A LATENT VARIABLE MODEL FOR SIMULTANEOUS DIMENSIONALITY REDUCTION AND CONNECTIVITY ESTIMATION

Ricardo Pio Monti[1] *and Aapo Hyvärinen*[1,2]

[1]Gatsby Computational Neuroscience Unit, University College London, UK
[2]Department of Computer Science and HIIT, University of Helsinki, Finland
{r.monti,a.hyvarinen}@ucl.ac.uk

ABSTRACT

Connectivity estimation is a fundamental problem in many areas of science. However, in the context of high-dimensional data it may be neither feasible nor useful to model the connectivities between all observed variables. Grouping variables into clusters or communities is a useful preprocessing step, but it is not clear how to do so optimally in view of connectivity estimation. A further practical problem is that we may have data from different classes (e.g. multiple subjects in an experiment), and we need to incorporate useful constraints about the similarities between the classes. In this abstract, we present a latent variable model to simultaneously address both of the aforementioned challenges. The model is essentially a factor analysis model where the factors (i.e., latent variables) are allowed to have arbitrary correlations. The associated factor loading matrix is constrained to express a community structure via the introduction of non-negativity and orthonormality constraints. Such constraints also allow us to prove the identifiability of the model, providing a clear interpretation for latent factors. Experimental results demonstrate the capabilities of the proposed model.

1. INTRODUCTION

Understanding the connectivity structure between observed variables is a fundamental problem in statistics and machine learning. Probabilistic methods are often based on estimation of the covariance matrix or it inverse. However, in practice we often have very high-dimensional data, and it may not be useful or feasible to estimate the connectivities between all of them. It is important to somehow reduce the number of variables so that the connectivity estimation is feasible, and furthermore, such reduction can greatly facilitate interpretation of the results. A relevant challenge is how such reduction in the number of variables can explicitly account for connectivity over latent variables whilst also accommodating data over multiple related classes (such as subjects in a biomedical setting).

In this work, we propose a latent variable model which is able to directly address the aforementioned issues[1]. The proposed model consists of a low dimensional set of latent variables in a factor analytic model. The associated factor loading matrix is shared across classes and constrained to be non-negative and orthonormal, thereby encoding module/community membership along its columns. In this manner, we may interpret the factors as activations in modules or communities. Importantly, and in contrast to almost all related models, the latent variables have full (i.e., non-diagonal) covariance structure which we term *latent connectivities*, giving the connectivity structure of the non-overlapping modules. We allow for the connectivity structure to vary across classes; however, the model is also applicable on data from a single class. Thus, we model both the grouping of variables and the connectivity between groups in a single probabilistic model. Our work is motivated by applications relating to fMRI data, where the estimation of "functional connectivity" networks are often modeled as covariance graphs. The modular structure of such networks, where regions cluster into non-overlapping modules, justifies the proposed model.

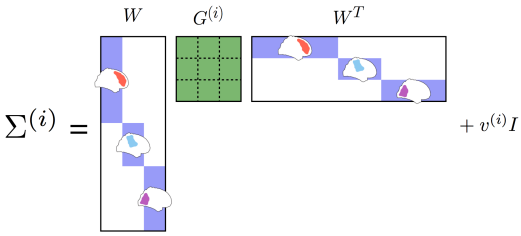

Figure 1: Visualization of the proposed covariance model. The factor loading matrix, W, is shared across classes and serves to denote membership into non-overlapping communities (in this example, communities encode brain modules). The *latent connectivity* across modules, parameterized by $G^{(i)}$, are allowed to vary across classes.

2. LATENT CONNECTIVITIES MODEL

In this section we describe a latent variable model to accurately find modules (communities, clusters) and model their connectivities, possibly across multiple related classes. We assume we have access to multivariate data over N distinct classes, but all results allow for the case $N = 1$ as well. For a given class i, we write $X^{(i)} \in \mathbb{R}^p$ to denote the p-dimensional observed random vector. The ith class is as-

[1]This abstract describes work presented in [1], where further details are provided.

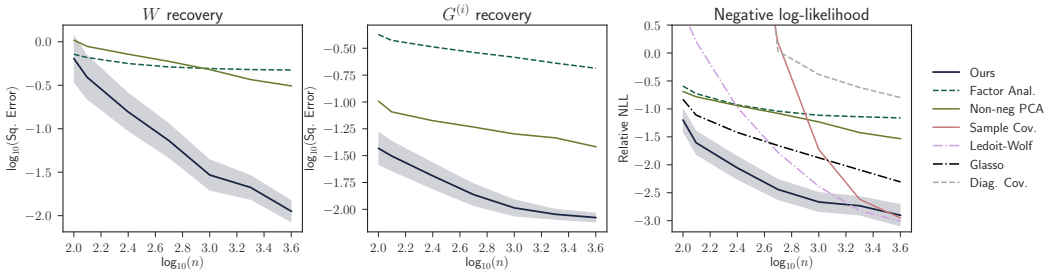

Figure 2: Simulated data results with $N = 10$ classes. Left and middle panels plot the mean squared error for the estimated loading and latent variable covariance matrices as a function of sample size, n. Right panels shows the mean negative log-likelihood for unseen data as a function of sample size, n. Shaded regions correspond to 95% error bars.

sociated with a k-dimensional latent vector, $Z^{(i)}$, which is related to observations, $X^{(i)}$, via a loading matrix $W \in \mathbb{R}^{p\times k}$. We note that the loading matrix is shared across all classes and will serve to encode module memberships across classes. We assume that the data for each class follows a stationary multivariate Gaussian distribution with zero mean and covariance $\Sigma^{(i)} \in \mathbb{R}^{p\times p}$. We further assume latent variables are Gaussian, such that:

$$Z^{(i)} \sim \mathcal{N}\left(0, G^{(i)}\right) \tag{1}$$

$$X^{(i)}|Z^{(i)} = z^{(i)} \sim \mathcal{N}\left(Wz^{(i)}, v^{(i)}I\right). \tag{2}$$

Traditional factor analysis or probabilistic PCA correspond to the special case where $G^{(i)}$ are diagonal. However, by allowing latent variables to have full (i.e., non-diagonal) covariance structure our model is able to capture low-rank connectivity structure as $\Sigma^{(i)} = WG^{(i)}W^T + v^{(i)}I$.

It follows that the loading matrix W serves to encode reproducible covariance structure which is present across all classes. Moreover, the loading matrix is constrained to be non-negative and orthonormal. This leads to a loading matrix with at most one non-zero entry per row. We may interpret the columns of W as encoding membership to k non-overlapping modules. Figure 1 provides an overview of the proposed model in the context of estimating brain connectivity networks. It is important to note that the introduction of marginally dependent latent variables is not possible in the context of traditional factor analysis, since the effects of factor connectivity and factor loadings cannot be distinguished. However, due to the non-negativity and orthonormality constraints on the loading matrix, it is possible to identify the latent connectivities in our model [1].

2.1. Parameter estimation

The parameters associated with the proposed model consist of the loading matrix, W, the latent variable covariances, $\{G^{(i)}\}$, and the observation noise, $\{v^{(i)}\}$. The standard approach to estimate such parameters would be maximum likelihood estimation. However, this results in an iterative algorithm where the computational cost of each parameter update is $\mathcal{O}(p^3)$. Instead, we propose to estimate parameters by score matching [2], leading to an algorithm with a computational cost of $\mathcal{O}(p^2k)$ per iteration. Non-negativity and orthogonality constraints are enforced via the introduction of Lagrange multipliers. Details are provided in [1].

2.2. Experimental results

Synthetic data was generated according to the model described in equations (1-2). The covariance structure for latent variables, $G^{(i)}$, was randomly generated. The dimensionality of observations and latent variables was set to $p = 50$ and $k = 5$ respectively. The number of observations per class, n, was allowed to vary. Data was generated in this manner for $N = 10$ distinct classes and each experiment was repeated 500 times. Results are provided in Figure 2, where we note that the proposed method is able to accurately recover both the loading matrix as well as latent variable connectivities. Alternative methods such as factor analysis and non-negative PCA perform poorly as they do not explicitly model the marginal dependency across latent variables. In terms of mean negative log-likelihood over unseen data, the proposed method out-performs alternative methods for small and moderate sample sizes.

3. CONCLUSION

We propose an extension of factor analysis which simultaneously performs a grouping of variables and estimates their *latent connectivities*. Experiments on synthetic data demonstrate the capabilities of the proposed method.

4. REFERENCES

[1] R. P. Monti and A. Hyvärinen, "A unified probabilistic model for learning latent factors and their connectivities from high-dimensonal data," *arXiv preprint arXiv:1805.09567. To Appear: 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[2] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–708, 2005.