HAL
archives-ouvertes.fr

Joint decompositions with flexible couplings

Rodrigo Cabral Farias, Jérémy Cohen, Christian Jutten, Pierre Comon

▶ **To cite this version:**

Rodrigo Cabral Farias, Jérémy Cohen, Christian Jutten, Pierre Comon. Joint decompositions with flexible couplings. 12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA 2015), Aug 2015, Liberec, Czech Republic. 10.1007/978-3-319-22482-414 . hal-01135920v2

HAL Id: hal-01135920
https://hal.archives-ouvertes.fr/hal-01135920v2

Submitted on 10 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Joint decompositions with flexible couplings

Rodrigo Cabral Farias*, Jérémy Emile Cohen, Christian Jutten, and Pierre Comon

GIPSA-Lab, UMR CNRS 5216, Grenoble Campus, 38400 Saint Martin d'Hères, France

Abstract. A Bayesian framework is proposed to define flexible coupling models for joint decompositions of data sets. Under this framework, a solution to the joint decomposition can be cast in terms of a maximum *a posteriori* estimator. Examples of joint posterior distributions are provided, including general Gaussian priors and non Gaussian coupling priors. Then simulations are reported and show the effectiveness of this approach to fuse information from data sets, which are inherently of different size due to different time resolution of the measurement devices.

Keywords: Tensor decompositions, coupled decompositions, data fusion, multimodal data.

1 Introduction

In some domains such as brain imaging, metabolomics and link prediction, different and diverse instrumentation and data gathering devices are used to collect information on some underlying phenomena. Since no device has a complete view of the phenomena, data fusion can be used to blend the different views provided by each device, thus allowing a broader understanding. It is then not surprising that multimodal data fusion, *i.e.* fusion of heterogeneous data, has become an important topic of research in these domains [21,2,10].

One way of defining a framework for multimodal data fusion is to state it as a problem of latent variable analysis. Variations of the hidden variables are supposed to explain most of the variations in the measured data sets. Since the data sets are considered to be different views of the same phenomena, a part of the hidden variables can be supposed to be related to each other. In a more simple way, we can say that the latent models are coupled through subsets of their variables. By exploiting this coupling in the joint estimation of the latent models, we expect that the information from one data set will help in the estimation of the latent variables related to the other data set.

Although the framework described above dates back to the coupled (or linked) decomposition model described in [11], it was reexplored in independent component analysis approaches [6,7,9] and repopularized recently in data

* This research was supported in part by the ERC Grants AdG-2013-320594 "DECODA" (R. Cabral Farias, J. E. Cohen and P. Comon) and AdG-2012-320684 "CHESS" (C. Jutten).

science in [17], where the problem of joint nonnonegative matrix factorization was considered under the constraint that one of the factors is shared by all matrices. In all these cases, the coupling occurs through equality constraints on latent factors. Following the work in [17], the framework of coupled tensor decompositions was revisited in [5,14], variations on this framework, such as tensor-matrix factorizations [3,1] and more general latent models [18]. Uniqueness properties and the development of algorithms for the exact coupled tensor decomposition problem are proposed in [19] and [20], while algorithms for the coupled tensor approximation problem under general cost functions are developed in [22,10].

A more flexible model for the coupling of the models has been proposed in [2]. Instead of considering equality constraints for the entire factors of a tensor model, only a few components are constrained. In [16] the problem of coupled nonnegative matrix factorization is considered and a flexible coupling is proposed by assuming that the shared components are similar in L_1 or L_2 sense and not equal. In this paper, we propose a generalization of the flexible models for joint decompositions above using a Bayesian approach. After presenting our framework, we present some examples of non trivial couplings and some simulation results. We show some results on the coupling of two tensor models where the coupled factors do not have the same size due to different sampling. This type of coupling cannot be dealt with the flexible models presented in [2] or [16]. Note also that this type of model appears naturally in multimodal data fusion since nothing guarantees that measurement devices of different nature will generate data sets with the same resolution.

In this paper the following notation is used: scalars and vectors are denoted by lower case x and bold lower case $\boldsymbol{x}$ letters respectively. Matrices are denoted by upper case bold letters $\boldsymbol{X}$, while tensors by calligraphic letters $\boldsymbol{\mathcal{X}}$. Elements of a given array are indicated by subscripts $\boldsymbol{\mathcal{X}}_{ijk}$. Vectorization of parameters is indicated by $\text{vec}(\cdot)$. The Kronecker product of two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is denoted by $\boldsymbol{X} \otimes \boldsymbol{Y}$, while the Khatri-Rao product (column-wise Kronecker product) by $\boldsymbol{X} \odot \boldsymbol{Y}$. The pseudo inverse is given by superscript $\dagger$, its side indicates the side of the pseudoinverse.

2 Coupled decompositions: hard and flexible approaches

2.1 Deterministic approach: hard measurements and coupling

Consider two arrays of measurements, $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$, which can be tensors of possibly different orders and dimensions. Arrays $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$ are related to two parametric models characterized by parameter arrays $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively.

For instance, if $\boldsymbol{\mathcal{Y}}$ is a matrix (a second order tensor) to be diagonalized, the model can be the SVD $\boldsymbol{\mathcal{Y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$, so that $\boldsymbol{\theta} = \text{vec}(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V})$. If $\boldsymbol{\mathcal{Y}}'$ is a third order tensor, its Canonical Polyadic (CP) decomposition [8], [13] writes $\boldsymbol{\mathcal{Y}}' = \left(\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}'\right)$, meaning that $\boldsymbol{\mathcal{Y}}'_{ijk} = \sum_{r=1}^{R'} \boldsymbol{A}'_{ir}\boldsymbol{B}'_{jr}\boldsymbol{C}'_{kr}$, and $\boldsymbol{\theta}' = \text{vec}(\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}')$.

In the case where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are not coupled, they can be obtained (non uniquely) by processing the data arrays separately. On the other hand, if they

are coupled then the data needs to be processed jointly (and parameters uniquely estimated). The common assumption on the coupling [1,3,18] is that some factors are equal, for example, $\boldsymbol{B} = \boldsymbol{V}$ and this leads to the problem of exact coupled decompositions [19], [20].

2.2 Classical estimation approach: flexible measurements and hard coupling

In general the relation between the measured data arrays $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$ and the underlying SVD and CP model is not an equality relation, since the measurement devices only generate an imperfect random response driven by the parametric model. Therefore, we are lead to solve an approximation problem instead of an exact decomposition problem.

By considering a probabilistic relation between $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\theta}$, we obtain a more flexible measurement model. This relation can be expressed in terms of the likelihood of the measurements $p(\boldsymbol{\mathcal{Y}}; \boldsymbol{\theta})$. If $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are uncoupled, then they can be estimated separately using maximum likelihood estimation (MLE), *i.e.*, $\arg\max_\theta \log p(\boldsymbol{\mathcal{Y}}; \boldsymbol{\theta})$. Note that depending on the likelihood function, we can obtain different objective functions in the approximation problem.

If $\boldsymbol{B} = \boldsymbol{V}$ and supposing that $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$ are independent, then the joint likelihood function factorizes $p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}'; \boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\mathcal{Y}}; \boldsymbol{\theta}) p(\boldsymbol{\mathcal{Y}}'; \boldsymbol{\theta}')$ and the approximation problem becomes a constrained MLE:

$$\begin{array}{cc} \text{maximize} & \log p(\boldsymbol{\mathcal{Y}}; \boldsymbol{\theta}) + \log p(\boldsymbol{\mathcal{Y}}'; \boldsymbol{\theta}') \\ \text{with respect to (w.r.t.)} & \boldsymbol{\theta},\, \boldsymbol{\theta}', \\ \text{subject to} & \boldsymbol{V} = \boldsymbol{B}. \end{array} \tag{1}$$

Different versions of this approach are presented in [17,22,10].

2.3 Bayesian estimation approach: flexible measurements and coupling

We can go one step further and consider that the coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ itself can be flexible. For example if we want to have $\boldsymbol{V} \approx \boldsymbol{B}$, but not equality, or even $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{B}$ for a transformation matrix $\boldsymbol{W}$ that is known only approximately. To formalize this we assume that the pair $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ is random and that we have at our disposal a joint probability distribution $p(\boldsymbol{\theta}, \boldsymbol{\theta}')$.

Maximum a posteriori estimator: since the pair $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ is random, we have to move from an MLE setting to a maximum *a posteriori* (MAP) setting[1]. The approximation setting under the MAP criterion becomes

$$\arg\max_{\boldsymbol{\theta}, \boldsymbol{\theta}'} p(\boldsymbol{\theta}, \boldsymbol{\theta}' | \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}') = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\theta}'} p(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}') = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\theta}'} \Upsilon(\boldsymbol{\theta}, \boldsymbol{\theta}'), \tag{2}$$

[1] We could also consider a minimum mean squared error setting but then we would need to evaluate $p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}')$ which is normally cumbersome.

where $\Upsilon(\boldsymbol{\theta}, \boldsymbol{\theta}') = -\log p(\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}')$. Conditioning on the parameters leads to a cost function that can be decomposed in a joint data likelihood term plus a term involving the coupling:

$$\Upsilon(\boldsymbol{\theta}, \boldsymbol{\theta}') = -\log p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}, \boldsymbol{\theta}') - \log p(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{3}$$

Hypotheses: in what follows, we state a few simplification hypotheses that we consider and the main hypotheses underlying the Bayesian approach:

H1 *Conditional independence of the data*: the data arrays $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$ are independent of $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ respectively, if they are conditioned on $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. We suppose also that they are conditionally independent between them. This results in the following $p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{Y}}', \boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})$ and $p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\mathcal{Y}}, \boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}')$.

H2 *Independence of uncoupled parameters*: all parameters except the coupled parameters are independent. In the joint SVD and CP case this means that $p(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{V}, \boldsymbol{B})p(\boldsymbol{U})p(\boldsymbol{\Sigma})p(\boldsymbol{A})p(\boldsymbol{C})$.

H3 *On the priors*: trivially, the joint distribution of the coupled parameters, *e.g.* $p(\boldsymbol{V}, \boldsymbol{B})$ needs to be known or, at least, one of the conditional distributions ($p(\boldsymbol{V}|\boldsymbol{B})$ or $p(\boldsymbol{B}|\boldsymbol{V})$). The marginal priors on the uncoupled and of the conditioning parameters ($p(\boldsymbol{U})$, ..., $p(\boldsymbol{C})$) are assumed either to be known or to be flat on some domain of definition.

H4 *Likelihoods*: the conditional probabilities (or likelihoods) $p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})$ and $p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}')$ are known, at least on their shape. In a MAP setting, this indirectly sets the weights which will be given to each data array in $\Upsilon\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)$.

Simplified MAP: under hypothesis **H1** $p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta})$. The simplified MAP estimator is given by minimizing a three-term cost function

$$\underset{\boldsymbol{\theta}, \boldsymbol{\theta}'}{\arg\min}\, \Upsilon(\boldsymbol{\theta}, \boldsymbol{\theta}') = -\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta}) - \log p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}') - \log p(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{4}$$

Hypotheses **H3** and **H4** are assumed so that all terms in this cost function are defined. **H2** is assumed to allow simplifications in the last term. In the next section we give many examples of possible joint decomposition problems and their objective functions under this setting.

3 Examples

In what follows we consider that the parametric models underlying the data arrays are two CP models $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ and $\left(\boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}'\right)$ with dimensions I,J,K and I',J',K' and number of components (*i.e.* number of matrix columns) R and R' respectively. And we consider that the coupling occurs between matrices $\boldsymbol{C}$ and $\boldsymbol{C'}$. We exploit this framework with two different examples: general joint Gaussian modeling of the parameters and non Gaussian couplings.

3.1 Joint Gaussian modeling

A general joint Gaussian model comprising coupled and uncoupled variables is given by the following expression:

$$\boldsymbol{M}\left[\boldsymbol{\theta}^{\top}\boldsymbol{\theta}'^{\top}\right]^{\top} = \boldsymbol{\Sigma}\boldsymbol{u} + \boldsymbol{\mu}, \tag{5}$$

where $\boldsymbol{M}$ is a matrix defining the structural relations between variables, $\boldsymbol{u}$ is a white Gaussian vector with zero mean and unit variances, $\boldsymbol{\Sigma}$ is a diagonal matrix of standard deviations and $\boldsymbol{\mu}$ is a constant vector. Observe that a condition for the pair $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ to define a joint Gaussian vector is the left invertibility of $\boldsymbol{M}$. Under this condition we have $\left[\boldsymbol{\theta}^{\top}\boldsymbol{\theta}'^{\top}\right]^{\top} \sim \mathcal{N}\{^{\dagger}\boldsymbol{M}\boldsymbol{\mu}, \boldsymbol{\Gamma}\}$, where $\boldsymbol{\Gamma} = (^{\dagger}\boldsymbol{M})\boldsymbol{\Sigma}\boldsymbol{\Sigma}(^{\dagger}\boldsymbol{M}^{\top})$ is the covariance matrix of the joint vector. The MAP objective function is

$$\Upsilon\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = -\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta}) - \log p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}') + \{[\boldsymbol{\theta}^{\top}\boldsymbol{\theta}'^{\top}] - \boldsymbol{\mu}^{\top}{}^{\dagger}\boldsymbol{M}^{\top}\}\boldsymbol{\Gamma}^{-1}\{[\boldsymbol{\theta}^{\top}\boldsymbol{\theta}'^{\top}]^{\top} - {}^{\dagger}\boldsymbol{M}\boldsymbol{\mu}\} \tag{6}$$

Below, we give a few examples of applications of this approach.

Shared components: a usual problem in multimodal data fusion is that some components are not present in all modalities, thus we have some components which are shared and some which are specific to each modality. Suppose $R = R' = 2$ in the coupled CP model and that the first component of $\boldsymbol{C}$ ($\boldsymbol{C}_1$) is approximately equal to the first component of $\boldsymbol{C}'$ ($\boldsymbol{C}'_1$). Supposing zero mean marginal Gaussian priors, we have

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & -\boldsymbol{I} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_{(u)} \\ \boldsymbol{\theta}_{(u')} \\ \hline \boldsymbol{C}_2 \\ \boldsymbol{C}'_2 \end{bmatrix} = \begin{bmatrix} \sigma_u^2\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_u'^2\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sigma_c^2\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \sigma_{\boldsymbol{C}'_2}^2\boldsymbol{I} \end{bmatrix} \boldsymbol{u}, \tag{7}$$

where $\boldsymbol{\theta}_{(u)}$, $\boldsymbol{\theta}_{(u')}$, σ_u^2 and $\sigma_u'^2$ are the uncoupled parameters and their respective variances, σ_c^2 is the variance of the couplings and $\sigma_{\boldsymbol{C}'_2}^2$ the prior variance of the $\boldsymbol{C}'_2$ elements.

Dynamical models: in some cases, more than two data arrays are present and, as a consequence, there is a large number of degrees of freedom on the definition of the couplings. If the data arrays are measured in time, then a natural coupling can be defined through a dynamical model. For example, if N instances of a CP model are measured in successive times and that $\text{vec}(\boldsymbol{C}^k)$ are related through a linear dynamic model with state transition matrices $\boldsymbol{E}_k$ and white gaussian vectors $\boldsymbol{u}_k$, then the k-th line of the joint model corresponding to the coupling

is given by

$$\left[\cdots \mathbf{0}\ \boldsymbol{I}\ -\boldsymbol{E}_k\ \mathbf{0}\cdots\right]\begin{bmatrix}\vdots\\ \mathrm{vec}(\boldsymbol{C}^k)\\ \mathrm{vec}(\boldsymbol{C}^{k-1})\\ \vdots\end{bmatrix}=\boldsymbol{u}_k. \tag{8}$$

Sampling a continuous function: another important problem in multimodal data fusion is related to sampling. Different measurement devices have different sampling frequencies or even different non uniform sampling grids. In some situations, the continuous functions being measured can be approximated by a common function $c(t)$. For two sampled vectors $\boldsymbol{c}$ and $\boldsymbol{c}'$, their relation with the continuous function can be obtained with an interpolation kernel (see the general description in [4])

$$c(t)\approx\sum_{k=1}^{K}\boldsymbol{c}_k h(t,t_k)\approx\sum_{k'=1}^{K'}\boldsymbol{c}'_k h'(t,t'_k), \tag{9}$$

for some kernels $h(\cdot,\cdot)$ and $h'(\cdot,\cdot)$ and for sampling times $\{t_k, k\in 1,\cdots,K\}$ and $\{t_{k'}, k'\in 1,\cdots,K'\}$. Therefore, we can impose a new common sampling grid of size L where both interpolations should match. This leads to the linear relations

$$\boldsymbol{H}\boldsymbol{c}\approx\boldsymbol{H}'\boldsymbol{c}', \tag{10}$$

where $\boldsymbol{H}_{lk}=h(t_l,t_k)$, $\boldsymbol{H}'_{lk'}=h(t_l,t_{k'})$ and $\{t_l, k\in 1,\cdots,L\}$. For example, in the coupled CP models, when $\boldsymbol{C}$ and $\boldsymbol{C}'$ have different dimensions due to different sampling rates, their coupling is given by (10). The approximation can then be rewritten in the joint Gaussian setting as

$$\left[\mathbf{0}\ \mathbf{diag}(\boldsymbol{H})\ -\mathbf{diag}(\boldsymbol{H}')\ \mathbf{0}\right]\begin{bmatrix}\vdots\\ \mathrm{vec}(\boldsymbol{C})\\ \mathrm{vec}(\boldsymbol{C}')\\ \vdots\end{bmatrix}=\boldsymbol{\Sigma}\boldsymbol{u}, \tag{11}$$

where $\mathrm{diag}(\boldsymbol{H})$ is a block diagonal matrix with repetitions of $\boldsymbol{H}$ on the diagonal.

3.2 Non Gaussian conditional coupling

Non trivial couplings between the factors $\boldsymbol{C}$ and $\boldsymbol{C}'$ can be considered by assuming that the coupling is given by a non Gaussian conditional distribution $p\left(\boldsymbol{C}|\boldsymbol{C}'\right)$.

Impulsive additive coupling: as a first example, we can consider that each element in $\boldsymbol{C}$ is a version of $\boldsymbol{C}'$ corrupted by independent and identically distributed (i.i.d.) impulsive noise:

$$\boldsymbol{C}_{ij}=\boldsymbol{C}_{ij}+\boldsymbol{V}_{i,j} \tag{12}$$

where $\boldsymbol{V}_{i,j}$ follows a Laplacian distribution $p(\boldsymbol{V}_{ij}) = (1/2\delta)\exp(-|\boldsymbol{V}_{i,j}|/\delta)$ with scale parameter δ or a Cauchy distribution $p(\boldsymbol{V}_{ij}) = 1/\{\pi\delta[1+(\boldsymbol{V_{ij}}/\delta)^2]\}$. Supposing that the priors are flat on a constraint set $\mathcal{C}$, the objective functions to be minimized are respectively

$$\begin{aligned}\Upsilon\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) = & \quad -\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta}) - \log p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}') + (1/2\delta)\|\boldsymbol{C}-\boldsymbol{C}'\|_1 \\ \Upsilon\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) = & -\log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta}) - \log p(\boldsymbol{\mathcal{Y}}'|\boldsymbol{\theta}') - \sum_{ij}^{IJ}\log\{1+[(\boldsymbol{C}_{ij}-\boldsymbol{C}'_{ij})/\delta]^2\},\end{aligned} \tag{13}$$

where $\|\cdot\|_1$ is the $\mathbf{L}_1$ norm. The first penalty was considered in [16] in a collective matrix factorization context. Both cost functions imply a sparse number of large discrepancies between $\boldsymbol{C}$ and $\boldsymbol{C}'$.

Positive general coupling: when $\boldsymbol{C}_{ij} > 0$ and $\boldsymbol{C}'_{ij} > 0$, an additive random coupling may not me be the best option, since to ensure positiveness the support of the additive term has to depend on the values of $\boldsymbol{C}'$, which is not realistic. Therefore, other alternatives naturally ensuring positiveness can be considered, for example the Tweedie's distribution [12]. Special cases of this distribution are the Poisson, Gamma and inverse-Gaussian distributions (the Gaussian distribution is a limit special case). In general, the PDF of the Tweedie's distribution has no analytical form, thus we cannot directly use it to write down a coupling term in the MAP objective function. However, if we consider that the coupling between $\boldsymbol{C}_{ij} > 0$ and $\boldsymbol{C}' > 0$ is strong (dispersion δ is small), then a saddle point approximation can be used [12]

$$p(\boldsymbol{C}_{ij}|\boldsymbol{C}'_{ij}) \approx (2\pi\delta^2\boldsymbol{C}'^{\beta}_{ij})^{-1/2}\exp[-d_\beta(\boldsymbol{C}_{i,j}|\boldsymbol{C}'_{i,j})/\delta^2] \tag{14}$$

where β is a shape parameter ($\beta = 1, 2, 3$ for the Poisson, Gamma and inverse-Gaussian distributions respectively) and d_β is the beta divergence [23]

$$d_\beta(\boldsymbol{C}_{i,j}|\boldsymbol{C}'_{i,j}) = \boldsymbol{C}'^{1-\beta}_{i,j}/[(1-\beta)(2-\beta)]\left[\boldsymbol{C}^{2-\beta}_{i,j}\boldsymbol{C}'^{\beta-1}_{i,j} - \boldsymbol{C}_{ij}(2-\beta) + \boldsymbol{C}'_{ij}(1-\beta)\right]. \tag{15}$$

Under this conditional distribution the coupling term in the MAP objective becomes $\sum_{ij}\left[(\beta/2)\log(C_{ij}) + (1/\delta)d_\beta(\boldsymbol{C}_{i,j}|\boldsymbol{C}'_{i,j})\right]$.

4 Gaussian problem with flat priors and alternating least squares (ALS)

From now on we will focus on the specific case of joint Gaussian modeling. Note that the joint decomposition can be found by minimizing (6). This can be done, for example, using an all-at-once minimization procedure such as a gradient algorithm.

If we consider that the priors are flat and that the coupling is of the form

$$\boldsymbol{G}\mathrm{vec}(\boldsymbol{C}) - \boldsymbol{G}'\mathrm{vec}(\boldsymbol{C}') = \frac{\boldsymbol{I}}{\sigma_c^2}\boldsymbol{u}, \tag{16}$$

where $\boldsymbol{G}$ and $\boldsymbol{G}'$ are two coupling matrices, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{u}$ is a white Gaussian vector and σ_c is a standard deviation related to the coupling intensity. Then the coupling term becomes a quadratic term on $\text{vec}(\boldsymbol{C})$ and $\text{vec}(\boldsymbol{C}')$. Moreover, if we suppose that the two tensors $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Y}}'$ are measured each with i.i.d. Gaussian noise with respective standard deviations σ_n and σ'_n. Then the objective function to be minimized is

$$\Upsilon = \frac{1}{\sigma_n^2}\|\boldsymbol{\mathcal{Y}}-(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})\|_F^2+\frac{1}{\sigma_n'^2}\|\boldsymbol{\mathcal{Y}}'-(\boldsymbol{A}',\boldsymbol{B}',\boldsymbol{C}')\|_F^2+\frac{1}{\sigma_c^2}\|\boldsymbol{G}\text{vec}(\boldsymbol{C})-\boldsymbol{G}'\text{vec}(\boldsymbol{C}')\|_F^2 \tag{17}$$

To minimize this function we can use an easy to implement algorithm, such as the alternating least squares (ALS) algorithm. Observe that the alternating updates for the uncoupled factors are simply the standard ALS steps for CP approximation, while for the coupled factors the updates are the solution of a joint least squares problem with a coupling term. The ALS procedure is the following, at iteration k:

Uncoupled Factors

$$\begin{array}{ll} \hat{\mathbf{A}}_k = \boldsymbol{Y}_{(1)}(\hat{\mathbf{C}}_{k-1}\odot\hat{\mathbf{B}}_{k-1})^\dagger & \hat{\mathbf{A}}'_k = \boldsymbol{Y}'_{(1)}(\hat{\mathbf{C}}'_{k-1}\odot\hat{\mathbf{B}}'_{k-1})^\dagger, \\ \hat{\mathbf{B}}_k = \boldsymbol{Y}_{(2)}(\hat{\mathbf{C}}_{k-1}\odot\hat{\mathbf{A}}_k)^\dagger & \hat{\mathbf{B}}'_k = \boldsymbol{Y}'_{(2)}(\hat{\mathbf{C}}'_{k-1}\odot\hat{\mathbf{A}}'_k)^\dagger, \end{array} \tag{18}$$

Coupled Factors

$$\begin{bmatrix}\text{vec}(\hat{\mathbf{C}}_k)\\ \text{vec}(\hat{\mathbf{C}}'_k)\end{bmatrix} = {}^{\dagger}\begin{bmatrix}\frac{(\boldsymbol{F}^\top\boldsymbol{F})\otimes\boldsymbol{I}}{\sigma_n^2}+\frac{\boldsymbol{G}^\top\boldsymbol{G}}{\sigma_c^2} & -\frac{\boldsymbol{G}^\top\boldsymbol{G}'}{\sigma_c^2}\\ -\frac{\boldsymbol{G}'^\top\boldsymbol{G}}{\sigma_c^2} & \frac{(\boldsymbol{F}'^\top\boldsymbol{F}')\otimes\boldsymbol{I}}{\sigma_n'^2}+\frac{\boldsymbol{G}'^\top\boldsymbol{G}'}{\sigma_c^2}\end{bmatrix}\begin{bmatrix}(\boldsymbol{F}\otimes\boldsymbol{I})\text{vec}\left(\boldsymbol{Y}_{(1)}\right)\\ (\boldsymbol{F}'\otimes\boldsymbol{I})\text{vec}\left(\boldsymbol{Y}_{(2)}\right)\end{bmatrix} \tag{19}$$

where $\boldsymbol{F}=\hat{\mathbf{B}}_k\odot\hat{\mathbf{A}}_k$, $\boldsymbol{F}'=\hat{\mathbf{B}}'_k\odot\hat{\mathbf{A}}'_k$.

5 Simulations

To show the effects of the flexible coupling of two CP models on approximation performance, we apply the ALS algorithm presented in Sec. 4 to three different types of coupling: direct coupling of the $\boldsymbol{C}$ factors, coupling of one component and coupling of $\boldsymbol{C}$ factors with different size through interpolation.

5.1 Similar factors

We start with a straightforward coupling model $\boldsymbol{G}=\boldsymbol{G}'=\boldsymbol{I}$. The two CP models are generated randomly with dimensions $I=I'=J=J'=K=K'=10$ and $R=R'=3$. The data array $\boldsymbol{\mathcal{Y}}'$ has low noise $\sigma_n=0.001$, while $\boldsymbol{\mathcal{Y}}$ is noisy $\sigma'_n=0.1$. We vary the coupling intensity $\frac{1}{\sigma_c}$ from 2 to 5000. The two CP models are first approximated separately (disregarding the coupling); in this case an all-at-once conjugate gradient algorithm is used. After convergence of the algorithm,

the columns of the resulting factors are permuted so that the components in the coupling match. The permuted factors are then used to initialize the ALS procedure described in Sec. 4. We simulate 50 times this procedure with different noise realizations and we evaluate the total mean squared error (MSE) on the $\boldsymbol{C}$ and $\boldsymbol{C}'$ factors. The results are shown in Fig. 1 .

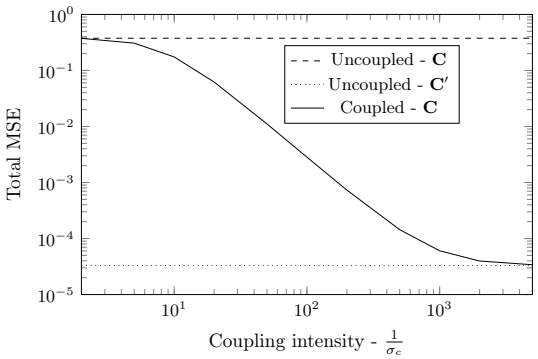

Fig. 1. Total MSE for the factors $\boldsymbol{C}$ and $\boldsymbol{C}'$ of the coupled CP model as a function of the coupling intensity $\frac{1}{\sigma_c}$. Results are shown for an algorithm that disregards the coupling and for the ALS algorithm presented in Sec. 4, which considers the coupling. The CP models are measured with different noise levels ($\sigma_n = 0.1$ and $\sigma'_n = 0.001$).

We can see that when the coupling is weak the MSE on $\boldsymbol{C}$ is close to its uncoupled MSE. By increasing the coupling intensity, the factor is estimated with a much better performance, since more information comes from the clean tensor through the coupling. The flexible coupling allows to assess the continuous transition between uncoupled models and exactly coupled models.

5.2 Shared component

We also simulate the case when only one component is shared between the models and the noise levels are similar. In this case $I = I' = J = J' = K = K' = 10$, $R = R' = 2$, $\boldsymbol{G} = \boldsymbol{G}' = [\boldsymbol{I}\ \boldsymbol{0}]$, $\sigma_c = 0.001$, $\sigma_n = 0.05$ and $\sigma'_n = 0.05$. The MSE on all elements of both $\boldsymbol{C}$ and $\boldsymbol{C}'$ are evaluated in a similar way as above. The MSE for the estimation of each element of the factors is shown in Fig. 2.

Note that in Fig. 1, we did not show the MSE for the clean factor, since it is too close to the uncoupled performance. Although the unshared component is not improved at all by the coupling, we can see in Fig. 2 that, when the noise levels are equivalent, the MSE is decreased for the shared components of both factors.

5.3 Different sampling rates

As a non trivial example of coupling we consider the case when $I = I' = J = J' = 10$, but with different sizes on the third mode $K = 37$ and $K' = 53$.

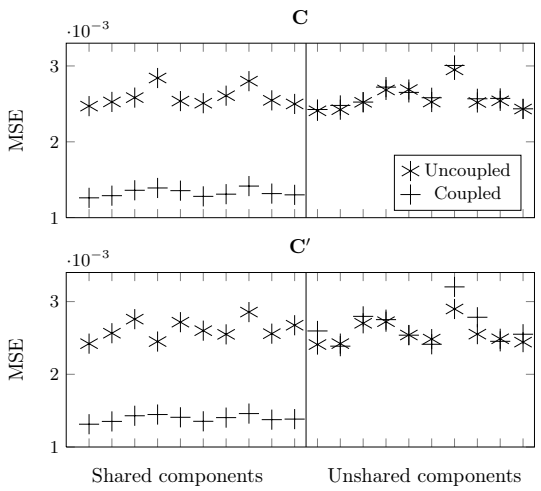

Fig. 2. MSE for the $\boldsymbol{C}$ and $\boldsymbol{C}'$ factors of the coupled CP models. The models have one similar component (a shared component) on their $\boldsymbol{C}$ factors while the other component is not similar.

We suppose that the components on $\boldsymbol{C}$ and $\boldsymbol{C}'$ are sampled versions of the same underlying continuous functions, however, the sampling periods to obtain the factors are different so that the elements in the factors cannot be similar, at least, in most of the points. Since the factors are not similar, we cannot apply the direct coupling model and we must use an interpolation approach as explained at the end of Sec. 3.

In this example we consider functions which are band limited and periodic. For an odd number of samples, the interpolation kernel is given by the Dirichlet kernel [15]

$$\boldsymbol{H}_{lk} = \frac{\sin\{K\pi[(l-1)T_i - (k-1)T]/[(L-1)T_i\}}{K\sin\{\pi[(l-1)T_i - (k-1)T]/[(L-1)T_i\}}, \tag{20}$$

where T is the original sampling period and T_i is the sampling period of the interpolation. As a consequence we have $\boldsymbol{G} = \boldsymbol{I} \otimes \boldsymbol{H}$ and $\boldsymbol{G}' = \boldsymbol{I} \otimes \boldsymbol{H}'$.

We simulate two random CP models with $R = R' = 3$. The components on the $\boldsymbol{C}$ factors are generated by sampling $c_r(t) = \sum_{i=1}^{3} \gamma_{ir} \sin(2\pi f_i t)$ where γ_{ir} are generated randomly and $f_1 = 2$, $f_2 = 2.5$, $f_3 = 3.5$. The sampling periods are $T = 1/9$ and $T' = 1/13$. An example of continuous-time component with its sampled points on different grids is shown in Fig. 3.

We fix $\sigma_n' = 0.1$, while σ_n varies from 0.001 to 0.1, so that the ratio σ_n'/σ_n varies in the interval $[0.1\,10]$. Since the signals are band limited and since we observe them over a finite time duration, interpolation can only approximate the continuous signal and it is necessary to set a nonzero σ_c even if the continuous signals are the same for both data sets. We set $L = 100$, $\sigma_c = 0.01$ and we evaluate the total MSE on the coupled factors in the same way as presented previously. The results are shown in Fig. 4. When the noise ratio increases, the total

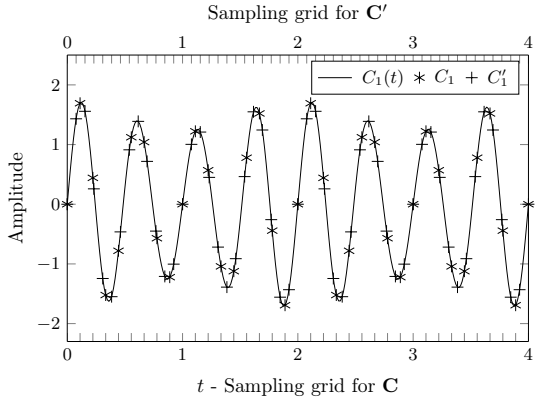

Fig. 3. Underlying continuous function $c_1(t)$ for the first components of the $\boldsymbol{C}$ factors and their corresponding sampled versions $\boldsymbol{c}_1$ and $\boldsymbol{c}_1'$ obtained with different sampling grids.

MSE for the uncoupled approach increases sharply, while the coupled approach has a smooth increase. This shows that even though the coupled factors are not similar, information can still be exchanged between them through interpolation.

As a last simulation, we consider that both data arrays are noisy $\sigma_n = \sigma_n' = 0.1$ and that the number of interpolation points for the coupling L is varied from 5 to 70. We set $\sigma_c = 0.001$ in the ALS algorithm. The total MSE is shown in Fig. 5.

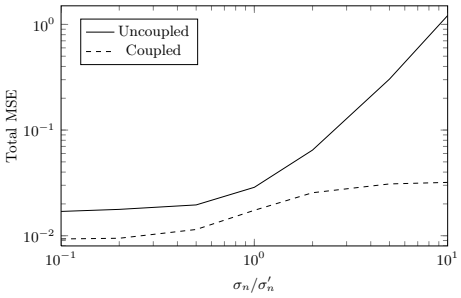

Fig. 4. Total MSE for the estimation of the $\boldsymbol{C}$ factors for different noise levels ratios σ_n/σ_n'. The CP models are coupled through interpolation.

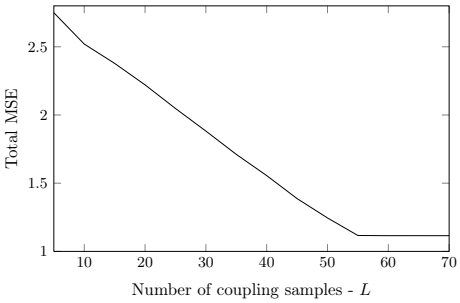

Fig. 5. Total MSE for the estimation of the $\boldsymbol{C}$ factors as a function of the number of interpolation samples.

Note that only the sampling time points $t = \{0, 1, 2, 3, 4\}$ coincide in the original factors. Thus in a standard coupling approach only these points can be coupled and the total MSE that we obtain is the first point in the curve. By increasing the number of interpolation points the information exchanged within

the model is larger and MSE decreases almost linearly. Above $L = 53$, only a small quantity of information can be exchanged because this is the maximum resolution present in the data and the total MSE curve becomes flat. Since the complexity of the joint factor update is dependent on the number of interpolation points, in practice when the dimensions are originally large, the choice of L depends on a trade off between MSE (large L) and complexity (small L).

6 Conclusions

Since the expression of a phenomenon can be different in different data sets it is clear that the link between factorizations of the data sets must be somehow flexible. To give a meaning to this flexibility we have proposed in this paper a Bayesian setting for the coupling between factors. Under this setting we can propose not only trivial flexible links between factors, *e.g.* an i.i.d. Gaussian model for the differences between factors, but also joint Gaussian models, sparse similarity models and nonnegative similarity models.

Through simulations, we have shown that the flexible coupling between factors allows to explore the entire range of possibilities between exactly coupled and uncoupled models. We have also shown that coupling allows not only to retrieve accurately a factor from noisy data by exploiting its coupling to another data set which has low noise, but also that if the two data sets are noisy then the accuracy on the estimation of both factorizations is increased. As an example of multimodal data fusion, we have presented the problem of fusing two data sets in which one dimension is different due to different sampling. Although the factors are almost completely different, the underlying hypothesis that they come from the some continuous-time function allows to exchange information between the data sets in an interpolated domain.

In the simulation examples we have focused on a joint Gaussian modeling for the couplings, in future work we can concentrate on non Gaussian couplings such as the Tweedie's coupling for nonnegative variables presented only briefly here. Moreover, since the CP approximation problem is an estimation problem, we can evaluate the Cramér-Rao bounds (CRB) for the coupled problem so that we can assess approximately the estimation performance without resorting to extensive simulation. In the hard coupling case a constrained CRB can be considered, while in the fully Bayesian and Bayesian with flat priors cases, the Bayesian and hybrid CRB can be considered. Finally, since there is an information flow from one data array to the other through the coupling, an interesting point for future research is to quantify and analyze this flow through the analysis of the mutual information between the data arrays and the factors.

References

1. E. Acar, T. G. Kolda, and D. M. Dunlavy. All-at-once optimization for coupled matrix and tensor factorizations. *arXiv preprint arXiv:1105.3422*, 2011.

2. E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro. Structure-revealing data fusion model with applications in metabolomics. In *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pages 6023–6026. IEEE, 2013.
3. E. Acar, M. A. Rasmussen, F. Savorani, T. Næs, and R. Bro. Understanding data fusion within the framework of coupled matrix and tensor factorizations. *Chemometr. Intell. Lab.*, 129:53–63, 2013.
4. A. Aldroubi and K. Gröchenig. Nonuniform sampling and reconstruction in shift-invariant spaces. *SIAM Review*, 43(4):585–620, 2001.
5. A. Banerjee, S. Basu, and S. Merugu. Multi-way clustering on relation graphs. In *SDM*, volume 7, pages 225–334. SIAM, 2007.
6. V. D. Calhoun, T. Adali, N. R. Giuliani, J. J. Pekar, K. A. Kiehl, and G. D. Pearlson. Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. *Hum. Brain Mapp.*, 27(1):47–62, 2006.
7. V. D. Calhoun, T. Adalı, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson. A method for multitask fMRI data fusion applied to schizophrenia. *Hum. Brain Mapp.*, 27(7):598–610, 2006.
8. P. Comon, X. Luciani, and A. L. F. de Almeida. Tensor decompositions, alternating least squares and other tales. *J. Chemometr.*, 23(7-8):393–405, 2009.
9. P. Comon and M. Rajih. Blind identification of under-determined mixtures based on the characteristic function. *Signal Process.*, 86(9):2271–2281, 2006.
10. B. Ermiş, E. Acar, and A. T. Cemgil. Link prediction via generalized coupled tensor factorisation. *arXiv preprint arXiv:1208.6231*, 2012.
11. R. A. Harshman and M. E. Lundy. Data preprocessing and the extended PARAFAC model. *Research Methods for Multimode Data Analysis*, pages 216–284, 1984.
12. B. Jørgensen. *The theory of exponential dispersion models and analysis of deviance.* Number 51. CNPQ, IMPA (Brazil), 1992.
13. T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.
14. Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Metafac: community discovery via relational hypergraph factorization. In *Proc. ACM SIGKDD Conf.*, pages 527–536. ACM, 2009.
15. E. Margolis and Y. C. Eldar. Nonuniform sampling of periodic bandlimited signals. *IEEE Trans. Sig. Process.*, 56(7):2728–2745, 2008.
16. N. Seichepine, S. Essid, C. Fevotte, and O. Cappé. Soft nonnegative matrix co-factorization. *IEEE Trans. Sig. Process.*, 62(22):5940–5949, Nov 2014.
17. A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proc. 14th ACM SIGKDD Conf.*, pages 650–658. ACM, 2008.
18. L. Sorber, M. Van Barel, and L. De Lathauwer. Structured data fusion. *Tech. Rep., KU Leuven (Belgium)*, pages 13–177, 2013.
19. M. Sørensen and L. De Lathauwer. Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms. Part I: Uniqueness. Technical report, KU Leuven (Belgium), 2013.
20. M. Sørensen, I. Domanov, D. Nion, and L. De Lathauwer. Coupled canonical polyadic decompositions and (coupled) decompositions in multilinear rank-$(L_{r,n}, L_{r,n}, 1)$ terms. Part II: Algorithms. Technical report, KU Leuven (Belgium), 2013.
21. J. Sui, T. Adali, Q. Yu, J. Chen, and V. D. Calhoun. A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Meth.*, 204(1):68–81, 2012.

22. K. Y. Yılmaz, A. T. Cemgil, and U. Simsekli. Generalised coupled tensor factorisation. In *Adv. Neural. Inf. Process. Syst.*, pages 2151–2159, 2011.
23. Y. K. Yilmaz and A. T. Cemgil. Alpha/beta divergences and tweedie models. *arXiv preprint arXiv:1209.4280*, 2012.