JMLR: Workshop and Conference Proceedings 25:1–14, 2012 Asian Conference on Machine Learning

Topographic Analysis of Correlated Components

Hiroaki Sasaki HSASAKI@CC.UEC.AC.JP
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan

Michael U. Gutmann MICHAEL.GUTMANN@HELSINKI.FI
Department of Mathematics and Statistics
Department of Computer Science
Helsinki Institute for Information Technology HIIT
University of Helsinki
Helsinki, Finland

Hayaru Shouno SHOUNO@UEC.AC.JP
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Department of Mathematics and Statistics
Department of Computer Science
Helsinki Institute for Information Technology HIIT
University of Helsinki
Helsinki, Finland

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Independent component analysis (ICA) is a method to estimate components which are as statistically independent as possible. However, in many practical applications, the estimated components are not independent. Recent variants of ICA have made use of such residual dependencies to estimate an ordering (topography) of the components. Like in ICA, the components in those variants are assumed to be uncorrelated, which might be a rather strict condition. In this paper, we address this shortcoming. We propose a generative model for the source where the components can have linear and higher order correlations, which generalizes models in use so far. Based on the model, we derive a method to estimate topographic representations. In numerical experiments on artificial data, the new method is shown to be more widely applicable than previously proposed extensions of ICA. We learn topographic representations for two kinds of real data sets: for outputs of simulated complex cells in the primary visual cortex and for text data.

Keywords: independent component analysis, topographic representation, higher order correlation, linear correlation, natural image statistics, natural language processing.

1. Introduction

A simple yet powerful approach to analyze some data $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\top$ is to decompose it into

$$\boldsymbol{x} = \mathbf{A}\boldsymbol{s}, \tag{1}$$

where $\mathbf{A}$ is an unknown mixing matrix and $\boldsymbol{s} = (s_1, s_2, \ldots, s_d)^\top$ is the source vector consisting of latent non-Gaussian random variables. A special instance of this generative model is independent component analysis (ICA) where the sources s_i are additionally assumed to be statistically independent (Hyvärinen and Oja, 2000). The goal of ICA and its related methods is to estimate $\mathbf{A}$ and $\boldsymbol{s}$ based on the observations of $\boldsymbol{x}$ only. The ICA model was shown to be identifiable up to the order, the signs and the scales of the components s_i (Comon, 1994). ICA has been applied to a wide range of fields such as computational neuroscience (Hyvärinen et al., 2009) or natural language processing (Kolenda et al., 2000; Honkela et al., 2010).

However, the estimated components may not be independent in many practical situations. Hence, one may want to relax the independence assumption of the s_i, and further analyze the relationship between the components. This was done in topographic ICA (TICA) where the sources s_i are allowed to have correlated variances ("energies") (Hyvärinen et al., 2001). The dependencies were further used to fix the order-indeterminacy of ICA: the sources s_i were ordered on a topographic grid such that close-by components had correlated variances while distant components were as statistically independent as possible. Related models were proposed by Osindero et al. (2006); Bach and Jordan (2003); Karklin and Lewicki (2005); Zoran and Weiss (2010).

In TICA, the sources s_i were constrained to be uncorrelated. However, linear correlations occur in many practical situations. One example is the outputs of co-linearly aligned Gabor filters for natural image inputs. Since natural images contain many long contours, the outputs of such Gabor filters are linearly correlated. Another practical situation occurs in MEG and EEG analysis where coherent sources can be linearly correlated due to neural interactions (Gómez-Herrero et al., 2008). As we will see in this paper, another example occurs in the analysis of text data.

In this paper, we propose a method to capture both linear and higher-order correlations of the components s_i. Like in TICA, we make use of the dependencies of the components, linear or not, to order them. We start in Section 2 with discussing the motivation for ordering the components, or as we also say, learning a topographic representation of $\boldsymbol{x}$. Then, we propose a model to generate sources s_i where neighboring components, for example s_i and s_{i+1}, have linear and higher-order correlations. This model contains ICA and TICA as special cases. Based on the model, we derive a simple objective function to estimate the mixing matrix $\mathbf{A}$ and the order of the components in (1). In Section 3, we use artificial data to verify that our objective function works as intended, and compare the performance of our method to ICA and TICA. We show that our new method, which we call correlated topographic analysis (CTA), is a generalization of TICA in terms of topographic estimations. In Section 4, we learn a topographic representation for two kinds of different data sets: outputs of simulated complex cells in the primary visual cortex and text data. Section 5 concludes the paper.

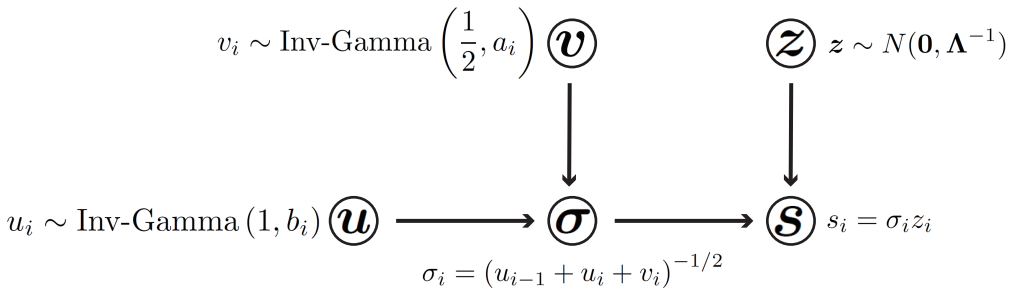

Figure 1: Generative model for the sources $\boldsymbol{s}$ appearing in the mixing model in (1). The generative model is discussed in Section 2.

2. Correlated Topographic Analysis

2.1. Motivation for Estimating Topographic Representations

By a topographic representation, we mean one in which the components s_i are ordered on a one- or two-dimensional lattice, such that components which are next to each other (or nearby) on that topographic lattice have special relationships to each other. In our case, the relationships are based on statistical dependencies.

The basic motivation for such representations is twofold. The foremost motivation is visualization. A topographic arrangement of the latent components s_i allows us to easily understand the relationships between them. A second motivation is that for natural data like images, sound, or text, the learned topography might be related to the cortical representation of the data. The reason is that, in order to minimize wiring length, neurons which tend to interact with each other are located near to each other (Hyvärinen et al., 2009).

2.2. The Generative Model for Sources

We consider here the situation where the sources $\boldsymbol{s}$ in (1) are generated according to

$$\boldsymbol{s} = \boldsymbol{\sigma} \odot \boldsymbol{z}, \tag{2}$$

where $\odot$ denotes element-wise multiplication. $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_d)^\top$ and $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)^\top$ in (2) are statistically independent. If the elements of $\boldsymbol{\sigma}$ are all positive and $\boldsymbol{z}$ is a multivariate Gaussian with mean $\mathbf{0}$, the marginal distributions of the sources $\boldsymbol{s}$ are super-Gaussian, *i.e.*, the sources are sparse (Hyvärinen et al., 2001). Their joint distribution depends on the correlation between the elements of $\boldsymbol{z}$ and the distribution of $\boldsymbol{\sigma}$. We can distinguish between the following four cases:

Case 1 If the elements of $\boldsymbol{z}$ are uncorrelated and the elements of $\boldsymbol{\sigma}$ are independent, the sources s_i are independent. This gives ICA with sparse sources.

Case 2 If the elements of $\boldsymbol{z}$ are uncorrelated and topographically nearby elements of $\boldsymbol{\sigma}$ are dependent, the sources are linearly uncorrelated but have correlated variances. This gives TICA.

Case 3 If topographically nearby elements of $\boldsymbol{z}$ are correlated and the elements of $\boldsymbol{\sigma}$ are independent, the sources can be made linearly but possibly very weakly higher-order correlated.

Case 4 If topographically nearby elements of $\boldsymbol{z}$ are correlated and the elements of $\boldsymbol{\sigma}$ are dependent as in case 2, the sources are both linearly and higher-order correlated. We call the estimation of the model in (1) with this type of dependencies correlated topographic analysis (CTA).

We see that the generative model in (2) can define sparse sources where each source may have dependencies within a certain neighborhood. These dependencies can be estimated from the data and used to order the components.

2.3. Approximation of the Likelihood and the Objective Function

In order to allow for dependencies in a neighborhood, we make, as shown in Figure 1, the following assumptions for $\boldsymbol{\sigma}$ and $\boldsymbol{z}$: For $\boldsymbol{z}$, we assume that the precision matrix $\boldsymbol{\Lambda}$ has a tridiagonal shape with a ringlike boundary, namely $z_{i\pm d} = z_i$, so that the distribution of $\boldsymbol{z}$ is

$$p(\boldsymbol{z};\boldsymbol{\lambda}) = \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{\Lambda}\boldsymbol{z}\right) = \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \prod_{i=1}^{d} \exp\left\{-\frac{1}{2}\left(z_i^2 + 2\lambda_i z_i z_{i+1}\right)\right\}, \tag{3}$$

For $\boldsymbol{\sigma}$, we assume that each σ_i has been created via

$$\sigma_i = (u_{i-1} + u_i + v_i)^{-1/2}, \tag{4}$$

where u_i and v_i are statistically independent nonnegative random variables. Nonzero u_i create energy correlations between the sources, like in TICA. The v_i account for source-specific variances. We assume inverse Gamma distributions for $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$p(\boldsymbol{v},\boldsymbol{u};\boldsymbol{a},\boldsymbol{b}) = \prod_{i=1}^{d} \sqrt{\frac{a_i}{2\pi}} v_i^{-3/2} \exp\left(-\frac{a_i}{2v_i}\right) \times \prod_{i=1}^{d} \frac{b_i}{2} u_i^{-2} \exp\left(-\frac{b_i}{2u_i}\right). \tag{5}$$

The a_i and b_i are positive scale parameters. If a scale parameter approaches zero, the corresponding variable converges (in distribution) to zero. For example, if $b_i \to 0$ for all i, the u_i approach zero, which decouples the σ_i from each other.

By inserting (2) and (4) into (3), the conditional distribution for $\boldsymbol{s}$ given $\boldsymbol{u}$ and $\boldsymbol{v}$ is

$$\begin{aligned} p(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{u};\boldsymbol{\lambda}) = \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \prod_{i=1}^{d} \sqrt{u_{i-1}+u_i+v_i} \exp\Bigg[-\frac{1}{2}\Big\{v_i s_i^2 + (s_i^2 + s_{i+1}^2)u_i \\ + 2\lambda_i \sqrt{(u_{i-1}+u_i+v_i)(u_i+u_{i+1}+v_{i+1})}\, s_i s_{i+1}\Big\}\Bigg]. \end{aligned} \tag{6}$$

Computation of the marginal of $\boldsymbol{s}$, that is integrating (6) with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$ using (5) as prior, is analytically intractable. We resort to two simple approximations,

$$\sqrt{u_{i-1}+u_i+v_i} \approx \sqrt{u_i}, \tag{7}$$

$$\sqrt{(u_{i-1}+u_i+v_i)(u_i+u_{i+1}+v_{i+1})} \approx u_i. \tag{8}$$

Both approximations above are similar to what has been done for TICA (Hyvärinen et al., 2001, Equation (3.7)). This gives the following approximation for (6),

$$p(\boldsymbol{s}|\boldsymbol{v},\boldsymbol{u};\boldsymbol{\lambda}) \approx \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \prod_{i=1}^{d} u_i^{1/2} \exp\left[-\frac{1}{2}\left\{s_i^2 v_i + \left(s_i^2 + s_{i+1}^2 + 2\lambda_i s_i s_{i+1}\right) u_i\right\}\right]. \tag{9}$$

Integrating out the u_i and v_i, we obtain the following approximative distribution for $\boldsymbol{s}$

$$\tilde{p}(\boldsymbol{s};\boldsymbol{\lambda},\boldsymbol{a},\boldsymbol{b}) \propto \prod_i \exp\left(-\sqrt{a_i}|s_i| - \sqrt{b_i}\sqrt{s_i^2 + s_{i+1}^2 + 2\lambda_i s_i s_{i+1}}\right). \tag{10}$$

We use the proportionality sign because we do not know the partition function which normalizes $\tilde{p}(\boldsymbol{s};\boldsymbol{\lambda},\boldsymbol{a},\boldsymbol{b})$.[1]

We discuss now the relation between (10) and the four cases outlined above. In the limit where $b_i \to 0$, $\tilde{p}$ becomes the distribution of independent Laplacian random variables, as often used in ICA with sparse sources (case 1). In the limit where $a_i \to 0$ and $\lambda_i = 0$ for all i, we obtain TICA (case 2). Case 3 is not explicitly covered by this model. However, we will show in the next section that CTA identifies its sources as well. In order to allow the components to have strong linear correlations and to be able to obtain closed-form solutions, we use the fixed values $a_i = b_i = 1$ and $\lambda_i = -1$, which give

$$\tilde{p}(\boldsymbol{s}) \propto \prod_{i=1}^{d} \exp\left(-|s_i| - |s_i - s_{i+1}|\right). \tag{11}$$

The distribution corresponds to case 4 with positively correlated sources. Note that this distribution is used as prior for the regression coefficients in the fused lasso (Tibshirani et al., 2005). Since in this paper, we are not interested in regression but in unsupervised learning, we do not explore this connection any further.

Using the distribution in (11) as prior for $\boldsymbol{s}$, we can compute the log-likelihood for $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d)^\top = \mathbf{A}^{-1}$. This gives the CTA objective function J,

$$J(\mathbf{W}) = J_1(\mathbf{W}) + J_2(\mathbf{W}), \tag{12}$$

$$J_1(\mathbf{W}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{d} |\boldsymbol{w}_i^\top \boldsymbol{x}(t)| + \log|\det \mathbf{W}|, \tag{13}$$

$$J_2(\mathbf{W}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{d} |\boldsymbol{w}_i^\top \boldsymbol{x}(t) - \boldsymbol{w}_{i+1}^\top \boldsymbol{x}(t)|. \tag{14}$$

The vector $\boldsymbol{x}(t)$ denotes the $t-$th observation of the data, $t = 1, 2, \ldots, T$. Note that J_1 is the log-likelihood for an ICA model and that J_2 is sensitive to the order and signs of the $\boldsymbol{w}_i$. For numerical reasons, we approximated $|\cdot|$ in two summations by $\log\cosh(\cdot)$ in all simulations in this paper.

1. Since $\tilde{p}(\boldsymbol{s};\boldsymbol{\lambda},\boldsymbol{a},\boldsymbol{b}) \leq \prod_i \exp\left(-\sqrt{a_i}|s_i|\right)$ we know, however, that $\tilde{p}$ is integrable so that the partition function exists.

3. Validation on Artificial Data and Optimization Procedure

In this section, we investigate whether the proposed objective function $J(\mathbf{W})$ can be used to estimate topographic representations. For that purpose, we generated sources $\boldsymbol{s}$ according to (2) for each of the discussed four cases. Then, we mixed them with a randomly chosen mixing matrix $\mathbf{A}$, and estimated the model by maximization of J. For comparison, we also applied ICA and TICA on the data. The dimension of the data and the number of samples are $d = 20$ and $T = 30'000$, respectively. We performed simulations for 100 randomly chosen mixing matrices, that is for 400 data sets in total.

3.1. Flow of the Optimization

For the estimation of $\mathbf{W}$ in CTA, we perform an optimization with three steps because preliminary results showed that CTA tends to get stuck in local maxima when we optimize $J(\mathbf{W})$ by a basic gradient method. The three steps are as follows:

Step 1 We optimize only $J_1(\mathbf{W})$ by the conjugate gradient method (Rasmussen, 2006) to get

$$\mathbf{W}^{(1)} = \arg\max_{\mathbf{W}} J_1(\mathbf{W}). \tag{15}$$

Step 2 With $\mathbf{W}^{(1)}$, we compute $\boldsymbol{s}^{(1)}(t) = \mathbf{W}^{(1)}\boldsymbol{x}(t)$. Then, the order and signs of $s_i^{(1)}$ are optimized by

$$\hat{\boldsymbol{k}}, \hat{\boldsymbol{c}} = \arg\max_{\boldsymbol{k},\boldsymbol{c}} \underbrace{-\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{d} \left| c_i s_{k_i}^{(1)}(t) - c_{i+1} s_{k_{i+1}}^{(1)}(t) \right|}_{J_2}.$$

Here, $\boldsymbol{k} = (k_1, \ldots, k_d)$ is an order vector with $k_i \in \{1, \ldots, d\}$ and $k_i \neq k_j$ for $j \neq i$, and $\boldsymbol{c} = (c_1 \ldots c_d)$ is a sign vector with $c_i \in \{-1, 1\}$. $\hat{\boldsymbol{k}}$ and $\hat{\boldsymbol{c}}$ give $\mathbf{W}^{(2)} = [\hat{c}_1 \boldsymbol{w}_{\hat{k}_1}^{(1)}, \hat{c}_2 \boldsymbol{w}_{\hat{k}_2}^{(1)}, \ldots, \hat{c}_d \boldsymbol{w}_{\hat{k}_d}^{(1)}]^\top$ where $\boldsymbol{w}_i^{(1)}$ is the $i-$ th row vector in $\mathbf{W}^{(1)}$.

Step 3 As in step 1, we optimize $J(\mathbf{W})$ by the conjugate gradient method (Rasmussen, 2006) using $\mathbf{W}^{(2)}$ as initial value and obtain the final result $\mathbf{W}^{(3)}$.

The purpose of the first two steps is to find better initial values for $\mathbf{W}$ before optimizing $J(\mathbf{W})$ in (12). In step 2, we solve a combinatorial optimization problem, for which we use a method based on dynamic programming (Bellman and Dreyfus, 1962). We omit the details here.

For the comparison with ICA, we perform only step 1. For the comparison with TICA, we perform all three steps but replace J_2 in step 2 and J in step 3 by

$$J_{tica}(\mathbf{W}) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{d} \sqrt{0.1 + (\boldsymbol{w}_i^\top \boldsymbol{x}(t))^2 + (\boldsymbol{w}_{i+1}^\top \boldsymbol{x}(t))^2} + \log|\det \mathbf{W}|, \tag{16}$$

where J_{tica} is one objective function for TICA proposed in (Hyvärinen et al., 2001, Equations (3.10) and (3.12)). For numerical stability, 0.1 is added in the square root. Step 2 is a bit simpler for TICA because we do not need to optimize with respect to the signs since J_{tica} is insensitive to them.

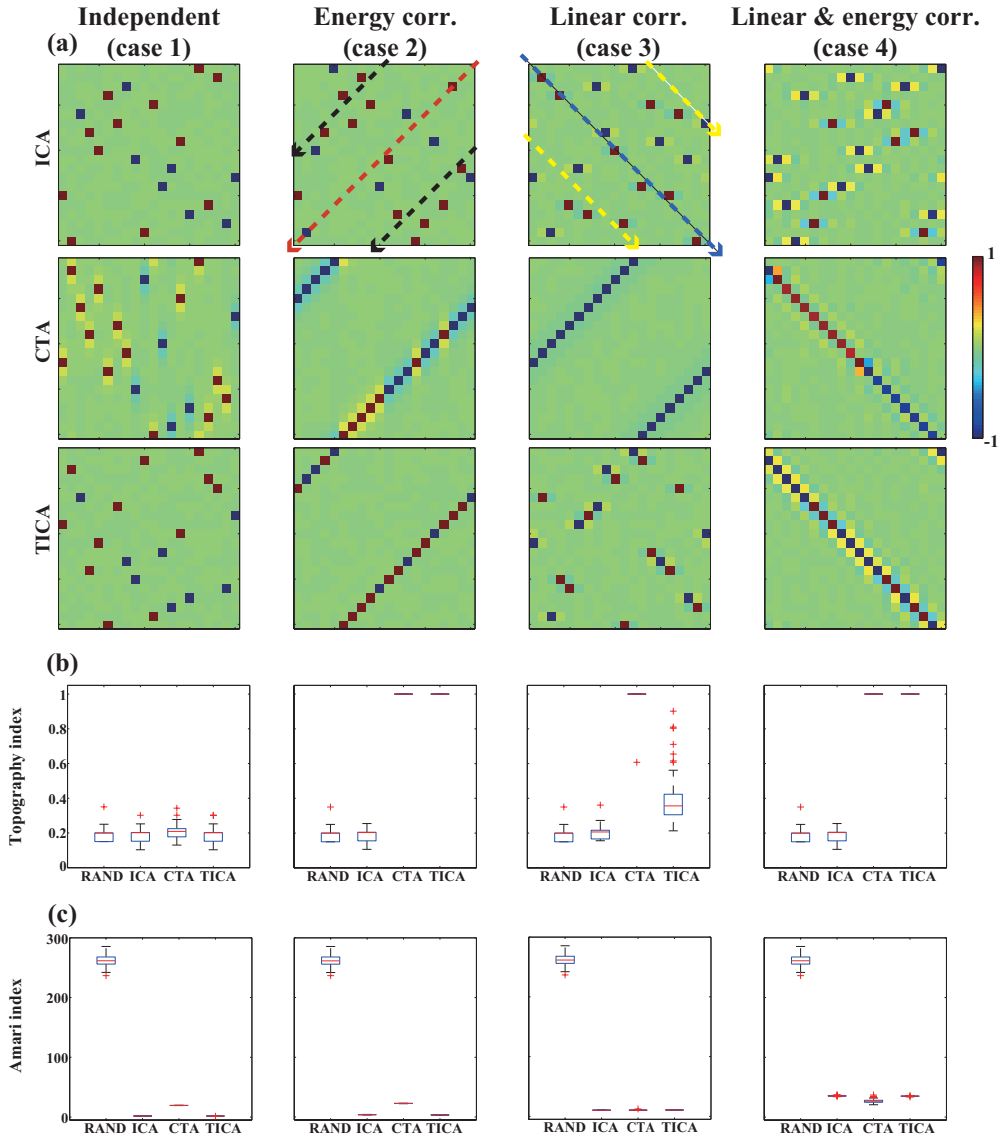

Figure 2: Validation on artificial data and comparison with ICA and TICA. From left to right, the data was created according to case 1 to 4 as in Section 2. (a) Examples of the performance matrices $\mathbf{P}$ for one mixing matrix. (b) The distribution of the topography index TI in (17) for 100 random estimation problems. (c) The distribution of the Amari index. See text body for a discussion of the results. The four differently colored diagonal paths in (a) are examples of summation paths traversed in the computation of TI.

3.2. Results

To measure the goodness of the results, we compute the performance matrix $\mathbf{P} = \mathbf{WA}$. If $\boldsymbol{s}$ is correctly identified and the topography is correctly estimated, $\mathbf{P}$ is a diagonal matrix or, because of the used ring-like boundary of the topography, a circularly shifted "diagonal" matrix.

Figure 2(a) shows the performance matrices for one of the 100 mixing matrices. For data with independent sources (case 1), no topography can be estimated because there is no statistical dependency in $\boldsymbol{s}$. For data with sources which have energy correlations only (case 2), both TICA and the proposed CTA give a $\mathbf{P}$ that is close to a shifted diagonal matrix. This means that the topography and the sources are correctly estimated. ICA, which is insensitive to energy correlations of the sources, is not able to find the topography. For data with linearly correlated sources (case 3), only CTA is able to estimate the topography correctly. The other methods do not preserve the topography in the estimated components. Furthermore, it seems that CTA solves also the sign-indeterminacy of ICA. For data with both linearly and energy correlated sources (case 4), both TICA and CTA perform well.

To quantify for all 100 mixing matrices how well the topography was estimated by the different methods, we used Amari index (AI) (Amari et al., 1996) and defined topography index (TI). For TI, like in the AI, we first normalized $\mathbf{P}$ in order to account for the scale indeterminacy in the ICA model. We did this by taking the absolute values of the elements of $\mathbf{P}$, and dividing each column and row of the resulting matrix by its maximal absolute value, giving us the two matrices $|\mathbf{P}|_1$ and $|\mathbf{P}|_2$. Then, we simply computed the largest sum along all possible diagonal paths through $|\mathbf{P}|_1$ and $|\mathbf{P}|_2$ giving us the numbers S_1 and S_2, respectively. Figure 2(a) shows four examples of diagonal paths taken. Note that they run both from left-to-right and from right-to-left. TI is finally given by

$$\mathrm{TI} = \frac{S_1 + S_2}{2d}, \tag{17}$$

where d is the dimension of the data. The best performance is obtained for TI=1.

The results from the 100 trials are summarized in Figure 2(b) in the form of boxplots. We show, as a baseline, also the topography index for random permutation matrices (labeled "RAND" in plot). We can see that the conclusions from the single example shown in Figure 2(a) generalize: For data with independent sources, no topography can be estimated, and all methods perform like the baseline. Our new method CTA performs well on all data sets which have dependencies. This is in contrast to TICA, which performs well only if energy correlations are present.

Figure 2(c) shows the distribution of the AI which measures how well the sources in the model (1) are identified but ignores the topographic arrangement (ordering) of the sources. As a baseline, we also show the AI for random matrices (labeled "RAND" in the plot). For data with sources bare of linear correlations (case 1 and 2), we see that ICA and TICA lead to a slightly better separation of the sources. A likely reason is that by setting in (10) $\lambda_i = -1$, CTA tries to look for sources which have linear correlations, which introduces some error if no such sources are present. On the other hand, if such sources are present, the separation results are at least as good as for ICA or TICA.

In summary, our simulations show that the sensitivity of CTA to linear correlations makes our method more widely applicable than TICA. If the data has no linearly corre-

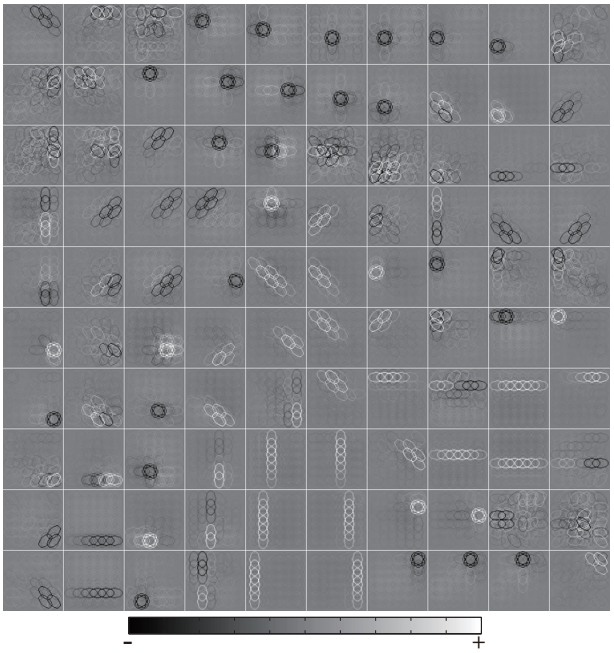

Figure 3: Estimated higher order basis from the outputs of complex cells for natural image inputs.

lated sources, this generalization of TICA comes at a cost of giving slightly less accurate separation results than the more specialized methods ICA or TICA.

4. Application to Real Data

In this section, CTA is applied to two kinds of real data: the outputs of simulated complex cells in the primary visual cortex when stimulated with natural images, and text data.

4.1. Outputs of Simulated Complex Cells

CTA was applied to the outputs of simulated complex cells in the primary visual cortex which are "stimulated" by natural images. Previously, ICA has been applied to such kind of data (Hyvärinen et al., 2005). The purpose here is to investigate what kind of topography could emerge between the learned higher-order features.

4.1.1. Methods

The outputs of the complex cells $\boldsymbol{x}$ are computed as

$$x'_k = \left(\sum_{x,y} W^o_k(x,y)I(x,y)\right)^2 + \left(\sum_{x,y} W^e_k(x,y)I(x,y)\right)^2,$$
$$x_k = \log(x'_k + 1.0), \tag{18}$$

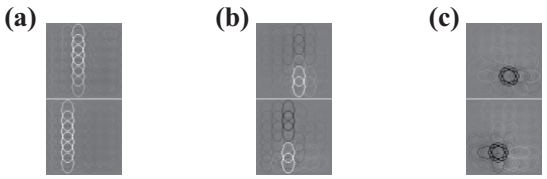

Figure 4: Three prominent features in Figure 3. (a) Long contours, (b) end-stopping and (c) star-like features.

where $I(x, y)$ is a 24×24 natural image patch,[2] and $W_k^o(x, y)$ and $W_k^e(x, y)$ are odd- and even-symmetric Gabor functions with the same parameters for spatial position, orientation and spatial frequency. We used $T = 100'000$ image patches. In this experiment, the complex cells are arranged on a two dimensional 6×6 grid. For each position, there are cells with four different orientations and one frequency band. The total number of complex cells is $6{\cdot}6{\cdot}4 = 144$. As preprocessing, first, the DC component of $\boldsymbol{x}$ is removed and then, whitening and dimensionality reduction are performed by PCA. We retained 100 dimensions.

In this experiment, we assume that the components s_i are arranged on a two dimensional topographic lattice and one component is dependent with nearby eight components (two horizontal, two vertical and four diagonal components). The objective function and the optimization method for this two dimensional lattice is a straightforward extension of the objective for the one dimensional lattice treated in Section 2.3 and Section 3.1. Therefore, we omit the details here.

4.1.2. Results

The map of the estimated higher order basis vectors is presented in Figure 3. In Figure 4, we highlight three prominent kinds of basis vectors: those forming long contours (Figure 4(a)), those with end-stopping behavior (Figure 4(b)), and star-like features (Figure 4(c)). In the map, the basis vectors forming long contours and the star-like features tend to be separated, and have a meaningful topography between themselves.

Next, we checked that the learned features are not artifacts due to the fixed complex cell model. For that purpose, we performed the same experiment again but with $I(x, y)$ being samples from the Gaussian distribution with mean $\mathbf{0}$ and covariance given by the covariance matrix of the natural images. The map of higher order basis vectors for this noise data is depicted in Figure 5. We see that star-like features are also present. However, there are much fewer long contours and no features with end-stopping behavior. Therefore, we suggest that long contours, end-stopping features, and their topography mainly reflect the properties of natural images.

4.2. Text Data

Here, we apply CTA to the analysis of a large text corpus. ICA has been applied before to text data: Kolenda and colleagues analyzed a set of documents and the terms they

2. We used the natural images of the imageica package. To compute $\boldsymbol{x}$, we used codes available in the contournet package. Both packages are available at http://www.cs.helsinki.fi/u/phoyer/software.html.

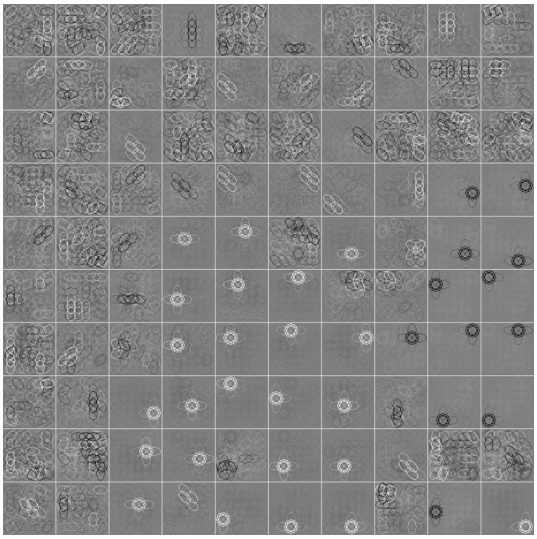

Figure 5: Estimated higher order basis from noise inputs.

contain (Kolenda et al., 2000). They found that ICA results in more easily interpretable topics which underlie the different documents than the more traditional latent semantic analysis. Honkela and colleagues analyzed words and the contexts in which they appear. Again, compared to latent semantic analysis, ICA produced more meaningful results where the latent sources reflected linguistic categories (Honkela et al., 2010). We apply here CTA to this kind of context-word data. The motivation is that the different latent categories may be often correlated so that the independence assumption in ICA is actually too strong. CTA, on the other hand, should be able to identify relationships between the latent categories.

4.2.1. Methods

We constructed the context-word data as in the literature (Honkela et al., 2010). The $T = 200'000$ most frequent words in $51'126$ larger English-language Wikipedia articles were selected. Then, a list of words which occurred two words before or after the selected words was compiled. From this list, the 1000 most frequent words formed the "contexts" words. From the selected and the context words, the joint frequency matrix $\mathbf{Y}$, of size $1000 \times 200'000$ was created. Finally, we obtained the context$\times$word data matrix $\mathbf{X} = (\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(T))$ by transforming the elements in $\mathbf{Y}$ as $x_i(j) = \log(y_{i,j} + 1.0)$.

As preprocessing, we perform centering to make the mean of each row of $\mathbf{X}$ zero and its variance one. Then, the data is whitened by PCA and the dimension reduced from 1000 to 60. We assume an one-dimensional topography and estimate $\mathbf{W}$ as described in Section 3.1. The estimation of the model in (1) allows us to represent the context $\times$ word matrix $\mathbf{X}$ as $\mathbf{X} = \mathbf{AS}$ where $\mathbf{S}$ is a $60 \times 200'000$ categories $\times$ word matrix. Note that in the context of the text data, we call a latent component a "category".

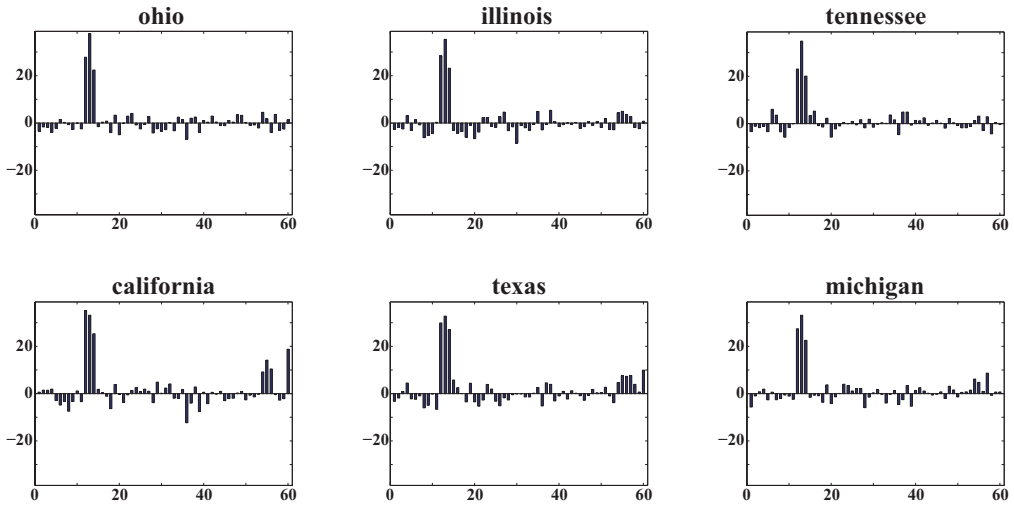

Figure 6: Latent representations of American states.

4.2.2. Results

Before analyzing the emerging topography, we show by example that, like ICA, CTA is able to extract meaningful categories. In the literature (Honkela et al., 2010), the learned categories were analyzed by showing the latent representations $\mathbf{S}_k$ (k-th column of the matrix $\mathbf{S}$) for different words k. The latent representations indicate to which extent the given word k "activates" the learned latent categories. Figure 6 shows that the latent representations $\mathbf{S}_k$ of American states are very similar. Likewise, colors, music styles, and nationalities for example were found to have very similar representations (results not shown).

CTA topographically arranges the latent categories so that related ones are next to each other. In Table 1, we show selected categories which were in several simulations robustly arranged next to each other. To visualize a latent category, we show the nine words which give the top nine activations for the given category (row of the matrix $\mathbf{S}$). The three categories in the leftmost panel are about time and numbers. More specifically, category $\mathbf{S}^7$ (row 7 of $\mathbf{S}$) is all about "units of time", $\mathbf{S}^8$ about "quantifiers" and $\mathbf{S}^9$ about roman numerals. The middle panel in the table shows that CTA placed American states and, in general, cities next to each other. The rightmost panel shows that media words were also topographically arranged.

5. Discussion and Conclusion

In this paper, we have proposed a method to estimate topographic representations. The proposed method, Correlated Topographic Analysis (CTA), is an extension of ICA where nearby latent components s_i are linearly and higher order correlated, while distant components are as statistically independent as possible. The related concept of structured sparsity has also been used in unsupervised learning (Mairal et al., 2011), but not to learn linearly correlated components like what we have done here.

Table 1: Three examples of a robust topographic ordering between three categories. Denoting the k-th row of the matrix $\mathbf{S}$ by $\mathbf{S}^k$, the words with the top nine absolute values of a $\mathbf{S}^k$ are shown. In the last columns, telev. abbreviate "television", and broadc.[1] and broadc.[2] denote "broadcast" and "broadcasting".

Ex.1: Time and numbers			Ex.2: States and cities			Ex.3: Media		
$\mathbf{S}^7$	$\mathbf{S}^8$	$\mathbf{S}^9$	$\mathbf{S}^{12}$	$\mathbf{S}^{13}$	$\mathbf{S}^{14}$	$\mathbf{S}^{28}$	$\mathbf{S}^{29}$	$\mathbf{S}^{30}$
weeks	few	3	california	ohio	philadelphia	comic	album	tv
months	several	6	texas	illinois	dublin	marvel	band's	telev.
month	various	32	angeles	tennessee	chicago	fantasy	pop	broadc.[1]
hours	numerous	4	florida	creek	boston	comics	albums	bbc
week	eight	16	illinois	country	georgia	fiction	solo	abc
days	mostly	13	ohio	michigan	los	batman	band	cbs
year	six	21	michigan	california	manchester	animated	rock	nbc
seven	two	8	washington	texas	texas	manga	songs	aired
five	four	23	minnesota	colorado	angeles	x-men	jazz	broadc.[2]

CTA itself was obtained by setting in the prior distribution for the sources, Equation (10), $a_i = b_i = 1$ and $\lambda_i = -1$. Ultimately, we would like to estimate these parameters instead of fixing them by hand. Estimating them is, however, difficult because we do not know the partition function so that we had to leave this endeavor to future work.

We showed that CTA is more widely applicable than TICA: Unlike TICA, CTA can estimate an ordering of components whose energy correlation is very weak (Figure 2). We have applied CTA to two different data sets. For outputs of simulated complex cells, CTA led to the emergence of a new representation where long contours and end-stopping features are topographically arranged. Past work found long contour features but they were not, in contrast to ours, topographically arranged (Hyvärinen et al., 2005; Hoyer and Hyvärinen, 2002). For text data, CTA identified, as ICA, latent linguistic categories. In addition, however, it allowed us to find relationships between them which are related to semantic similarities.

Acknowledgments

H. Sasaki was supported by Grant-in-Aid for JSPS Fellows. H. Shouno was partly supported by Grand-in-Aid for Scientific Research (C) 21500214 and on Innovative Areas, 21103008, MEXT, Japan. A. Hyvärinen and M. U. Gutmann were supported by the Academy of Finland (CoE in Algorithmic Data Analysis, CoE in Inverse Problems Research, and Computational Science Program). The authors wish to thank Timo Honkela and Jaakko J. Väyrynen for providing us the text data, and to thank Shunji Satoh and Jun-ichiro Hirayama for their helpful discussion.

References

S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763, 1996.

F.R. Bach and M.I. Jordan. Beyond independent components: trees and clusters. *Journal of Machine Learning Research*, 4:1205–1233, 2003.

R.E. Bellman and S.E. Dreyfus. *Applied dynamic programming*. Princeton University Press, 1962.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J.L. Cantero. Measuring directional coupling between EEG sources. *Neuroimage*, 43(3):497–508, 2008.

T. Honkela, A. Hyvärinen, and J.J. Väyrynen. WordICA-emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16 (03):277–308, 2010.

P.O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, 42(12):1593–1605, 2002.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. ISSN 0893-6080.

A. Hyvärinen, P.O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001. ISSN 0899-7667.

A. Hyvärinen, M. Gutmann, and P.O. Hoyer. Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, 6(1):12, 2005.

A. Hyvärinen, J. Hurri, and P.O. Hoyer. *Natural Image Statistics: A probabilistic approach to early computational vision*, volume 39. Springer-Verlag New York Inc, 2009.

Y. Karklin and M.S. Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, 17(2):397–423, 2005.

T. Kolenda, L.K. Hansen, and S. Sigurdsson. Independent components in text. In *Advances in Independent Component Analysis*, pages 229–250, 2000.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

S. Osindero, M. Welling, and G.E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2006.

C.E. Rasmussen. Conjugate gradient algorithm, version 2006-09-08. 2006.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

D. Zoran and Y. Weiss. The "tree-dependent components" of natural images are edge filters. In *Advances in Neural Information Processing Systems*, volume 22, pages 508–514. MIT Press, Cambridge, MA, 2010.