PACM IDeAS Seminar

Joint Independent Subspace Analysis November 4, 2015

Lecturer: Dana Lahat Scribe: Kiran Vodrahalli

Contents

1 Overview

This talk gives an overview of blind source separation with various statistical independence assumptions, generalizing ICA to learning subspaces of low-rank instead of just rank 1 subspaces.

2 Introduction

My talk is about a topic that many of you have heard of, blind source separation. I will explain blind source separation and some more recent methods maybe less known for doing blind source separation. The difference of this talk from other talks is that we will try to put an emphasis on the data fusion aspect of blind source separation.

In natural settings, we take measurements and use sensors. These can be our eyes or ears, or technological sensors. Usually, natural setups record not only the phenomenon of interest but also additional contributions. These can be partitioned into **sources of interest** or **noise**.

A general model to look at this is as follows: We can approximate $x = f(z)$ where $z = \{z_1, \cdots, z_V\}$ where these are signals, parameters or other latent phenomena. f represents the transformation, i.e. channel effects. We are interested where z is unknown and cannot be observed directly without the intermediate transformation f, which may also not be known. We call the unknowns latent variables.

We define the inverse problem to estimate as precisely as possible z and f given x. We would also like to find the simplest explanatory model (done in exploratory research). Finally, if the dimension of measurements is large, we would like to recover the sallest size of z that best explains the observations. This can be regarded as compression, which is useful in large-scale data analysis.

3 Methods

3.1 Model driven methods

We rely on an explicit realistic model of the underyling processes, which are successful if the model is right. These are not always the best choice for the data since sometimes the underlying model of the signals or the transformation between signals and sensors is too complicated and we don't know them. They can vary rapidly and so on.

3.2 Data driven approaches

You make very few assumptions on the data/model and use the simplest models possible, where simple means "use linear relationships", avoid priors and model-dependent parameteres. Sparsity, non-negativity, smoothness, statistical independence and so on are examples of assumptions we make.

In some sense, a data drive approach is self-cntained, in the sense that it relies only on observations and their assumed model. Thus, they are known as "blind" methods. If we want to draw knowledge about the latent variables, we must constrain the degrees of freedom so that the problem is actually solveable with a unique solution (well-posed).

4 What is blind source separation?

4.1 A simple linear model

We take $x = f(z)$ is represented by

$$x_j(t) = \sum_{i=1}^{N} a_{ji} s_i(t)$$

where $f = \{a_1, \cdots, a_N\}$; $z = \{s_1, \cdots, s_N\}$. The x are measurements and the a are mixes. As a matrix,

$$X = AS^T = \sum_{i=1}^{N} \mathbf{a}_i \mathbf{s}_i^T$$

Each of these low-rank terms represents a different underlying phenomenon of interest. In BSS, we want to factorize this set of observations into a sum of rank 1 terms. In the more general case, we would like to represent as a sum of low-rank terms with observable factors.

This factorization is not unique in general since $X = AS^T = (AZ^{-1})(ZS^T)$ for any invertible Z. Uniqueness is necessary to achieve interpretability (attach physical meaning to output). If we simply order the low rank terms and within scaling factors, we can write

$$X = \sum_{i=1}^{n} \mathbf{a}_k \lambda^{-1} \cdot \lambda \mathbf{s}_i^T$$

Think of this as reducing the general invertibility of Z into a specifically diagonal matrix, which has some useful meaning.

5 Uniqueness and Diversity

Definition 5.1. Diversity is any type of constraint or assumption on underyling variables that helps achieve essential uniqueness; for instance, orthogonality, sparsity, nonegativity, statistical independence. Sometimes you need more than one type of diversity. At the very least, each reduces number of degrees of freedom in the model.

Statistical independent means the columns of S are independent random variables, with A deterministic. Then this gives rise to independent components analysis (ICA). In some cases, it makes sense, in other cases, statistical independence does not make sense.

6 Blind source separation with statistical independence

We have $\mathbf{x}(t) = A\mathbf{s}(t)$, and assume real values and $A \in \mathbb{R}^{M\times M}$ is invertible, and that the number of sensors and latent sources is M. The number of latent sources is given by nature, and the number of sensors is given by technology and so on. We assume that the sensors/sources matrix is square for convenience. We also make the assumtion that the covariance matrix of sources is diagonal. Thus we define $X = \mathbf{E}[xx^T] = ASA^T$, and $x(t) = \sum_{i=1}^{M} \mathbf{a}_k\lambda^{-1} \cdot \lambda\mathbf{s}_i^T$ is a sum or statistically independent rank 1 terms where λs account for scale ambiguity.

If we count the number of degrees of freedom, just using second-order statistics and assuming stationarity, we see that it is not enough to constrain the model. The three most well-known/used aspects of ICA is to use higher order statistics. For instance, non-gaussinity, non-stationarity, non-whiteness (temporal correlation between samples if sources are stationary). Here a novel fourth approach is to use data fusion. The idea is to obtain the desired uniqueness if we properly exploits links between data sets. We can separate sources where the first three methods cannot.

6.1 Non-stationarity or time correlation in stationary signals

We can look at a set of blind source separation mixtures. The mixing process is the same over these problems. Basically we have $X^{(k)} = AS^{(k)}A^T$ where $S^{(k)} = \mathbf{E}[s(t)s(t)^T]$ and the $S^{(k)}$ are all distinct. This is basically joint diagonalization. We took two types of measurements and fused them using the common parameter which is the mixing mapping A.

6.2 Independent Vector Analysis

Each $x^{(k)} = \sum_{i=1}^{M} a_i^{(k)}s_i^{(k)}$. We assume that there are dependencies between all $s_1^{(k)}$, $s_2^{(k)}$, and so on. In this model, we would like to estimate these dependencies. We want covariance matrices and the cross-covariances between datasets. We allow more flexible links between datasets. We call this generalized joint diagonalization with a set of mixing matrices on the left and on the right. It can be proved that this model provides sufficient constraints for unique decomposition. We have

$$X^{(k,l)} = A^{(k)}S^{(k,l)}A^l$$

where $S^{(k,l)}$ is diagonal. Once we can exploit the statistical dependencies, automatically it guarantees a fixed frame of reference for the mixtures.

6.3 Sum of low-rank terms

We can relax sum of rank 1 terms to sum of low-rank terms. Natural signals sometimes have more complex and rich properties, and we would like to represent these natural sources and signals within

a subspace. For computational simplicity, let us assume invertibility and square. thus we write

$$x(t) = \sum_{i=1}^{N} A_i s_i(t) = \sum_{i=1}^{N} (A_i Z_i^{-1}) \cdot (Z s_i(t)) = \sum_{i=1}^{N} x_i(t)$$

where the sum terms are rank m_i. How do you know the m_i? This is just a flexibility: you still have to choose these. In practice, you have to use trial and error and see which model gives you the best results. You need to do cross-validation. We now have a joint block diagonalization problem instead (since we no longer have all rank 1 terms, we get blocks along diagonal). Here we are estimating subspaces of dimension greater than one. Thus we write

$$X^{(k,l)} = A^{(k)} S^{(k,l)} A^{(l)}$$

where $S^{(k,l)}$ is **block diagonal**. This just means that we have

$$\begin{bmatrix} S_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & S_{NN} \end{bmatrix}$$

where $S_{ii} \in \mathbb{R}^{m_i \times m_i}$. Again recall that these m_i are hyperparameters you can choose before hand (more general than ICA where $m_i = 1$ for all i).

Here note that we are only using second-order statistics assumptions. We claim that IF it is true that the sources are independent, then the algorithm will do the correct thing and give you independence in the higher moments. However, if the assumption does not hold, then you will do no better than using second-order statistics (your higher order moments won't have any guarantee on independence etc). If you have good reasons to assume your data has statistical independence, then these methods may be good to use. If the assumptions don't hold, then you may have very poor performance, and you should always validate.

Another way to think of this is as follows: Take the covariance matrix for the data and then permute to get a block diagonal structure: It's a better way to organize the data.

6.4 fMRI and EEG application

We are recording same signals from the same brain, but fMRI may have a different representation from the representation of EEG. This gives us some **flexibility** in using different subspaces, though the signal are still related.

6.5 Optimization

For dependence across mixtures, you can do joint block diagonalization. If the data is Gaussian, we kind of the maximum likelihood, then we obtain minimal mean square error for JISA.

6.6 Uniqueness and identifiability

We have proved that JISA is generally unique up to specific cases: If there is an equivalence between latent sources, then this model is blind to the difference betwen them and is blind to the difference between them. We may know that a certain element in each modality is dependent on another specific element in another modality. Regardless of this, we can still exploit links between links between datasets for a joint factorization of the whole model.