternating and all-at-once approaches).

Once the gradient, $\nabla f_W$, is computed, we then use the Nonlinear Conjugate Gradient (NCG) method with Hestenes-Steifel updates [12] and the Moré-Thuente line search as implemented in the Poblano Toolbox [15].

III. EXPERIMENTS AND RESULTS

In this section, performance of the proposed approach in terms of capturing the underlying sparse patterns in coupled data sets, is demonstrated using both simulated and real data.

A. Simulated Data

The goal of simulations is two-fold: (i) to demonstrate that underlying sparse factors used to generate coupled data sets can be accurately captured using the proposed model/algorithm (ii) to study the sensitivity of the proposed approach to different parameter values.

1) Experimental Set-up: We generate coupled matrices, $X \in \mathbb{R}^{I \times J}$ and $Y \in \mathbb{R}^{I \times K}$ computed as $X = AB^T$ and $Y = AC^T$, where $A \in \mathbb{R}^{I \times R}$ has entries randomly drawn from the standard normal distribution; matrices $B \in \mathbb{R}^{J \times R}$ and $C \in \mathbb{R}^{K \times R}$, similarly, have entries randomly drawn from the standard normal but $S\%$ of the entries in each column of $B$ and $C$ is set to zero to have sparse factors. Columns of $A$, $B$ and $C$ are normalized to unit norm.

We then add noise to $X$ and $Y$ to form coupled noisy matrices, i.e., $X_{\text{noisy}} = X + \eta \cdot N_{\text{ext}} || X ||$ and $Y_{\text{noisy}} = Y + \eta \cdot N_{\text{ext}} || Y ||$, where entries of $N_1 \in \mathbb{R}^{I \times J}$ and $N_2 \in \mathbb{R}^{I \times K}$ are randomly drawn from the standard normal.

In order to assess the performance of CMF-SPOPT in terms of capturing the underlying sparse patterns, we generate data sets with (i) sparsity levels: $S \in \{30, 50, 70\}$, (ii) noise levels: $\eta = 0.1, 0.5$, and (iii) sizes: $(I, J, K) \in \{(20, 30, 40), (20, 300, 400), (20, 3000, 4000)\}$. We use $R = 2$ as the number of components.

Once coupled matrices are generated, CMF-SPOPT is used to capture $\hat{A} \in \mathbb{R}^{I \times R_{\text{ext}}}$, $\hat{B} \in \mathbb{R}^{J \times R_{\text{ext}}}$ and $\hat{C} \in \mathbb{R}^{K \times R_{\text{ext}}}$ for different values of penalty parameters: $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5\}$. $R_{\text{ext}}$ indicates the number of extracted components.

We compare the extracted matrices $\hat{B}$ and $\hat{C}$ with the original sparse matrices $B$ and $C$ used to generate the coupled data, in terms of sparsity patterns. For instance, the first column of $B$, $b_1$, is compared with the matching column of $\hat{B}$, $\hat{b}_1$. If a nonzero in $b_1$ corresponds to a nonzero in $\hat{b}_1$, then it is a true-positive; if a zero in $b_1$ corresponds to a nonzero in $\hat{b}_1$, it is a false-positive.

As stopping conditions, CMF-SPOPT uses the relative change in function value (set to $10^{-10}$) and the 2-norm of the gradient divided by the number of entries in the gradient (set to $10^{-10}$). For initialization, we use multiple random starts and choose the run with the minimum function value.

2) Results: CMF-SPOPT can capture the underlying sparse patterns accurately for varying levels of sparsity; in particular, the recovery is perfect for higher sparsity. We illustrate the performance of CMF-SPOPT in terms of

Due to the permutation ambiguity, we look for the best permutation to match the columns.
true-positive rates (TPR) and false-positive rates (FPR) for different sparsity levels in Figure 1. The best performance, i.e., exact recovery of the underlying sparsity patterns, corresponds to TPR=1 and FPR =0. The top and bottom rows of Figure 1(a) show the performance of CMF-SPOPT in terms of capturing the sparsity pattern of the first column of B and C, respectively, for sparsity level $S = 30$. We observe that underlying patterns can be captured accurately but not perfectly as the best FPR values are around 0.1 - 0.2 with corresponding TPR values around 0.8-0.9. However, for higher sparsity, underlying sparsity patterns can be perfectly captured (Figure 1(b)). For all sparsity levels, the best performance is achieved for $\alpha = 0.1$ and $\lambda = 0.1$. Here, we set $(I, J, K) = (20, 30, 40)$, $S = 50$, and $R_{ext} = 2$, and present the average performance on 15 different sets of data.

CMF-SPOPT performs well in terms of capturing the underlying sparse patterns even at high amounts of noise. Figure 2 shows the performance of CMF-SPOPT at different noise levels. While TPR is high and FPR is low for low noise level, i.e., $\eta = 0.1$, with increasing noise we observe the degradation in performance. However, TPR is still high and FPR is low when $\eta = 0.5$. Here, we set $(I, J, K) = (20, 30, 40)$, $S = 50$, and $R_{ext} = 2$, and again report the average performance on 15 sets of data.

As we change data set sizes, best performing penalty parameters change drastically. Figure 3 shows the performance of CMF-SPOPT for varying sizes of coupled data sets for $S = 50$, $\eta = 0.5$, and $R_{ext} = 2$. We observe that for small number of dimensions in the variables mode, i.e., small values of $J$ and $K$, $\alpha = 0.1$ and $\lambda = 0.1$ can accurately capture the sparse factors in B and C. As $J$ and $K$ increase, though, higher $\alpha$ and lower $\lambda$ values become effective.

We have only reported the results for the first component of B and C. Results for the second component are similar and omitted here. Also note that matrix factorizations have
rotational ambiguity; in other words, they can capture the factor matrices uniquely only up to a rotation. For certain combination of penalty values, the factor matrices, though, are uniquely captured by CMF-SPOPT, i.e., unique up to scaling and permutation. Reported TPR and FPR values correspond to those cases, where we can uniquely capture the factor matrices up to scaling and permutation.

Finally, we show that CMF-SPOPT is robust to the selection of the component number. Here, we generate data using $R = 2$ but fit the model using $R_{ext} \in \{2, 3, 4\}$. We set $\eta = 0.5$, $S = 50$, $\lambda = \alpha = 0.1$, $(I, J, K) = (20, 30, 40)$. Table I shows the weight of each coupled component, calculated as follows: We can rewrite $X = \hat{A} \hat{B}^T$ and $Y = \hat{A} \hat{C}^T$ as $X = \sum_{r=1}^{R} \beta_r \hat{a}_r \hat{b}_r^T$ and $Y = \sum_{r=1}^{R} \gamma_r \hat{a}_r \hat{c}_r^T$, where $\beta_r$ and $\gamma_r$ are the weights of component $r$ in $X$ and $Y$, respectively, and $\| \hat{a}_r \| = \| \hat{b}_r \| = \| \hat{c}_r \| = 1$, for $r = 1, 2, ..., R$. We define the weight of a coupled component $r$ as $\sigma_r = \beta_r + \gamma_r$. Similarly, when $\hat{A}$, $\hat{B}$ and $\hat{C}$ are extracted using CMF-SPOPT, columns are normalized and $\hat{\sigma}_r$ is computed. Table I shows that when there are two common components, i.e., $R = 2$, and data sets are overfactored using $R_{ext} = 3, 4$, weights of the extra components are 0. Besides, sparsity patterns of common components are still accurately captured as indicated by high TPR and low FPR.

In summary, simulation studies demonstrate that CMF-SPOPT is quite effective in terms of capturing the underlying sparse patterns in coupled data; however, we also observe that the method is sensitive to penalty parameter values.

### B. Metabolomics Data Analysis

Next, we use CMF-SPOPT to jointly analyze metabolomics data measured using different analytical techniques and identify potential markers for apple intake.

1) Data: The data consists of blood samples collected from a group of rats, which was part of a study on the effect of apple feeding on colon carcinogenesis [16]. Here, we use the samples from forty-six male Fisher 344 rats (5-8 weeks old) obtained from Charles River (Sulzfeld, Germany). After one week of adaptation on a purified diet, the animals were randomized to two experimental groups: fed either the same purified diet (group 1: Apple 0) or the purified diet added 10 g raw whole apple (group 2: Apple 10) for 13 weeks. At the end of the study, rats were sacrificed after an overnight fasting (16hrs). Animal experiments were carried out under the supervision of the Danish National Agency for Protection of Experimental Animals.

The rat plasma samples were analyzed by untargeted liquid chromatography - time-of-flight (LC-QTOF) mass spectrometry [17] and NMR [18]. In LC-MS analysis, raw data is converted into a feature set, where each feature is denoted by the mass over charge (m/z) ratio and a retention time (see [17] for details). In NMR analysis, the spectra were preprocessed (see [18] for details) and then converted into a set of peaks using an in-house automated peak detection algorithm. We also have a third data set containing Total...
cholesterol (chol), low density cholesterol (LDL), very low density cholesterol (VLDL) and high density cholesterol (HDL) lipoproteins (computed based on the NMR data [18]) and triacylglycerol (TG) concentrations (measured using the rat plasma samples). In summary, our data can be represented using the following three matrices:

- \( X \in \mathbb{R}^{I \times J} \) of type samples by features corresponding to LC-MS data, where \( I = 46 \) and \( J = 1086 \).
- \( Y \in \mathbb{R}^{I \times K} \) of type samples by chemical shifts corresponding to NMR measurements, where \( K = 115 \).
- \( Z \in \mathbb{R}^{I \times M} \) of type samples by quality variables corresponding to quality measurements, where \( M = 9 \). Matrix \( Z \) has missing entries.

2) Model: Based on the formulation in (4), we jointly analyze \( X, Y \) and \( Z \) by minimizing the following objective:

\[
\begin{align*}
    f_w(A, B, C, D) &= \left\| X - AB^T \right\|^2 + \left\| Y - AC^T \right\|^2 + \left\| W \ast (Z - AD^T) \right\|^2 \\
    &+ \lambda \sum_{r=1}^{R} \sum_{j=1}^{J} \sqrt{b_{jr}^2 + \epsilon} + \lambda \sum_{r=1}^{R} \sum_{k=1}^{K} \sqrt{c_{kr}^2 + \epsilon} \\
    &+ \lambda \sum_{r=1}^{R} \sum_{m=1}^{M} \sqrt{d_{mr}^2 + \epsilon} + \alpha \sum_{r=1}^{R} \left\| a_r \right\|^2
\end{align*}
\]

and extract the factor matrices \( A \in \mathbb{R}^{I \times R} \), \( B \in \mathbb{R}^{J \times R} \), \( C \in \mathbb{R}^{K \times R} \) and \( D \in \mathbb{R}^{M \times R} \) corresponding to the samples, features, chemical shifts and quality variables, respectively. Using simulation data of similar sizes (with sparsity levels of \( S = 50 \) and \( S = 70 \)), best performing penalty parameter values are determined as \( \lambda = 0.01 \) and \( \alpha = 0.1 \).

3) Results: Before discussing the sparse patterns captured using CMF-SPOPT, we first illustrate the factors extracted using the Singular Value Decomposition (SVD) of matrix \( X \). SVD decomposes \( X \) as \( X = U \Sigma V^T \), where \( U \) and \( V \) are orthogonal matrices corresponding to the left and right singular vectors, respectively, and \( \Sigma \) is a diagonal matrix with singular values on the diagonal. Figure 4(a) shows the scatter plot of \( u_1 \) vs. \( u_7 \) demonstrating that two apple groups can be almost separated using the seventh left singular vector. The goal in metabolomics studies is often to understand the reason for the separation; in other words, the metabolites responsible for the separation. Therefore, we plot the seventh right singular vector in Figure 4(b) to identify the significant features. However, capturing the significant features is difficult since this vector is dense.

In Figure 5, we illustrate the performance of CMF-SPOPT in terms of apple group separation by coupled analysis of \( X, Y \) and \( Z \). The scatter plot of \( a_1 \) vs. \( a_3 \) in Figure 5(a) shows that the first component can almost separate the two groups. In Figure 5, we can see the sparse patterns, i.e., \( b_1, c_1 \) and \( d_1 \), responsible for this separation. Unlike Figure 4(b), we can clearly identify the significant features in Figure 5(b). Through coupled analysis, we also get the sparse patterns relevant to apple groups in each data set. Results illustrated in Figure 5 are based on a 5-component CMF-SPOPT model, i.e., \( R = 5 \). If we decrease \( R \), none of the components can separate the apple groups. For \( R = 5 \) and \( R = 6 \), we get almost the exact same component for apple separation. As \( R \) increases, we lose the component responsible for the separation. In real data, unlike simulation studies, there are both common and uncommon components in coupled data sets; therefore, robustness of CMF-SPOPT to overfactoring (shown in Table I) is not enough to deal with the problem of determining \( R \). In this study, since we are interested in apple markers, we use a component number that captures the apple group separation.

In order to make sure that the sparse pattern in Figure 5(b) is really meaningful, we form a small matrix, \( \bar{X} \in \mathbb{R}^{I \times L} \), using only the features identified in \( b_1 \), where \( L = 14 \), and check the separation achieved by its SVD. We observe that using only 14 out of 1086 features, we can still separate the apple groups (results not shown but separation is similar to Figure 5(a)); therefore, these features are potential candidates for markers of apple intake.

We further study the sparse patterns captured by CMF-SPOPT from a biological perspective. Metabolites identified in the sparse patterns are shown in Figure 5(b) and Figure 5(c). Some of these have been verified by chemical standards while some of them are tentative identifications, further to be explored. Based on the identifications, we find that patterns in Figure 5 are related to apple-induced changes in the endogenous metabolism. These changes include an increase in circulating branched-chain and aromatic amino acids, an increase in circulating glycerol- and choline-containing lipids, a decrease in corticosteroids and possibly in androgens, and a decrease in lactate, hypoxanthin and free fatty acids. Several elements of this pattern indicate that the transition from the postprandial to the fasting state was delayed in apple-fed rats, with a slower increase in lactate and free fatty acids and a slower loss of amino acids and lipids from the blood. Moreover, apple feeding seems to
suppress the increase in corticosteroids and possibly also of androgens with fasting in support of an effect on genes involved in steroid metabolism that we have observed in these rats.

Using CMF-SPOPT, we were able to extract meaningful sparse patterns from LC-MS and NMR complementing each other and describing apple-induced changes in the metabolism.

IV. RELATED WORK

Simultaneous analysis of multiple matrices dates back to one of the earliest models aiming to capture the common variation in data sets, i.e., Canonical Correlation Analysis (CCA) [19]. CCA looks for the patterns in each data set that correlate well and it is, in that sense, different from coupled matrix factorization. This difference has been illustrated in a recent metabolomics study [20].

More in line with the formulation in (1), Levin [21] studied simultaneous factorization of Gramian matrices. Similarly, in signal processing, joint diagonalization of symmetric and Hermitian matrices has been a topic of interest [22]. Furthermore, principal component analysis of multiple matrices has been widely studied in chemometrics using various models, some with clear objective functions while some are based on heuristic multi-level approaches [5]. Badea [23] extended the formulation in (1) to simultaneous nonnegative matrix factorizations by extracting nonnegative factor matrices. Another line of work related to simultaneous matrix factorization is Generalized SVD and its extension to multiple matrices [24].

With the increasing interest in the analysis of multi-relational data, Singh and Gordon [8] and Long et al. [7] studied Collective Matrix Factorization for joint factorization of matrices. We can also consider tensor factorizations as simultaneous factorization of multiple matrices (see a recent survey for various tensor models [25]).

While coupled matrix factorization has been widely studied in many disciplines, a recent study by Deun et al. [26] is the only study that enforces sparsity on the factors within the coupled matrix factorization framework, to the best of our knowledge. This work considers various penalty schemes such as the lasso, elastic net, group lasso, etc., and it is the most related to what we propose, or more specifically to (2). The main differences are (i) we do not enforce orthogonality constraints on factor matrix $A$, as in [26], (ii) while alternating least squares is used in [26], we use an all-at-once approach solving a smooth approximation of the objective in (2), and (iii) we extend our formulation to joint analysis of incomplete data as in (4).

V. CONCLUSIONS

While we can collect huge amounts of data using different platforms in metabolomics, we are still lacking the data mining tools for the fusion and analysis of these data sets. In this paper, we have formulated data fusion as a coupled matrix factorization model with penalties to enforce sparsity with a goal of capturing the underlying sparse patterns in coupled data sets. We have also discussed the extension of the proposed model to coupled analysis of incomplete data. In order to fit the model to coupled data sets, we have developed a gradient-based optimization algorithm solving for all factor matrices simultaneously. Using numerical experiments on simulated data, effectiveness of the proposed
approach in terms of capturing the underlying sparse patterns is demonstrated. We have also illustrated the usefulness of the proposed method in a metabolomics application, where potential markers for apple intake are identified through coupled analysis of LC-MS and NMR data. The main limitation of our formulation is to impose the same level of sparsity on different data sets. We plan to extend our model to different levels of sparsity in coupled data sets; in other words, to use different $\lambda$ values for different matrices. This may require reformulation of the model in order to deal with the scaling ambiguity problem.

VI. ACKNOWLEDGMENTS

This work is funded by the Danish Council for Independent Research — Technology and Production Sciences and Sapere Aude Program under the projects 11-116328 and 11-120947.

REFERENCES


