Multi-tensor factorization

Suleiman A. Khan 1, Eemeli Leppäaho 1, Samuel Kaski 1,2

December 16, 2014

1 Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University
2 Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland

Abstract

We introduce Bayesian multi-tensor factorization, a model that is the first Bayesian formulation for joint factorization of multiple matrices and tensors. The research problem generalizes the joint matrix-tensor factorization problem to arbitrary sets of tensors of any depth, including matrices, can be interpreted as unsupervised multi-view learning from multiple tensors, and can be generalized to relax the usual trilinear tensor factorization assumptions. The result is a factorization of the set of tensors into factors shared by any subsets of the tensors, and factors private to individual tensors. We demonstrate the performance against existing baselines in multiple tensor factorization tasks in structural toxicogenomics and functional neuroimaging.

1 Introduction

Matrix and tensor factorization methods have been studied for data analysis and machine learning for decades. These methods decompose a single data set into a low dimensional representation of factors that explain the variation in it. With linked data sets emerging, joint factorization of multiple data sources is now gaining significant attention.

Joint factorization of multiple matrices integrates information from multiple coupled data sets. It decomposes them into underlying latent components or factors, taking advantage of the common structure between all of them. For the simplest case of two paired matrices, canonical correlation analysis finds latent variables that capture the shared variation explaining them (Hotelling 1936; Hardoon et al 2004; Bach and Jordan 2005). While canonical correlation analysis searches for common patterns between two data matrices, its straightforward extensions have limited applicability in multiple coupled matrices. Recently, a multi-view method called group factor analysis (GFA; Virtanen et al 2012), has been presented for decomposing multiple paired matrices. GFA decomposes multiple cou-
pled matrices identifying both the variation patterns shared between some of the data sets as well as those specific to each.

Tensor factorizations have also been considered as a means of analyzing multiple matrices by coupling them together as slabs of a tensor. These factorizations, however, are more general and are able to take advantage of the natural tensor structure of the data. A host of low-dimensional tensor factorization methods have been proposed earlier (see Kolda and Bader (2009) for a review). The most well-known are the CANDECOMP/PARAFAC (CP, Carroll and Chang (1970); Harshman (1970)) and the Tucker family of models (Tucker 1966; Kiers 1991). CP assumes a trilinear structure in the data and is easier to interpret, while the Tucker family defines more generic models for complex interactions.

However, neither the tensor factorization nor the joint matrix factorization is able to factorize mixed and partially linked data sets. Recently, fusion of partially coupled data sets has been discussed, for example to predict the values in a tensor with side information from a matrix, or vice versa. For example, in Acar et al (2013b), used metabolomics data of fluorescence emission × excitation measurements and NMR recordings of several human blood samples to form a coupled tensor and a matrix, to demonstrate that joint factorization outperforms individual factorization. The concept of such multi-block decompositions was originally introduced by Smilde et al (2000), and proposed by Harshman and Lundy (1994), though the recent formulation by Acar et al (2011, 2013b) has brought coupled matrix tensor factorizations to practical use.

We call this general research problem multi-tensor factorization (MTF) and present the first Bayesian formulation for matrix-tensor factorization, that simultaneously extends to be the first joint factorization of multi-view tensors and matrices. We also present the first generalized formulation of multi-tensor factorization to arbitrary tensors and introduce a relaxed low-dimensional decomposition that allows the tensor to factorize flexibly. Our model decomposes multiple co-occurring matrices and tensors into a set of underlying factors that can be shared between any subset of them, with an intrinsic solution for automatic model selection. Finally, we demonstrate the use of the method in novel coupled matrix-tensor factorization applications, including structural toxicogenomics and stimulus-response prediction in neuroimaging.

The rest of the paper is structured as follows: In Section 2 we start by formulating the special case of a single matrix and single tensor factorization, inferring components that are shared between both of them, or are specific to either one. In Section 3 we present our Bayesian model that extends to multiple paired tensors and matrices. In Section 4 we introduce an extension of our new Bayesian solution of Section 3 that automatically tunes the decomposition structure for the data. We propose a generic formulation in Section 5 and discuss the special cases and related works in 6. We validate the performance of our models in various technical experiments in Section 7 and demonstrate their applicability in a neuroimaging stimulus-response relationship study and in a novel structural toxicogenomics setting in Section 8. We conclude with discussion in Section 9.
Figure 1: CP factorization (left) is analogous to matrix factorization (right), each matrix slab of the tensor is just scaled by the corresponding value of each component in an additional $U$.

**Notations:** We denote a tensor as $\mathcal{X}$, a matrix as $X$, vector $x$ and a scalar as $x$. The Mode-1 product $\times_1$ between a tensor $\mathcal{A} \in \mathbb{R}^{K \times D_1 \times 1}$ and a matrix $B \in \mathbb{R}^{N \times K}$ is the projected tensor $(\mathcal{A} \times_1 B) \in \mathbb{R}^{N \times D_1 \times 1}$, that reshapes the first mode of the tensor. A Mode-2 product $\times_2$ similarly reshapes the 2nd mode. A reshaped Khatri Rao product $\odot$ of two matrices, $A \in \mathbb{R}^{N \times K}$ and $C \in \mathbb{R}^{L \times K}$, is the “column-wise matched” outer product of the K vector-pairs that results in the tensor $(A \odot C) \in \mathbb{R}^{N \times K \times L}$. The outer product of two vectors is denoted by $\circ$ and the element-wise product by $\ast$. The *order* of a tensor is the total number of axes, modes or ways in the tensors, while tensor *rank* is the smallest number of rank-1 component tensors that generate it. For notational simplicity we present the models for third order tensors only; this includes matrices, for which the dimension of the third mode is one. As matrices are also tensors they can be equivalently represented as $\mathcal{X} \in \mathbb{R}^{N \times K \times 1}$ where the third dimension size is 1.

## 2 Matrix tensor factorization

We formulate the joint matrix tensor factorization problem as the identification of a combined low-dimensional representation of the matrix $\mathcal{X}^{(1)} \in \mathbb{R}^{N \times D_1 \times 1}$ and the tensor $\mathcal{X}^{(2)} \in \mathbb{R}^{N \times D_2 \times L}$ such that each underlying factor is either shared by both the matrix and the tensor, or is private to one of them. The matrices and tensors can jointly be referred to as different *views* of the data, analogously as used in multi-view learning. The shared factors represent variation that is common in both the views, while specific components capture the *view-specific* variation.

The joint factorization can be defined, with a common set of low-dimensional latent variables $Z \in \mathbb{R}^{N \times K}$, matrix-specific loadings $\mathcal{W}^{(1)} \in \mathbb{R}^{K \times D_1 \times 1}$ and tensor-specific loadings $\mathcal{W}^{(2)} \in \mathbb{R}^{K \times D_2 \times L}$, for each view $t$ as

$$
\mathcal{X}^{(t)} = \mathcal{W}^{(t)} \times_1 Z + \epsilon^{(t)},
$$

where $\epsilon^{(1)} \in \mathbb{R}^{N \times D_1 \times 1}$ is a matrix representing noise, while $\epsilon^{(2)} \in \mathbb{R}^{N \times D_2 \times L}$ is the noise tensor. The factorization of the tensor $\mathcal{X}^{(2)}$ into $\mathcal{W}^{(2)} \times_1 Z$, where $\mathcal{W}^{(2)}$ is unconstrained, is equivalent to Tucker-1 factorization (Kiers, 1991), which is analogous to matrix factorization of a matricized tensor. This factorization still
has a huge number of parameters, and the loadings $W^{(2)}$ can be further factorized to model tensorial interactions. A popular choice is the CP-type formulation which captures trilinear relationships (Fig. 1) and has the advantage of being easier to interpret. CP is equivalent to a sum of rank-1 tensors where each rank-1 tensor is the outer product of vector loadings in all modes. This rank-1 component decomposition of CP make it the choice in most studies. For other useful properties of CP, like the so-called intrinsic axis property from parallel proportional profiles, see Cattell (1944) and Harshman (1970).

Assuming a CP-type decomposition, the loading tensor $W^{(2)}$ is factorized into a tensor product of two latent variable matrices $U^{(2)} \in \mathbb{R}^{L_2 \times K}$ and $W^{(2)} \in \mathbb{R}^{D_2 \times K}$. Noting that matrix factorization is a special case of CP, both of them can be expressed in the same way, reformulating the joint decomposition as

$$X^{(t)} = (Z \odot U^{(t)}) \times_2 W^{(t)} + \epsilon^{(t)},$$

where the factorization of matrix $X^{(1)}$ corresponds to Eq. 2 with $U^{(1)} \in 1^{1 \times K}$.

A key property of our joint factorization is to have factors that can be shared by both the matrix and the tensor, or be specific to either one of them. This can be achieved by imposing a group-sparse prior on the view-specific loadings $W^{(t)}$, similar to that in Virtanen et al (2012). The group-sparse prior controls which of the $k$ latent variables are active (i.e., non-zero) in each view. A component active in both views is said to be shared between them, while the component active in one captures variation specific to that particular view. This formulation allows the matrix and tensor to be decomposed comprehensively, while identifying the common and specific patterns.

### 3 Multi-tensor factorization

We present the first Bayesian treatment of matrix tensor factorization while simultaneously extending it to collections of multi-view matrices and tensors, coupled on a common set of samples. The proposed factorization framework is very general, and could as well go by the names collective or collaborative matrix-tensor factorization, or other similar names. Acknowledging that matrices are two-dimensional tensors, we formulate the model family in a simpler but equivalent way as multi-tensor factorization.

Fig. 2 illustrates the MTF problem for one matrix and two tensors. The samples couple one mode across the collection, and two modes for the tensors. The task now is to perform a joint decomposition of the matrix and tensors, distinguishing also between the shared, and private components. Assuming an underlying CP decomposition for the tensor as in Eq. 2 we perform an unsupervised joint factor analysis and CP-type decomposition of the matrices and tensors, respectively. The joint decomposition is characterized by (i) $Z$, a common set of latent variables between all the views (matrices and tensors), (ii) $U$, the latent variables that model the third mode common to the two tensor views only ($X^{(2)}$, $X^{(3)}$) and (iii) $W^{(t)}$, ...
Figure 2: Multi-tensor factorization projects a set of coupled matrices and tensors into a joint low-dimensional space. In this case, one matrix $X^1$ and two tensors $X^2$ and $X^3$ are decomposed into a common set of latent variables $Z$ and view-specific projection matrices $W^{1,2,3}$, with the two tensor decompositions having an additional joint set of latent variables $U$ for the 3rd mode. The $W$’s control the activity of each component in each view, with black representing a component active in a view, white being switched off. A component active (black) in two or more views captures common patterns of variation.

the view-specific loadings that control which patterns from $Z$ and $U$ are reflected in each view.

This factorization can be seen as a joint FA-CP decomposition where variation patterns can be shared between matrices and tensors, or be specific to each. There are two main characteristics that drive this. First, the decomposition of all the matrices and tensors is coupled with the latent variables $Z$ that capture the common response patterns, enabling the model to capture dependencies between all the views for learning a better factorization. Second, the decomposition allows each factor to be active in any combination of matrices and tensors. This gives the formulation the ability to capture the underlying dependencies between all or some of the data views as well as segregates them from the variation that is specific to just one view, often interpretable as (structured) noise. We learn these dependencies between the views in a fully data-driven fashion, automatically inferring the exact number and nature of each type of dependency.

Formally, we define multi-tensor factorization for the collection of $t = (1, \ldots, T)$ paired tensors and matrices, $X^{(1)}, X^{(2)}, \ldots, X^{(T)} \in \mathbb{R}^{N \times D_t \times L_t}$, each referred to as a data view. The central assumption, which will be relaxed later in Section 5, is that all the views are coupled in one common mode and the tensors in two. Assuming normal distributions and conjugate priors, the generative model underlying the joint matrix-tensor factorization can be expressed as

$$x_{n,:,:}^{(t)} \sim \mathcal{N}\left(W^{(t)}(z_n * u_n^{(t)}), \frac{1}{\tau_t}I\right)$$

$$Z, U^{(1)} \sim \mathcal{N}(0, I)$$

$$U^{(2)} \sim 1_{1 \times K}$$

$$w_{d,k}^{(t)} \sim h_{t,k} \mathcal{N}\left(0, (\alpha_{d,k}^{(t)})^{-1}\right) + (1 - h_{t,k})\delta_0$$

$$h_{t,k} \sim \text{Bernoulli}(\pi_k)$$

(3)
\[ \pi_k \sim \text{Beta}(a^\pi, b^\pi) \]
\[ \alpha_{d,k}^{(t)} \sim \text{Gamma}(a^\alpha, b^\alpha) \]
\[ \tau^{(t)} \sim \text{Gamma}(a^\tau, b^\tau) , \]

where the indicator variable \( \beta_t \) equals 1 for tensors and 2 for matrices.

This formulation is similar to Eq. 2 with the separate noise precision \( \tau_t \) for each view. For a matrix view, \( \tau_t \) samples the noise matrix, and a tensor for a tensor view. The Gaussian latent variables \( Z \) are shared across all views, while the indicator variable \( \beta_t \) controls the third mode projections \( U \) to be random variables only for tensors, and constant unity vector for matrices. The \( U \) hence depend on the matrix views only indirectly (via \( Z \)).

The binary variable \( h_{t,k} \) controls which components are active in each view by switching the \( w_{t,k}^{(t)} \) on or off. This is achieved via the spike and slab prior which samples from a two-part distribution (Mitchell and Beauchamp [1988]). We center the spike at zero (\( \delta_0 \)) allowing the components to be shut down, while the slab is sampled from an element-wise formulation of the ARD prior (Neal [1996]), \( \alpha_{d,k}^{(t)} \) enabling active components with additional feature level sparsity. This way the \( h_{t,k} \) effectively govern the sharing of components across all the \( T \) views, irrespective if they are matrices or tensors. The view-specific loading matrices, \( W^{(t)} \), capture active patterns in each data view, while containing zeros for all inactive components, as illustrated with the white and black patterns in Fig. 2.

The learning of \( h_{t,k} \) in a data-driven way automatically gives the algorithm the power to segregate between components that are shared between the matrices and the tensors from those specific to only one of them. This is achieved in an unbiased fashion by placing an uninformative beta-Bernoulli prior on \( h_{t,k} \) (default parameters being \( a^\pi = b^\pi = 1 \)). The formulation also allows the model to learn the total cardinality of each dataset, as well as of all the data sets combined. This is accomplished by setting \( K \) to a large enough value that for a few \( k \), \( h_{t,k} \) goes to zero in all \( t \) views. Such components will be referred to as empty components, and the presence of empty components indicates that \( K \) was large enough to model the data. The effective cardinality of the data set collection is then \( K \) minus the number of empty components. The element-wise sparsity within each active component \( (\alpha_{d,k}^{(t)}) \) helps in better regularizing the solution. Inference of the model posterior was performed using Gibbs sampling.

### 4 Relaxed multi-tensor factorization

The trilinear CP structure places a strong assumption on the factorization of a tensor, namely that the projection weights in the second dimension are identical for each slab (3rd mode) with just a varying scale (\( U \), Fig. 1). While in many applications this decomposition has been shown to be useful (see e.g. Latchoumane et al [2012]), the assumption may not follow exactly the structure in all datasets. For
Figure 3: **Left:** An illustration of joint factorization of a matrix and a tensor. **Right:** Generation of the projection matrices corresponding to the tensor, $\mathbf{W}^{(2)}$, is shown in more detail ($D_2 = 5$ and $K = 2$). MTF does a trilinear CP decomposition, whereas rMTF allows deviation from this, governed by a precision parameter $\lambda$. Moderate $\lambda$ values induce a relaxed trilinear factorization, or equivalently a multiple matrix factorization, where some information is shared between the slabs of $\mathbf{W}^{(2)}$. High values correspond to a trilinear factorization and low values to multiple matrix factorization.

a more general, relaxed trilinear factorization, we present a novel formulation that allows MTF to capture variation that may not be strictly trilinear. More specifically, the relaxed formulation aims to decompose the tensor $\mathbf{W}^{(t)}$ of Eq. 1 in a flexible fashion, allowing identification of trilinear structure (characteristic of CP factorization), along with structure that is specific to each slab of the tensor (bilinear, matrix structure); but also structures that are along a continuum between the two extremes, as shown in Fig. [3].

We formulate a Bayesian solution for the relaxed MTF (rMTF) problem, allowing flexible factorization. The tensors $\mathbf{X}^{t=(1:T)}$ are factorized jointly, capturing factors that are shared between all, some or one of the views. The formulation is relaxed in the sense that the tensors $\mathbf{X}^{(t)}$ have a hierarchical decomposition that allows the model to flexibly tune between the generic matrix factorization and the trilinear CP factorization of the tensors. The distributional assumptions of our model are:

$$
\mathbf{x}^{(t)}_{n:l} \sim \mathcal{N} \left( \mathbf{W}^{(t)}_l \mathbf{z}_n, \frac{1}{r^{(t)}_l} \mathbf{I} \right)
$$

$$
\mathbf{w}^{(t)}_{i,k} \sim h^{(t)}_{i,k} \mathcal{N} \left( u^{(\beta)}_{i,k} v^{(\alpha)}_k, \lambda^{-1} \mathbf{I} \right) + \left( 1 - h^{(t)}_{i,k} \right) \delta_0
$$

$$
\mathbf{Z}, \mathbf{U}^{(1)} \sim \mathcal{N}(0, \mathbf{I})
$$

$$
\mathbf{U}^{(2)} = \mathbf{1}_{1 \times K}
$$

$$
\nu^{(t)}_{i,k} \sim \mathcal{N} \left( 0, \left( \alpha_{d,k}^{(t)} \right)^{-1} \right)
$$

$$
\beta_{i,k} \sim \text{Bernoulli}(\pi_k)
$$

$$
\pi_k \sim \text{Beta}(\alpha^\pi, \beta^\pi)
$$

$$
\lambda_1 \sim \text{Gamma}(a^\lambda, b^\lambda)
$$

$$
\lambda_2 = \infty
$$
Figure 4: Tensor collections with arbitrary pairing between the modes can be assembled into a single large sparse tensor. If the first and the second mode contain groups $\{g_1, \ldots, g_G\}$ and the third mode groups $\{f_1, \ldots, f_F\}$, the whole data collection forms a $\sum_{i=1}^{G} |g_i| \times \sum_{i=1}^{G} |g_i| \times \sum_{j=1}^{F} |f_j|$ tensor. An illustration of the two first modes of this tensor is presented on the right. A group sparse factorization for this tensor amounts to a MTF generalized to arbitrarily paired tensor collections.

$$
\alpha_{d,k}^\gamma \sim \text{Gamma}(a^\alpha, b^\alpha)
$$
$$
\tau_{l}^{(t)} \sim \text{Gamma}(a^\tau, b^\tau)
$$

The key aspect of the model structure is the parameter $\lambda$ describing the similarity of the tensor slabs. The difference to regular MTF is illustrated in Fig. 3-Right. High $\lambda$ values indicate a factorization that is primarily driven by the low-rank CP structure and is effectively trilinear, whereas low values correspond to highly parameterized bilinear multiple matrix factorization. In the tensor formulation, Tucker-1 factorization closely resembles the bilinear multiple matrix factorization, and is considered the least restrictive form of tensor factorization (Kiers, 1991). By default, $\lambda$ will have a relatively uninformative prior ($a^\lambda = b^\lambda = 1$), enabling data driven inference of the tensor structure.

Alternatively, informative priors can be considered as well, if there indeed is some prior information about the structure of the tensors. Additionally, specifying $\lambda_k$ for each component or $\lambda_l$ for each tensor slab would allow learning interesting information about the data structure, namely how strongly each component or slab, respectively, is associated to the trilinear tensor structure. The inference for rMTF is performed with Gibbs sampling.

5 Generalized multi-tensor factorization

The notion of coupled matrices and tensors can be formulated for all modes of all tensors, allowing decomposition of arbitrarily coupled data sets, for investigation
of factors shared and specific to each. To this end, we formulate the general problem of multi-tensor factorization, for data collections consisting of matrices and tensors paired in any user-defined fashion.

The joint factorization task in such multi-mode blocks can be framed as identification of a low-dimensional representation for each of the data modes. This is enabled by the observation that the distinction between samples and dimensions vanishes, as all modes of the data become analogous. Fig. 4 illustrates the formulation with an example of two matrices \(X^{(2)} \in \mathbb{R}^{D_1 \times D_2 \times 1}, X^{(3)} \in \mathbb{R}^{D_4 \times D_5 \times L_4}\), paired in a non-trivial way. The task in this case is to find \(K\) factors to represent each of the dimension blocks \(g_1, g_2, g_3, g_4, f_1, f_4\), while capturing the common as well as distinct activity patterns that link the data sets \(X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}\) together. To solve the task, we represent the entire data collection as a symmetric tensor \(\hat{X} \in \mathbb{R}^{\sum D_i \times \sum D_i \times \sum L_i}\), and the \(K\) factors as the low dimensional tensor \(\mathcal{W} \in \mathbb{R}^{\sum D_i \times K \times \sum L_i}\), which has a strict block-structure, active only for regions corresponding to data sets being modeled.

Formally, for \(m\) data sets collected into the tensor \(\hat{X}\) the model is

\[
\hat{X}_{i,:,:} \sim \mathcal{W}_{i,:,:} \mathcal{W}_{i,:,:}^T,
\]

where the block-structure \(\{g_1, g_2, g_3, g_4, f_1, f_4\}\) is imposed by the binary variable \(h_{b,k,l}\) for \(k \in 1 \ldots K, l \in 1 \ldots L\) via a spike and slab prior. The relaxed formulation of Section 4 is embedded by assuming that the \(\sum L_i\) slabs of \(\mathcal{W}\) are drawn from the mean matrix \(\mathcal{V} \in \mathbb{R}^{\sum D_i \times K}\) as

\[
w_{d,k,l} \sim h_{b_d,k,l} \mathcal{N}(v_{d,k},u_{l,k},\lambda^{-1}_{\beta_l}) + (1 - h_{b_d,k,l})\delta_0
\]

\[
U \sim \mathcal{N}(0,1)
\]

\[
v_{d,k} \sim \mathcal{N}(0, (\alpha_{d,k})^{-1})
\]

where \(b_d\) denotes which group feature \(d\) belongs to, \(\beta_l\) denotes whether slab \(l\) belongs to a tensor (\(\beta_l = 1\)) or a matrix (\(\beta_l = 2\)) and the other priors remain unchanged.

The binary variable \(h_{b_d,k,l}\) learns in which group each component is active, producing the block-component activation plot that extends also to slabs for tensor data sets. The \(\lambda\) again controls the balance between the trilinear and Tucker-1 structure in the data. Model specification is completed by assuming normal distributions for \(\hat{X}, V\) and a data view specific noise precision \(\tau^{(t)}\).

The key characteristic here is that group sparsity controls the activation of each latent block-component pair instead of the data set-component pair; therefore a component’s contribution in a data set can be switched off in multiple ways. For example, in matrix \(X^{(2)}\), the component \(k\) can be switched off if either \(h_{1,k,1} = 0\) or \(h_{3,k,1} = 0\). For tensors, the switching notion extends to each of the \(L_i\) slabs. This specification makes the model fully flexible and allows components with all possible sharing and specificity patterns to be learned, given enough regularization.
6 Related work

The MTF problem and our solution for it are related to several matrix and tensor factorization techniques. In the following we discuss existing techniques that solve special cases of the multi-tensor factorization problem, and relate to our work.

For a tensor coupled with one or more matrices, our MTF model can be seen as a Bayesian coupled matrix-tensor factorization (CMTF) method, which can additionally automatically infer the number and type of the components in the data, and enforce feature-level sparsity for improved regularization and interpretability. In this line of work, ours is closest to the non-probabilistic CMTF of Acar et al (2011, 2013a,b). They assumed an underlying CP decomposition for tensors too, and used a gradient-based least squares optimization approach. In their recent work, Acar et al (2013a, 2014) enforced an $l_1$ penalty on the components assuming they can be shared or specific to data sets. However, unlike ours, they still required the data cardinality to be pre-specified. Determining the cardinality of tensors has been considered a challenging problem (Kolda and Bader, 2009), and our method presents an intrinsic solution for this. Researchers have also used matrices as side information sources to tensors in CMTF to show improved factorization quality (Narita et al, 2012), while others have used generalized linear models for solving coupled link prediction and audio processing tasks (Yılmaz et al, 2011; Ermis et al, 2013). Recently, solutions have been presented for speeding up the computation of coupled matrix tensor factorization algorithms on big data (Beutel et al, 2014; Papalexakis et al, 2014). However, none of the methods has been extended for multi-view matrix tensor coupling.

When all the tensors have $L_t = 1$ and they are paired in the first mode, our framework reduces to the group factor analysis (GFA) problem presented in (Virtanen et al, 2012). GFA has been generalized to allow pairings between arbitrary data modes under the name collective matrix factorization (CMF) (Klami et al, 2014).

In tensor factorization research, a multi-view problem was recently studied under the name of multi-view tensor factorization (Khan and Kaski, 2014). The goal there was to perform a joint CP decomposition of multiple tensors to find dependencies between data sets. This method can be seen as a special case of our model, when all data views are only tensors of the same order, paired in two modes and assume a strict CP-type factorization.

7 Technical demonstration

In this section we demonstrate the proposed MTF methods on artificial data. This is done in comparison to multi-view matrix factorization method group factor analysis (GFA) (Virtanen et al, 2012) by transforming the tensors $X^{(t)} \in \mathbb{R}^{N \times D_t \times L_t}$ into $L_t$ matrices $X^{(1)} , X^{(2)} \ldots X^{(L_t)} \in \mathbb{R}^{N \times D_t}$, one for each slab of the tensor. In this setting, GFA corresponds to a joint matrix and Tucker-1 tensor factoriza-
Figure 5: Illustration of MTF inference on simulated data when the projections of the tensor are given the shape of a sine wave (“True”). The solid curve denotes the first slab and dashed the second. **Top:** MTF infers the correct shape more accurately than multi-view matrix factorization (GFA). **Bottom:** MTF cannot compensate for tensor slabs deviating from the common signal, whereas the relaxed version (rMTF) can, detecting the true parameters more accurately.

7.1 Visual example

We start with an experiment that illustrates the properties of MTF. We generated $N = 300$ data samples from MTF (Eq. (3)), one matrix view with $D_1 = 50$ and one tensor view with $D_2 = 50$ and $L = 30$. A total of 11 components were used to generate the data: 1 fully shared, 2 specific to the matrix and 8 to the tensor. The fully shared component was given the shape of a sine wave in the second dimension ($W$), as shown in Fig. 5 (top, black curves) for the first 25 features of two tensor slabs. MTF is able to detect this structure considerably more accurately than GFA, which assumes a bilinear structure in the data instead of the trilinear tensor. The inferred posterior means are shown in Fig. 5 (top). MTF detected the cardinality and component activation correctly, while GFA returned four shared components, out of which the one closest to the true parameters is shown.

To illustrate the advantages of relaxed MTF, we add some distortion to the sine wave; the distortion is $w^p$, where $p = \{0.5, 1.5\}$ for the first two tensor slabs, and an evenly spaced series between $p = \{0.3, 1.7\}$ for the rest. The weights corre-
Figure 6: A simulation study showing the performance (lower is better) as a function of the proportion of trilinearity (vs. bilinearity) in the data; 0 corresponds to (bilinear) multi-view matrix factorization and 1 to perfectly trilinear matrix tensor factorization. As expected, MTF has poor performance when the data do not match the modeling assumptions (on the left), and top-level when they do (on the right). Matrix factorization method GFA is the ideal model when the data have close to bilinear structure, and relaxed MTF is generally close to the better one of these two. It suffers somewhat from having a more general parametrization, but is the most accurate model in the mid-region (from 0.6 to 0.75). Null refers to predicting with the feature-wise mean of the training data.

7.2 Spectrum from bilinear to trilinear factorization

In this experiment we evaluate the performance of MTF and rMTF on the continuum between multi-view matrix factorization and matrix tensor factorization. The trilinear tensor factorization for slab \( l \) of tensor \( t \) is of the form \((Z \odot u_l) \times_2 W^{(l)}\), whereas the bilinear multi-view matrix factorization can be presented as \(ZW^{(l)}\), where the matrix \(W^{(l)}\) is a priori independent from all the other data views. We studied the case where neither of these assumptions is correct, but the true factorization is between the assumptions of the two models. For this, we generated \( N = 20 \) samples of data: one matrix with \( D_1 = 20 \) and one tensor with \( D_2 = 20 \) and \( L = 9 \). The data set was generated with one fully shared component, 2 specific components for the matrix and 8 for the tensor. The generative model used was a weighted sum of the bilinear and trilinear factorizations. The performance of the MTF models can be seen in Fig. 6; the performance was quantified with the pre-
diction RMSE on a left-out test data set with \( N = 100 \), predicting one unobserved tensor slab from the others.

8 Applications

In this section we demonstrate the proposed MTF methods on two applications: functional neuroimaging and structural toxicogenomics. To illustrate the strengths of the new methods, we compare them with tensor factorization methods that are most closely related to them. In particular, we compare with coupled matrix tensor factorization (Acar et al. 2013b), which decomposes a tensor along with a coupled matrix as side information. The available implementation uses CP as the underlying factorization for the tensor, as does our MTF. Additionally, we compare against an asymmetric version of coupled matrix tensor factorization (ACMTF) (Acar et al. 2013a, 2014), which allows both private and shared components in the data collection. CMTF and ACMTF are the closest existing tensor baselines and are non-probabilistic formulations. We also compare our method to a multi-view matrix factorization method group factor analysis (GFA) (Virta et al. 2012) by transforming the tensors \( X^{(t)} \in \mathbb{R}^{N \times D_t \times L_t} \) into \( L_t \) matrices \( X^{(1)}, X^{(2)}, \ldots X^{(L_t)} \in \mathbb{R}^{N \times D_t} \), one for each slab of the tensor, as this corresponds to a joint matrix and Tucker-1 tensor factorization.

The model complexity was determined in a data-driven way, that is \( K \) was set high enough so that some of the inferred components were shut down. The model parameters \( a^\pi, b^\pi, a^\lambda, b^\lambda \) were initialized to 1 to represent uninformative priors. Feature level sparsity was assumed with parameters \( a^\alpha, b^\alpha \) set to \( 10^{-3} \), while high noise in the data was accounted by initializing the noise hyperparameters \( a^{\tau}, b^{\tau} \) for a single-to-noise ratio of 1. All remaining model parameters were learned. CMTF and ACMTF were run with \( K \) values inferred from MTF as they are unable to learn \( K \). ACMTF was run with the author recommended sparsity setting of \( 10^{-3} \) (Acar et al. 2013a). For all the models, the predictions for missing data were averaged over 7 independent sampling chains/runs to obtain robust findings.

8.1 Functional neuroimaging

A key task in many neuroimaging studies is to find the response related to a stimulus. This is an interesting problem in natural stimulation and multi-subject settings in particular. MTF can be applied in this scenario directly, as the stimulus can generally be represented with a matrix of \( N \) samples (time points) and \( D_1 \) features, whereas the imaging measurements are a tensor with \( N \) samples, dimension \( D_2 \) (e.g. MEG channels) and depth \( L \) (subjects). We analyzed a data set presented by Koskinen and Seppälä (2014), where \( L = 9 \) subjects (one out of ten omitted due to unsuccessful recordings) listened to an auditory book for 60 minutes, while being measured in a magnetoencephalography (MEG) device. In this context, analysis...
with multi-matrix factorization methods (with subjects regarded as different data views) would assume that the subjects \textit{a priori} do not have any shared information. MTF, on the other hand, aims to decompose the data such that latent time series (components) have equal feature weights for all the subjects, just scaled differently. Although the imaging device is the same for all the subjects, they will not share either the exactly same brain structure or functional responses. This makes rMTF a promising model for neuroimaging applications.

The data set was preprocessed in a similar fashion as in \cite{Koskinen2014}. Namely, the 60 minutes of MEG recordings were wavelet-transformed with central frequency 0.54, decreasing the sample size to $N = 28547$. The recordings were preprocessed with the signal-space-separation (SSS) method \cite{Laulu2004} and furthermore with PCA (jointly for all the subjects) to reduce the dimensionality from 204 (MEG channels) to the number of degrees of freedom left after the SSS procedure ($D_2 = 70$). As there is a delay in brain responses corresponding to the stimulus, the mel-frequency cepstrum coefficients (MFCC, computed with Matlab toolbox \textit{voicebox}) describing the power spectrum of the auditory stimulus ($D_1 = 13$) were shifted to have maximal correlation with the response, and then downsampled and wavelet-transformed to match the MEG recordings.

We inferred the matrix-tensor decompositions with the two methods introduced in this paper and the three comparison methods. The decomposition was inferred from the $n$ first measurements, and the models were then used to predict all the later MEG measurements given the later audio. Relaxed MTF and GFA were run with $K = 500$, leaving empty components with every training sample size. For MTF, even $K = 700$ (larger than $\sum D_m$) was not sufficient, suggesting the data do not fully fit the strong CP assumptions, and we used a stronger regularizing prior

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{\textbf{Left:} The mean squared errors (MSE) on predicting (later) brain responses to stimuli as a function of (earlier) training data (x-axis). Predicting mean of training data results in MSE of 1. \textbf{Middle:} The feature weights, averaged over subjects, of a robust rMTF component shown in the MEG channel space (\cite{Hansen2010}). Activations are focused around the auditory areas. \textbf{Right:} The averaged feature weights of an unstable GFA component shown in the MEG channel space.}
\end{figure}
\[ \alpha^\pi = \frac{1}{\beta^\pi} = 10^{-3} \] and peaked noise prior for SNR of 1; ending up with around 500 active components). Due to memory requirements ACMTF was run with only \( K = 70 \), using over 10GB of RAM.

Relaxed MTF learned the stimulus-response relationship most accurately for a wide range of training data sizes (Fig. 7), whereas the trilinear factorization of MTF seems to be too strict and the multi-matrix factorization of GFA too flexible. CMTF showed similar performance as MTF once it was given enough training samples. The sparse solution of ACMTF did not deviate from null prediction even though different sparsity parameters and convergence criteria were tested. Relaxed MTF was significantly superior to all the other methods with at least \( p < 10^{-5} \) (t-test for 10 independent runs). For practical use on small data sets it is an important finding that rMTF learns as accurate a stimulus-response relationship as GFA with roughly half of the training set size. As overly long neuroimaging experiments tend to cause decreased signal to noise ratio (Hansen et al, 2010), relaxed MTF can offer significant benefit in this area.

The strongest finding of the factorization models in this application is the brain response to the energy of the speech signal. Of the 13 acoustic MFCC features, 9 are highly dependent on the signal energy and hence two-peaked, corresponding to words and breaks between the words. A robust component found in all the rMTF chains had similar two-peaked structure, and was found to be active in the auditory areas of the brain (Fig. 7, middle). With enough samples, GFA was able to detect this component robustly as well, but it produced more unstable components present in individual sampling chains only. These components had no clear structure in the MEG channels, as shown in Fig. 7 and are hence likely to be artifacts explaining noise in the recordings. No other robust components were found with rMTF, likely because we analyzed the relationship between the stimulus and the brain response at only one time lag. Various brain responses occur at different lags; in this experiment we focused on the initial response, simple auditory processing of the heard sound. For a more thorough neuroscientific analysis it would be important to take the temporal nature of the events more fully into account.

### 8.2 Structural toxicogenomics

We next analyzed a novel drug toxicity response problem, where the tensors arise naturally when gene expression responses of multiple drugs are measured for multiple diseases (different cancers) across the genes. The data contain three views, the structural descriptors of drugs (a matrix), measurement of post-treatment gene expression (a tensor), and drug toxicity (a tensor) as shown in Fig. 8.

In this setting, MTF can be used to answer two key questions: (1) which parts of the responses are specific to individual types of cancer and which occur across cancers, and which of these responses are related to known structural properties of the drugs; and (2) can we use the links from gene expression responses along with structural properties of drugs, to predict toxicity of an unseen drug.

For the first problem, the response patterns of drugs, if uncovered, can help
understand the mechanisms of toxicity (Hartung et al. 2012). The identification of links from structural properties of drugs (a matrix) to the gene/toxic responses (tensors) opens up the opportunity for drug-designers to better understand the functional effects of drugs’ structural fragments. As the existing CMTF methods do not benefit from multi-view tensors, it is interesting to explore how well our method predicts unseen responses by using multiple side-information sources and the structure in the data.

The data set contained three views (Fig. 8). The first contained the structural descriptors of \( N = 73 \) drugs. The descriptors known as functional connectivity fingerprints FCFP4 represent the presence or absence of a structural fragment in each compound. For this data set the drugs are described by \( D_1 = 290 \) small fragments, forming a matrix of 73 drugs by 290 fragments. The second view contains the post-treatment differential gene expression responses \( D_2 = 1106 \) of \( N = 73 \) drugs, as measured over multiple diseases or here cancer-types, \( L = 3 \). The third view contained the corresponding drug sensitivity measurements, \( D_3 = 3 \). The

Figure 8: Structural toxicogenomics data set with a matrix of drug descriptors paired with a tensor of gene expression responses of different diseases to the drugs, and the corresponding toxicity profiles.

Figure 9: Component activity plot for the structural-toxicogenomics data set. The model found 3 interesting components active in all the views, capturing patterns shared by structures of drugs, their disease-specific gene expression responses, and the toxicity measurements. The components active in the drug structures and gene expression only capture non-toxic responses of drugs.
Figure 10: An illustrative example: Component 1 captures the well-known heat-shock protein response. The model identifies the common structural descriptors of the drugs (left) as drivers of the biological (middle) and toxic (right) responses. The responses are shown for the top genes and all toxicity variables, and top drugs for all the three cell lines (rows), along with the $Z$, $U$ and $W$ loadings (shown at the ticks). The top genes and drugs are identified as those with highest loadings in $W$ and $Z$ respectively. This component links structures of the heatshock protein inhibitor drugs (circled) with strong upregulation of the heatshock protein genes (red) to high toxicity (green).

two tensors are paired with the common identity of $N = 73$ drugs and $L = 3$ cancer types, while the drug structure matrix is paired with the tensors on the common set of $N = 73$ drugs. The gene expression data were obtained from the ConnectivityMap \cite{Lamb2006} that contained response measurements of three different cancers: Blood Cancer, Breast Cancer and Prostate Cancer. The data were processed so that gene expression values represent up (positive) or down (negative) regulation from the untreated (base) level. Strongly regulated genes were selected, resulting in $D_2 = 1106$. The Structural descriptors (FCFP4) of the drugs were computed using the Pipeline Pilot software \cite{http://accelrys.com/products/pipeline-pilot/} by Accelrys. The drug screen data for the three cancer types were obtained from the NCI-60 database \cite{Shoemaker2006}, measuring toxic effects of drug treatments via three different criteria: GI50 (50% growth inhibition), IC50 (50% lethal concentration) and TGI (total growth inhibition). The data were processed to represent the drug concentration used in the connectivity map to be positive (when toxic) and negative indicating non-toxic.

MTF resulted in 3 components shared between the structural descriptors, the
Table 1: Structural toxicogenomics: toxicity prediction of an unseen drug given its structural descriptors and genomic responses. Average Prediction RMSE is given over the entire set of drugs, with lower values signifying better performance. MTF outperforms GFA, CMTF, ACMTF significantly with t-test p-values $< 10^{-3}$.

|        | MTF  | rMTF | CMTF | ACMTF | GFA  |
|--------|------|------|------|-------|------|
| Mean   | 0.583| 0.619| 0.717| 0.771 | 0.696|
| StdError| 0.053| 0.054| 0.069| 0.073 | 0.066|

gene expression and toxicity views, revealing that some patterns are indeed shared (Fig. 9). Model complexity was again selected, as in the previous section, by assigning $K$ large enough, here $K = 30$, such that then sparsity prior shuts some of them off. The 3 shared components form hypotheses about underlying biological processes that characterize toxic responses of drugs, and we find all three of them to be well linked to either established biology or potentially novel findings.

The first component captures the well-known heatshock protein response, of the three HSP90 inhibitor drugs (Fig. [10] left). The response is characterized by a strong upregulation of heatshock genes in all three cancers (Fig. [10] middle) and the corresponding high toxicity indications in GI50 (Fig. [10] right). The component identifies similarities between the three close structurally analogous drugs, which is in line with knowledge that the drugs directly bind to the HSP90 protein (Stebbins et al, 1997). The heatshock protein inhibition response has already been well studied for treatment of cancers (Kamal et al, 2003), evaluating its potential therapeutic efficacy. This trilinear MTF component could have been important in revealing the response, if the mechanism had not already been discovered.

The second component captures DNA damage response of several structurally similar cardiac glycoside drugs and a structurally different drug, bisacodyl, which is a laxative. Interestingly, our component found the response of bisacodyl to be specific to only one of the cancer types. The link of bisacodyl with cardiac glycosides has very recently been found (Iorio et al, 2010), but the possible cancer specificity, which comes out naturally with our approach, is new.

The third and final shared component captures a common response of protein synthesis inhibitors along with an anti-metabolite (8-azaguanine) drug. Interestingly, the response is specific to two of the three cancer types, namely blood and prostate. With 8-azaguanine having been used in blood cancer before (Colsky et al, 1955), our component opens up an interesting opportunity for its exploration in prostate cancer.

This application demonstrated the model’s ability to identify both established and novel links between multiple views. Systematic studies along these lines could be very valuable in precision medicine for targeting specific disease types, and in drug design when tailoring drugs to match a desired response profile.
We next evaluated the model’s ability to predict the toxicity response of a new drug, by using the gene expression data (tensor) along with the structural descriptors (matrix). Both are used as side information sources, coupled in a multi-view setting. This is done by modeling the dependencies between all the observed data sets and then using the learned dependencies to predict the toxicity response of a new drug, given the side information sources. The entire toxicity slab (shaded white in Fig. 8) for each drug is predicted using its gene expression and structural descriptors. We compared with the existing methods as baselines, as they are not defined for multi-view coupling. CMTF and ACMTF were run by transforming the gene expression tensor into matrices, one for each of the $L = 3$ slabs, and GFA was run by doing the same for both the gene expression and toxicity tensors. We performed leave one out prediction and report the average prediction error of unseen drugs (RMSE) in Table 1. The results demonstrate that MTF predicts drug toxicity of unseen drugs significantly better, confirming that it solves the task well, for which it was designed.

9 Discussion

We introduced the problem of multi-tensor factorization (MTF) and as its special case the first Bayesian formulation of joint matrix-tensor factorization, extending the former formulation further to multiple sets of paired matrices and tensors. Our model decomposes the data views into factors that are shared between all or some of the views, and those that are specific to each. It also learns the total number and type of the factors automatically for each data collection.

We simultaneously extended our novel formulation to explore a relaxed underlying tensor factorization problem, automatically moving between the CP-type of a trilinear model and a generalized variant of a Tucker-1 type of decomposition. The CP and the Tucker-1 decompositions fall out as a special case of the relaxed variant. This is important as Tucker-1, in particular, is suitable when data sets have minimal trilinear structure, while the CP-type trilinear decomposition in this paper has the advantage of being interpretable analogously to the matrix factorizations more familiar to most analysts.

We validated the models’ performances in identifying the correct components on simulated data, and illustrated that the relaxed factorization performs well when the structure of the data is unknown or not strictly CP or Tucker-1 type. The models’ performances were then demonstrated on a new structural toxicogenomics problem and on stimulus-response relationship analysis in a neuroimaging experiment, yielding interpretable findings matching with expected effects, recent discoveries, and potential new innovations. The experiments indicated that taking the appropriate structure of the data into account makes the results both more accurate and easily interpretable.

Our work opens up the opportunity for novel applications and integrative studies of diverse and partially coupled data views, both for predictive and feature
identification purposes.

Acknowledgments

This work was financially supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, grant no 251170; and ChemBio project, grant no 140057), the Department of Information and Computer Science (Aalto University) and the Finnish Graduate School in Computational Sciences (FICS). We would like to thank Miika Koskinen for his help with the neuroscience application. We acknowledge the computational resources of Aalto Science-IT project.

References

Acar E, Kolda TG, Dunlavy DM (2011) All-at-once optimization for coupled matrix and tensor factorizations. arXiv preprint arXiv:11053422

Acar E, Lawaetz AJ, Rasmussen MA, Bro R (2013a) Structure-revealing data fusion model with applications in metabolomics. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, IEEE, pp 6023–6026

Acar E, Rasmussen MA, Savorani F, Naes T, Bro R (2013b) Understanding data fusion within the framework of coupled matrix and tensor factorizations. Chemometrics and Intelligent Laboratory Systems 129:53–63

Acar E, Papalexakis E, Grdeniz G, Rasmussen M, Lawaetz A, Nilsson M, Bro R (2014) Structure-revealing data fusion. BMC Bioinformatics 15(1):239

Bach FR, Jordan MI (2005) A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley

Beutel A, Kumar A, Papalexakis EE, Talukdar PP, Faloutsos C, Xing EP (2014) Flexifact: Scalable flexible factorization of coupled tensors on hadoop. In: Mohammed Z, Zoran O, Pang Ning T, Arindam B, Chandrika K, Srinivasan P (eds) SIAM International Conference on Data Mining, pp 109–117

Carroll JD, Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. Psychometrika 35(3):283–319

Cattell RB (1944) Parallel proportional profiles and other principles for determining the choice of factors by rotation. Psychometrika 9(4):267–283

Colsky J, Meiselas LE, Rosen SJ, Schulman I (1955) Response of patients with leukemia to 8-azaguanine. Blood 10(5):482–492
Ermis B, Acar E, Cemgil AT (2013) Link prediction in heterogeneous data via generalized coupled tensor factorization. Data Mining and Knowledge Discovery pp 1–34

Hansen P, Kringelbach M, Salmelin R (2010) MEG: An introduction to methods. Oxford University Press

Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12):2639–2664

Harshman RA (1970) Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis. UCLA Working Papers in Phonetics 16:1–84

Harshman RA, Lundy ME (1994) Parafac: Parallel factor analysis. Computational Statistics & Data Analysis 18(1):39–72

Hartung T, Vliet EV, Jaworska J, Bonilla L, Skinner N, Thomas R (2012) Food for thought ... systems toxicology. ALTEX 29(2):119–128

Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3):321–377

Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, et al (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. Proceedings of the National Academy of Sciences 107(33):14,621–14,626

Kamal A, et al (2003) A high-affinity conformation of Hsp90 confers tumour selectivity on Hsp90 inhibitors. Nature 425(6956):407–410

Khan SA, Kaski S (2014) Bayesian multi-view tensor factorization. In: Calders T, Esposito F, Hüllermeier E, Meo R (eds) Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2014, Springer Berlin Heidelberg, pp 656–671

Kiers HA (1991) Hierarchical relations among three-way methods. Psychometrika 56(3):449–470

Klami A, Bouchard G, Tripathi A (2014) Group-sparse embeddings in collective matrix factorization. In: International Conference on Learning Representations

Kolda T, Bader B (2009) Tensor decompositions and applications. SIAM Review 51(3):455–500

Koskinen M, Seppä M (2014) Uncovering cortical MEG responses to listened audiobook stories. NeuroImage 100:263–270
Lamb J, et al (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science 313(5795):1929–1935

Latchoumane CFV, Vialatte FB, Solé-Casals J, Maurice M, Wimalaratna SR, Hudson N, Jeong J, Cichocki A (2012) Multiway array decomposition analysis of EEGs in alzheimer’s disease. Journal of Neuroscience Methods 207(1):41–50

Mitchell TJ, Beauchamp JJ (1988) Bayesian variable selection in linear regression. Journal of the American Statistical Association 83(404):1023–1032

Narita A, Hayashi K, Tomioka R, Kashima H (2012) Tensor factorization using auxiliary information. Data Mining and Knowledge Discovery 25(2):298–324

Neal RM (1996) Bayesian learning for neural networks. Springer-Verlag

Papalexakis EE, Mitchell TM, Sidiropoulos ND, Faloutsos C, Talukdar PP, Murphy B (2014) Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x. In: Mohammed Z, Zoran O, Pang Ning T, Arindam B, Chandrika K, Srinivasan P (eds) SIAM International Conference on Data Mining, pp 118–126

Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer 6(10):813–823

Smilde AK, Westerhuis JA, Boque R (2000) Multiway multiblock component and covariates regression models. Journal of Chemometrics 14(3):301–331

Stebbins CE, Russo AA, Schneider C, Rosen N, Hartl FU, Pavletich NP (1997) Crystal structure of an Hsp90–geldanamycin complex: targeting of a protein chaperone by an antitumor agent. Cell 89(2):239–250

Taulu S, Kajola M, Simola J (2004) Suppression of interference and artifacts by the signal space separation method. Brain topography 16(4):269–275

Tucker LR (1966) Some mathematical notes on three-mode factor analysis. Psychometrika 31(3):279–311

Virtanen S, Klama A, Khan SA, Kaski S (2012) Bayesian group factor analysis. In: Lawrence N, Girolami M (eds) Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, pp 1269–1277

Yilmaz KY, Cemgil AT, Simsekli U (2011) Generalised coupled tensor factorisation. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds) Advances in Neural Information Processing Systems 24, pp 2151–2159