SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling

Jun-ichiro Hirayama [1 2] **Aapo Hyvärinen** [3 4] **Motoaki Kawanabe** [2 1]

Abstract

We present a novel probabilistic framework for a hierarchical extension of independent component analysis (ICA), with a particular motivation in neuroscientific data analysis and modeling. The framework incorporates a general subspace pooling with linear ICA-like layers stacked recursively. Unlike related previous models, our generative model is fully tractable: both the likelihood and the posterior estimates of latent variables can readily be computed with analytically simple formulae. The model is particularly simple in the case of complex-valued data since the pooling can be reduced to taking the modulus of complex numbers. Experiments on electroencephalography (EEG) and natural images demonstrate the validity of the method.

1. Introduction

Linear component analysis and pooling are two fundamental concepts of unsupervised representation or feature learning on continuous-valued data. The basic method for linear decomposition is independent component analysis (ICA) (Hyvärinen et al., 2001b) or sparse coding. Pooling originates from computational models of "complex cells" in the visual cortex (Hubel & Wiesel, 1962; Adelson & Bergen, 1985), which typically takes the sum of squares of components or neuronal outputs (L_2-pooling) to achieve invariances in higher features. The combination of the two concepts have so far found many applications, including advanced image recognition by deep neural networks.

In the present study, we focus on applications of great current interest related to neuroscience/engineering, such as electro- or magneto-encephalography (EEG/MEG) signal analysis and natural image statistics (Hyvärinen et al., 2009). Related previous studies have longly attempted to combine together linear component analysis and pooling and further built them up to hierarchical probabilistic models (Hyvärinen & Hoyer, 2000; Hyvärinen et al., 2001a; Valpola et al., 2004; Karklin & Lewicki, 2005; Shan et al., 2006; Onton & Makeig, 2009; Cadieu & Olshausen, 2012; Hirayama et al., 2015; Hosoya & Hyvärinen, 2015), among which one of the earliest combination of ICA with pooling was independent subspace analysis (ISA) (Hyvärinen & Hoyer, 2000). A more general energy-based modeling (EBM) framework (e.g., Osindero et al., 2006; Salakhutdinov & Hinton, 2009; Köster & Hyvärinen, 2010; Ngiam et al., 2011) has also been popularly used. These developments were somewhat parallel with the rise of general unsupervised deep learning techniques (see, e.g., Bengio et al., 2013, for review), while those neuroscientific applications specifically seek simple explanations and interpretations of data, and even two- or three-layer architectures have been of practical relevance.

However, related previous models were highly intractable, and they necessarily resorted to approximative or non-conventional methods, e.g., for learning and inference. Such a lack of theoretical transparency, as well as the computational difficulties associated, has hindered their extensive applications and further developments. Specifically, hierarchical generative models usually need approximations or numerical methods to evaluate the posterior estimates on latent variables or the likelihood (Bengio et al., 2013); EBM may avoid approximate posterior computation, while being still hampered by an intractable partition function (normalizing constant) to compute the likelihood. Conventional maximum likelihood (ML) estimation is thus not easily applicable in both types of models. In practice, simply stacking together ICA/ISA models trained layerwise, or with fixed lower layers, has often been used as an alternative (e.g., Shan et al., 2006; Onton & Makeig, 2009; Le et al., 2011; Cadieu & Olshausen, 2012; Hosoya & Hyvärinen, 2015) although its theoretical underpinning is rather unclear.

Here, we present a simple, *fully tractable* statistical framework for a hierarchical extension of ICA with an intrinsic pooling mechanism. We will refer to the framework as

[1]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan [2]Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan [3]Department of Computer Science and HIIT, University of Helsinki, Finland [4]Gatsby Computational Neuroscience Unit, University College London, UK. Correspondence to: Jun-ichiro Hirayama <jun-ichiro.hirayama@riken.jp>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

SPLICE, abbreviating *stacked pooling and linear components estimation*. By fully tractable, we mean that both the posterior estimates on latent variables and the likelihood function associated are given by simple (computable) analytical formula without resorting to any approximations. In this sense, both general hierarchical generative models and EBMs have only a limited tractability.

Our SPLICE extends ISA so that the subspaces may be dependent of each other via higher layers' latent variables; the layers can in principle be stacked recursively without violating the full tractability of the model. In the present study, we specifically introduce the basic framework and a practical learning scheme which combines a layerwise ICA pretraining with an unsupervised finetuning of the entire layers by non-approximate ML. As a proof of concept, we also demonstrate the method with EEG and natural images, as commonly targeted in related previous studies. The method turns out to have interesting connections to neural networks, which we also briefly discuss below.

2. Proposed Method

2.1. First-Layer Model

We begin with formulating the generative model for our SPLICE. Denote by $\mathbf{x}_t$ observed data vectors ($t = 1, 2, \ldots, n$), either real- or complex-valued, consisting of d entries x_{it}. Each of the d entries is given by a linear combination of the same number of unknown (first-layer) components or *sources*, collectively denoted as source vector $\mathbf{s}_t$. Here, we consider the fundamental case where $\mathbf{x}_t$ and $\mathbf{s}_t$ are independently and identically distributed (i.i.d.). Omitting sample index t for notational simplicity, we write

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

where the coefficient matrix $\mathbf{A}$, called mixing matrix, is square and assumed to be invertible; the inverse $\mathbf{W} := \mathbf{A}^{-1}$ is called demixing matrix. For convenience, we assume without loss of generality that $\mathbf{x}$ and $\mathbf{s}$ are zero-mean, by subtracting the sample mean from original data vectors.

Like ISA, we divide the d first-layer sources into m groups without overlapping, and denote by $\mathbf{s}_{[j]}$ the vector consisting of the d_j sources in the j-th group ($d = \sum_{j=1}^{m} d_j$). Hence $\mathbf{s}_{[j]}$ represents a d_j-dimensional subspace in the original data space, spanned by the corresponding columns in $\mathbf{A}$. Unlike ISA, however, the m source vectors $\mathbf{s}_{[j]}$ may be dependent of each other in our generative model.

2.2. Second-Layer Model

To extend the model to multiple layers by modeling the dependencies between the subspaces, we introduce an additional (second) layer on the top of the above ISA-like first layer model. Note that we don't count the pooling as a separate layer, so what we call the second layer is called the third layer in some previous work.

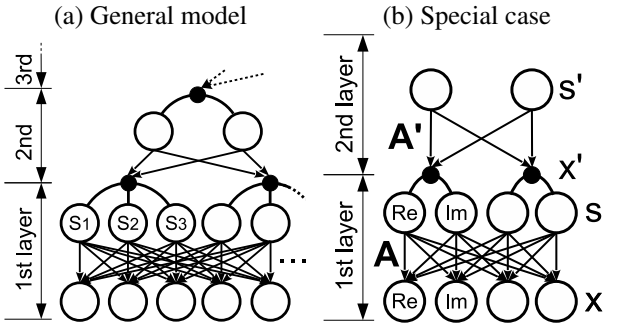

Figure 1: Generative model of SPLICE: (a) A higher layer directly gives the squared L_2-norms of lower sources s within each subspace. (b) An important special case having one complex source s per subspace.

Specifically, we assume that source vectors $\mathbf{s}_{[j]}$ may be dependent of each other in their "powers" or "energies," as typically quantified by their (squared) L_2-norms $\|\mathbf{s}_{[j]}\|^2$. We model the dependency using a linear mixing of additional (second-layer) sources with pointwise nonlinearities:

$$\|\mathbf{s}_{[j]}\|^2 = F_j^{-1}([\mathbf{A}'\mathbf{s}']_j), \quad j = 1, 2, \ldots, m, \tag{2}$$

where a monotonic link function $F_j : [0, \infty) \to \mathbb{R}$ maps (nonnegative) squared norms into real values, with its inverse denoted by F_j^{-1}; $\mathbf{A}'$ and $\mathbf{s}'$ are invertible mixing matrix and source vector (and $\mathbf{W}' := \mathbf{A}'^{-1}$ is demixing matrix) of second layer, and $[\cdot]_j$ denotes j-th entry of a vector. For later convenience, we denote by $\mathbf{x}' := \mathbf{A}'\mathbf{s}'$ (with entries x'_j) the "observed" data vector for the second layer.

To fully specify the generative model, in the present study, we simply put a sphericality assumption on every $\mathbf{s}_{[j]}$; i.e., we assume that the corresponding normalized vector

$$\mathbf{u}_{[j]} := \mathbf{s}_{[j]} / \|\mathbf{s}_{[j]}\|, \tag{3}$$

for every j, is uniformly distributed on the unit hypersphere, independently of any other random variables.

2.3. Third Layer and Beyond, and a Special Case

An intriguing fact with our model is that it can in principle be extended with any number of layers (Fig. 1(a)), up to the limit of subspace partitioning. This can be done by recursively stacking a higher layer (2) to generate the lower layer's subspace norms, with the lower layer appropriately partitioned into subspaces. Note that in the second and further layers, complex-valued variables may not be useful at least in our current i.i.d. setting. At the top layer, we simply assume that the sources are mutually independent and

non-Gaussian; throughout our experiments in Section 4, we used a typical prior, $p(s) = (1/2)\mathrm{sech}((\pi/2)s)$, which corresponds to the conventional $\tanh$ nonlinearity in ICA.

On the other hand, if one's goal is primarily to give a simple explanation of data, adding extra layers might overcomplicate the model. In fact, a simplified special case of our model (Fig. 1(b)), having only two layers with one complex-valued source per subspace (i.e., $|s_j|^2 = F_j^{-1}([\mathbf{A}'\mathbf{s}']_j)$), may already have a high practical relevance in the context of neuroscientific data analysis and modeling (e.g., Onton & Makeig, 2009; Cadieu & Olshausen, 2012; Hirayama et al., 2015). Then the squares $|s_j|^2$ and arguments $\arg s_j$ specifically represent the power (squared amplitude) and phase of an oscillatory source signal, where the sphericality assumption reduces to the circularity of s_j, i.e., the phase is uniform, and is independent of the power.

2.4. Choice of Intermediate Nonlinearity F_j

The true forms of F_j in (2) are usually unknown and ideally they would be learned from the data by either parametric or nonparametric methods. However, in practice, it is presumably sufficient that they are fixed, as a first approximation, so that the computational costs can be reduced.

Logarithm Conventionally, one typical option is the logarithm (e.g., Valpola et al., 2004; Karklin & Lewicki, 2005; Cadieu & Olshausen, 2012), i.e.,

$$F_j(q) = \ln(\lambda_j q), \tag{4}$$

where λ_j is a nonnegative scale parameter. The scale parameter λ_j can in fact be arbitrarily chosen, because one cannot determine the true scales of the sources due to the inherent scaling ambiguity as in ICA. To avoid this ambiguity, we specifically set λ_j so that

$$F_j(1) = 0. \tag{5}$$

Gaussianization Another popular choice in the literature is (radial) Gaussianization (Chen & Gopinath, 2001; Shan et al., 2006; Lyu & Simoncelli, 2009), generally given by

$$F_j(q) = \Phi^{-1}(\Psi_j(\lambda_j q)), \tag{6}$$

where Φ and Ψ_j denote the cumulative distribution function (cdf) of standard Gaussian distribution and that of a certain distribution over $[0, \infty)$, respectively. Gaussianization originally had no generative interpretation but we may use the principle as an intuitive "adversarial" definition of F_j, i.e., any non-Gaussianity in the data comes from the non-Gaussianity of the second-layer sources s'_k; one may simply set the cdf $\Psi_j(\lambda_j(\cdot))$ of $q_j = \|\mathbf{s}_{[j]}\|^2$ as chi-squared with an appropriate degrees of freedom, so that $\mathbf{s}_{[j]}$ is Gaussian when the second layer x'_j is (standard) Gaussian. The scale parameter λ_j can be fixed in the same manner as above. Figure 2 illustrates the forms of this type of F and F^{-1} for complex sources when $d_j = 1$.

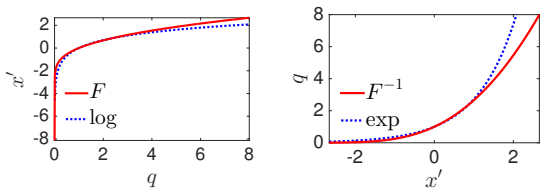

Figure 2: Forms of Gaussianization-based nonlinearity F (left panel) and its inverse F^{-1} (right) for $\mathbf{s}_{[j]} \in \mathbb{C}$, $d_j = 1$.

2.5. Properties of the Model

Now we will show that the hierarchical probabilistic model formulated above is in fact fully tractable. To ease exposition, we will give the result only for the two-layer case, but the generalization with more layers is straightforward.

2.5.1. The PDF is Analytically Normalized

First, we show that the probability density function (pdf) of observed data vector $\mathbf{x}$, associated with our model, is analytically normalized, i.e., the density has a unit sum without any intractable normalizing constant. Thus, the likelihood of our model can easily be evaluated without approximative or numerical techniques. The lack of this desirable property has long been an obstacle in hierarchical generative or energy-based modeling combined with pooling.

To derive the pdf, first observe that the linear map $\mathbf{s} \mapsto \mathbf{x}$ implies that $p(\mathbf{x}) = p_{\mathbf{s}}(\mathbf{W}\mathbf{x})|\det \mathbf{W}|^c$, where $p_{\mathbf{s}}(\cdot)$ denotes the pdf of first-layer source vector $\mathbf{s}$ and $c = 1$ and 2 for real- and complex cases, respectively (c.f. Adali & Haykin, 2010, sec. 1). Then, the following theorem explicitly relates $p_{\mathbf{s}}(\cdot)$ to the second layer's pdf of $\|\mathbf{s}_{[j]}\|^2$:

Theorem 1. *Denote by* $\mathbf{q}$ *the vector having* $q_j := \|\mathbf{s}_{[j]}\|^2$ *in the* j*-th entry. Assume that 1)* $\mathbf{q}$ *has the pdf given by* $p_{\mathbf{q}}(q_1, \ldots, q_m)$ *and 2) unit vectors* $\mathbf{u}_j := \mathbf{s}_{[j]}/\|\mathbf{s}_{[j]}\|$ *are independent of* $\mathbf{q}$ *and uniformly distributed in (the Cartesian products of) the corresponding unit hyperspheres. Then,*

$$p_{\mathbf{s}}(\mathbf{s}) = p_{\mathbf{q}}(\|\mathbf{s}_{[1]}\|^2, \ldots, \|\mathbf{s}_{[m]}\|^2) \prod_j \kappa_j(\|\mathbf{s}_{[j]}\|^2), \tag{7}$$

where $\kappa_j(q_j) = q_j^{1-d_j/2}\Gamma(d_j/2)\pi^{-d_j/2}$ *(if* $\mathbf{s}$ *is real) or* $q_j^{1-d_j}\Gamma(d_j)\pi^{-d_j}$ *(if complex).*

The proof is given in Supplementary Material A. Theorem 1 is a generalization of the result on single spherically-symmetric random vector (e.g., Ollila et al., 2012).

The formula (7) in fact holds for any probabilistic model $p_{\mathbf{q}}(\cdot)$ of the second layer. Our SPLICE specifically introduces the model (2) which resembles "post-nonlinear"

ICA (Taleb & Jutten, 1999), implying that

$$p_{\mathbf{q}}(\mathbf{q}) = \prod_{k=1}^{d} \exp\bigl(-H_k(\mathbf{w}'_k \cdot F(\mathbf{q}))\bigr) \prod_{j=1}^{d} |f_j(q_j)| |\det \mathbf{W}'|,$$

where $H_k(s'_k) = -\ln p(s'_k)$ are fixed functions that correspond to any typical choice of non-Gaussian prior such that $p(\mathbf{s}') = \prod_k p(s'_k)$. We denote the entrywise mapping F_j collectively by $F : \mathbb{R}_+^d \to \mathbb{R}^d$ and the first derivatives of F_j by f_j. We also denote by $\mathbf{w}'_k$ the k-th transposed row of $\mathbf{W}'$ and by a dot operator the standard inner product.

Taken together, we eventually obtain the pdf of $\mathbf{x}$, which is analytically normalized:

$$p(\mathbf{x}) = \prod_{k=1}^{d} \exp\bigl(-H_k(\sum_j w'_{kj} F_j(\|\mathbf{W}_{[j]}\mathbf{x}\|^2))\bigr) |\det \mathbf{W}'| \times \prod_{j=1}^{d} \exp(-G_j(\|\mathbf{W}_{[j]}\mathbf{x}\|^2)) |\det \mathbf{W}|^c, \quad (8)$$

where $\mathbf{W}_{[j]}$ consists of only the d_j rows in $\mathbf{W}$ so that $\mathbf{s}_{[j]} = \mathbf{W}_{[j]}\mathbf{x}$, and $G_j(q) := -\ln |f_j(q)| - \ln \kappa_j(q)$.

2.5.2. Exact Posterior Estimate via Pooling

Second, we see that our model also allows a simple analytical estimate on latent variables, which was in fact already implied by the above development. The inverse process can readily be obtained for the two linear layers by $\mathbf{s} = \mathbf{W}\mathbf{x}$ and $\mathbf{s}' = \mathbf{W}'\mathbf{x}'$, since we assumed that both $\mathbf{A}$ and $\mathbf{A}'$ are invertible. The remaining part that links the first layer $\mathbf{s}$ and the second layer $\mathbf{x}'$ is also readily given from Eq. (2) as

$$x'_j = F_j(\|\mathbf{s}_{[j]}\|^2) = F_j(\sum_{i \in \mathcal{S}_j} |s_i|^2), \quad (9)$$

for every j, where $\mathcal{S}_j$ denotes the index set for the sources belonging to the j-the subspace. Hence the overall transformation from observed $\mathbf{x}$ to the top-level representation $\mathbf{s}'$ is given by an analytically very simple form. Although this consequence is almost obvious from the definition of our model, this is still remarkable since previous hierarchical generative models usually did not possess such a tractability which in fact partly led to the invention of EBM.

Note that the relation (9) essentially implements an L_2-pooling operation. Interestingly, the two demixing layers, interleaved by the pooling layer, constitute a simplified multilayer neural network with linear neurons. Thus, one may also view SPLICE as a principled framework for unsupervised learning of a multilayer neural network with pooling layers.

We remark that the framework of SPLICE can even be extended with other ingredients of neural networks without violating the full analytical tractability, which will be an interesting open topic for future study. For example, nonlinear activation functions (e.g., leaky rectified linear unit) and other types of pooling (e.g., L_p-pooling, by introducing $s \mapsto |s|^{\frac{p}{2}} \mathrm{sign}(s)$) can readily be incorporated at least if the extra nonlinearity is bijective by itself and the associated Jacobian determinant is analytically tractable.

2.6. Learning by Maximum Likelihood (SPLICE-ML)

Next we develop the method for parameter estimation (learning) in our generative model. The analytically simple form of the pdf (8) makes conventional maximum likelihood (ML) estimation readily applicable, which theoretically has a number of desirable properties. To obtain an ML estimate, we simply minimize the sample average of the corresponding loss function, $L := -\ln p(\mathbf{x}) + \text{const.}$, given for the two-layer case by

$$L = \sum_{k=1}^{m} H_k\bigl(\sum_j w'_{kj} F_j(\|\mathbf{W}_{[j]}\mathbf{x}\|^2)\bigr) - \ln |\det \mathbf{W}'| + \sum_{j=1}^{m} G_j\bigl(\|\mathbf{W}_{[j]}\mathbf{x}\|^2\bigr) - c \ln |\det \mathbf{W}|. \quad (10)$$

From the neural network or EBM viewpoint, Eq. (10) is interesting since the loss is associated with not only the output (the top) but also the intermediate (lower) layers. Such a layer-specific loss has seldom been used in the literature.

In practice, the log-determinant terms in (10) may lead to a computational difficulty due to the costly matrix inversion when evaluating the gradient. Fortunately, the popular (stochastic) natural gradient method (Amari et al., 1996) is readily applicable for our model just like ICA, which can eliminate the need for matrix inversion. In our experiment (Section 4), however, we actually used the limited-memory BFGS quasi-Newton method (Schmidt, 2005) with an explicit matrix inversion, as it converged empirically faster in our setting (results not shown).

Since the objective function is not convex, the optimization needs to start with a good initial estimate not to stack with poor local optima. We use the following two-step approach that resembles a typical pretraining-finetuning strategy in the deep learning literature. That is, we first perform a layerwise learning developed below (SPLICE-LW) and then optimize the likelihood of the entire layers (SPLICE-ML) by starting from the layerwise solution. Note that both steps are unsupervised, in contrast to typical finetuning strategies in deep neural networks which are supervised.

2.7. Practical Layerwise Learning (SPLICE-LW)

Our unsupervised layerwise learning scheme combines ICA with an adaptive subspace partitioning for pool-

ing. A similar approach has previously been studied for ISA (Szabó et al., 2012, see also Hosoya & Hyvärinen, 2016). Specifically, we first perform an ICA to estimate the sources s_{it} up to their permutation, and then solve a simple optimization problem (see below) to assign the sources into a preset number of subspaces (except for the special case of $d_j \equiv 1$), which adaptively partition the data space into subspaces. The input to the upper layer can then be computed by (9), for which ICA is applied again. For a general number of layers, the procedure is recursively applied.

The idea of the adaptive subspace partitioning scheme is that our model implies that the sources' correlation-in-squares $\omega_{ij} := \mathrm{corr}(|s_i|^2, |s_j|^2)$ $(i \neq j)$ are constant γ_{kl} if $s_i \in \mathcal{S}_k$ and $s_j \in \mathcal{S}_l$ because of the L_2-sphericality within each subspace; matrix $\mathbf{\Omega} = (\omega_{ij})$ thus has a block structure after an appropriate permutation of rows and columns. This observation leads to a simple objective function $\|\mathbf{\Omega} - \mathbf{Z}^\mathsf{T}\mathbf{\Gamma}\mathbf{Z}\|^2$ (with Frobenius norm) to be minimized with respect to $\mathbf{Z} = (z_{ki})$ and $\mathbf{\Gamma} = (\gamma_{kl})$, where $\mathbf{Z} \in \{0,1\}^{m\times d}$ is a subspace assignment matrix such that $z_{ki} = 1$ if and only if source i belongs to subspace k.

The problem further reduces to an equivalent maximization of $\|\widetilde{\mathbf{Z}}\mathbf{\Omega}\widetilde{\mathbf{Z}}^\mathsf{T}\|^2$ with respect to $\widetilde{\mathbf{Z}} := (\mathbf{Z}\mathbf{Z}^\mathsf{T})^{-1/2}\mathbf{Z}$ (Supplementary Material B), where $\widetilde{\mathbf{Z}}$ is necessarily nonnegative and $\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\mathsf{T} = \mathbf{I}$. We thus borrow the idea of orthogonal relaxation from spectral clustering (Yu & Shi, 2003, see also Ding et al., 2006), and alternatively solve

$$\max_{\mathbf{V}} \|\mathbf{V}\mathbf{\Omega}\mathbf{V}^\mathsf{T}\|^2 \quad \text{s.t.} \quad \mathbf{V}\mathbf{V}^\mathsf{T} = \mathbf{I}, \quad \mathbf{V} \in \mathbb{R}_+^{m\times d}, \quad (11)$$

which previously appeared in a rather different context (Hirayama et al., 2016). Due to the joint orthogonality and nonnegativity, solution $\mathbf{V}$ has at most one nonzero entry in each column, which readily gives the subspace assignment. In our experiment below, we used an alternating projected gradient algorithm to solve (11), which empirically worked very well for natural images (Section 4.3).

3. Related Work

3.1. Natural Image Statistics and EEG/MEG Analysis

Our primary motivation for the development of a simple and tractable hierarchical probabilistic model is in data analysis and modeling related to neuroscience/engineering. In fact, many preceding studies on natural image statistics and EEG/MEG data analysis developed hierarchical extensions of ICA, while their intractability has hindered extensive applications or further developments.

A conventional approach was to make the second layer explain the dependency in variances of first-layer sources (Hyvärinen & Hoyer, 2000; Hyvärinen et al., 2001a; Valpola et al., 2004; Karklin & Lewicki, 2005; Zhang & Hyvärinen, 2010; Hirayama et al., 2015), e.g., $s_j \sim \mathcal{N}(0, F_j^{-1}(\sum_k a'_{jk} s'_k))$ under a conditional Gaussianity assumption. However, this approach usually leads to intractability in learning and inference except for restricted special cases (Hyvärinen & Hoyer, 2000; Hirayama et al., 2015). Alternatively, two recent studies modeled $|s_j|^2$ rather than the variance (Cadieu & Olshausen, 2012; Hirayama & Hyvärinen, 2012), but neither of them implemented subspace pooling with a full tractability.

In practice, many previous studies have rather preferred the layerwise learning strategy, which stacks together ICA/ISA models trained layerwise, or apply ICA upon fixed lower layers (Shan et al., 2006; Onton & Makeig, 2009; Cadieu & Olshausen, 2012; Hosoya & Hyvärinen, 2015); the stacked ISA strategy has also been developed in other application field (Le et al., 2011). Our new development may give a theoretical basis for the previous layerwise approach and also provides a principled unsupervised finetuning method.

3.2. Energy-Based Modeling (EBM)

Another line of research on natural image statistics have used the energy-based modeling (EBM) strategy (e.g., Osindero et al., 2006; Salakhutdinov & Hinton, 2009; Köster & Hyvärinen, 2010; Ngiam et al., 2011) instead of the hierarchical generative approach. However, EBM suffers from computational difficulties related to an intractable partition function, as well as limited interpretability since there are no independent latent variables.

To compare our model with EBM in Section 4, we introduce an EBM with deterministic hidden units that corresponds to our SPLICE in the two-layer setting. The associated loss function, i.e., the negative log-pdf up to irrelevant additive terms, is given by

$$L_{\mathrm{EBM}} = \sum_k H_k\left(\sum_j w'_{kj} F_j\left(\|\mathbf{W}_{[j]}\mathbf{x}\|^2\right)\right) + \ln \mathcal{Z}, \quad (12)$$

where $\mathcal{Z}(\mathbf{W}, \mathbf{W}')$ is the partition function to ensure $\int p(\mathbf{x})\mathrm{d}\mathbf{x} = 1$. We emphasize that partition function $\mathcal{Z}$ in the EBM is intractable while it is simple and tractable in SPLICE. Moreover, the H_k now lacks the connection to the prior pdf $p(s'_k)$ of independent sources; hence, $s'_k := \sum_j w'_{kj} F_j\left(\|\mathbf{W}_{[j]}\mathbf{x}\|^2\right)$ are generally not independent in EBM, which is a clear distinction from SPLICE.

3.3. Nonlinear ICA and Deep Generative Models

From a more general perspective, another important type of hierarchical extension of ICA is nonlinear ICA using multilayer neural networks (Almeida, 2003; Dinh et al., 2014; Hyvärinen & Morioka, 2017b). The difference from SPLICE (or other related models) is that the theory of nonlinear ICA basically assumes a bijectivity between ob-

servations and (nonlinear) independent components. Furthermore, the general nonlinear ICA model is not identifiable (Hyvärinen & Pajunen, 1999) (for an alternative approach, see Hyvärinen & Morioka, 2017b;a). In contrast, simulations below indicate that our model is identifiable, although we don't have a formal proof.

In a related context, some authors (Deco & Brauer, 1995; Dinh et al., 2014) have pointed out and addressed the computational difficulty associated with the Jacobian determinant of the multilayer neural network. Fortunately, SPLICE explicitly decomposes the Jacobian determinant into analytically tractable terms (Eq. (8)), and the popular natural gradient technique for ICA can further simplify the computation (Section 2.6).

Recently, several new techniques have been made available for learning and inference in general-purpose hierarchical (deep) generative models on continuous data, such as variational/autoencoder methods and non-classical learning principles (e.g., Kingma & Welling, 2014; Kingma et al., 2014; Goodfellow et al., 2014; Rezende & Mohamed, 2015; Hyvärinen & Morioka, 2017b;a). These developments mainly seek a *computational* tractability of learning and inference, maintaining the complexity (representation capability) of the model. In contrast, our SPLICE rather reduces the complexity (while keeping the essence) of the model to achieve the *analytical* tractability as well as the interpretability, with a particular emphasis on the tractable pooling. The motivations, as well as the target applications, are therefore quite distinct between the two approaches.

4. Experimental Results

In this section, we demonstrate our SPLICE in a simulation study and with two motivating real datasets.

4.1. Synthetic Data

First, we examined the important special case of SPLICE (Fig. 1(b)) with a synthetic dataset, and further with a real EEG dataset (Section 4.2). The goal was to demonstrate the validity of the basic concept of SPLICE (i.e., combining pooling and linear layers in a fully tractable manner) as well as its practical relevance in exploratory signal analysis.

The two-layer model assumes that both observed $\mathbf{x}$ and first-layer source vectors $\mathbf{s}$ are complex-valued, where the pooling operation reduces to taking the squared modulus of each scalar source variable; the adaptive subspace partitioning (Section 2.7) was not necessary in this basic case. SPLICE-LW consecutively performed real and complex-valued versions of FastICA (Hyvärinen, 1999; Bingham & Hyvärinen, 2000). SPLICE-ML used the SPLICE-LW solution as the initial estimate. For comparison, we also trained the EBM (12) by noise-contrastive estimation (Gutmann & Hyvärinen, 2012) (EBM-NCE) using SPLICE-LW as the initial estimate. We generated the reference (noise) dataset for NCE from multivariate Gaussian of the same mean and covariance as the original, having ten times larger sample size. All methods used Gaussianization-based F_j (Fig. 2), specifically with $\Psi(q) = 1 - \exp(-\lambda q)$ (i.e., the cdf of Exponential distribution) and $\lambda = -\ln(1 - \Phi(0))$.

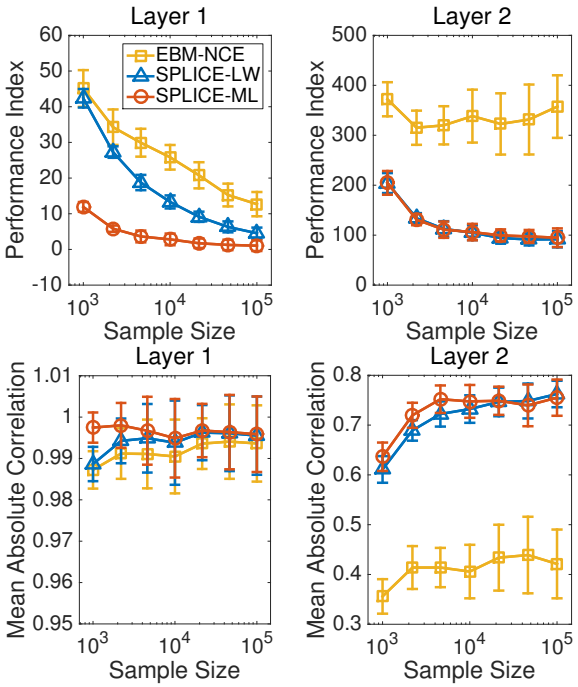

Figure 3: Synthetic Data: performance index (Amari et al., 1996) (top; smaller is better) and mean absolute correlations of true and estimated sources (bottom; larger is better) in each of the two layers (left: Layer 1; right: Layer 2). The plot shows the mean and standard deviation of 24 runs.

A simple simulation was performed to compare the basic performance of the three methods for blind source separation in each layer (Fig. 3). The dataset consisted of 30-dimensional complex-valued vectors $\mathbf{x}_t$ which we synthesized by our generative model. We generated the top-layer sources s'_k from t-distribution of the three degrees of freedom and every entry in $\mathbf{A}'$ and (the real and complex parts of) $\mathbf{A}$ uniformly in $[-1, 1]$. For simplicity, we used the same Gaussianization-based F^{-1} to generate the data. As seen in the figure, both SPLICE-LW and SPLICE-ML clearly outperformed EBM in Layer 2, which well corresponds to the lack of prior independence in EBM (Section 3.2). In Layer 1, all the methods seem to have correctly recovered the first-layer sources, as indicated by the high mean absolute correlations, while the two SPLICE methods exhibited improvements in accuracy particularly with smaller sample sizes. The consistent improvements (or non-degradation) by SPLICE-ML over SPLICE-LW also

demonstrated the effect of unsupervised finetuning.

4.2. EEG Data

To further demonstrate the applicability to exploratory analysis of neuroimaging signals, we applied the same methods to a publicly available EEG datasets: Datasets 1 (Blankertz et al., 2007) from the BCI competition IV (http://www.bbci.de/competition/iv/). The data were measured in four human subjects during a number of trials of a two-class cued motor imagery task; see Supplementary Material C for the details of data preprocessing. We eventually obtained the complex-valued data vectors $\mathbf{x}_t \in \mathbb{C}^{1845}$ by concatenating 41 sensor channels' complex time-frequency spectra (45 points within 8-30Hz; typical α and β bands) at every time points indexed by $t = 1, 2, \ldots, 4000$. As a preprocessing, PCA reduced the dimensionality with the 99% of total variance kept.

The idea for the analysis was that the amplitude $|s_j|$ of oscillatory EEG sources might couple together to represent higher-order information, in particular, that associated with the ongoing task states (i.e., imagery of two different motor modalities like left and right hands). To verify this, we evaluated individual second-layer sources s'_k, obtained in the unsupervised manner by each of the three methods, in terms of their relevance to discriminating the task states.

Specifically, we calculated AUC (area under the ROC curve) as the relevance measure, by regarding the within-trial average of every s'_k as a single discriminant score (Fig. 4). We also evaluated the similar score on another dataset (provided originally for the evaluation purpose in the competition) by transferring the same model without modification (Supplementary Material C). For both datasets, the increase of the fraction of high-AUC components s'_k by the two SPLICE methods is evident by the heavier upper tails as well as the Q-Q plots above the straight lines; the effect of finetuning was unclear in this result. The result implies that SPLICE may effectively discover task-related functional couplings of source amplitudes, which will be practically useful to enhance further explorations of data or help consolidating new hypotheses.

4.3. Natural Images

Finally, we demonstrate the validity of our general SPLICE model (Fig. 1(a)), using (real-valued) natural images obtained from ImageNet10K (Deng et al., 2010). We followed (Hosoya & Hyvärinen, 2015) for basic preprocessing. Image patches were of 32×32 pixels, with the pixel values in each patch normalized to have zero mean and unit variance. The dimensionality was then reduced to $d = 200$ by PCA. The logarithmic nonlinearity F was commonly used in both SPLICE and EBM.

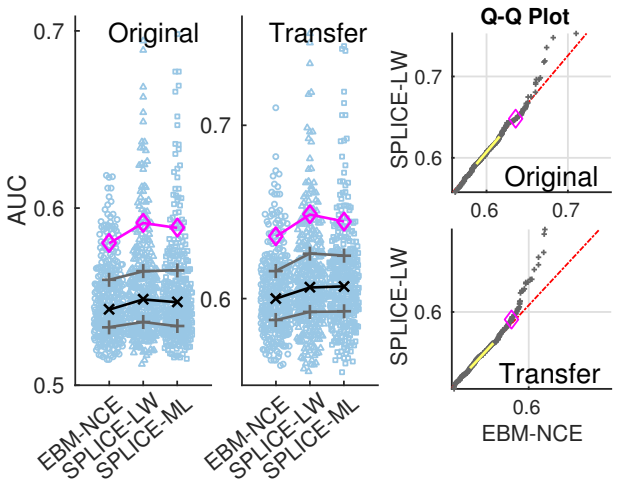

Figure 4: EEG Data: relevance of individual second-layer components s'_k to the task state. Left: distributions of AUC scores by the three methods ('$\times$': median, '+': 1st & 3rd quartiles, '$\diamond$': 90-percentile) for original and transfer data. Right: quantile-quantile (Q-Q) plots between SPLICE-LW and EBM-NCE (yellow line connects 1st and 3rd quartiles, extended by red dashed line; '$\diamond$': 90-percentile).

We first quantitatively compared two-layer SPLICE and EBM (as well as a single-layer ICA) by their test log-likelihood evaluated with 10-fold cross-validation (CV), for two different sample sizes n (Table 1). The number of second-layer sources was commonly set as $m = 50$. SPLICE-LW initialized the model parameters and subspace partitioning in both SPLICE-ML and EBM-NCE. The likelihood of EBM was numerically evaluated with hybrid annealed importance sampling (AIS) (Ngiam et al., 2011; Sohl-Dickstein & Culpepper, 2012) [1]. For a reference, we also give another result of SPLICE when its partition function ($= 1$) was evaluated numerically by AIS. The table indicates a superior performance of SPLICE-ML, followed by SPLICE-LW, as compared to ICA or EBM-NCE. The AIS was very accurate in this result. The low performance of EBM could be attributed to either the model difference or the relative inefficiency (in sample size) of NCE as compared to the (quasi-)ML estimators of SPLICE-ML/LW.

The two-layer SPLICE model can be considered a tractable and generative counterpart to existing models in natural image statistics (e.g., Gutmann & Hyvärinen, 2013; Hosoya & Hyvärinen, 2015). Note that our second layer actually corresponds to their third layer as we do not count the intermediate pooling layer. In fact, our model qualitatively reproduced local spatial pooling of first-layer linear features (Fig. 5(a)) as well as various types of excitatory/inhibitory

[1] We obtained the Matlab code from https://github.com/Sohl-Dickstein/Hamiltonian-Annealed-Importance-Sampling. We set the number of intermediate distributions as 10^7.

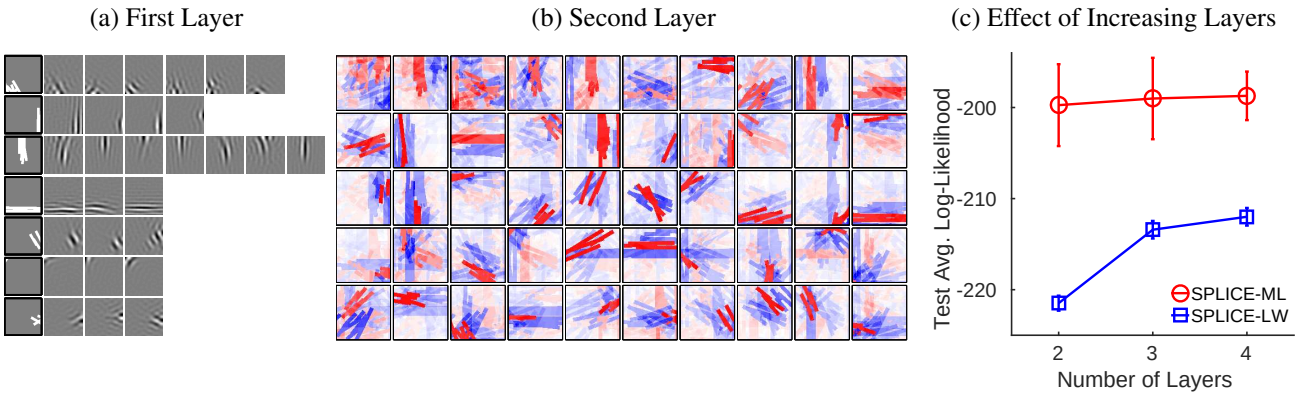

Figure 5: Natural Images: visualizations and the effect of increasing layers. (a) Examples of first layer's feature pooling learned by two-layer SPLICE-ML ($n = 200,000$; one pool per row, randomly selected out of 50). In each row, the left-most panel superimposes oriented bars that "iconify" (Gutmann & Hyvärinen, 2012) the pooled Gabor-like feature detectors $\mathbf{w}$ illustrated on its right. (b) All the 50 second-layer feature detectors $\mathbf{w}'_k$. Each panel visualizes a weighted superposition (using weights w'_{kj}, separately within positive and negative signs) of the 50 binarized iconified images of pooled first-layer features (i.e., the left-most panels in (a)). Red and blue corresponds to positive and negative signs and their thickness to absolute values (normalized in each sign to span the color ranges). The global sign of each $\mathbf{w}'_k$ (originally indefinite) was chosen so that its maximum absolute entry was positive. (c) Test average log-likelihood computed for both SPLICE-LW and ML with the numbers of layers 2, 3 and 4. The mean and standard deviations by 10-fold cross validation are shown.

Table 1: Natural Images: comparison with ICA and EBM. Test average log-likelihood evaluated by 10-fold cross validation (mean±SD) for different sample sizes n.

Method	Test Avg. Log-Likelihood (10-fold CV) $n = 100,000$	$n = 200,000$
ICA	- 270.09 ± 0.96	- 271.63 ± 1.40
SPLICE-LW	- 222.21 ± 1.04	-221.18 ± 0.86
SPLICE-ML	- 200.77 ± 6.53	-199.67 ± 6.46
SPLICE-ML*	-200.76 ± 6.54	-199.68 ± 6.47
EBM-NCE*	-342.12 ± 17.36	-330.04 ± 20.02

(∗: with a numerically estimated partition function)

patterns on pooled first-layer features (Fig. 5(b)) which theoretically models the properties of cortical neurons in the visual area V2 (Hosoya & Hyvärinen, 2015).

Finally, we examined the effect of increasing the number of layers (Fig. 5(c), $n = 200,000$). The three-layer SPLICE model included an additional pooling layer reducing from 50 to 10 inputs to the third-layer ICA; the four-layer model further added the pooling and ICA layers which reduces the dimensionality from 10 to 2. A remarkable increase of test likelihood by adding the layers is seen in SPLICE-LW; a weak increase was seen but not completely evident in SPLICE-ML. We therefore conclude that adding layers was effective to compensate errors in the pretrained lower layers, while only two layers (but not one layer; see ICA in Table 1) may be almost sufficient to represent the data when combined with finetuning. This result could be accounted for by either misspecification of third- and fourth-layer models (e.g., the type of nonlinearity F or the number of subspaces) or limited statistical regularities that can be present in small image patches.

5. Conclusion

We introduced SPLICE, a novel hierarchical extension of ICA with an intrinsic pooling mechanism. The striking feature of SPLICE is that the model is fully tractable, i.e., both the posterior estimates on latent variables and the associated pdf or likelihood can be computed with simple analytical formulae. The conceptual simplicity of SPLICE, as well as its approximation-free nature, will be particularly useful in exploratory data analysis and modeling, for example, in neuroscientific contexts. As a proof-of-concept, we demonstrated the applicability of the method with EEG and natural image patches.

So far, the development of hierarchical or deep generative models has been hampered by intractability and the ensuing computational difficulties. Intriguingly, SPLICE relies only on conventional principles for statistical modeling and estimation, and it can easily be extended with an arbitrary number of layers as well as other typical ingredients of multilayer neural networks. We hope our developments will open up new avenues for applications of hierarchical probabilistic models in unsupervised representation learning of continuous-valued data.

Acknowledgments

This work was partially supported by a contract with the National Institute of Information and Communica-

tions Technology entitled, Development of network dynamics modeling methods for human brain data simulation systems, Strategic International Collaborative Research Program (SICORP) from Japan Science and Technology Agency (JST), and KAKENHI 15H02759 from Japan Society for the Promotion of Science (JSPS). A.H. was funded by the Academy of Finland (Centre-of-Excellence in Inverse Problems Research).

References

Adali, T. and Haykin, S. *Adaptive Signal Processing: Next Generation Solutions*. Wiley-IEEE Press, 2010.

Adelson, E. H. and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, 1985.

Almeida, L. B. MISEP–linear and nonlinear ICA based on mutual information. *Journal of Machine Learning Research*, 4:1297–1318, 2003.

Amari, S., Cichocki, A., and Yang, H. H. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 8 (NIPS1995)*, pp. 757–763, 1996.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 35 (8):1798–1828, 2013.

Bingham, E. and Hyvärinen, A. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, 2000.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., and Curio, G. The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects. *NeuroImage*, 37(2):539–550, 2007.

Cadieu, C. F. and Olshausen, B. A. Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, 24(4):827–866, 2012.

Chen, S. S. and Gopinath, R. A. Gaussianization. In *Advances in Neural Information Processing Systems 13 (NIPS2000)*, pp. 423–429, 2001.

Deco, G. and Brauer, W. Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures. *Neural Networks*, 8:525–535, 1995.

Deng, J., Berg, A. C., Li, K., and Fei-Fei, Li. What does classifying more than 10,000 image categories tell us? In *Computer Vision – ECCV2010*, pp. 71–84, 2010.

Ding, C., Li, T., Peng, W., and Park, H. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD2006)*, pp. 126–135, 2006.

Dinh, L., Krueger, D., and Bengio, Y. NICE: Non-linear independent components estimation. arXiv:1410.8516, 2014.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Gutmann, M. U. and Hyvärinen, A. A three-layer model of natural image statistics. *Journal of Physiology-Paris*, 2013.

Hirayama, J. and Hyvärinen, A. Structural equations and divisive normalization for energy-dependent component analysis. In *Advances in Neural Information Processing Systems 24 (NIPS2011)*, pp. 1872–1880, 2012.

Hirayama, J., Ogawa, T., and Hyvärinen, A. Unifying blind separation and clustering for resting-state EEG/MEG functional connectivity analysis. *Neural Compution*, 27 (7):1373–1404, 2015.

Hirayama, J., Hyvärinen, A., Kiviniemi, V., Kawanabe, M., and Yamashita, O. Characterizing variability of modular brain connectivity with constrained principal component analysis. *PLoS ONE*, 11(12):e0168180, 2016.

Hosoya, H. and Hyvärinen, A. A hierarchical statistical model of natural images explains tuning properties in V2. *The Journal of Neuroscience*, 35(29):10412–10428, 2015.

Hosoya, H. and Hyvärinen, A. Learning visual spatial pooling by strong PCA dimension reduction. *Neural Computation*, 82:1–16, 2016.

Hubel, D. H. and Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1):106–154, 1962.

Hyvärinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Hyvärinen, A. and Hoyer, P. O. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *Proc. Artificial Intelligence and Statistics (AISTATS2017)*, 2017a.

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS2016)*, 2017b.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

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

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. John Wiley & Sons, 2001b.

Hyvärinen, A., Hurri, J., and Hoyer, P. O. *Natural Image Statistics – A probabilistic approach to early computational vision*. Springer-Verlag, 2009.

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

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pp. 3581–3589, 2014.

Köster, U. and Hyvärinen, A. A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, 22:2308–2333, 2010.

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pp. 3361–3368, 2011. ISBN 978-1-4577-0394-2.

Lyu, S. and Simoncelli, E. P. Nonlinear extraction of 'Independent Components' of natural images using radial Gaussianization. *Neural Computation*, 21(6):1485–1519, 2009.

Ngiam, J., Chen, Z., Koh, P., and Ng., A. Y. Learning deep energy models. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML)*, pp. 1105–1112, 2011.

Ollila, E., Tyler, D. E., Koivunen, V., and Poor, H. V. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, 2012.

Onton, J. and Makeig, S. High-frequency broadband modulations of electroencephalographic spectra. *Frontiers in Neuroscience*, 159:99–120, 2009.

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

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML2015)*, volume 37, pp. 1530–1538, 2015.

Salakhutdinov, R. and Hinton, G. Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR: W&CP*, pp. 448–455, 2009.

Schmidt, M. minFunc: unconstrained differentiable multivariate optimization in Matlab. http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html, 2005.

Shan, H., Zhang, L., and Cottrell, G. W. Recursive ICA. In *Advances in Neural Information Processing Systems*, volume 19, pp. 1273–1280, 2006.

Sohl-Dickstein, J. and Culpepper, B. J. Hamiltonian annealed importance sampling for partition function estimation. arXiv:1205.1925, 2012.

Szabó, Z., Póczos, B., and Lörincz, A. Separation theorem for independent subspace analysis and its consequences. *Pattern Recognition*, 45(4):1782–1791, 2012.

Taleb, A. and Jutten, C. Source separation in post-nonlinear mixtures. *IEEE Transactions on signal processing*, 47 (10):2807–2820, 1999.

Valpola, H., Harva, M., and Karhunen, J. Hierarchical models of variance sources. *Signal Processing*, 84(2): 267–282, 2004.

Yu, S. X. and Shi, J. Multiclass spectral clustering. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV2003)*, pp. 313–319, 2003.

Zhang, K. and Hyvärinen, A. Source separation and higher-order causal analysis of MEG and EEG. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pp. 709–716, 2010.