# Sparsity and Independence: Balancing Two Objectives in Optimization for Source Separation with Application to fMRI Analysis

Zois Boukouvalas[a], Yuri Levin-Schwartz[b], Vince D. Calhoun[c,d], and Tülay Adalı[b]

[a] *University of Maryland, Baltimore County, Dept. of Mathematics and Statistics, Baltimore, MD 21250*
[b] *University of Maryland, Baltimore County, Dept. of CSEE, Baltimore, MD 21250*
[c] *The Mind Research Network, Albuquerque, NM*
[d] *Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM*

## Abstract

Because of its wide applicability in various disciplines, blind source separation (BSS), has been an active area of research. For a given dataset, BSS provides useful decompositions under minimum assumptions typically by making use of statistical properties—types of diversity—of the data. Two popular types of diversity that have proven useful for many applications are *statistical independence* and *sparsity*. Although many methods have been proposed for the solution of the BSS problem that take either the statistical independence or the sparsity of the data into account, there is no unified method that can take into account both types of diversity simultaneously. In this work, we provide a mathematical framework that enables direct control over the influence of these two types of diversity and apply the proposed framework to the development of an effective ICA algorithm that can jointly exploit independence and sparsity. In addition, due to its importance in biomedical applications, we propose a new model reproducibility framework for the evaluation of the proposed algorithm. Using simulated functional magnetic resonance imaging (fMRI) data, we study the trade-offs between the use of sparsity versus independence in terms of the separation accuracy and reproducibility of the algorithm and provide guidance on how to balance these two objectives in real world applications where the ground truth is not available.

*Keywords:* Diversity, independent component analysis, sparsity, reproducibility.

[*]This work was supported in part by the NSF grants 1631838, 1539067, and 1618551, and the NIH-NIBIB grant R01EB020407.

## 1. Introduction

Blind source separation (BSS) is an active area of research in statistical signal processing due to its numerous applications, including, analysis of medical imaging data, wireless communications, and image processing. The objective of BSS methods is to decompose a set observations into the product of a mixing matrix and a matrix of latent sources. However, without the exploitation of any prior knowledge about the data, most typically its statistical properties—types of diversity—the matrix factorization problem is ill-posed. Two of the most popular forms of diversity that have proven useful in many practical applications and enable unique solutions up to scaling and permutation ambiguities are independence [1, 2, 3, 4, 5, 6, 7, 8] and sparsity [9, 10, 11, 12, 9].

A powerful method that solely relies on the independence of the sources is independent component analysis (ICA) [1, 2]. ICA provides a unique decomposition such that the sources are statistically independent subject to only scaling and permutation ambiguities. In contrast, methods such as dictionary learning (DL) [9] and sparse component analysis (SCA) [13, 14], take the sparsity of the sources directly into account, yielding decompositions where the estimated components are as sparse as possible, subject to the same permutation and scaling ambiguities as ICA, however with uniqueness guarantees only under specific conditions [9].

Though all of the above methods work well when their underlying assumptions are satisfied, neither of the methods exploit both independence and sparsity under a unified framework. In order to take advantage of these two forms of diversity, jointly, many *ad hoc* methods have been proposed, such as by selecting a density model that favors sparse distributions in ICA as noted in [1] or by using sparsity transformations following ICA [15]. Although selecting the source distribution would allow the ICA model to enjoy the desirable large sample properties of the maximum likelihood (ML) formulation [1, 2], the model would be limited to a specific type of sparse distribution [1]. Additionally, sparsity transformations are an indirect way of imposing sparsity and do not allow a direct way of controlling independence versus sparsity.

In this work, we present a mathematical framework that enables taking both independence and sparsity into account in an efficient manner. We incorporate sparsity through the use of a decoupled ICA cost function, penalized by an $\ell^1$ regularization term, thus, enabling a direct exploitation of sparsity for each source individually. We use ICA by entropy bound minimization (ICA-EBM) [16], a flexible yet parameter–free algorithm that effectively maximizes independence, as the underlying ICA algorithm for demonstrating the application of the proposed framework. The new algorithm we develop, the SparseICA-EBM algorithm, inherits all the advantages of ICA-EBM, namely its flexibility, though with enhanced performance due to the exploitation of sparsity and enables direct control over the degree to which independence and sparsity are emphasized. Although, estimation accuracy is an effective metric to evaluate the separation power of a BSS algorithm, in many applications such as the analysis of fMRI data, model reproducibility is an important performance metric. Its importance derives from its ability to reveal how consistently an algorithm can produce similar estimated sources across different sets of data that are supposed to have come from the same distribution, such as different scans of the same subject. Thus, we propose a new model reproducibility framework to evaluate the consistency of SparseICA-EBM, and using simulated fMRI data, we study the impact of the regularization parameters on the reproducibility of the results as well as on the estimation accuracy. This enables us to understand the trade-off between those two objectives in the ICA optimization framework and provides a guideline for parameter selection when ground truth is not available.

The remainder of this paper is organized as follows. In Section 2, we provide a brief background on the relevant BSS methods, ICA, SCA, and DL. Section 3, provides the mathematical development of SparseICA-EBM as well as the pseudo-code of the main part of the proposed algorithm. In Section 4, we describe the data generation as well as the evaluation metrics. In Section 5, we present the experimental results for SparseICA-EBM. The conclusions and future research directions are presented in Section 6.

## 2. Mathematical Background

For a given observation matrix $\mathbf{X} \in \mathbb{R}^{M\times V}$, the noiseless BSS generative model is given by

$$\mathbf{X} = \mathbf{A}\mathbf{S}^\top, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{M\times N}$ is the mixing matrix and $\mathbf{S} \in \mathbb{R}^{V\times N}$ is the matrix that contains the source signals. The matrix decomposition in (1) is an ill-posed problem, since for any invertible matrix $\mathbf{T} \in \mathbb{R}^{N\times N}$, it always holds that

$$\mathbf{X} = \mathbf{A}\mathbf{S}^\top = (\mathbf{A}\mathbf{T})(\mathbf{T}^{-1}\mathbf{S}^\top). \tag{2}$$

However, by the exploitation of different types of diversity, we can achieve unique decompositions up to only scaling and permutation ambiguities. Two types of diversity that have been used in many applications are statistical independence and sparsity.

### *2.1. Independent Component Analysis*

One of the most widely used methods for solving the BSS problem (1) is ICA and its basic assumption is that the source signals are statistically independent. Therefore, by rewriting (1) using the random vector notation, we have

$$\mathbf{x}(v) = \mathbf{A}\mathbf{s}(v), \quad v = 1, \ldots, V, \tag{3}$$

where $v$ is the sample index, $\mathbf{s}(v) \in \mathbb{R}^N$ are the unknown source signals, and $\mathbf{x}(v) \in \mathbb{R}^M$ are the mixtures. A common case in many applications is the overdetermined one ($M > N$), which can be reduced to the case where $M = N$ using dimensionality reduction following principal component analysis (PCA). Since the sources $s_n(v),\ 1 \le n \le N$ in $\mathbf{s}(v) = [s_1(v), \ldots, s_N(v)]^\top$ are assumed to be statistically independent, the goal is to estimate a demixing matrix $\mathbf{W} \in$

$\mathbb{R}^{N\times N}$ to yield maximally independent source estimates $\mathbf{y}(v) = \mathbf{W}\mathbf{x}(v)$. Due to its large sample size optimality properties, the maximum likelihood (ML) can serve as the objective function for ICA and is given by

$$\begin{aligned}\mathcal{L}(\mathbf{W}) &= \frac{1}{V}\sum_{v=1}^{V}\sum_{n=1}^{N}\log p(\mathbf{w}_n^\top \mathbf{x}(v)) + \log|\det(\mathbf{W})| \\ &\approx E\left\{\sum_{n=1}^{N}\log p(\mathbf{w}_n^\top \mathbf{x})\right\} + \log|\det(\mathbf{W})|, \end{aligned} \tag{4}$$

where $p(\mathbf{w}_n^\top \mathbf{x})$ is the probability density function (PDF) of the estimated random variable $y_n = \mathbf{w}_n^\top \mathbf{x}$. The approximation in (4) is obtained by the mean ergodic theorem under the assumption that the samples are independent and identical distributed (i.i.d). It has been shown that maximization of (4) is equivalent to the minimization of the mutual information (MI), as long as the assumed model PDF matches the true latent source PDF [1]. Mutual information, which is defined as the Kullback-Leibler (KL)-distance between the joint source density and the product of the marginal estimated source densities, is given by

$$\begin{aligned}J_{ICA}(\mathbf{W}) &= E\left\{-\log\left[\frac{p_{s_1}(y_1)p_{s_2}(y_2)\cdots p_{s_N}(y_N)}{p_{s_1 s_2\ldots s_N}(y_1, y_2, \ldots, y_N)}\right]\right\} \\ &= E\left\{-\sum_{n=1}^{N}\log p_{s_n}(y_n)\right\} + E\left\{\log p_{\mathbf{s}}(\mathbf{y})\right\} \\ &= \sum_{n=1}^{N} H(y_n) - H(\mathbf{y}) \\ &= \sum_{n=1}^{N} H(y_n) - \log|\det(\mathbf{W})| - H(\mathbf{x}), \end{aligned} \tag{5}$$

where the terms $H(y_n)$, $H(\mathbf{x})$, and $H(\mathbf{y})$ are the (differential) entropy of the source estimates, the mixtures, and the estimated random vector $\mathbf{y}$ respectively. Note that the term $H(\mathbf{x})$ is independent of $\mathbf{W}$ and can be treated as a constant during the optimization procedure. For a more detailed discussion of ICA and different types of ICA algorithms that lie under the maximum likelihood umbrella we refer the reader to [1, 17].

### *2.2. Dictionary Learning and Sparse Component Analysis*

Though ICA has proven useful in many practical applications, generally independence is not the only form of diversity inherent to the sources in (1). One of the most popular forms of diversity to exploit is sparsity and BSS methods that exploit solely the sparsity of the sources include DL and SCA.

By assuming that the observations can be expressed as sparse combinations of a dictionary $\mathbf{\Phi}$, DL seeks to estimate both the dictionary and the collection of weight vectors, $\mathbf{S}$, generally through an alternating estimation procedure. The cost for this task is given by

$$\min_{\mathbf{\Phi},\mathbf{S}} ||\mathbf{X} - \mathbf{\Phi}\mathbf{S}||_F^2 + \lambda||\mathbf{S}||_{1,1}, \tag{6}$$

where $||\mathbf{S}||_{1,1} = \sum_{i=1}^{M}\sum_{j=1}^{N}|s_{ij}|$ and $\lambda$ is the regularization parameter. Different DL algorithms include those based on probabilistic learning methods, learning methods based on clustering, among others [9]. For a more detailed review of DL and its applications, we refer the reader to [9, 18].

A related method to DL that exploits solely sparsity is SCA. If $\mathbf{\Phi} \in \mathbb{R}^{K\times V}$ denotes a dictionary matrix, whose rows are called the atoms, then at the first step of SCA, $\mathbf{\Phi}$ is applied to the mixture matrix $\mathbf{X}$, to obtain $\mathbf{C_x} \in \mathbb{R}^{P\times K}$. In such a case, the column vectors $\mathbf{C_x}(k)$ $k = 1, \dots K$, form the scatter plot $\{\mathbf{C_x}(k)\}_{k=1}^{K}$. If the dictionary has been selected properly, *i.e.*, has as sparse a representation of the data as possible, the elements of $\{\mathbf{C_x}(k)\}_{k=1}^{K}$ are almost aligned with the columns of the mixing matrix. In the second step, the mixing matrix $\mathbf{A}$ needs to be estimated by $\{\mathbf{C_x}(k)\}_{k=1}^{K}$. Thus, under the assumption that at most one source contributes to each point of the scatter plot, clustering techniques can be used to estimate $\hat{\mathbf{A}}$. The third step consists of the estimation of the source representations that can be denoted as $\mathbf{C_s} \in \mathbb{R}^{P\times K}$, due to the sparsifying transformation, $\mathbf{C_x} = \mathbf{X}\mathbf{\Phi}^\top$, that has been applied to the mixture matrix at the first step of SCA. Each column of $\mathbf{C_s}$ can be estimated through the minimization problem

$$\hat{\mathbf{C}}_\mathbf{s}(k) = \arg \min_{\mathbf{c}|\mathbf{C_x}(k)=\hat{\mathbf{A}}\mathbf{c}} ||\mathbf{c}||_1, \tag{7}$$

where $\mathbf{c}$ is the vector that needs to be minimized such as $\mathbf{C_x}(k) = \hat{\mathbf{A}}\mathbf{c}$ and the solution of the minimization problem gives an estimate of the $k$th column of $\mathbf{C_s}$. The final step consists of reconstructing the sources by $\mathbf{y} = \mathbf{C_s}\mathbf{\Phi}$, when the initial dictionary matrix is orthogonal. For a more detailed discussion of SCA, we refer the reader to [13].

Although ICA, DL, and SCA have their own justifications in terms of the diversity that they exploit, the differences among these methods do not facilitate transformation from one method to another, thus making it difficult to balance these two different forms of diversity, independence and sparsity. Specifically, ICA is based on the assumption that the sources are statistically independent, while DL or SCA assumes that the sources are sparse. The main contribution of our work is to develop a new framework that enables a translation between the two objectives of sparsity and independence and exploration of the trade-offs between emphasizing one over the other. We apply this framework to the development of an effective ICA algorithm that can jointly exploit both independence and sparsity.

## 3. Sparse Independent Component Analysis

### 3.1. Cost Function

Classically, sparsity is measured using the $\ell^0$ norm, and is defined as the number of non-zero coefficients from a vector $\mathbf{u} \in \mathbb{R}^V$

$$||\mathbf{u}||_0 = \#\{u_i \neq 0; i = 1, \dots V\}. \tag{8}$$

Although the incorporation of (8) into the ICA framework is the most direct way to impose sparsity on the ICA cost function, the $\ell^0$ norm is computationally intractable. On the other hand, the $\ell^1$ norm, defined as the sum of the absolute values of a vector's coefficients, has served as a computationally efficient sparsity regularizer see *e.g.*, [19, 20, 21]. For this reason, we propose a direct way to promote sparsity into the ICA model through the addition of an $\ell^1$ regularization term to the ICA cost function. The addition of this term is expected to improve separation performance beyond what is achieved solely through the maximization of independence when the underlying sources are truly sparse.

However, it is difficult to balance the contribution of sparsity for each of the individual sources while optimizing (5), due to the $\log|\det(\mathbf{W})|$ term. This issue can be avoided by expressing (5) and its gradient as a sequence of equations, where each equation is written with respect to each row $\mathbf{w}_n,\ n = 1, \dots, N$ of the demixing matrix $\mathbf{W}$. Therefore, by using this decoupling approach [16, 22], the sequence of MI cost functions is given by

$$J_{ICA}(\mathbf{w}_n) = H(y_n) - \log\left|\mathbf{h}_n^\top \mathbf{w}_n\right| - C_n,\ n = 1, \dots, N, \tag{9}$$

where $\mathbf{h}_n$ is a unit vector that is perpendicular to all row vectors of $\mathbf{W}$ except $\mathbf{w}_n$ and each $C_n$ is a constant that contains all the terms that are independent of $\mathbf{w}_n$. Therefore, using (9), the proposed sequence of cost functions that take both independence and sparsity of each individual source into account is given by

$$J(\mathbf{w}_n) = J_{ICA}(\mathbf{w}_n) + \lambda_n f(y_n),\ n = 1, \dots, N, \tag{10}$$

where $f(y_n) = ||y_n||_1$ is the regularization term and $\lambda_n$ is the sparsity parameter for $n = 1, \dots N$. Note, that with a slight abuse of notation in (10), we treat $y_n$ as a vector where each coordinate corresponds to a sample drawn from the random variable $y_n$. The $\ell^1$ norm is a non-differentiable function, so it is replaced by the the sum of multi-quadratic functions [23], given by

$$f(y_n) = \lim_{\epsilon_n \to 0} \sum_{v=1}^{V} \sqrt{y_{n_v}^2 + \epsilon_n}, \tag{11}$$

where $\epsilon_n$ is the smoothing parameter.

### *3.2. Algorithmic Development*

ICA by entropy bound minimization (ICA-EBM) is a flexible and parameter-free algorithm that can maximize independence in an efficient manner through the use of four measuring functions matching: unimodal or bimodal, symmetric or skewed, heavy-tailed or not heavy-tailed distributions [16]. It is due to this flexibility and ability to effectively maximize independence that ICA-EBM serves as the algorithm for the direct integration of (10).

The gradient of (10) with respect to (w.r.t) $\mathbf{w}_n$ is given by

$$\frac{\partial}{\partial \mathbf{w}_n} J(\mathbf{w}_n) = \frac{\partial J_{ICA}(\mathbf{w}_n)}{\partial \mathbf{w}_n} + \lambda_n \lim_{\epsilon_n \to 0} \sum_{v=1}^{V} \frac{y_{n_v}}{\sqrt{y_{n_v}^2 + \epsilon_n}} \mathbf{x}, \tag{12}$$

where

$$\frac{\partial J_{ICA}(\mathbf{w}_n)}{\partial \mathbf{w}_n} = -E\left\{ \frac{\partial \log p(y_n)}{\partial y_n} \mathbf{x} \right\} - \frac{\mathbf{h}_n}{\mathbf{h}_n^\top \mathbf{w}_n},$$

and $p(y_n)$, can be adaptively determined for each estimated source independently. We refer to this new ICA algorithm as SparseICA-EBM. For better convergence properties, we follow the technique in [16] and define the domain of our cost function to be the unit sphere in $\mathbb{R}^N$. By using the projection transformation onto the tangent hyperplane of the unit sphere at the point $\mathbf{w}_n$, the normalized gradient of our cost function is given by

$$\mathbf{u}_n = \mathbf{P}_n(\mathbf{w}_n) \frac{\partial J(\mathbf{w}_n)}{\partial \mathbf{w}_n}, \tag{13}$$

where $\mathbf{P}_n(\mathbf{w}_n) = \mathbf{I} - \mathbf{w}_n \mathbf{w}_n^T$ and $||\mathbf{w}_n|| = 1$.

In order to achieve fast convergence, SparseICA-EBM has been implemented using three stages. First, FastICA [24] is performed on the mixtures, generating an initial estimate of the demixing matrix $\mathbf{W}$. This estimate is further refined through the performance of orthogonal ICA using (10). The final stage consists of the application of non-orthogonal ICA using the estimated $\mathbf{W}$ obtained from the previous stage. The pseudo-code description of the non-orthogonal ICA stage is presented in Algorithm 1.

The term $J(\mathbf{W})$ introduced in Algorithm 1, is given by

$$J(\mathbf{W}) = \sum_{n=1}^{N} H(y_n) - \log|\det(\mathbf{W})| + \sum_{n=1}^{N} \lambda_n ||y_n||_1. \tag{14}$$

and is computed after the estimation of each $\mathbf{w}_n$ for each ICA iteration. To calculate the vector $\mathbf{h}_n$, we introduce the matrix $\mathbf{W}_n = [\mathbf{w}_1, \ldots, \mathbf{w}_{n-1}, \mathbf{w}_{n+1}, \ldots, \mathbf{w}_N]^\top$. Then, $\mathbf{h}_n$ is obtained as

$$\mathbf{h}_n = \frac{\mathbf{r} - \mathbf{W}_n^\top \mathbf{Q}_n^{-1} \mathbf{W}_n \mathbf{r}}{||\mathbf{r} - \mathbf{W}_n^\top \mathbf{Q}_n^{-1} \mathbf{W}_n \mathbf{r}||}, \tag{15}$$

where $\mathbf{r}$ is an arbitrary vector of size $N \times 1$ that is not orthogonal to $\mathbf{w}_n$ and $\mathbf{Q}_n = \mathbf{W}_n \mathbf{W}_n^\top$.

**Algorithm 1** SparseICA-EBM

```
1: Input: $\mathbf{X} \in \mathbb{R}^{N \times V}, \mathbf{W}_{\text{init}}, \lambda_n, \epsilon_n$
2: for $n = 1{:}N$ do
3:     Compute $\mathbf{h}_n$, orthogonal to $\mathbf{w}_i$ for all $i \neq n$
4:     Calculate the derivative $\frac{\partial J(\mathbf{w}_n)}{\partial \mathbf{w}_n}$ using (12)
5:     Project the gradient onto the unit sphere using (13)
6:     $(\mathbf{w}_n)^{\text{new}} \leftarrow (\mathbf{w}_n)^{\text{old}} - \gamma \mathbf{u}_n$
7: end
8: Repeat steps 2 through 7 until convergence in $J(\mathbf{W})$ or until the maximum
   number of iterations is exceeded
9: Output: $\mathbf{W}$
```

The proposed SparseICA-EBM algorithm not only provides flexible density matching but also yields solutions with variable levels of sparsity, through manual selection of $\lambda_n$ and $\epsilon_n$.

## 4. Evaluation Methods and Data Generation

For a BSS algorithm to be useful in real world applications, it must be able to efficiently extract the latent sources and do so consistently. Consequently, motivated by [25], to evaluate our proposed model, we consider two different metrics of performance. The first is in terms of its separation power, *i.e.*, its ability to accurately extract the latent sources, and the second is in terms of its reproducibility, *i.e.*, the consistency of the solutions across different datasets and runs. Such metrics are especially important in applications such as the analysis of fMRI data, since if sources are extracted incorrectly, the conclusion may be flawed, for instance leading to improper identification of biomarkers, *i.e.*, spatial patterns, of disease. Generally when using ICA on fMRI data, the estimated components tend to have sparse distributions [26], motivating the study of the synergy between independence and sparsity. Therefore, using these two measures of performance, we explore the trade-offs between the use of sparsity versus independence, through fMRI data, and provide a guidance on how to balance these two objectives in real world applications where the ground truth is not available.

This investigation is performed through the generation of simulated fMRI data using SimTB [27], which enables flexible generation of fMRI–like datasets under a model of spatio-temporal separability. To study the effect of independence against that of sparsity, we generate 10 datasets, each representing a different subject with 20 sources, for three different scenarios each with different levels of noise. The three scenarios are shown in Fig. 1 and consist of the cases where all sources are very sparse with little to no spatial overlap, a mixture of very sparse and less sparse sources again with little to no spatial overlap, and very sparse as well as less sparse sources with an increased amount of spatial overlap. The sparsity and the degree of overlap of the original sources is controlled by adjusting the SimTB parameter value that controls the “spread” of the sources. Note that when we decrease the spread of each individual source the sparsity of this particular source is decreased. This comes from the definition of a sparse distribution which is one for which most of the energy is contained in only a few of the coefficients [28]. The additive noise is Rician distributed and has energy specified by the contrast-to-noise ratio (CNR) defined as the ratio of the temporal standard deviation of the true signal divided by the temporal standard deviation of the noise [27]. Each source is a 100 × 100 image and the length of the experiment is 260 samples, meaning that simulated $\mathbf{X}$ is of dimension $260 \times 10^4$.

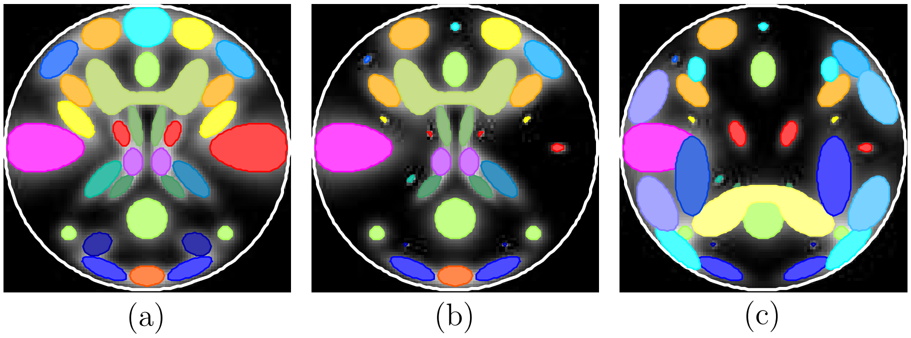


Figure 1: Simulated fMRI-like components for the three different scenarios. Note that each color indicates a different component. The scenarios are (a) all sources are very sparse with no spatial overlap, (b) a mixture of very sparse and less sparse sources and no spatial overlap, (c) very sparse as well as less sparse sources with a certain degree of spatial overlap.

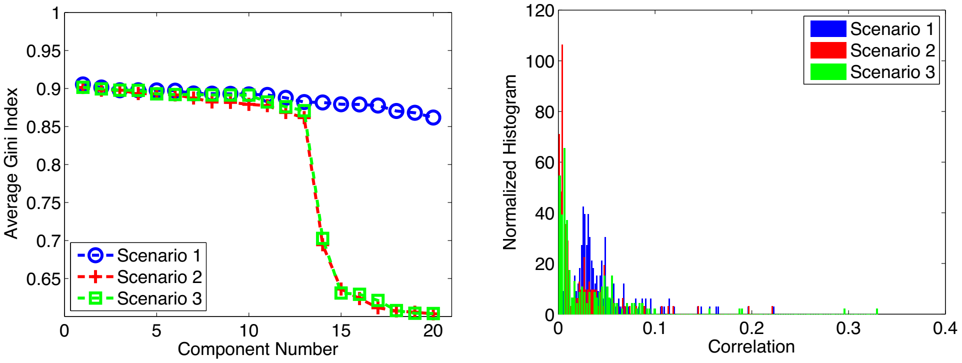
To verify the sparse nature of the sources for each of the three different scenarios, we measure the sparsity level of each source across all subjects, using the Gini index, defined as [28]

$$S(\mathbf{u}) = 1 - 2\sum_{v=1}^{V} \frac{u^{(v)}}{||\mathbf{u}||_1} \left( \frac{V - v + 1/2}{V} \right), \tag{16}$$

where $u^{(1)} \leq u^{(2)} \leq \cdots \leq u^{(V)}$ are the ordered coordinates of the vector $\mathbf{u} \in \mathbb{R}^V$. Note from (16) that the Gini index is normalized, with 1 corresponding to very sparse sources while 0 to dense sources. The average Gini indices for the 20 sources and for the three different scenarios are summarized in Fig. 2 (a). Additionally, we compute the average correlation across subjects and display the distribution of the values in Fig. 2 (b). Note that the mean and standard deviation of the pairwise source correlations are: $0.044 \pm 0.034, 0.022 \pm 0.031$, and $0.03 \pm 0.043$, respectively.

### *4.1. Balancing Independence and Sparsity*

Since the ground truth is available for our simulated sources, we evaluate the performance of SparseICA-EBM in terms of its separation power, using the



(a) (b)

Figure 2: (a) Average Gini index (b) distribution of the correlation values of the 20 latent sources for the three different scenarios. The Gini Index is normalized, with 1 corresponding to very sparse sources while 0 to dense sources.

average absolute value of the correlation between the true and the estimated sources. Thus, for the first part of our study, we evaluate the correlation coefficient between the true and the estimated spatial maps as a function of $\lambda_n$ and $\epsilon_n$. Since $\lambda_n$ controls the degree to which sparsity is emphasized over independence in SparseICA-EBM, we would like to visualize the behavior of the algorithm when we relax the independence assumption for each of the three groups and for different levels of noise.

The first step in processing the fMRI-like data consists of the application of PCA to each dataset, individually. Since 20 sources are generated for each dataset, the dimension of each dataset is temporally reduced to 20. After dimension reduction, we apply SparseICA-EBM to each dataset. After SparseICA-EBM has been applied to each subject's data, we pair the extracted components with the true latent sources. In the case where more than one estimated component is paired with a single true source, we use the Bertsekas algorithm [29], an iterative method that maximizes a given cost in a bipartite graph, to find the best assignment.

### *4.2. Model Reproducibility*

Since besides estimation accuracy it is important for a BSS algorithm to consistently produce similar results, we also study the reproducibility of the SparseICA-EBM as a function of the sparsity parameter $\lambda_n$ and the smoothing parameter $\epsilon_n$. Motivated by the the nonparametric, prediction, activation, influence, reproducibility, resampling (NPAIRS) framework in neuroimaging [25], we split the original dataset into two, and perform separate analyses on each of the sub-datasets and study the similarity of the two sets of resulting separated sources. Since selecting certain rows of $\mathbf{X}$ is equivalent to sub-sampling the corresponding rows of $\mathbf{A}$ multiplied by the source matrix $\mathbf{S}$, the similarity of the estimated sources is a good measure of the reproducibility of the proposed algorithm. A graphical illustration of this approach is presented in Fig. 3.

For this analysis, we split the mixture matrix $\mathbf{X}$, defined as the collection of all realizations of $\mathbf{x}(v)$, into two submatrices by selecting every other row of $\mathbf{X}$ creating $\mathbf{X}_1$ and $\mathbf{X}_2$, for each of the subjects. We apply PCA to each $\mathbf{X}_1$ and $\mathbf{X}_2$ for each subject and reduce their dimension to 20. After dimension reduction, we apply SparseICA-EBM to each reduced dataset. After SparseICA-EBM has been applied to the reduced submatrices, we pair the extracted components from the first submatrix with the extracted components from the second submatrix for each subject. In the case of multiple assignments, we again use the Bertsekas algorithm to determine the optimal assignment. We measure how close the pairs of estimated components are using the average absolute value of the correlation across subjects.

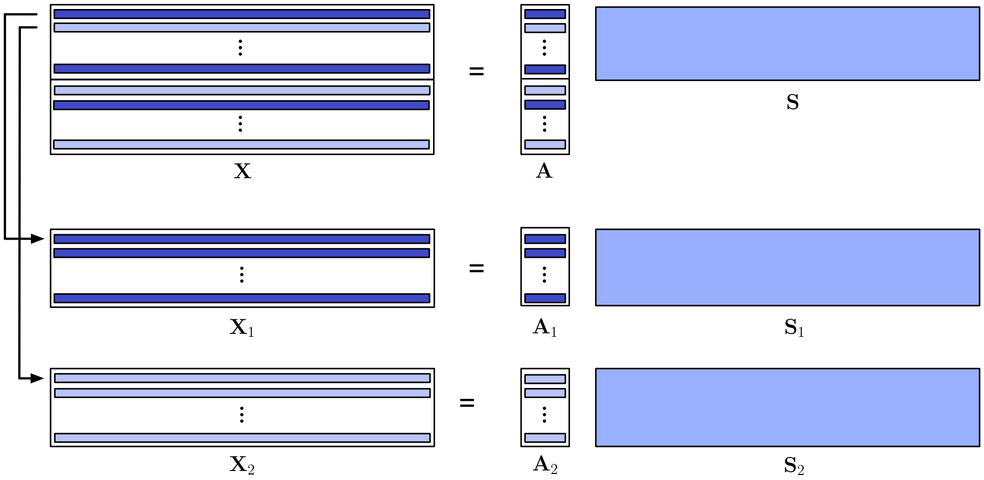


Figure 3: Visualization of subsampling method used to split the observation matrix in order to evaluate the reproducibility of the model. Note that under this reproducibility framework $\mathbf{S}_1 \cong \mathbf{S}_2 \cong \mathbf{S}$.

## 5. Experimental Results

Fig. 4 displays the average spatial correlation between the true and the estimated components as a function of the two key parameters for SparseICA-EBM, the regularization parameter $\lambda_n$ and the smoothing parameter $\epsilon_n$. Fig. 5 displays the average spatial correlation between the estimated components generated when applying SparseICA-EBM on the first half and the other half of the data as a function of $\lambda_n$ and $\epsilon_n$. For both figures the first column shows the results where data have been generated with no noise and the second column when noise has CNR = 1. For each noise level, we show the behavior and the reproducibility of the algorithm for the three different scenarios as described in the previous section. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF), for more information see hpcf.umbc.edu. Note that since the effectiveness of a BSS algorithm depends on both its accuracy and its consistency, in the two sets of figures that we present, we seek to find values of $\lambda_n$ and $\epsilon_n$ for which we obtain high source reconstruction accuracy as well as high reproducibility.

From Fig. 4(a) and (d), we observe that when the original sources do not have significant overlaps and all of them are characterized as very sparse, high values of $\lambda_n$ and $\epsilon_n$ produce higher average spatial correlation values, with the gain decreasing when noise level increases. This result shows that, in the case of truly sparse and independent sources, promoting sparsity within an ICA framework improves performance, since we effectively exploit another form of diversity, *i.e.*, property of the sources. In Fig. 4(b), we observe that, when some of the sources are sparse and some are less sparse, for high values of $\epsilon_n$, SparseICA-EBM with sparsity enforced, *i.e.*, high values of $\lambda_n$, provides better results, than with small values of $\lambda_n$, since only a third of the total sources are less sparse, thus the performance is dominated by the extraction of the sources that are sparse. From Fig. 4(e), SparseICA-EBM with only independence enforced, *i.e.*, small values of $\lambda_n$, and SparseICA-EBM with sparsity enforced and high values of $\epsilon_n$ provide similar separation performance, since the additive noise destroys the sparse nature of the data. Finally from Fig. 4(c) and (f), SparseICA-EBM with high values of $\lambda_n$ and $\epsilon_n$ provides similar results to SparseICA-EBM with low values of $\lambda_n$.

From Fig. 5(a) and (d), we observe that when the original sources do not overlap and all of them are characterized as very sparse, high values of $\lambda_n$ and for almost all values of $\epsilon_n$ the results are highly reproducible. Thus, for these cases, SparseICA-EBM produces both accurate and consistent results for large values of $\lambda_n$ and $\epsilon_n$. A similar trend can be observed in Fig. 5(b), where some of the sources are very sparse and some are less sparse. Fig. 5(e), for all values of $\lambda_n$ and $\epsilon_n$, SparseICA-EBM is becoming always consistent. Fig. 5(c), shows that, for some intermediate values of $\epsilon_n$, we have high reproducibility. Finally, in Fig. 5(f), SparseICA-EBM shows nearly identical results except for high values of $\lambda_n$ and $\epsilon_n$.

Based on Figs. 4 and 5, we can draw several interesting conclusions regarding the behavior of SparseICA-EBM, also can note few points for the selection of its parameters when we are working with real fMRI data. Since our goal is to have both high performance and high reproducibility, we observe that for the first and second scenarios where component overlaps are limited, sufficiently high values of $\lambda_n$, *i.e.,* in the interval $(10^{-2}, 10^4)$, as well as sufficiently high values of $\epsilon_n$, *i.e.,* in the interval $(0.5, 10)$, will produce sparse and smooth sources consistently. Moreover, for scenario 1, SparseICA-EBM with very small $\lambda_n$ is robust to noise. For the third scenario and when the values of $\lambda_n$ are small, SparseICA-EBM has relatively high performance. Therefore, for real world applications where all or a majority of sources can be assumed to be sparse high values of $\lambda_n$ and $\epsilon_n$ are expected to provide reasonable results, consistently. However for the case where overlaps are likely, by emphasizing both independence and sparsity in the optimization procedure will produce better overall performance.

An additional point worth noting, is that from Fig. 5(a), we observe a significant drop in reproducibility for $\lambda_n = 10^{-3}$. The reason for this is likely due to the fact that the SparseICA-EBM cost function consists of an independence term that is described by the negative of the ICA maximum likelihood function and a sparsity term that is described by the $\ell^1$ norm, the second term in the cost function. Since the contribution of sparsity is weighted by the parameter $\lambda_n$ and the optimal solutions of the two terms are not necessarily the same, changing the value of $\lambda_n$ affects the overall solution space each time SpaceICA-EBM is applied to $\mathbf{X}_1$ and $\mathbf{X}_2$. Numerical experiments have shown that for this data the two terms contribute almost equally in the optimization procedure when $\lambda_n = 10^{-3}$. This situation expands the solution space, resulting in more local optima, and thus, when SpaceICA-EBM is applied to $\mathbf{X}_1$ and $\mathbf{X}_2$ separately, it yields pairs of estimated components that correlate less with each other. The performance drop observed in Fig. 5(a) starts to disappear in the rest of the figures for which noise is introduced to the data, since noise destroys sparsity, or for scenarios where we manually reduce the sparsity of the original sources. Since in these two cases sparsity is insufficient to fully extract the sources, the solution space of the second term in the cost function is close to being flat. This results in a joint cost surface with fewer local minima and therefore better correlation between the two sets of estimated components.

## 6. Conclusion

Methods that exploit sparsity and independence have proven useful in many applications. This motivates the development of a method that can effectively take into account both types of diversity. In this work, we propose a new mathematical framework that enables direct control over the influence that independence and sparsity have on the result and use this framework to generate a powerful algorithm that takes both sparsity and independence into account. We explore the trade-offs between emphasizing these two objectives for different scenarios of simulated fMRI data and provide a guideline on the parameter selection for fMRI analysis when the ground truth is not available. Our results indicate that careful selection of the regularization parameters under certain scenarios will increase the quality of the final extracted sources enabling meaningful interpretations for fMRI analysis.

Our work motivates several interesting directions of further research, such as the development of automated techniques for parameter selection when the ground truth is not available. Additionally, the development of a technique that adaptively updates $\lambda_n$ and $\epsilon_n$ for each source would significantly increase the separation performance and improve the quality of the final extracted sources, especially when sources have different levels of sparsity. Finally, the study of the effect on the algorithm using different approximations of the $\ell^1$ norm would be of high interest.

## References

[1] T. Adalı, M. Anderson, G.-S. Fu, Diversity in Independent Component and Vector Analyses: Identifiability, algorithms, and applications in medical imaging, IEEE Signal Processing Magazine 31 (3) (2014) 18–33. `doi:10.1109/MSP.2014.2300511`.

[2] P. Comon, C. Jutten, Handbook of Blind Source Separation: Independent Component Analysis and Applications, 1st Edition, Academic Press, 2010.

[3] A. J. Bell, T. J. Sejnowski, The independent components of natural scenes are edge filters, Vision research 37 (23) (1997) 3327–3338.

[4] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4) (2000) 411–430.

[5] M. J. Mckeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, T. J. Sejnowski, Analysis of fMRI data by blind separation into independent spatial components, Human Brain Mapping 6 (3) (1998) 160–188. `doi:10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1`.
URL `http://dx.doi.org/10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1`

[6] T. Ristaniemi, J. Joutsensalo, On the performance of blind source separation in cdma downlink, in: Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA99), 1999, pp. 437–441.

[7] A. D. Back, A. S. Weigend, A first application of independent component analysis to extracting structure from stock returns, International journal of neural systems 8 (04) (1997) 473–484.

[8] I. Lee, T. Kim, T.-W. Lee, Fast fixed-point independent vector analysis algorithms for convolutive blind source separation, Signal Processing 87 (8) (2007) 1859–1871.

[9] I. Tosic, P. Frossard, Dictionary learning, IEEE Signal Processing Magazine 28 (2) (2011) 27–38. `doi:10.1109/MSP.2010.939537`.

[10] R. Gribonval, Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture, in: Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, Vol. 3, IEEE, 2002, pp. III–3057.

[11] M. Zibulevsky, B. A. Pearlmutter, Blind source separation by sparse decomposition in a signal dictionary, Neural computation 13 (4) (2001) 863–882.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 689–696.

[13] R. Gribonval, S. Lesage, A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges, in: ESANN'06 proceedings-14th European Symposium on Artificial Neural Networks, d-side publi., 2006, pp. 323–330.

[14] P. Bofill, M. Zibulevsky, Underdetermined blind source separation using sparse representations, Signal processing 81 (11) (2001) 2353–2362.

[15] S. Ma, X. L. Li, N. M. Correa, T. Adalı, V. D. Calhoun, Independent subspace analysis with prior information for fMRI data, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 1922–1925. `doi:10.1109/ICASSP.2010.5495320`.

[16] X.-L. Li, T. Adalı, Independent component analysis by entropy bound minimization, Signal Processing, IEEE Transactions on 58 (10) (2010) 5151–5164.

[17] A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural computation 7 (6) (1995) 1129–1159.

[18] X. Zhao, G. Zhou, W. Dai, W. Wang, Blind source separation based on dictionary learning: A singularity-aware approach, in: Blind Source Separation, Springer, 2014, pp. 39–59.

[19] E. J. Candes, T. Tao, Decoding by linear programming, IEEE transactions on information theory 51 (12) (2005) 4203–4215.

[20] M. Schmidt, G. Fung, R. Rosales, Fast optimization methods for $\ell^1$ regularization: A comparative study and two new approaches, in: European Conference on Machine Learning, Springer, 2007, pp. 286–297.

[21] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[22] M. Anderson, G. S. Fu, R. Phlypo, T. Adali, Independent vector analysis, the kotz distribution, and performance bounds, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3243–3247. `doi:10.1109/ICASSP.2013.6638257`.

[23] S.-I. Lee, H. Lee, P. Abbeel, A. Y. Ng, Efficient $\ell^1$ regularized logistic regression, in: Proceedings of the National Conference on Artificial Intelligence, Vol. 21, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 401.

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

[25] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, D. Rottenberg, The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework, NeuroImage 15 (4) (2002) 747–771.

[26] V. D. Calhoun, T. Adalı, Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery, IEEE Reviews in Biomedical Engineering 5 (2012) 60–73. `doi:10.1109/RBME.2012.2211076`.

[27] E. B. Erhardt, E. A. Allen, Y. Wei, T. Eichele, V. D. Calhoun, SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability, Neuroimage 59 (4) (2012) 4160–4167.

[28] N. Hurley, S. Rickard, Comparing measures of sparsity, IEEE Transactions on Information Theory 55 (10) (2009) 4723–4741.

[29] D. P. Bertsekas, The auction algorithm: A distributed relaxation method for the assignment problem, Annals of operations research 14 (1) (1988) 105–123.

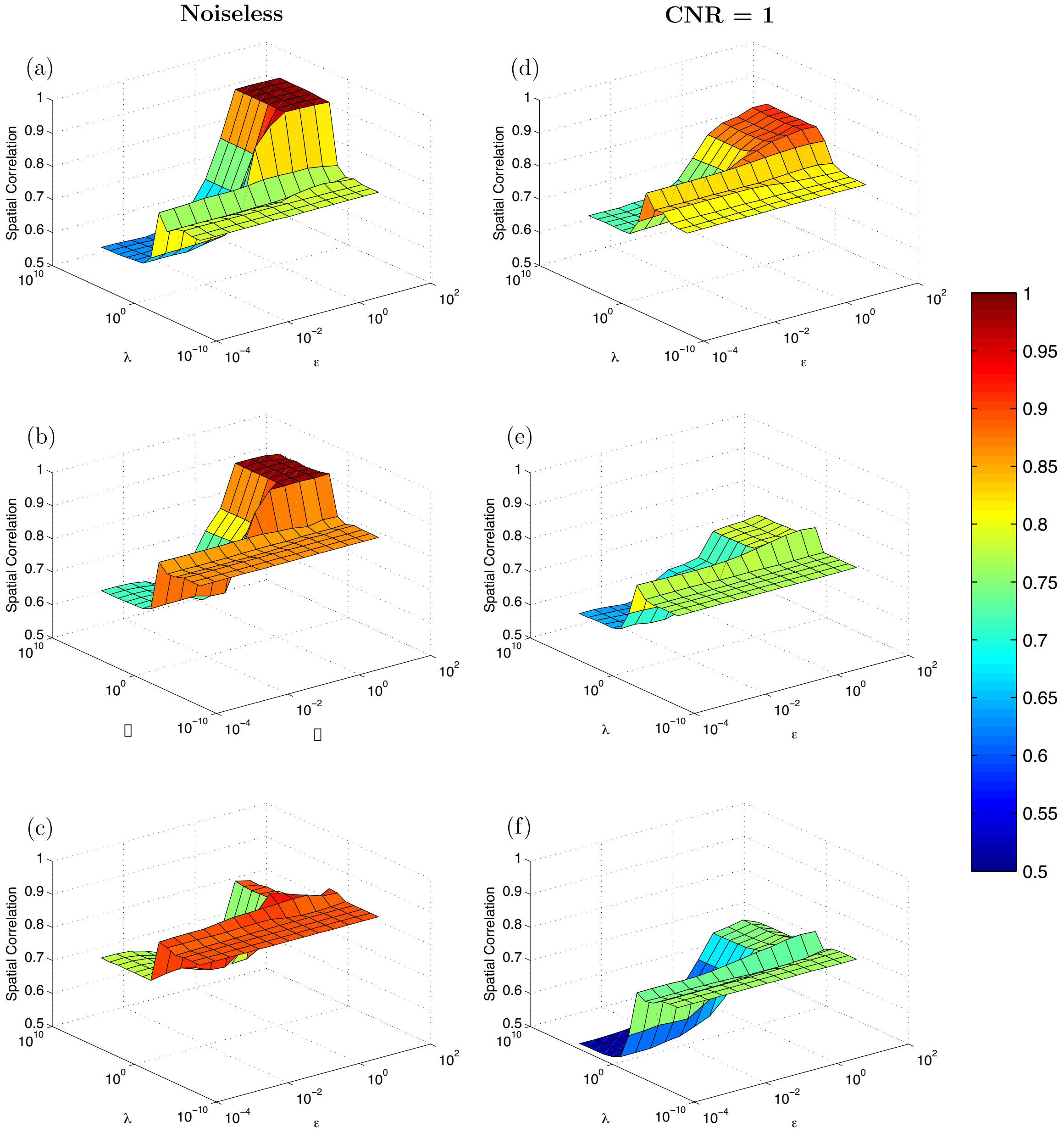


Figure 4: Spatial correlation of the true and the estimated sources as a function of $\lambda_n$ and $\epsilon_n$ for different CNR values: (a)-(c) is the noiseless case and (d)-(f) have a CNR of 1. Plots (a) and (d) are from scenario 1, all sparse sources. Plots (b) and (e) are from scenario 2, some sparse and some less sparse sources with no overlap. Plots (c) and (f) are from scenario 3, some sparse sources and some less sparse sources with some degree of overlap. The results are the average of 128 runs.

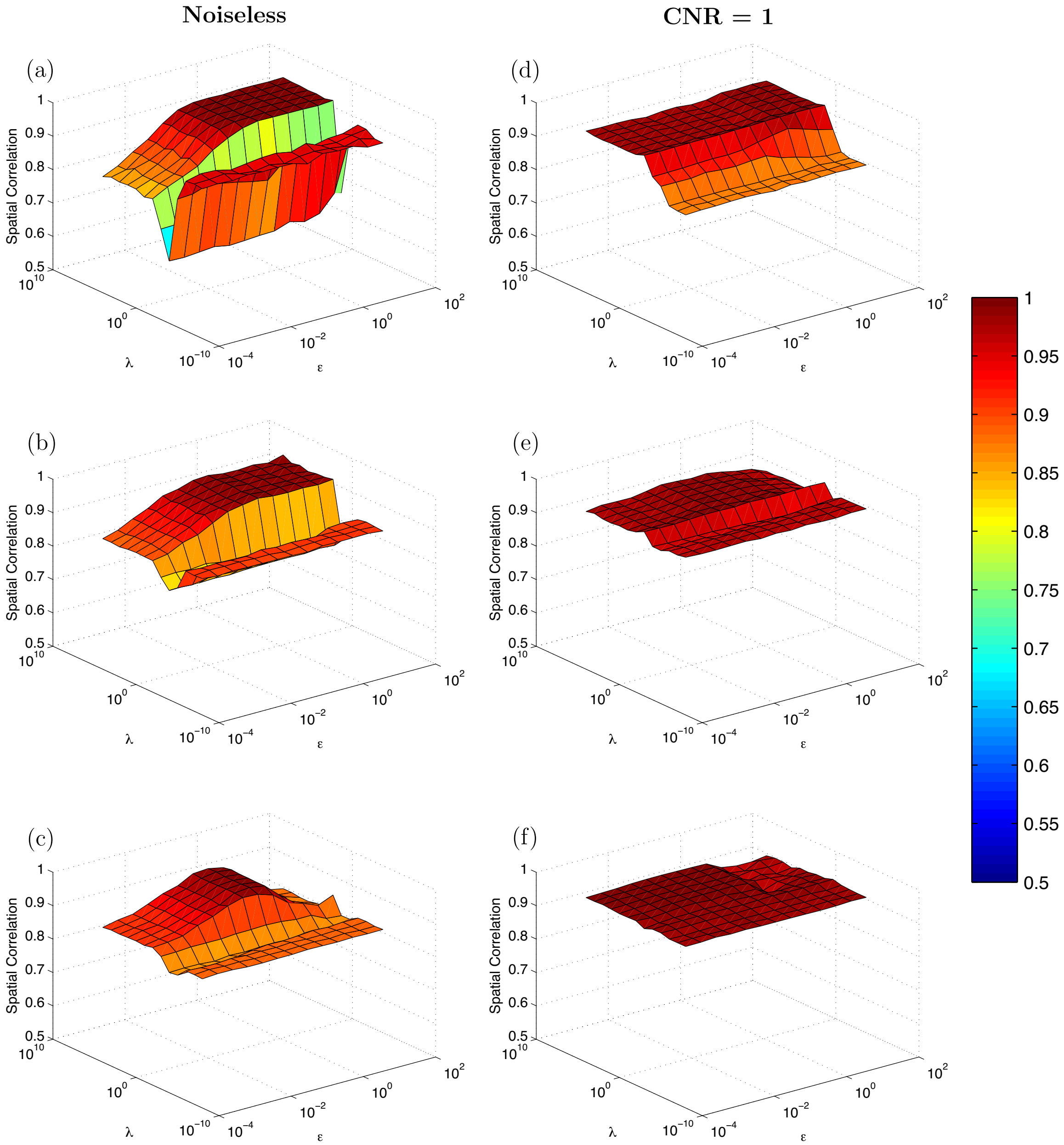


Figure 5: Spatial correlation of the estimated components generated when applying SparseICA-EBM on the two halves of the data as a function of $\lambda_n$ and $\epsilon_n$ for different CNR values: (a)-(c) is the noiseless case and (d)-(f) have a CNR of 1. Plots (a) and (d) are from scenario 1, all sparse sources. Plots (b) and (e) are from scenario 2, some sparse and some less sparse sources with no overlap. Plots (c) and (f) are from scenario 3, some sparse sources and some less sparse sources with some degree of overlap. The results are the average of 64 runs.