Group Factor Analysis

Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski

Abstract—Factor analysis provides linear factors that describe relationships between individual variables of a data set. We extend this classical formulation into linear factors that describe relationships between groups of variables, where each group represents either a set of related variables or a data set. The model also naturally extends canonical correlation analysis to more than two sets, in a way that is more flexible than previous extensions. Our solution is formulated as variational inference of a latent variable model with structural sparsity, and it consists of two hierarchical levels: The higher level models the relationships between the groups, whereas the lower models the observed variables given the higher level. We show that the resulting solution solves the group factor analysis problem accurately, outperforming alternative factor analysis based solutions as well as more straightforward implementations of group factor analysis. The method is demonstrated on two life science data sets, one on brain activation and the other on systems biology, illustrating its applicability to the analysis of different types of high-dimensional data sources.

Index Terms—factor analysis, multi-view learning, probabilistic algorithms, structured sparsity

1 Introduction

Factor analysis (FA) is one of the cornerstones of data analysis, the tool of choice for capturing and understanding linear relationships between variables [1]. It provides a set of K factors, each explaining dependencies between some of the features in a vectorial data sample $\mathbf{y}_i \in \mathbb{R}^D$ based on the model

$$\mathbf{y}_i = \sum_{k=1}^{K} z_{i,k}\mathbf{w}_k + \boldsymbol{\epsilon}_i,$$

where $z_{i,k}$ is the value of the kth unobserved factor, $\mathbf{w}_k \in \mathbb{R}^D$ contains its loadings, and $\boldsymbol{\epsilon}_i$ is Gaussian noise. To correctly capture the relationships, we need to assume a diagonal noise covariance with free variance for each of the variables. If the noise model was more flexible, having non-diagonal covariance, it would allow describing some of the relationships as noise. On the other hand, forcing the variances to be equal would imply that heteroscedastic noise would need to be explained as factors, reducing the model to probabilistic PCA [2].

- *A. Klami and S. Kaski are with Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki. E-mail: arto.klami@hiit.fi, samuel.kaski@hiit.fi*
- *S. Virtanen, E. Leppäaho and S. Kaski are with Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University. E-mail: seppo.j.virtanen@aalto.fi,eemeli.leppaaho@aalto.fi*

Building on our preliminary conference paper [3], we generalize factor analysis to a novel problem formulation of *group factor analysis* (GFA), where the task is to explain relationships between groups of variables. We retain the linear-Gaussian family of FA, but modify the model so that each factor now describes dependencies between some of the feature groups instead of individual variables. Again the choice of residual noise is crucial: it needs to be flexible enough to model everything that is not a true relationship between two variable groups, but restricted enough so that all actual relationships will be modeled as individual factors. For FA these requirements were easily satisfied by assuming independent variance for each dimension. For GFA more elaborate constructions are needed, but the same basic idea applies.

From another perspective, GFA extends multi-battery factor analysis (MBFA), introduced by McDonald [4] and Browne [5] as a generalization of inter-battery factor analysis (IBFA) [6], [7] to more than two variable groups. MBFA is a factor analysis model for multiple co-occurring data sets, or, equivalently, for a vectorial data sample whose variables have been split into groups. It includes a set of factors that model the relationships between all variables, as well as separate sets of factors explaining away the noise in each of the variable groups. These group-specific factor sets are sufficiently flexible for modeling all variation within each group. However, each of the remaining factors is assumed to describe relationships between *all* of the groups, which is not sufficient for providing interpretable factors that reveal the relationships between the data sets as will be explained below. Nevertheless, the MBFA models are useful tools for multi-source data analysis, illustrated by the fact that the problem has been re-discovered in machine learning literature several times; see Section 4 for more details.

To solve the GFA problem, we need to have also factors that describe relationships between subsets of the groups. This makes the solutions to the problem both more flexible and more interpretable than MBFA. For example, a strong factor tying two groups while being independent of the other groups can then be explicitly modeled as such. The MBFA-based models would, falsely, reveal such a factor as one that is shared by all groups. Alternatively, they would need to, again incorrectly, split them into multiple group-specific ones.

In recent years, the need for the GFA solution has been identified by several authors, under different ter-

minology. Jia et al. learned sparse matrix factorization by convex optimization [8], and Van Deun et al. used group-lasso penalty to constrain the factors of a simultaneous component analysis (SCA) model [9]. Various Bayesian techniques have also been proposed for learning shared and individual subspaces of multiple data sources [3], [10], [11], [12].

In this work we lay the foundation for future development of GFA solutions, by properly defining the problem setup and terminology. We also present a general solution outline and show that the solutions mentioned above are all instances of the same basic approach; they all learn structured sparse FA models with varying techniques for obtaining group-wise sparsity for the factor loadings. We then propose a novel GFA solution that does not make a strong simplifying assumption shared by all the previous approaches. They all assume that we can independently infer, for each factor-group pair, whether that factor describes variation related to that group, whereas our solution explicitly models also these associations with an additional linear model. In brief, our model hence consists of two linear hierarchical levels. The first models the relationships between the groups, and the latter models the observed data given the output of the higher level. Alternatively, it can be viewed as a direct generalization of [3] with a more advanced structured sparsity prior making it possible to reduce the degrees of freedom in the model when needed.

Before delving into the details on how we solve the GFA problem, we introduce some general application scenarios. The model is useful for analyzing multi-view setups where we have several data sets with co-occurring samples. The variables can be grouped according to the data sets: all variables in one set belong to one group etc. Then GFA explains relationships between data sources, and for two data sets it equals the problem of canonical correlation analysis (CCA; see [13] for a recent overview from a probabilistic perspective). Alternatively, each group could contain a collection of variables chosen to represent a multi-dimensional concept, such as cognitive abilities of a subject, which cannot be summarized with a single feature. Then GFA could be used for associating cognitive abilities with other multi-dimensional concepts. The groups can also represent a meaningful partitioning of larger data sets; we present two practical examples of this kind of a setup. In one example we split a high-dimensional feature vector over the human genome into subsets according to functional pathways to describe drug responses, and in the other example we split magnetic resonance images of the human brain into local regions to study relationships between brain areas.

2 Group factor analysis

2.1 Problem formulation

The group factor analysis problem is as follows: Assume a collection of observations $\mathbf{y}_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$ collected in a data matrix $\mathbf{Y} \in \mathbb{R}^{N \times D}$, and a disjoint partition of the D variables into M groups $\{G_m\}$. The GFA task is to find a set of K factors that describe $\mathbf{Y}$ so that relationships between the groups can be separated from relationships within the groups. For notational simplicity, assume that the first D_1 variables correspond to the first group G_1, the following D_2 variables to G_2, and so on. Then we can write $\mathbf{Y} = [\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}]$, where $\mathbf{X}^{(m)}$ is a subset of the data corresponding to G_m. We use $\mathbf{x}_i^{(m)}$ to denote the ith sample (row) of $\mathbf{X}^{(m)}$. Throughout this paper we use the superscript $^{(m)}$ to denote variables related to the mth group or data set.

2.2 General solution

A general solution to the GFA problem can be formulated as a joint factor model for the observed data sets. The model for the mth group of the ith sample is

$$\mathbf{x}_i^{(m)} \sim \mathcal{N}(\mathbf{W}^{(m)\top}\mathbf{z}_i, \tau_m^{-1}\mathbf{I}), \tag{1}$$

where $\mathbf{W}^{(m)^\top} = [\mathbf{w}_1^{(m)}, \ldots, \mathbf{w}_K^{(m)}]$, $\mathbf{z}_i \in \mathbb{R}^K$, and τ_m is noise precision. Equivalently, we can directly write $\mathbf{y}_i = \mathbf{W}^\top\mathbf{z}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ is Gaussian noise with diagonal covariance but separate variance for each group, by denoting $\mathbf{W} = [\mathbf{W}^{(1)}, ..., \mathbf{W}^{(M)}]$.

To make the factors interpretable in the GFA-sense, that is, to describe relationships between the groups, we need to make $\mathbf{W}$ sparse so that it satisfies the following properties (for a visual illustration see Figure 1):

1) Some factors are private to each group, so that $\mathbf{w}_k^{(m)} \neq 0$ only for one m. These factors explain away the variation independent of all the other groups, and play the role of residual noise in regular FA.
2) The rest of the factors describe relationships between some arbitrary subset of the groups; they are non-zero for those groups and zero for the others.

A trivial solution would explicitly split the factors into separate sets so that there would be one set of factors for each possible subset of the groups (including the singletons and the set of all groups). This can be done for small M; for example Klami and Kaski proposed such a model for $M = 2$ [14] and Gupta et al. formulated the model for general M but ran experiments only with $M = 3$ [10]. Due to the exponential number of subsets, these approaches cannot generalize to large M.

A better approach is to associate the projection matrix $\mathbf{W}$ with a structural sparsity prior that encourages solutions that satisfy the necessary properties. This strategy was first presented for $M = 2$ by Virtanen et al. [15], and extended for general M independently by several authors [3], [11], [12]. Despite technical differences in how the structural sparsity is obtained, all of these approaches can be seen as special instances of our GFA solution principle. Also the non-Bayesian models that can be used to solve the GFA problem follow the same principle [8], [9].

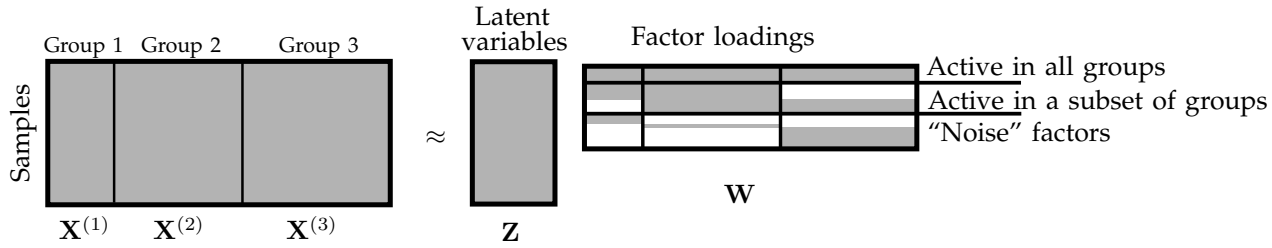

Fig. 1. Illustration of the group factor setup for three groups. The model learns a linear factorization of a data matrix $\mathbf{Y} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{X}^{(3)}]$ whose features have been split into (here) three groups, so that the factor loadings $\mathbf{W}$ are group-wise sparse. The model then automatically learns which factors describe dependencies between either all of the groups or a subset of them, and which describe structured noise specific to each group. The sparsity structure of $\mathbf{W}$ is here represented by coloring; the white areas correspond to zeros whereas the gray areas are non-zero.

3 Model

We propose a novel GFA solution that is another instantiation of the general approach described above. The technical novelty is in a more advanced structural sparsity prior which takes into account possible dependencies between the groups, instead of assuming the group-factor activations to be *a priori* independent as in the earlier solutions. The model can also be interpreted as a two-level model that uses one level to model association strengths between individual groups and the other level to model the observations given the association strength. This interpretation clarifies the conceptual novelty, explicating how the new structural sparsity prior has an intuitive interpretation.

The generative model is the one given in (1) coupled with suitable priors. For $\mathbf{z}_i$ we use the unit Gaussian prior $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and for the noise precisions τ_m we employ a gamma prior with both shape and rate parameters set to 10^{-14}; the model is fairly insensitive to these hyperparameters. To find a GFA solution these standard choices need to be complemented with structured sparse priors for $\mathbf{W}$, described next.

3.1 Sparsity prior

We denote by $\alpha_{m,k}$ the inverse strength of association between the mth group and the kth factor, and directly interpret it as the precision parameter of the prior distribution for $\mathbf{w}_k^{(m)}$, the projection mapping the kth factor to the observed variables in the mth group. That is, we assume the prior

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \prod_{d=1}^{D_m} \mathcal{N}(\mathbf{w}_{k,d}^{(m)}|0, \alpha_{m,k}^{-1}).$$

The same prior was used in our preliminary work [3], where we drew $\alpha_{m,k}$ independently from a flat gamma prior to implement group-wise extension to automatic relevance determination (ARD).

Here we replace the independent draws with a linear model for $\boldsymbol{\alpha}$ to explicitly model the association strengths between group-factor pairs. Since the entries correspond to precisions for the second level projections, we model them in the log-space as

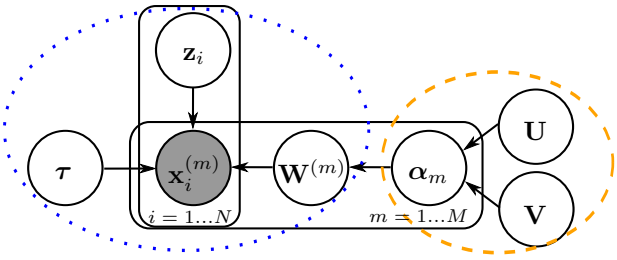

Fig. 2. Plate diagram of group factor analysis. The observation model, used also by earlier GFA solutions, is highlighted by the blue dotted region, whereas the novel low-rank model for the group-factor associations is indicated by the orange dashed region.

$$\log \boldsymbol{\alpha} = \mathbf{U}\mathbf{V}^\top + \boldsymbol{\mu}_u \mathbf{1}^\top + \mathbf{1}\boldsymbol{\mu}_v^\top, \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{M \times R}$ and $\mathbf{V} \in \mathbb{R}^{K \times R}$. The vectors $\boldsymbol{\mu}_u \in \mathbb{R}^M$ and $\boldsymbol{\mu}_v \in \mathbb{R}^K$ model the mean profiles. Here R is the rank of the linear model, and typically $R \ll \min(M, K)$ so that we get a low-rank decomposition for the association strengths, obtained by element-wise exponentiation $\boldsymbol{\alpha} = \exp(\mathbf{U}\mathbf{V}^\top + \boldsymbol{\mu}_u \mathbf{1}^\top + \mathbf{1}\boldsymbol{\mu}_v^\top)$. Finally, we place an element-wise normal prior for the matrices $\mathbf{U}$ and $\mathbf{V}$ with zero mean and precision set to a fixed constant $\lambda = 0.1$; extensions to further hierarchical priors would also be tractable if needed. The resulting GFA model is visualized as a plate diagram in Figure 2, highlighting the two levels of the model.

The motivation for modeling the $\alpha_{m,k}$ instead of assuming them independent comes from the original modeling task of GFA. The goal is to understand the relationships between the groups, and hence we should explicitly model them. The earlier models with independent priors assume that the groups are independent, which is unlikely to hold in practical applications. Our model, in turn, directly represents correlations between the group activation profiles.

An alternative formulation for correlated groups would directly draw $\log \boldsymbol{\alpha}_m$ from a multivariate distri-

bution, such as multivariate normal [16]. However, specifying the correlations for such a model would require $M(M-1)/2$ parameters, making the approach infeasible for large M. Since modeling the correlations is expected to be the most useful for large number of groups, it is clearly beneficial to use the low-rank model that requires only $(M+K)\times(R+1)$ parameters.

3.2 Interpretation

As mentioned above, the model can be interpreted in two alternative ways. The straightforward interpretation is that of a factor analysis model for the D observed variables, with a structural sparsity prior for making the projections implement the GFA properties. This viewpoint illustrates the relationship between the earlier Bayesian solutions for GFA [3], [11], [12]; they follow the same general approach presented in Section 2.2, but our sparsity prior is more advanced.

Perhaps the more interesting interpretation is to consider (2) as the primary model. Then the entries of $\boldsymbol{\alpha}$ are considered as unobserved data describing the groups; $\mathbf{U}$ are the factor loadings and $\mathbf{V}$ provide the latent factors for the groups. The mapping from $\boldsymbol{\alpha}$ to the observations, parameterized by $\mathbf{Z}$ and $\mathbf{W}$, is then merely a set of nuisance parameters. From this perspective, the earlier models presented for the GFA problem are very simple. They do not assume any structure between the groups, but instead draw the association strengths independently. Their results will naturally still reveal such associations, but not as well as the proposed model that models them explicitly.

As we later empirically demonstrate, the rank R of the group association level can typically be very low even for very high-dimensional data collections with a large number of groups. This makes it possible to visually study the associations between the groups, for example via a scatter plot of the columns of $\mathbf{U}$ for $R=2$. We discuss approaches for determining the value of R for practical applications in Section 5.4.

3.3 Inference

For inference, we use mean-field variational approximation. We approximate the posterior with a factorized distribution

$$q(\Theta) = q(\mathbf{Z})q(\mathbf{W})q(\boldsymbol{\tau})q(\mathbf{U})q(\mathbf{V}),$$

where $\Theta = \{\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}, \mathbf{U}, \mathbf{V}\}$, and find the approximation that minimizes the Kullback-Leibler divergence from $q(\Theta)$ to $p(\Theta|\mathbf{Y})$. Equivalently, this corresponds to estimating the marginal likelihood $p(\mathbf{Y})$ with maximal lower bound,

$$\log p(\mathbf{Y}) \geq L(\Theta) = \int q(\Theta)\log\frac{p(\Theta,\mathbf{Y})}{q(\Theta)}.$$

The mean-field algorithm proceeds by updating each of the terms in turn, always finding the parameters that maximize the expected lower bound $L(\Theta)$, given all the other factorized distributions. Since the model uses conjugate priors for everything except $\mathbf{U}$ and $\mathbf{V}$, the updates for most parameters are straightforward and match those of other FA models, for example [13]. The terms $q(\mathbf{U})$ and $q(\mathbf{V})$ are more complex and hence we derive the updates for that part in detail below. The VB inference is summarized in Algorithm 1, and the parameters of the variational distributions can be found in the Appendix. An open-source implementation of the model in R is available as part of the CCAGFA package in CRAN (http://cran.r-project.org/package=CCAGFA).

Algorithm 1 VB inference of GFA

Input: initialized $q(\mathbf{W}), q(\mathbf{Z}), q(\boldsymbol{\tau})$, and either $q(\boldsymbol{\alpha})$ or $\mathbf{U}$ and $\mathbf{V}$.
while not converged **do**
  Check for empty factors to be removed
  $q(\mathbf{W}) \leftarrow \prod_{m=1}^{M}\prod_{j=1}^{D_m}\mathcal{N}(\mathbf{w}_{:,j}^{(m)}|\mathbf{m}_{m,j}^{(w)}, \boldsymbol{\Sigma}_m^{(w)})$
  $q(\mathbf{Z}) \leftarrow \prod_{i=1}^{N}\mathcal{N}(\mathbf{z}_i|\mathbf{m}_i^{(z)}, \boldsymbol{\Sigma}^{(z)})$
  if full-rank GFA ($R = \min(M,K)$) **then**
    $q(\boldsymbol{\alpha}) \leftarrow \prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{G}(a_{m,k}^{\alpha}, b_{m,k}^{\alpha})$
  else
    $\mathbf{U},\mathbf{V} \leftarrow \arg\max_{\mathbf{U},\mathbf{V}} L(\Theta)_{\mathbf{U},\mathbf{V}}$
    $\langle\boldsymbol{\alpha}\rangle \leftarrow \exp(\mathbf{U}\mathbf{V}^\top)$
  end if
  $q(\boldsymbol{\tau}) \leftarrow \prod_{m=1}^{M}\mathcal{G}(\tau_m|a_m^{\tau}, b_m^{\tau})$
end while

For $q(\mathbf{U})$ and $q(\mathbf{V})$ we use fixed-form distributions, that is, we choose point distributions $q(\mathbf{U}) = \delta_{\mathbf{U}}$ and $q(\mathbf{V}) = \delta_{\mathbf{V}}$, and optimize the lower bound numerically[1]. The bound as a function of $\mathbf{U}$ and $\mathbf{V}$ is given by

$$L(\Theta)_{\mathbf{U},\mathbf{V}} = \sum_{m,k}\left(D_m\mathbf{u}_m^\top\mathbf{v}_k - \langle\mathbf{W}^{(m)}\mathbf{W}^{(m)^\top}\rangle_{k,k}e^{\mathbf{u}_m^\top\mathbf{v}_k}\right) + 2\log p(\mathbf{U},\mathbf{V}), \quad (3)$$

where $\langle\mathbf{W}^{(m)}\mathbf{W}^{(m)^\top}\rangle$ denotes the second moment of $\mathbf{W}^{(m)}$ with respect to $q(\mathbf{W}^{(m)})$, and $\log p(\mathbf{U},\mathbf{V})$ collects the prior terms affecting the factorization. Since the parameters $\mathbf{U}$ and $\mathbf{V}$ are highly coupled, we optimize (3) jointly with second order approximate gradient method (L-BFGS), using the gradients

$$\frac{\delta L}{\delta\mathbf{U}} = \mathbf{A}\mathbf{V} + \frac{\delta\log p(\mathbf{U},\mathbf{V})}{\delta\mathbf{U}}, \quad \frac{\delta L}{\delta\mathbf{V}} = \mathbf{A}^\top\mathbf{U} + \frac{\delta\log p(\mathbf{U},\mathbf{V})}{\delta\mathbf{V}},$$

where $\mathbf{A} = \mathbf{D}\mathbf{1}^\top - \exp(\mathbf{U}\mathbf{V}^\top)$. Full variational inference over these parameters would also be possible [18], but we did not consider it necessary for the practical applications.

An interesting special case is obtained when $R = \min(M,K)$. Then the factorization is not low-rank, but instead we can find a unique optimal solution for (3) as

$$\alpha_{m,k} = \frac{D_m}{\langle\mathbf{W}^{(m)}\mathbf{W}^{(m)^\top}\rangle_{k,k}},$$

1. To keep the notation clean we assume the $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_v$ have been appended as parts of $\mathbf{U}$ and $\mathbf{V}$, respectively.

assuming λ is negligibly small. This is identical to the variational update for a model that draws $\alpha_{m,k}$ from a gamma prior with the parameters approaching zero (the uniform distribution limit of gamma). This is the prior used by some of the earlier GFA solutions [3], [12], and hence we get the earlier models as special cases of ours.

The inference scales linearly in M, D and N, and has cubic complexity with respect to K. In practice, it is easily applicable for large data sets as long as K is reasonable (at most hundreds). However, during inference empty factors may occur and in this case they can be removed from the model to speed up and stabilize computation[2].

3.4 Predictive inference

Even though the GFA model is in this work formulated primarily as a tool for exploring relationships between variable groups, it can readily be used also as a predictive model. In this prediction setting new (test) samples are observed for a subset of groups and the task is to predict unobserved groups based on observed data.

For simplicity of presentation, we assume that only the mth group is unobserved. Then the missing data are represented by the predictive distribution $p(\mathbf{X}^{(m)^*}|\mathbf{Y}^{-(m)^*})$, where $\mathbf{Y}^{-(m)^*}$ denotes partially observed test data consisting of all the other groups. However, this distribution involves marginalization over both $\mathbf{W}$ and $\mathbf{Z}$ that is analytically intractable and hence we need to approximate it. In particular, given $\mathbf{Y}^{-(m)^*}$, $q(\mathbf{W})$ and $q(\boldsymbol{\tau})$, we learn the approximate posterior distribution for the latent variables $q(\mathbf{Z}^*)$ corresponding to $\mathbf{Y}^{-(m)^*}$ and approximate the mean of the predictive distribution as

$$\begin{aligned}\langle \mathbf{X}^{(m)^*}|\mathbf{Y}^{-(m)^*}\rangle &= \langle \mathbf{Z}^*\mathbf{W}^{(m)}\rangle_{q(\mathbf{W}^{(m)}),q(\mathbf{Z}^*)} \\ &= \mathbf{Y}^{-(m)^*}\mathbf{T}\langle \mathbf{W}^{-(m)\top}\rangle \boldsymbol{\Sigma}^{-1}\langle \mathbf{W}^{(m)}\rangle, \quad (4)\end{aligned}$$

where $\mathbf{T} = \mathrm{diag}(\{\langle\tau_j\rangle \mathbf{I}_{D_j}\}_{j\neq m})$ and $\boldsymbol{\Sigma} = \mathbf{I}_K + \sum_{j\neq m}\langle\tau_j\rangle\langle \mathbf{W}^{(j)}\mathbf{W}^{(j)^\top}\rangle$. In the experiments we use this mean value for prediction.

4 Related work

The GFA problem and our solution for it are closely related to several matrix factorization and factor analysis techniques. In the following, the related work is discussed from two perspectives. First we cover other techniques for solving the group factor analysis problem or its special cases. Then we proceed to relate the proposed solution to multiple regression, which is a specific use-case for GFA.

4.1 Factor models for multiple groups

For a single group, $M = 1$, the model is equivalent to Bayesian principal component analysis [2], [17]; all of the factors are active for the one group and they describe the variation with linear components. We can also implement sparse FA with the model, by setting $M = D$ so that each group has only one variable. The residual noise has independent variance for each variable, and the projections become sparse because of the ARD prior.

For two groups, $M = 2$, the model is equivalent to Bayesian CCA and inter-battery factor analysis [13]; some factors model the correlations whereas some describe the residual noise within either group.

Most multi-set extensions of CCA, however, are not equivalent to our model. For example, Archambeau et al. [19] and Deleus et al. [20] extend CCA for $M > 2$, but instead of GFA they solve the more limited problem of multiple-battery factor analysis [4], [5]. The MBFA models provide one set of factors that describe the relationships between *all* groups, and then model the variation specific to each group either with a free covariance matrix or a separate set of factors for that group. Besides the multi-set extensions of CCA, also the probabilistic interpretation of sparse matrix factorization [21], and the JIVE model for integrated analysis of multiple data types [22], [23] belong to the family of MBFA models. These models differ in their priors, parameterization and inference, but are all conceptually equivalent.

In recent years, a number of authors have independently proposed solutions for the GFA problem. They all follow the general solution outline presented in Section 2.2 with varying techniques for obtaining the group-wise sparsity. Common to all of them is that they do not explicitly model the relationships between the groups, but instead assume that the choice of whether a factor describes variation in one particular group can be made independently for all factor-group pairs. This holds for the non-Bayesian solutions of multi-view sparse matrix factorizations [8] and the group lasso penalty variant of SCA [9], as well as for the earlier Bayesian GFA models [3], [11], [12]. Compared to these, our model explicitly describes the relationships between the groups, which helps especially for setups with a large number of groups. Finally, we get the sparsity priors of [3] and [12] as special cases of our model.

4.2 Factor regression and group-sparse regression

The GFA problem can also be related to supervised learning, by considering one of the groups as dependent variables and the others as explanatory variables. For just one group of dependent variables (that is, $M = 2$ in total), GFA is most closely related to a setting called factor regression [24]. It shares the goal of learning a set of latent factors that are useful for predicting one group from the other. For more recent advances of factor regression models, see [25], [26]. By splitting the explanatory variables into multiple groups, GFA provides a group-wise sparse alternative for these models. Assuming the split corresponds to meaningful prior information on the structure of the explanatory variables, this will usually

2. We remove the kth factor from the model if $c_k = \sum_{i=1}^{N}\langle \mathbf{z}_{i,k}\rangle^2/N < 10^{-7}$.

(as demonstrated in experiments) reduce overfitting by allowing the model to ignore group-specific variation in predictions.

Other models using group-wise sparsity for regression have also been presented, most notably group lasso [27], [28] that uses a group norm for regularizing linear regression. Compared to GFA, lasso lacks the advantages of factor regression; for multiple output cases it predicts each variable independently, instead of learning a latent representation that captures the relationships between the inputs and outputs. GFA has the further advantage that it learns the predictive models for not only all variables but in fact for all groups at the same time. Given a GFA solution one can make predictions for arbitrary subsets of the groups given another subset, instead of needing to specify in advance the split into explanatory and dependent variables.

5 Technical demonstration

In this section we demonstrate the proposed GFA model on artificial data. To illustrate the strengths of the proposed method we compare it with Bayesian implementations of the most closely related factor models, always using a variational approximation for inference and ARD for complexity control also for the competing methods. In particular, we compare against the regular factor analysis (FA) and its sparse version (sFA) to show that one should not completely ignore the group information. We also compare against MBFA, to demonstrate the importance of modeling also relationships between subsets of the groups, and against the GFA solution of [3], obtained as a special case of the proposed model by setting $R = \min(M, K)$, as an example of a method that makes the group-factor activity decisions independently for each pair. For MBFA we use two different implementations depending on the setup; for low-dimensional data we model the group-specific variation with full-rank covariance, whereas for high-dimensional data we use a separate set of factors for each group; see Klami et al. [13] for discussion on these two alternatives for the special case of $M = 2$.

We also compare against SCA [9], a non-Bayesian solution for the GFA problem using group lasso penalty, using the same initial K as for the Bayesian models, with 5-fold cross-validation for the the group lasso regularization parameter. For predictive inference we compute point estimates for the latent variables of the test samples.

5.1 Evaluation

For evaluating the quality of the models we use an indirect measure of predictive accuracy for left-out groups, based on the intuition that if a factor analysis model is able to make accurate predictions it must have learned the correct structure. In particular, given M groups we will always compute the test data predictive error for each of the groups one at a time, using the rest of the groups as explanatory variables.

Since we use a regression task for measuring the quality, we will also compare GFA against alternative standard solutions for multiple output regression, in addition to the alternative factor models mentioned in the previous section. In particular, we will show comparison results with group lasso [28] and simple regularized multiple output linear regression (MLR) model that ignores the group structure. For MLR the prediction is obtained as $\mathbf{X}^{(m)^*} = \mathbf{Y}^{-(m)^*}\mathbf{B}$, where the weight matrix is given by $\mathbf{B} = (\mathbf{Y}^{-(m)^\top}\mathbf{Y}^{-(m)} + \gamma\mathbf{I})^{-1}\mathbf{Y}^{-(m)^\top}\mathbf{X}^{(m)}$, and γ is a regularization parameter. For this model we learn a separate model for each choice of the dependent variable groups, which results in M-fold increase in computational cost compared to GFA that learns all tasks simultaneously. Furthermore, we validate for the regularization parameters via 10-fold cross-validation within the training set, which further increases the computational load.

5.2 Beyond factor analysis and MBFA

We generated $N = 100$ samples from a GFA model with $D = 30$ split into $M = 3$ equally-sized groups. The true generative model had $K = 7$ factors, including one factor specific for each group as well as for each possible subset between the groups. Figure 3 shows the true model as well as the solutions found by the proposed GFA model (using $R = \min(M, K) = 3$), MBFA, and both regular and sparse FA. The proposed model finds the correct structure, whereas MBFA suggests spurious correlations; each factor describing correlation between just two of the groups is falsely indicating activity also in the third one, according to the MBFA specification. The regular FA result is simply considerably noisier overall, while sparse FA suffers from a few false factor loadings. For this simple demonstration SCA learns the same result as GFA, after manually optimizing the thresholding of component activity. For all methods we set K to a sufficiently large number, allowing ARD to prune out unnecessary components, chose the best solution by comparing the lower bounds of 10 runs with random initializations, and for illustrative purposes ordered the components based on their similarity (cosine) with the true ones.

The conclusion of this experiment is that the proposed method indeed solves the GFA problem, whereas the MBFA and FA solutions do not provide as intuitive and interpretable factors. For this data GFA with $R < 3$ (not shown) fails to unveil the underlying (full-rank) structure. Instead, the loadings lie between those of GFA and FA of Figure 3, which is understandable since FA is closely related to GFA with $R = 0$.

5.3 Performance for several groups

Next we studied how well the proposed solution scales up for more groups. We generated $N = 30$ samples from a GFA model with $K = 18$ true factors. We used $D_m = 7$

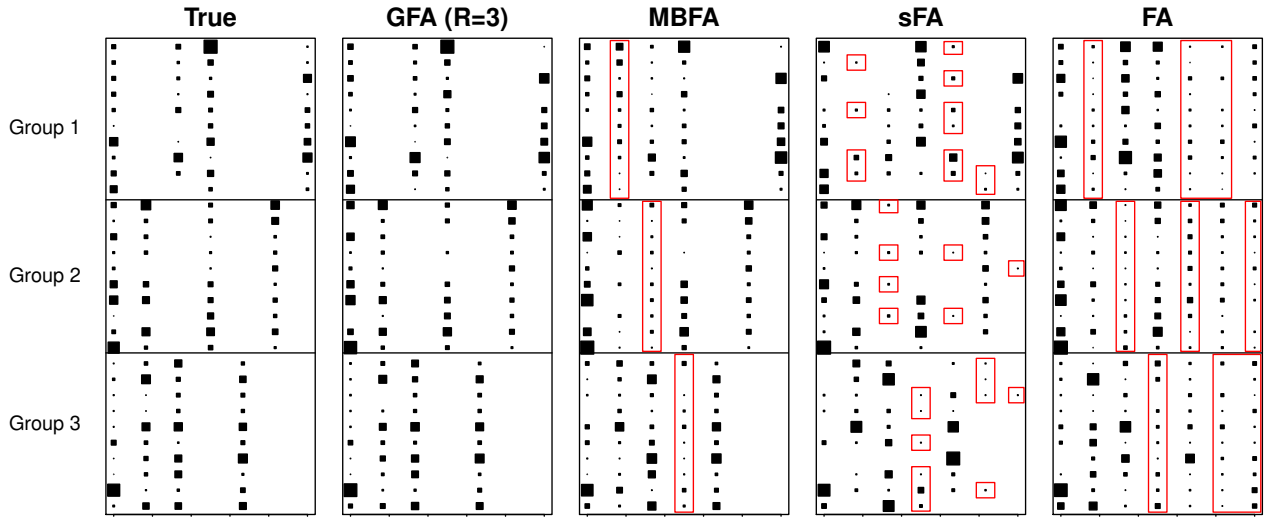

Fig. 3. The projections $\mathbf{W}$ for the true generating model, the proposed GFA solution (using full rank, $R = 3$), and comparison methods multiple-battery factor analysis (MBFA), sparse factor analysis (sFA), and regular factor analysis (FA). GFA finds the true solution, whereas the other methods report spurious correlations (red boxes). The three sub-blocks in each matrix correspond to the three groups of variables, and the areas of the black patches indicate the absolute values of the corresponding weights.

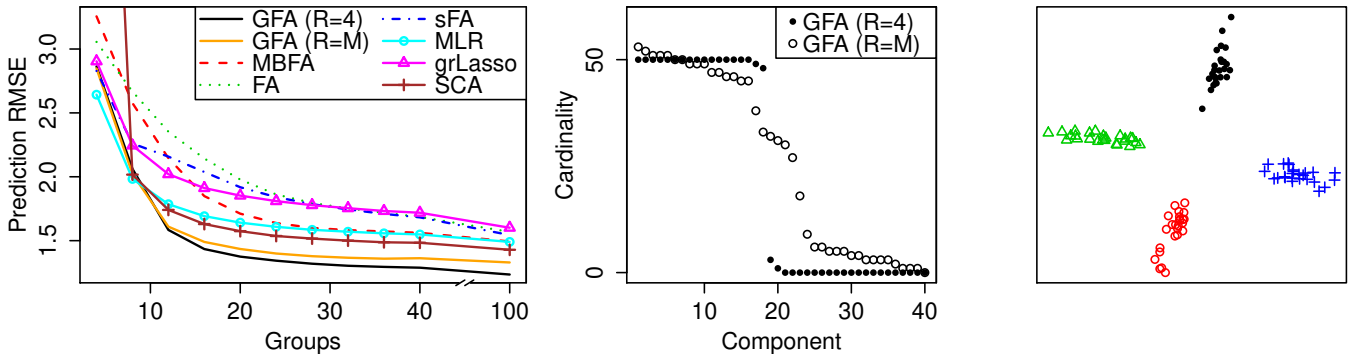

Fig. 4. **Left**: GFA is considerably better in modeling the relationships between the groups compared to MBFA, FA and SCA; the y-axis shows the average prediction error for a new group on the test data, providing an indirect quality measure. For large M explicitly modeling the relationships between the groups (R=4) helps compared to earlier solutions that make the activity decisions independently ($R = \min(M, K)$, here R=M). GFA also outperforms standard regression models in the prediction task, with the exception of MLR for very low dimensional data. **Middle**: For $M = 100$ the $R = 4$ solution correctly finds the $K = 18$ factors that are all active in 50 groups, whereas the earlier solution splits some of the factors into several ones. **Right**: Scatter-plot of matrix $\mathbf{U}$ for a R=2 solution, illustrating clearly the underlying structure of the data. Here the symbols and colors indicate the ground truth types of the groups that were not available for the algorithm during learning.

variables for each group and let the number of groups grow from $M = 4$ to $M = 100$. In other words, the total dimensionality of the data grew from $D = 28$ to $D = 700$. The variable groups were divided into four types of equal size, so that the groups within one type had the same association strengths for all factors. This implies a low-rank structure for the associations.

For validation we used average leave-one-group-out prediction error, further averaged over 50 independent data sets. Figure 4 (left) shows that the proposed model outperforms the other FA methods by a wide margin. For small M the difference between the $R = 4$ and $R = \min(M, K)$ solutions are negligible, since a small number of factor-group association strengths can just as well be selected independently. For large M, however, explicitly modeling the relationships pays off and results in better predictions. The prediction errors for GFA models with rank R between 2 and 10 are very similar, and hence for clarity, only one (R=4) is shown in Figure 4. GFA also outperforms SCA, an alternative group-sparse factor model, for all cases except $M = 8$, for which the two methods are equal. For cases with only 4 or 8 groups multiple linear regression is the most accurate method, but for all other cases, when the total dimensionality

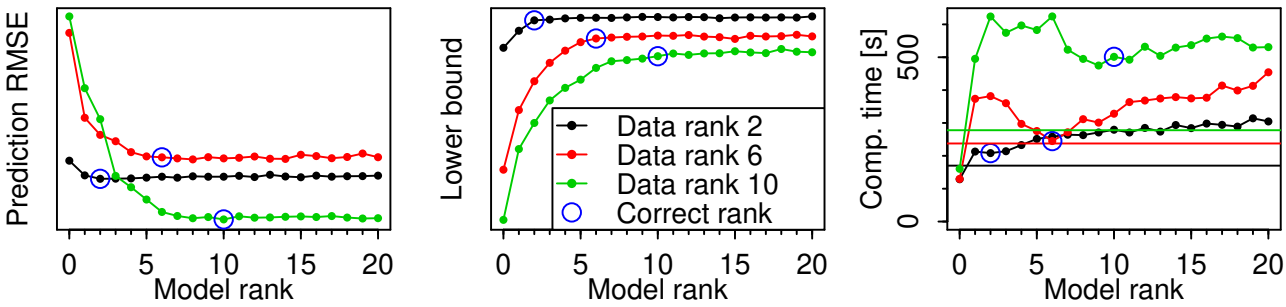

Fig. 5. The 5-fold cross-validation prediction performances (**left**), lower bounds (**middle**) and computation times (**right**) for three different artificial data sets as a function of the model rank R. Both approaches provide a reliable way for choosing the rank R. In the illustrated example the chosen model ranks using cross-validation are 2, 8, and 10 (correct: 2, 6 and 10). By monitoring the lower bounds the true ranks, used to generate the data sets, can be approximately detected from the figure. For clarity, the vertical positions of prediction and lower bound curves were here altered, retaining the relative changes. The **right** subplot illustrates the computational cost as a function of the rank R, compared against the full-rank solution shown as horizontal lines.

increases, GFA clearly outperforms also these supervised methods.

Besides showing better prediction accuracy, the low-rank solution (here $R = 4$) captures the underlying group structure better than the naive model of $R = \min(M, K)$. Figure 4 (middle) compares the two by plotting the number of groups active in each factor for the case with $M = 100$. There are $K = 18$ factors that should each be active in 50 groups. With $R = 4$ we find almost exactly the correct solutions, whereas the naive model finds $K = 40$ factors, many of which it believes to be shared by only $5 - 40$ groups. In other words, it has split some real factors into multiple ones, finding a local optimum, due to making all the activity decisions independently. For illustrative purposes, the components were considered active if the corresponding $\boldsymbol{\alpha}$-values were below 10. The cardinality plot is sensitive to the threshold, but the key result stays the same regardless of the threshold: inferring the component activities independently results in a less accurate model.

Finally, Figure 4 (right) illustrates the relationships between the groups as a scatter plot of the latent factors for the $R = 2$ solution, revealing clearly the four types of variable groups.

5.4 Choosing the model rank

GFA contains a free parameter R and this value needs to be specified by the user. Acknowledging that choosing the correct model order is in general a difficult problem, we resort to demonstrating two practical approaches that seem to work well for the proposed model. The first choice is L-fold cross-validation within the training set, using the predictive performance for left-out groups as the validation criterion.

A computationally more efficient alternative is to use the 'elbow'-principle for the variational lower bounds $L(\Theta)$ computed for different choices of R. The bounds improve monotonically[3] as a function of R, but for typical data sets the growth rate rapidly diminishes after the correct rank, producing an 'elbow'.

We tested both of these principles for 50 independent artificial data sets generated from the GFA model with parameters $N = 50$, $K = 30$, $M = 50$ and $D_m = 10$, for three different data ranks: $R = \{2, 6, 10\}$, representing the kinds of values one would typically expect for real applications. The prediction and lower bound curves as a function of model rank are shown for a representative example in Figure 5. In the 5-fold cross-validation the correct rank was found correctly with over half of the data sets, and the most common error was overestimating the rank by one. The computationally lighter 'elbow'-principle allows the analyst to choose roughly the correct rank by simply comparing the lower bounds.

The rank R influences also the computation time, as illustrated in Figure 5. The computation time increases roughly linearly as a function of R, but for ranks smaller than the optimal the algorithm requires more iterations for convergence which slows it down. In this experiment, the low-rank model is slower than the special case updating $\boldsymbol{\alpha}$ in closed form, but only by a small factor. In the real data experiments reported in Sections 6 and 7 the low-rank variant was slightly faster to compute for all the tested values of R; the relative time spent on updating $\boldsymbol{\alpha}$ becomes small for larger data, and the low-rank models prune out excess factors faster.

In the remaining experiments we do not explicitly select a particular value for R, but instead present the results for a range of different values. This is done to illustrate the relative insensitivity of the model for the precise choice; for both real-world applications a wide range of values outperform the alternative models and also the special case of $R = \min(M, K)$, implying that picking exactly the right rank is not crucial.

3. Given that the constant terms in the priors of $\mathbf{U}$ and $\mathbf{V}$ are ignored.

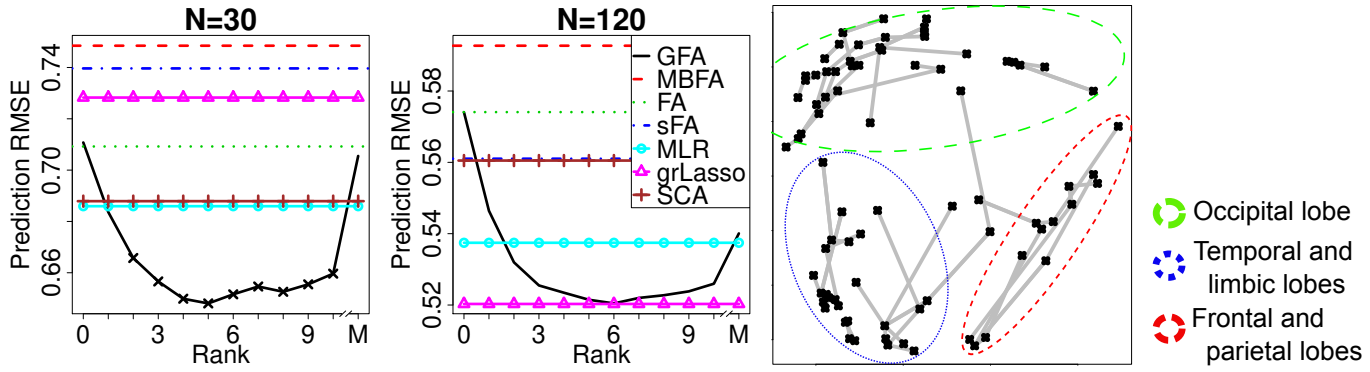

Fig. 6. **Left and middle:** Predictive errors (smaller is better) for $N = 30$ and $N = 120$ training samples. For $N = 30$ the proposed model outperforms the competitors for almost all ranks (the crosses indicate statistically significant difference compared to all competitors), with $R = 5$ providing the best results. For $N = 120$, GFA for ranks $3 - 10$ is significantly better than alternative factor models and MLR, equaling the predictive performance of group lasso for correctly chosen rank R. **Right:** Illustration of the group factor loadings for $R = 2$. The model automatically discovers three groups of ROIs that can be identified as the occipital lobe, temporal and limbic lobes, and frontal and parietal lobes.

6 Analysis of Brain Regions

Functional MRI experiments are commonly used to study the interactions (or connectivity) between brain regions of interest (ROIs) [29], [30]. One way to learn these interactions is based on calculating correlations between individual fMRI BOLD (blood-oxygen-level dependent) signals using PCA or FA models [31].

We applied GFA to brain imaging data to analyze connections between multiple brain ROIs, using fMRI data recorded during natural stimulation with a piece of music [32], for $N = 240$ time points covering two seconds each. We have 11 subjects, and we computed a separate GFA model for each, providing 11 independent experiments. We chose $M = 81$ ROIs, some of which are lateralized, using the FSL software package [33], based on the Harvard-Oxford cortical and subcortical structural atlases [34] and the probabilistic cerebellar atlas [35]. Further, we divided the voxels in these regions to $\sum_m D_m = 676$ local uniformly distributed supervoxels by spatial averaging. In the end, each ROI was represented on average by eight such supervoxels and each supervoxel contained on average 120 voxels. These ROIs and their corresponding dimensionalities are given in Supplementary material available at http://research.ics.aalto.fi/mi/papers/GFAsupplementary.pdf.

For quantitative comparison we again use leave-one-group-out prediction, predicting the activity of each ROI based on the others for unseen test data. We set aside half of the data as test set and train the models varying the amount of training data for $K = 100$. The prediction errors are given in Figure 6, averaged over the ROIs and 11 subjects. The proposed solution outperforms (Wilcoxon signed-rank test, $p < 10^{-6}$) all other factor models for a wide range of ranks, from $R = 2$ to $R = 10$, and in particular is also clearly better than the special case with $R = \min(M, K)$ [3]. For these ranks, with $N = 30$ training samples, GFA also outperforms all the supervised regression models, whereas for a larger training set, $N = 120$, group lasso provides comparable accuracy.

Figure 6 shows also a visualization of $\mathbf{U}$ for $R = 2$, averaged over all subjects using all observations. Since $\mathbf{U}$ is not identifiable, before averaging we projected each $\mathbf{U}$ to a common coordinate system. Each dot is one ROI and the lines connect each ROI to its spatially closest neighbor (minimum Euclidean distance of supervoxels between the corresponding ROIs) to reveal that the model has learned interesting structure despite not knowing anything about the anatomy of the brain. Further inspection reveals that the model partitions visual areas, frontal areas, and auditory areas as separate clusters. Note that these results are a demonstration of the model's capability of discovering structure between groups; for a serious attempt to discover functional or anatomic connectivity further tuning that takes into account the properties of fMRI and the anatomy of the brain should be done.

7 Analysis of Drug Responses

Both chemical descriptors and biological responses can be used to analyze the properties of drugs. However, the shared variation between these two views may reveal more of the drug properties than either one of the views independently [36]. We applied GFA to the drug response data of [37], consisting of the responses of $N = 684$ drugs when applied to three cancer cell lines (*HL60*, *MCF7* and *PC3*). For the analysis, we selected the genes found in CP: Biocarta gene sets[4], where the genes are grouped according to functional pathway information. We preprocessed the data as Khan et al.

4. http://www.broadinstitute.org/gsea/msigdb/collections.jsp

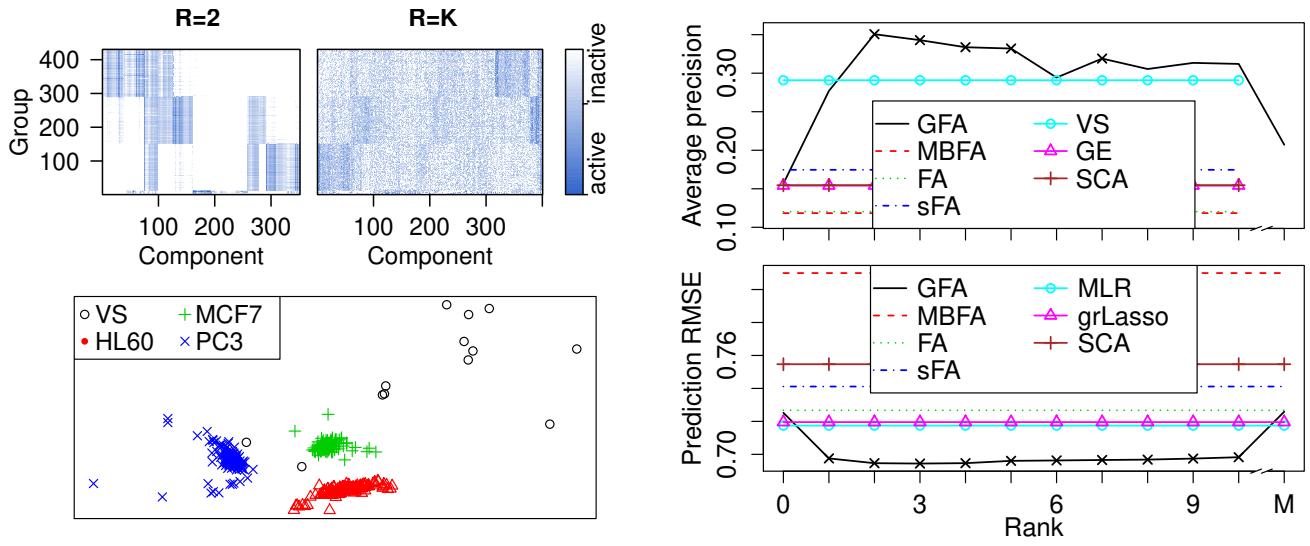

Fig. 7. **Top left:** The group-factor associations of GFA for $R = 2$ and $R = K$ illustrate how the low-rank model reveals much clearer structure. The first 13 groups are the chemical descriptors, followed by 139 feature groups (pathways) for each of the 3 cell lines. **Top right:** Drug retrieval accuracies (higher is better, the crosses indicate statistical significance); the GFA solution is clearly the best for a wide range of ranks, and it is the only method outperforming retrieval based on the chemical descriptors. Here GE denotes raw gene expression of all three cell lines and VS corresponds to the chemical descriptors. Note that the accuracy of SCA is almost identical to GE, which makes the two lines hard to distinguish. **Bottom left:** The group factor loadings $\mathbf{U}$ for $R = 2$, showing how the model is able to separate the three cell lines and the chemical descriptors almost perfectly. **Bottom right:** Predictive errors (smaller is better) for the drug response data. GFA with ranks $R = 1, ..., 10$ outperforms the comparison methods significantly.

[36] and left duplicate genes only in the largest groups, removing groups with less than two genes left. As in [36], we augmented the data by adding 76 chemical descriptors (*VolSurf*) of the drugs, here as separate variable groups. Hence the whole data contain $M = 430$ groups: 13 contain chemical descriptors of the drugs, whereas 139 groups describe the response in each of the three cell lines. The feature groups and their dimensionalities are listed in the Supplementary material available at http://research.ics.aalto.fi/mi/papers/GFAsupplementary.pdf. The total data dimensionality is $D = 3172$.

Figure 7 illustrates GFA results with $K = 400$ on this data (a component amount high enough for GFA with $R < 10$). Already with $R = 2$ the model partitions the groups almost perfectly into four different clusters (bottom left subfigure), one for each of the three cell lines and one for the chemical descriptors. That is, the model effectively learns the underlying structure it has no information about. It also reveals clearer factor-group association structure (top left subfigure) compared to the earlier solution with $R = M$ [3]. With $R = 3$ the four different clusters can be perfectly separated.

Next we quantitatively validated the performance of the different solutions, using a drug retrieval task [36]. Using one of the drugs as a query, we ranked the remaining drugs based on their similarity with the query, and used an external database of drug functions for assessing the retrieval results. By comparing the similarity in the latent space of the models, we can indirectly evaluate the quality of the representations the models have learned. It has been shown that the chemical VolSurf descriptors capture drug functions significantly better than raw gene expression data for this data [36], and hence we computed the similarity measures of all the models based on factors that are active in at least one chemical descriptor group. For this purpose we thresholded the activities by regarding components with $\boldsymbol{\alpha} < 100$ as active; the results were not very sensitive for the exact threshold level. The retrieval can be quantified by measuring the average precision [38], further averaged over all the drugs (separate retrieval tasks), which is a measure that gives most weight to the drugs that are retrieved first. Figure 7 (top right) shows that the proposed solution again outperforms all of the competing methods for all ranks above zero (Wilcoxon signed-rank test, $p < 10^{-6}$), and for $R = 2$ to $R = 5$ and $R = 7$ it significantly ($p < 0.05$) outperforms also the chemical descriptors that are considerably more accurate than any of the competing methods. The shown retrieval accuracies are based on the ten most similar drugs, but the results are consistent for sets of several sizes.

In addition to the retrieval task, we measured the performance of the models in a leave-one-group-out prediction task with $N = 616$. The average errors are shown in Figure 7 (bottom right). GFA with $R = 1, ..., 10$ outperforms all the competing models significantly.

8 Discussion

Joint analysis of multiple data sets is one of the trends in machine learning, and integrated factor analysis of multiple real-valued matrices is one of the prototypical scenarios for that task. In recent years multiple authors have re-discovered the multiple-battery factor analysis (MBFA) task originating from the early works in statistics [4], [5], [6], [7], calling it either multi-set CCA [19], [20], or simply as a model for integrated analysis of multiple data sources [22], [23]. Despite varying technical details, all of these models can be seen as FA models with two sets of factors: one set describes dependencies between all of the variable groups, whereas the other set describes, or explains away, variation specific to each group.

The group factor analysis problem formulated in this article, extending the preliminary treatment in [3], differs from the MBFA models in one crucial aspect. Instead of only modeling relationships between *all* of the groups, we also introduce factors that model relationships between any subset of them. While some other recent works [8], [9], [11], [12] have also addressed the same problem, in this paper the GFA setup is for the first time introduced explicitly, putting it into its statistical context. We described a general solution principle that covers the earlier solutions, identifying the structural sparsity prior as the key element. We then presented a more advanced sparsity prior that results in a novel GFA solution: Instead of choosing the activities of each group-factor pair independently, we explicitly model the relationships between the groups with another linear layer. Our model hence directly provides factor loadings also between the groups themselves, which was exactly the original motivation for the GFA problem. Our preliminary model [3] is a special case with *a priori* independent loadings.

We showed, using artificial data, how the GFA problem and solution differ from the MBFA-problem and classical FA. We also demonstrated that, especially for a large number of groups or data sets, it pays off to explicitly model the relationships between the groups. Finally, we applied the model on two real-world exploratory analysis scenarios in life sciences. We demonstrated that the model is applicable to connectivity analysis of fMRI data, as well as for revealing structure shared by structural description of drugs and their response in multiple cell lines. These demonstrations illustrated the kinds of setups the GFA is applicable for, but should not be considered as detailed analyses of the specific application problems.

Besides showing that the proposed model solves the GFA problem considerably better than the alternatives MBFA, FA and SCA [9], the empirical experiments revealed that there is a qualitative difference between the proposed model having the more advanced structural sparsity prior and the earlier GFA solutions such as [3]. Even though the earlier models also solve the GFA problem reasonably well, they are outperformed by supervised regression models in predictive tasks. The proposed solution with a low-rank model for the group association strengths is clearly more accurate in prediction tasks and, at least for small training sets, outperforms also dedicated regression models trained specifically to predict the missing groups. This is a strong result for a model that does not know in advance which groups correspond to explanatory variables and which to the dependent variables, but that instead learns a single model for all possible choices simultaneously.

The model presented here is limited to scenarios where each training sample is fully observed. Support for missing observations could be added using the fully factorized variational approximation used for PCA and collective matrix factorization with missing data [17], [39]. A similar approach can also be used for semi-paired setups where some samples are available only for some groups [40], by filling in the remaining groups by missing observations. Empirical comparisons on these are left for future work. Another possible direction for future work concerns more justified inference for the rank parameter R; even though the experiments here suggest that the method is robust to the choice, the method would be more easily applicable if it was selected automatically.

Acknowledgment

We thank the Academy of Finland (grant numbers 140057, 266969, and 251170; Finnish Centre of Excellence in Computational Inference Research COIN), the aivoAALTO project of Aalto University, and Digile ICT SHOK (D2I programme) for funding the research. We would also like to thank Suleiman A. Khan for his help with the biological application, and Enrico Glerean for his help with the neuroscience application. We acknowledge the computational resources provided by Aalto Science-IT project.

References

[1] L. Thurstone, "Multiple factor analysis," *Psychological Review*, vol. 38, no. 5, pp. 406–427, 1931.

[2] C. M. Bishop, "Bayesian PCA," *Advances in Neural Information Processing Systems*, vol. 11, pp. 382–388, 1999.

[3] S. Virtanen, A. Klami, S. A. Khan, and S. Kaski, "Bayesian group factor analysis," *Proc. 15th Int. Conf. Artificial Intelligence and Statistics*, pp. 1269–1277, 2012.

[4] R. McDonald, "Three common factor models for groups of variables," *Psychometrika*, vol. 37, no. 1, pp. 173–178, 1970.

[5] M. Browne, "Factor analysis of multiple batteries by maximum likelihood," *British J. Mathematical and Statistical Psychology*, vol. 33, no. 2, pp. 184–199, 1980.

[6] L. R. Tucker, "An inter-battery method of factor analysis," *Psychometrika*, vol. 23, no. 2, pp. 111–136, 1958.

[7] M. W. Browne, "The maximum-likelihood solution in inter-battery factor analysis," *British J. Mathematical and Statistical Psychology*, vol. 32, no. 1, pp. 75–86, 1979.

[8] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," *Advances in Neural Information Processing Systems*, vol. 23, pp. 982–990, 2010.

[9] K. V. Deun, T. F. Wilderjans, R. A. v. Berg, A. Antoniadis, and I. V. Mechelen, "A flexible framework for sparse simultaneous component based data integration," *BMC Bioinformatics*, vol. 12, no. 1, pp. 448, 2011.

[10] S. K. Gupta, D. Phung, B. Adams, and S. Venkatesh, "A Bayesian framework for learning shared and individual subspaces from multiple data sources," *Proc. Advances in Knowledge Discovery and Data Mining, 15th Pacific-Asia Conf., PAKDD*, pp. 136–147, 2011.

[11] S. K. Gupta, D. Phung, and S. Venkatesh, "A Bayesian nonparametric joint factor model for learning shared and individual subspaces from multiple data sources," *Proc. 12th SIAM Int. Conf. Data Mining*, pp. 200–211, 2012.

[12] A. Damianou, C. Ek, M. Titsias, and N. Lawrence, "Manifold relevance determination," *Proc. 29th Int. Conf. Machine Learning*, pp. 145–152, 2012.

[13] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *J. Machine Learning Research*, vol. 14, no. 1, pp. 899–937, 2013.

[14] A. Klami and S. Kaski, "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, vol. 72, no. 1, pp. 39–46, 2008.

[15] S. Virtanen, A. Klami, and S. Kaski, "Bayesian CCA via group sparsity," *Proc. 28th Int. Conf. Machine Learning*, pp. 457–464, 2011.

[16] J. Aitchison, "The statistical analysis of compositional data," *J. Royal Statistical Society. Series B (Methodological)*, vol. 44, no. 2, pp. 139-177, 1982.

[17] A. Ilin, and T. Raiko, "Practical approaches to principal component analysis in the presence of missing data," *J. Machine Learning Research*, vol. 11, pp. 1957-2000, 2010.

[18] O. Dikmen and C. Févotte, "Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2267–2275, 2011.

[19] C. Archambeau and F. Bach, "Sparse probabilistic projections," *Advances in Neural Information Processing Systems*, vol. 21, pp. 73–80, 2009.

[20] F. Deleus and M. V. Hulle, "Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis," *J. Neuroscience Methods*, vol. 197, no. 1, pp. 143–157, 2011.

[21] X. Qu and X. Chen, "Sparse structured probabilistic projections for factorized latent spaces," *Proc. 20th ACM Int. Conf. Information and Knowledge Management*, pp. 1389–1394, 2011.

[22] E. Lock, K. Hoadley, J. Marron, and A. Nobel, "Joint and individual variation explained (JIVE) for integrated analysis of multiple datatypes," *Annals of Applied Statistics*, vol. 7, no. 1, pp. 523–542, 2013.

[23] P. Ray, L. Zheng, Y. Wang, J. Lucas, D. Dunson, and L. Carin, "Bayesian joint analysis of heterogeneous data," Duke Univ., Dept. of Elec. and Comp. Eng., Tech. Rep., 2013.

[24] M. West, "Bayesian factor regression models in the large p, small n paradigm," *Bayesian Statistics*, vol. 7, pp. 733–742, 2003.

[25] N. Chen, J. Zun, and E. Xing, "Predictive subspace learning for multi-view data: a large margin approach," *Advances in Neural Information Processing Systems*, pp.361–369, 2010.

[26] P. Rai, and H. Daumé, "The Infinite Hierarchical Factor Regression Model," *Advances in Neural Information Processing Systems*, vol. 21, pp. 1321–1328, 2009.

[27] M. Yuan, and Y. Li, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[28] P. Breheny, and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Statistics and Computing*, pp.1–15, 2013.

[29] K. J. Friston, "Functional and effective connectivity in neuroimaging: a synthesis," *Human Brain Mapping*, vol. 2, no. (1-2), pp. 56–78, 1994.

[30] S. Smith et al., "Correspondence of the brains functional architecture during activation and rest," *Proc. National Academy of Sciences (PNAS)*, vol. 106, no. 31, pp. 13040–13045, 2009.

[31] K. Friston, C. Frith, P. Liddle, and R. Frackowiak, "Functional connectivity: the principal-component analysis of large (PET) data sets,"*J. Cerebral Blood Flow and Metabolism*, vol. 13, no. 1, pp. 5–14, 1993.

[32] V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico, "Large-scale brain networks emerge from dynamic processing if musical timbre, key and rhythm," *NeuroImage*, vol. 59, no. 4, pp. 3677–3689, 2012.

[33] M. Jenkinson, C. Beckmann, T. Behrens, M. Woolrich, and S. Smith, "FSL," *NeuroImage*, vol. 62, no. 2, pp. 782–790, 2012.

[34] R. Desikan, F. Segonne, B. Fischl, B. Quinn, B. Dickerson, D. Blacker, R. Buckner, A. Dale, R. Maguire, B. Hyman, M. Albert, and R. Killiany, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 980–986, 2006.

[35] J. Diedrichsen, J. Balster, E. Cussans, and N. Ramnani, "A probabilistic MR atlas of the human cerebellum," *NeuroImage*, vol. 46, no. 1, pp. 39–46, 2009.

[36] S. A. Khan, A. Faisal, J. P. Mpindi, J. A. Parkkinen, T. Kalliokoski, A. Poso, O. P. Kallioniemi, K. Wennerberg, and S. Kaski, "Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs," *BMC Bioinformatics*, vol. 13, no. 112, pp. 1–15, 2012.

[37] J. Lamb, E. Crawford, D. Peck, J. Modell, I. Blat, M. Wrobel, J. Lerner, J. Brunet, A. Subramanian, K. Ross *et al.*, "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[38] R. Baeza-Yates, B. Ribeiro-Neto, "Modern information retrieval," ACM Press New York, 1999.

[39] A. Klami, G. Bouchard, and A. Tripathi, "Group-sparse embeddings in collective matrix factorization," *Proc. Int. Conf. Learning Representations*, 2014.

[40] X. Chen, S. Chen, H. Xue, and X. Zhou, "A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data," *Pattern Recognition*, vol. 45, no. 5, pp. 2005–2018, 2012.

APPENDIX

The latent variables are updated as $q(\mathbf{Z}) = \prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|\mathbf{m}_i^{(z)}, \mathbf{\Sigma}^{(z)})$, where

$$\mathbf{\Sigma}^{(z)} = \left(\mathbf{I}_k + \sum_{m=1}^{M} \langle \tau_m \rangle \langle \mathbf{W}^{(m)} \mathbf{W}^{(m)\top} \rangle \right)^{-1}$$

$$\mathbf{m}_i^{(z)} = \sum_{m=1}^{M} \mathbf{\Sigma}^{(z)} \langle \mathbf{W}^{(m)} \rangle \langle \tau_m \rangle \mathbf{x}_i^{(m)}.$$

The projection matrices are updated as $q(\mathbf{W}) = \prod_{m=1}^{M} \prod_{j=1}^{D_m} \mathcal{N}(\mathbf{w}_{:,j}^{(m)}|\mathbf{m}_{m,j}^{(w)}, \mathbf{\Sigma}_m^{(w)})$, where

$$\mathbf{\Sigma}_m^{(w)} = \left(\langle \tau_m \rangle \sum_{i=1}^{N} \langle \mathbf{z}_i \mathbf{z}_i^\top \rangle + \langle \overline{\overline{\boldsymbol{\alpha}}}_m \rangle \right)^{-1}$$

$$\mathbf{m}_{m,j}^{(w)} = \mathbf{\Sigma}_m^{(w)} \langle \tau_m \rangle \left(\sum_{i=1}^{N} x_{ij}^{(m)} \langle \mathbf{z}_i \rangle \right),$$

and $\overline{\overline{\boldsymbol{\alpha}}}_m$ is the mth row of $\boldsymbol{\alpha}$ transformed into a diagonal $K \times K$ matrix.

Noise precision $q(\boldsymbol{\tau}) = \prod_{m=1}^{M} \mathcal{G}(\tau_m|a_m^\tau, b_m^\tau)$ parameters are updated as

$$a_m^\tau = a^\tau + \frac{D_m N}{2}$$

$$b_m^\tau = b^\tau + \frac{1}{2} \sum_{i=1}^{N} \left\langle (\mathbf{x}_i^{(m)} - \mathbf{W}^{(m)\top} \mathbf{z}_i)^2 \right\rangle.$$

Finally, for the low-rank model, $\boldsymbol{\alpha} = e^{\mathbf{U}\mathbf{V}^\top + \boldsymbol{\mu}_u \mathbf{1}^\top + \mathbf{1}\boldsymbol{\mu}_v^\top}$ is updated by optimizing the lower bound numerically. The bound as a function of $\mathbf{U}$ and $\mathbf{V}$ is given by

$$\sum_{m,k} D_m \log(\alpha_{m,k}) - \langle \mathbf{W}^{(m)} \mathbf{W}^{(m)^\top} \rangle_{k,k} \alpha_{m,k}$$
$$- \lambda(\text{tr}(\mathbf{U}^\top \mathbf{U}) + \text{tr}(\mathbf{V}^\top \mathbf{V})).$$

The gradients w.r.t. the cost function are given as

$$\frac{\delta L}{\delta \mathbf{U}} = \mathbf{A}\mathbf{V} + \lambda \mathbf{U}, \qquad \frac{\delta L}{\delta \boldsymbol{\mu}_v} = \mathbf{A}\mathbf{1},$$
$$\frac{\delta L}{\delta \mathbf{V}} = \mathbf{A}^\top \mathbf{U} + \lambda \mathbf{V}, \qquad \frac{\delta L}{\delta \boldsymbol{\mu}_v} = \mathbf{A}^\top \mathbf{1},$$

where $\mathbf{A} = \mathbf{D}\mathbf{1}^\top - \exp(\mathbf{U}\mathbf{V}^\top + \boldsymbol{\mu}_u \mathbf{1}^\top + \mathbf{1}\boldsymbol{\mu}_v^\top)$.

With full rank [3] the ARD parameters are updated as $q(\boldsymbol{\alpha}) = \mathcal{G}(a_m^\alpha, b_{mk}^\alpha)$, where

$$a_m^\alpha = a^\alpha + \frac{D_m}{2}$$
$$b_{mk}^\alpha = b^\alpha + \frac{\mathbf{w}_k^{(m)\top} \mathbf{w}_k^{(m)}}{2}.$$