PLOS ONE

RESEARCH ARTICLE

# An algorithm for separation of mixed sparse and Gaussian sources

**Ameya Akkalkotkar**[1], **Kevin Scott Brown**[1,2,3,4]*

**1** Department of Chemical and Biomolecular Engineering, University of Connecticut, Storrs, CT, United States of America, **2** Department of Biomedical Engineering, University of Connecticut, Storrs, CT, United States of America, **3** Departments of Physics, and Marine Sciences, University of Connecticut, Storrs, CT, United States of America, **4** Institute for Systems Genomics and CT Institute for the Brain & Cognitive Sciences, Storrs, CT, United States of America

* kevin.s.brown@uconn.edu





**Citation:** Akkalkotkar A, Brown KS (2017) An algorithm for separation of mixed sparse and Gaussian sources. PLoS ONE 12(4): e0175775. https://doi.org/10.1371/journal.pone.0175775

**Editor:** Boris Podobnik, University of Rijeka, CROATIA

**Received:** October 7, 2016

**Accepted:** March 12, 2017

**Published:** April 17, 2017



**Data Availability Statement:** A python package (miprest) that implements the mixed PCA/ICA method described in this manuscript is publicly available at https://github.com/thelahunginjeet/miprest. The miprest package depends on author python packages maintained by the senior author; these will be automatically installed when the miprest package is installed. Installation instructions are contained in the README file distributed with miprest.

**Funding:** K.S.B. thanks the Office of the Vice President for Research at the University of Connecticut's Scholarship Facilitation Fund Award

## Abstract

Independent component analysis (ICA) is a ubiquitous method for decomposing complex signal mixtures into a small set of statistically independent source signals. However, in cases in which the signal mixture consists of both nongaussian and Gaussian sources, the Gaussian sources will not be recoverable by ICA and will pollute estimates of the nongaussian sources. Therefore, it is desirable to have methods for mixed ICA/PCA which can separate mixtures of Gaussian and nongaussian sources. For mixtures of purely Gaussian sources, principal component analysis (PCA) can provide a basis for the Gaussian subspace. We introduce a new method for mixed ICA/PCA which we call **M**ixed **I**CA/**P**CA via **Re**producibility **St**ability (MIPReSt). Our method uses a repeated estimations technique to rank sources by reproducibility, combined with decomposition of multiple subsamplings of the original data matrix. These multiple decompositions allow us to assess component stability as the size of the data matrix changes, which can be used to determinine the dimension of the nongaussian subspace in a mixture. We demonstrate the utility of MIPReSt for signal mixtures consisting of simulated sources and real-word (speech) sources, as well as mixture of unknown composition.

## Introduction

Trying to infer underlying source signals present in a complex signal mixture is a ubiquitous problem in signal processing with applications across science and engineering. The classic example is the so-called "cocktail party problem," in which the goal is to recover the voices of individuals speaking simultaneously using recordings from ambient microphones placed throughout the room [1]. In most cases, very little information about the underlying source signals is known; algorithms to attempt to solve this problem go under the heading of blind source separation [2]. Independent Component Analysis (ICA) is a blind source separation method that uses statistical independence of the sources as a criterion for solving the unmixing problem. The sources and mixing coefficients produced by ICA when multiplied together recover the data matrix. This is similar to the kind of matrix decomposition obtained via

(http://research.uconn.edu) for support of this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared no competing interests exist.

principal component analysis (PCA) [3], which is not surprising given that ICA reduces to PCA if the assumption of statistical independence of the sources is relaxed to the weaker condition of linear decorrelation. PCA is so widely used that is has been reinvented multiple times under different names (empirical orthogonal function analysis [4], the Karhunen-Loeve transform [5], and proper orthogonal decomposition [6]). ICA is used in many application domains [7, 8], particularly in neuroimaging, in which the goal is to decompose electroencephalographic (EEG) data in temporally independent sources [9] and functional magnetic resonance imaging (fMRI) data into spatially independent brain networks [10]. Two of the most common algorithms for ICA are maximum-likelihood [11] and minimization of the between-component mutal information [2]. Another frequently used neural network method (infomax [12, 13]) is equivalent to maximum-likelihood [14].

Unfortunately, once Gaussian sources are mixed with nongaussian sources ICA encounters problems. The unmixing matrix loses uniqueness because of the rotational invariance of the Gaussian subspace; with only nongaussian sources uniqueness is preserved [15]. Once two or more Gaussian sources are present in the signal mixture ICA can no longer separate those sources, and ignoring these sources in the ICA model will result in spurious sparse sources. This sphericity problem led Woods et al. [15] to propose a model for mixed ICA/PCA. They maximize an explicit likelihood model that incorporates supergaussian, subgaussian, and Gaussian sources and use cross-validation to determine the appropriate number of components of each kind. The method performs well but with a huge computational burden. Cross-validation alone is computationally expensive, and multiple likelihood maximizations are required for each model. A combinatorially large number of models must be evaluated, making this method difficult to use on the kinds of high-dimensional mixtures common in many application domains. Concerns about computational efficiency in ICA calculations have made FastICA [16], a fast fixed-point algorithm for ICA, an extremely popular method for source separation. Another of its strengths, relative to explicit likelihood maximization, is the fact that it can relatively easily separate mixtures of sources with both positive (subgaussian) and negative (supergaussian) kurtosis, without having to specify in advance how many of each are likely present.

RAICAR (Ranking and Averaging Independent Component Analysis by Reproducibility) [17] is an ICA method that uses repeated FastICA realizations to rank and select components based on their reproducibility, a measure of realization-to-realization consistency for a particular extracted source. It forms the basis for BICAR [18, 19], an ICA-based algorithm for multiresolution spatiotemporal data fusion. ICA decomposition of Gaussian mixtures produces purely spurious sparse components that do not stably converge as sample size increases. This property suggests that what reproducibility may be measuring is the degree to which a particular extracted component is part of a Gaussian subspace. This led Woods et al. [15] to speculate that a technique like RAICAR could be used to provide information for model selection in ICA. Unfortunately, as we will show, reproducibility alone is insufficient in determining how many Gaussian components are present in a complex signal mixture. However, component reproducibility along with a measure of reproducibility fluctuations across extractions from many random subsamples of the signal mixture matrix *can* identify the number of Gaussian and nongaussian sources in real mixtures.

In what follows, we describe a new algoritm for mixed ICA/PCA which we call MIPReSt: **M**ixed **I**CA/**P**CA via **Re**producibility **St**ability. Our method has RAICAR at its core, which allows it to take advantage of the speed and distributional flexibility of FastICA. While RAICAR itself requires multiple ICA runs, this number of decompositions does not grow with the size of the data matrix and FastICA is much faster than likelihood maximization. We demonstrate the performance of our algorithm on simulated mixtures of statistical sources, mixtures of real speech signals, and the famous Iris data of R.A. Fisher [20].

## Methods

### Algorithm

**RAICAR.** MIPReSt has at its core the RAICAR algorithm itself [17]; we use a modified version described previously [18, 19]. Briefly, in RAICAR the data matrix $X$ is subjected to a $K$-fold FastICA [16] decomposition; each of the $K$ FastICA realizations begins with random unmixing matrices (orthogonalized matrices of random Gaussian elements) and the same number of sources $N_s$ are extracted in each. Matrices of source-source correlation coefficients are produced for each of the $K(K-1)/2$ pairs of realizations, and components are grouped according to similarity across realizations. This re-sorts the estimated sources and mixing matrices from $K$ sets of size $N_s$ into $N_s$ sets of size $K$. The average inter-group cross-correlation among the $K$ sources in each of the $N_s$ groups (either over all sources [18] or thresholded [17]) produces a value $R$, the reproducibility, which can be used to rank sources in descending order of statistical robustness.

**MIPReSt.** In MIPReSt, the RAICAR algorithm becomes one step in a multi-step pipeline (see Fig 1). The key additional step is to perform RAICAR many times over many decimated versions of the original data. As discussed in the introduction, we expect the Gaussian subspace to be randomly oriented from subsample to subsample. We typically use multiple twofold decimations only; results are similiar for multiple decimations of increased order (twofold, fourfold, eightfold, etc.) or multiple fourfold decimations alone. From each RAICAR run, we obtain ranked reproducibility values for all sources. We use these to compute another quantity $\delta_{ij}$ that measures decimation-to-decimation variability in the reproducibility, computed as

$$\delta_{ij} = \left| R\left(S_j^i\right) - R\left(S_j^0\right) \right|. \tag{1}$$

Here, $i$ indexes realizations, with a superscript of 0 indicating reproducibility values obtained from a RAICAR decomposition applied to the parent (non-decimated) data matrix. $j$ indexes sources, of which there will be $N$ in all RAICAR runs.

Once we have used the decimation results to identify the dimension $N'$ of the sparse subspace, we project it out of the parent data matrix via

$$\tilde{X} = X - AS, \tag{2}$$

in which $A$ (size $N \times N'$) and $S$ (size $N' \times p$) are the portions of the mixing matrix and sources arising from the RAICAR decomposition of the *parent* data matrix. The matrix $\tilde{X}$ now consists of only Gaussian components. The dimension of this residual Gaussian subspace will be no more than $N - N'$. A basis for the Gaussian subspace can be obtained via singular value decomposition of $\tilde{X}$, as described below.

**Basis for the Gaussian subspace.** Estimating the dimension of the Gaussian subspace is equivalent to deciding how many principal components to retain when performing PCA on the data matrix [3]. There are a large number of proposed PCA stopping rules, both heuristic and statistical, which have been reviewed and compared elsewhere [21, 22]. No clear consensus yet exists as to which (if any) rule is superior, likely because the ability of a particular rule to stop at the correct number of true components depends on the correlation structure in the data and size of the data matrix [22]. We therefore compare dimensionality estimates from six different rules, described below. Unless otherwise noted we begin with a singular value decomposition of the sample covariance matrix $C = YY^T/(p-1)$, where $Y$ is a row-centered version of $\tilde{X}$ and $p$ is the column dimension of $\tilde{X}$. Eigenvalues $\lambda_1, \ldots, \lambda_n$ are assumed ordered from largest to smallest.

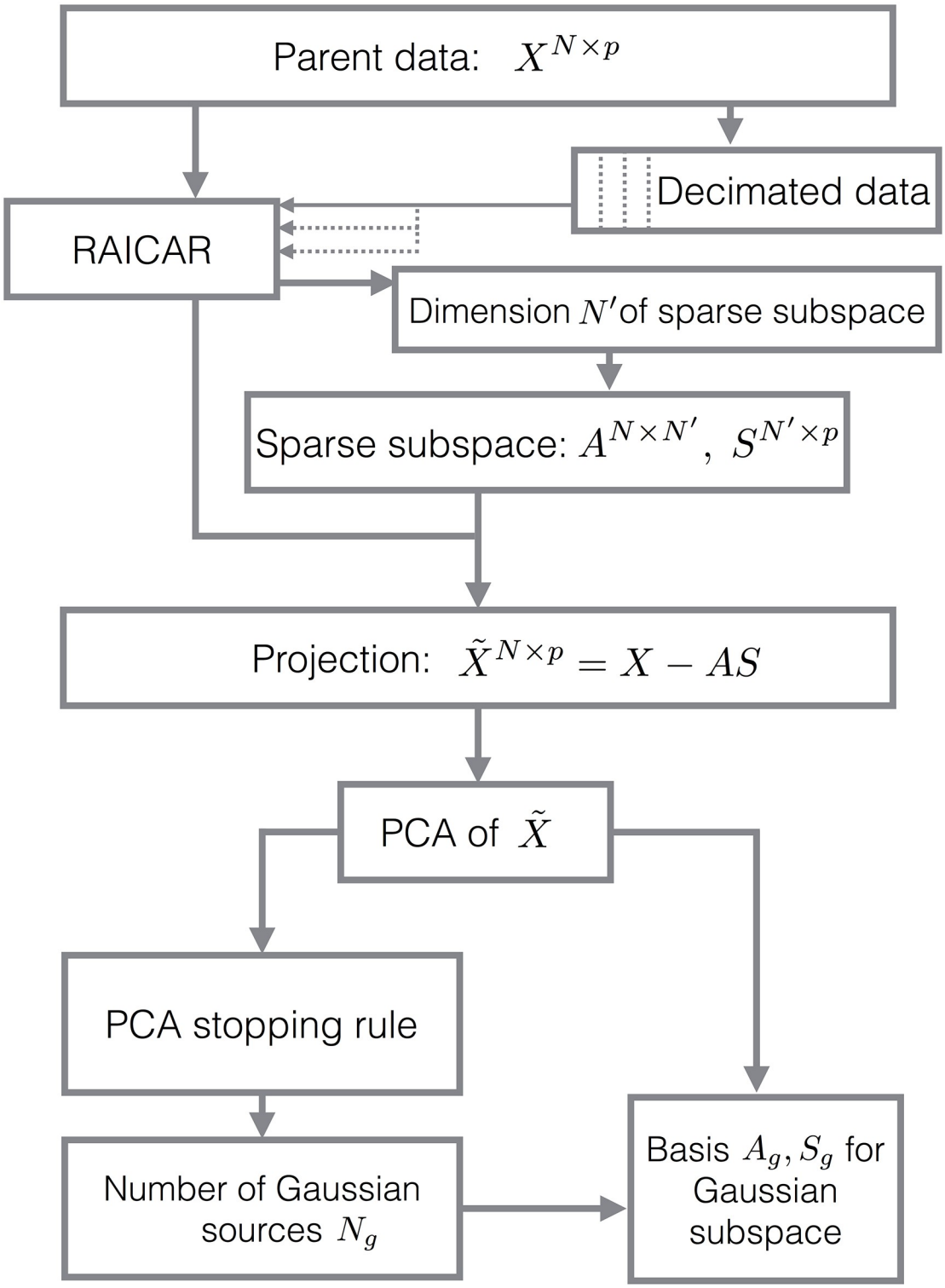


**Fig 1. Schematic for MIPReSt.** MIPReSt runs the RAICAR algorithm on both the original data matrix *X* and many random subsamples of smaller column dimension. Comparison of the reproducibilities from the original data and the random subsamples determines the size of the sparse subspace. After projecting that subspace out of *X*, singular value decomposition $\tilde{X}$, along with an eigenvalue selection rule, produces both the dimension of the Gaussian subspace and a basis for that subspace. (See Methods for details.).

https://doi.org/10.1371/journal.pone.0175775.g001

**Kaiser-Guttman (KG) Criterion** The Kaiser-Guttman selection rule [23] is one of the simplest and most widely-used rules, despite its known shortcomings [24]. To use Kaiser-Guttman, calculate $\bar{\lambda} = (1/n)\sum_i \lambda_i$ and keep all components for which $\lambda_i > \bar{\lambda}$.

**Joliffe-Modified KG** Joliffe has proposed a modification of the KG criterion which accounts for sampling variance [3]. In Joliffe's method, retain all components for which $\lambda_i > 0.7\bar{\lambda}$.

**Broken Stick Model** The broken stick model began as a resource distribution model in ecology [25], and was only later applied to eigenvalue selection in PCA by associating the resource to apportion with the total variance in the data [26, 27]. Broken stick partitions the unit interval into $n$ subintervals of random length, using $n - 1$ division points uniformly sampled in [0, 1]. If the subintervals are arranged in order of largest to smallest, then the expected value for the length of the $k^{\text{th}}$ subinterval is

$$l_k = \frac{1}{n}\sum_{i=k}^{n}\frac{1}{i}. \tag{3}$$

To use the broken stick model for eigenvalue selection, first transform $C$'s eigenvalues to $f_k = \lambda_k / \Sigma\lambda_k$ and then compare $f_k$ to the values $l_k$ from the broken stick distribution. Component $k$ is retained if $f_k$ is greater than $l_k$.

**Information Dimension** Information dimension is a heuristic measure of the number of "informative" modes in PCA. Full details and motivation can be found elsewhere [21]. Briefly, it begins by converting eigenvalues to "probabilities" via $p_k = \lambda_k/\Sigma_k \lambda_k$. These probabilities are then used to calculate a normalized entropy $\tilde{H} = -\sum_k p_k \log_2 p_k / \log_2 N$, where $N$ is the row (or column) dimension of the covariance matrix. The information dimension $n_0$ of the data is computed as $n_0 = N^{\tilde{H}}$.

**Parallel Analysis (PA)** Horn's parallel analysis criterion [28] compares the observed eigenvalues to the eigenvalues obtained from random matrices consisting of standard Gaussian random variables. First, standardize each variable in $\tilde{X}$ so $C$ becomes the correlation (and not covariance) matrix. Then generate $10^3$ matrices of the same dimensions as $\tilde{X}$ with $N(0, 1)$ entries. Obtain critical values using a predetermined significance level, and stop retaining components once the real data eigenvalues drop below the critical values. We use a 95% significance level to calculate critical values.

**Random Lambda** This method is a permutation test for each eigenvalue/component [29]. The values within $\tilde{X}$ are randomly shuffled 999 times and eigenvalues are computed each time. A permutation $p$-value is computed via $p = (n + 1)/1000$, where $n$ is equal to the number of times a random eigenvalue was larger than its corresponding data value. We then discard any components for which $p > 0.05$.

Once the dimension of the Gaussian subspace has been computed, a basis (set of sources) for the Gaussian subspace can be obtained by projection of the data matrix onto the subspace spanned by the retained eigenvectors. A python package for MIPReSt will be available on Github (https://github.com/thelahunginjeet); it depends on other packages which are also all available at the same location.

## Data

We used three types of data: simulated sources, speech signals extracted from public-domain audiobooks, and the famous Fisher's Iris data [20]. Gaussian sources were always sampled

**Table 1. Simulated sparse sources used in this study.**

| Name | Distribution | Type |
|---|---|---|
| Inverse Cosh | $\left(2\cosh\left(\frac{\pi x}{2}\right)\right)^{-1}$ | super |
| Laplace | $\frac{1}{\sqrt{2}}e^{-\sqrt{2}\lvert x\rvert}$ | super |
| Logistic | $\frac{\pi}{4\sqrt{3}}\,\mathrm{sech}^2\left(\frac{\pi x}{2\sqrt{3}}\right)$ | super |
| Exponential ArcSinh | $\frac{1}{\sqrt{2\pi\alpha^2(1+x/\alpha)^2}}\exp\left(-\mathrm{arcsinh}^2\left(\frac{x}{\alpha}\right)/2\right)$ | super |
| Double Cosh | $\frac{1}{\sqrt{e\pi}}e^{-x^2/2}\cosh\left(x\sqrt{2}\right)$ | sub |
| Exponential Sinh | $\sqrt{\frac{1+\sinh^2(x)}{2\pi}}\exp\left(-\frac{\sinh^2(x)}{2}\right)$ | sub |
| Generalized Gaussian | $\frac{\beta}{2\Gamma(1/\beta)}e^{-\lvert x\rvert^\beta}$ | super for $0<\beta<2$, sub for $\beta>2$ |

https://doi.org/10.1371/journal.pone.0175775.t001

from standard unit normal distributions, specifically

$$p(x) = \sqrt{\frac{1}{2\pi}}e^{-x^2/2}. \tag{4}$$

**Simulated sources.** We used several different distributions to generate subgaussian and supergaussian sources (see Table 1); some of these distributions have been used as test data previously [15]. All sparse sources were either generated from distributions with zero mean and unit variance or were standardized after construction.

**Speech.** We obtained five mp3s of public domain audiobooks from librivox.com [18]. The works used were "History of the Peloponnesian War", Book 5 by Thucydides; "Flatland" by Edwin A. Abbott; "The Adventures of Huckleberry Finn" by Mark Twain; The "Confessions" of St. Augustine; and "Moby Dick" by Herman Melville. These audiobooks were converted from stereo to single channel (mono) when appropriate, and then downsampled to 2.75 kHz. Examples of supergaussian and subgaussian distributions and a histogram of a random five second audiobook segment are shown in Fig 2.

**Iris data.** R.A. Fisher's famous Iris dataset [20] is available for download at the UCI Machine Learning Repository [30]. With the exception of ignoring the class labels in the data file no additional processing (beyond standard ICA preprocessing) of this data was peformed.

## Results

### Full rank extraction

As we discussed in the introduction, the Gaussian subspace in a mixture of sparse and Gaussian signals should randomly orient as the number of samples increases. To motivate the MIPReSt algorithm, we performed the following numerical experiment. We generated a five-dimensional signal mixture consisting of one supergaussian source (Inverse Cosh), one subgaussian source (Double Cosh), and three Gaussian sources (see Table 1 for these super- and subgaussian distributions). Each source consisted of $5 \times 10^5$ standardized iid samples, and we mixed them using a random $5 \times 5$ orthogonal matrix. We then randomly subsampled this parent signal mixture to obtain signal mixtues consisting of between $5 \times 10^3$ and $5 \times 10^5$ samples. We applied the RAICAR algorithm to each of the signal mixtures, and we calculated reproducibilities for each RAICAR source in every mixture. In order to match each of the five RAICAR sources to a unique known input source to which it was most similar, we solved the linear assignment problem [31] using Munkres' version of the Hungarian algorithm [32]. The cross

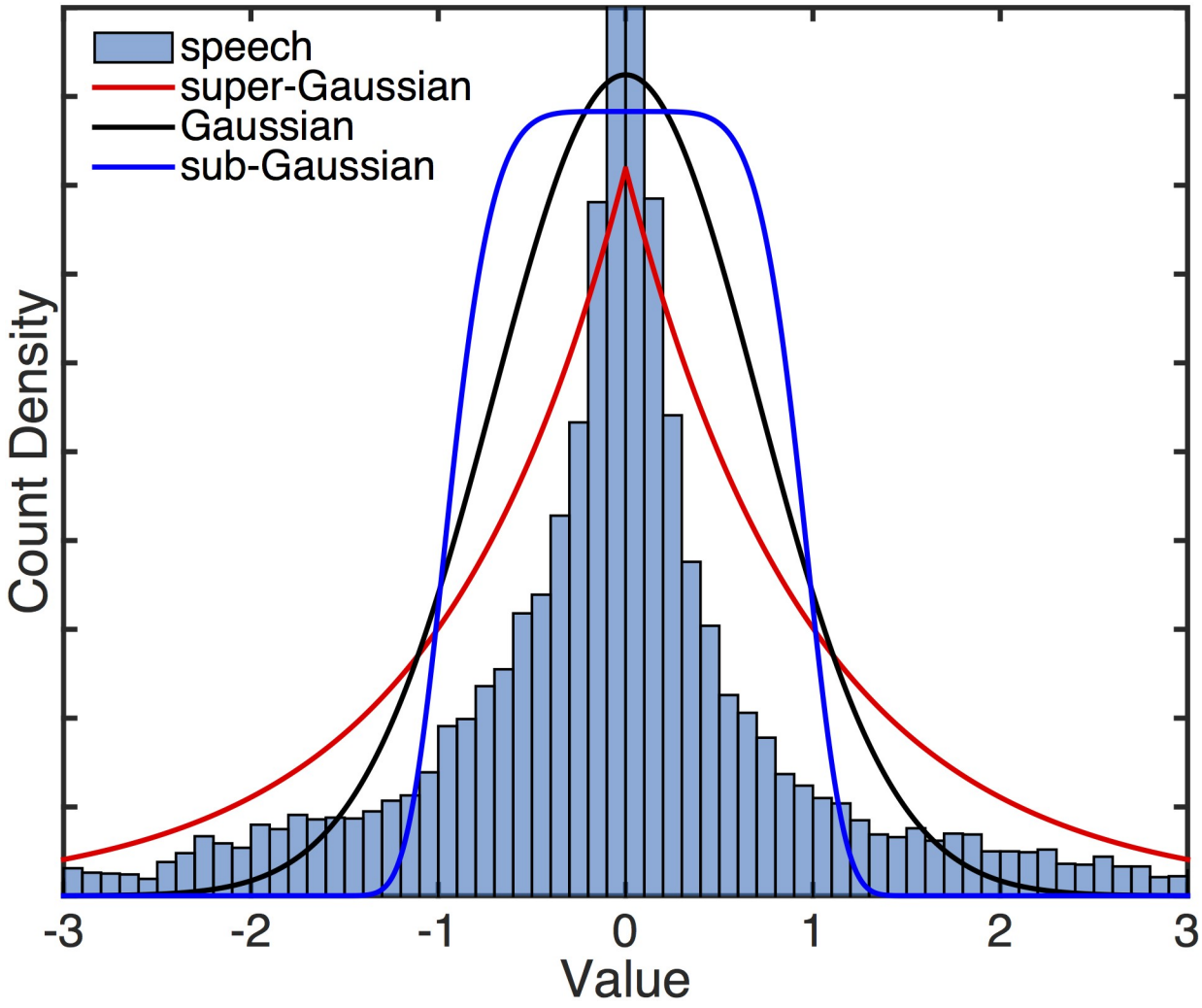


**Fig 2. Examples of super- and subgaussian sources.** Shown here are histograms for a Gaussian source (black), a subgaussian source (the generalized Gaussian), and a supergaussian source (Laplace). Also shown is a histogram for one of the speech signals used in this study. The speech signal is far more leptokurtic than the Laplace source; without truncating the *y*-axis the massive spike near zero of the speech signal obscures the shapes of the other distributions.

https://doi.org/10.1371/journal.pone.0175775.g002

correlations between the RAICAR sources and the known sources were used as the basis of the cost matrix for the assignment problem.

Fig 3 shows the results of these calculations. In each set of five reproducibilities, one set per decimated data set, the blue point corresponds to the best assignment match to the known supergaussian source, the red point corresponds to the best match to the subgaussian source, and all the Gaussian sources are shown in black. The inset shows the RAICAR sources from the parent data set of $5 \times 10^5$ samples plotted against their best assigment match. There are several things of note in this figure. First, RAICAR finds that the nongaussian subspace is highly, and usually perfectly, reproducible even at far more modest sample sizes than in the parent data. Secondly, the Gaussian subspace does orient randomly, as shown by the fluctuating Gaussian reproduciblities. Third, the Gaussians sometimes have extremely high reproducibility, which indicates that *reproducibility alone cannot discriminate*

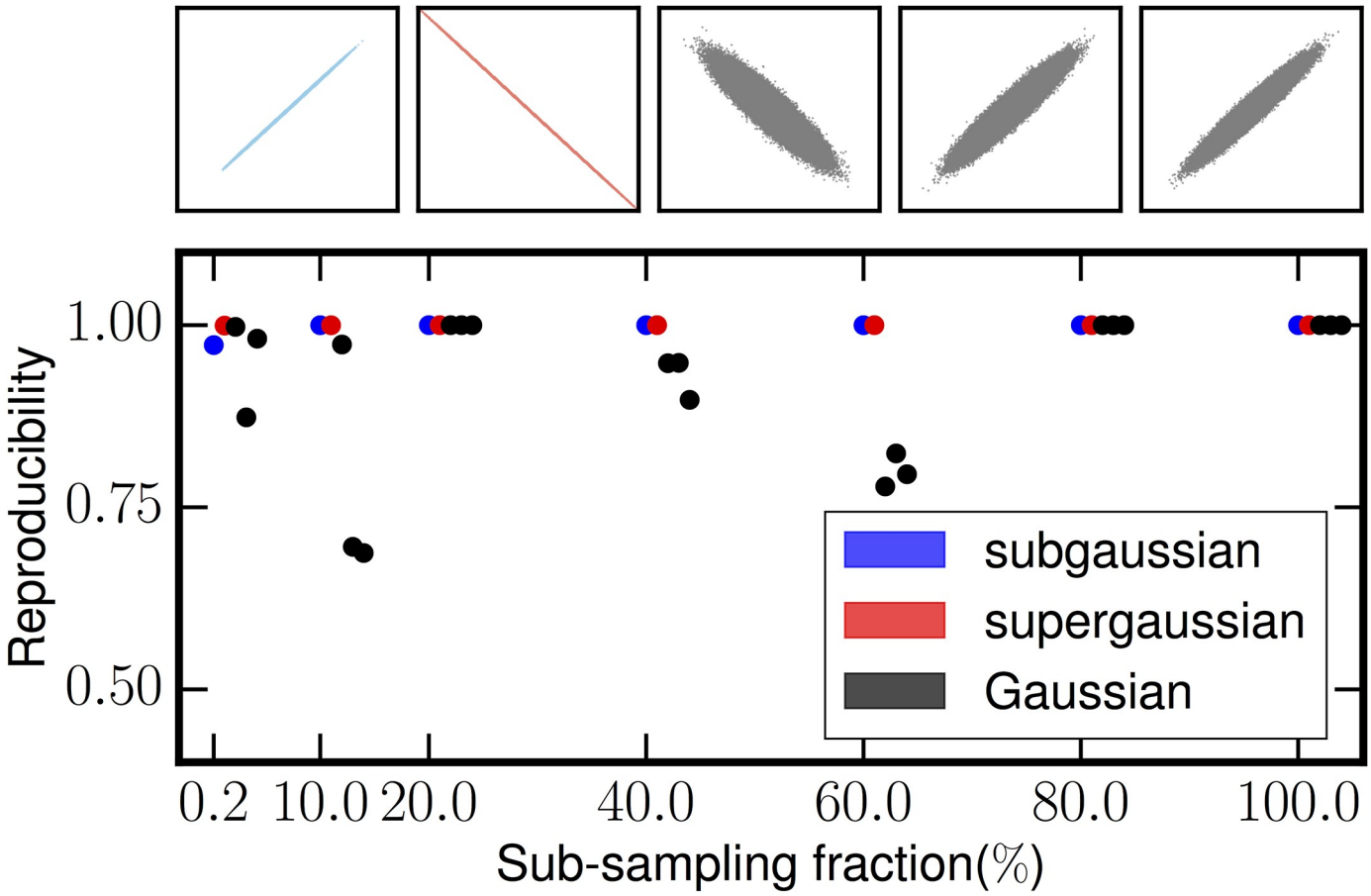


**Fig 3. Full rank extraction.** We constructed a simulated data matrix with five sources: one supergaussian, one subgaussian, and three Gaussian sources. The simulated data matrix had $5 \times 10^5$ samples. The main panel shows the results of RAICAR extractions at different levels of decimation, including the parent data. The best assignment match to the supergaussian source is shown in blue and to the subgaussian source in red. While the Gaussian sources may sometimes have extremely high reproducibility, they show poor stability when the data is decimated, in constrast to the sparse sources. The top panel shows scatter plots of the estimated sources from the parent data against their best assignment match; the sparse sources are recovered perfectly by RAICAR.

https://doi.org/10.1371/journal.pone.0175775.g003

*the Gaussian subspace from the nongaussian subspace*. These three observations motivate the core of the MIPReSt algorithm. We perform RAICAR on many subsampled versions of the input data. We monitor not only the distribution of reproducibilities $R$, but also the decimation-to-decimation variations in reproducibility $\delta_{ij}$ defined in Eq 1. True sparse sources will tend to have uniformly high $R$ and low $\delta_{ij}$, while Gaussian sources may sometimes have high $R$ but also larger $\delta_{ij}$.

## Overextraction

In most cases, the total signal dimension—the total number of true sources of any kind—is unknown and must be either estimated from the data matrix or guessed. We have previously found that repeated estimation methods like RAICAR and BICAR [18, 19] are relatively robust to overestimation of the data dimension. For example, if the signal mixture is seven dimensional but the number of true sources is five, extracting seven sources yields two sources (which we will call spurious) with extremely low reproducibility. This suggests a simple protocol when confronted with a real signal mixture: extract as many sources as possible, up to the

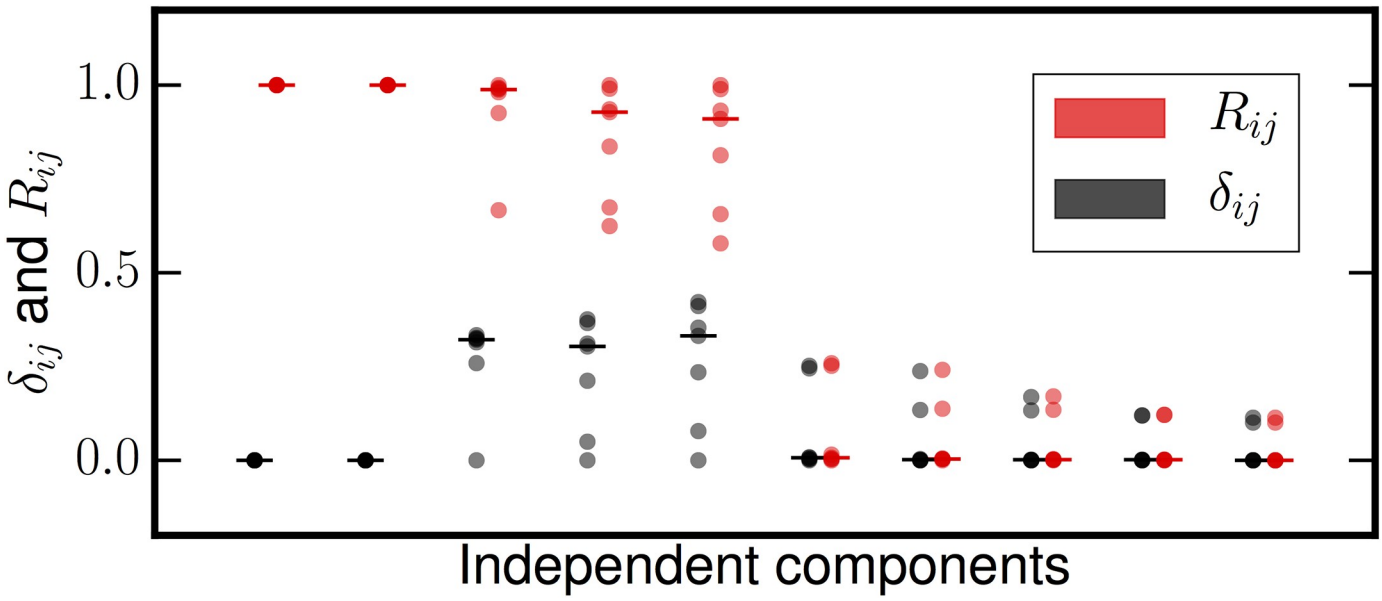


**Fig 4. Reproducibility ($R$) and reproducibility fluctuations ($\delta_{ij}$) from overextraction.** Only five sources (Gaussian or otherwise) are present, but the mixture dimension is ten. Horizontal bars are located at the median value. There are clearly three groups of sources here. Two sources (the recovered sparse sources) have near-perfect $R$ that does not fluctuate from decimation-to-decimation. Three sources have occasionally high reproducibility, but also significant $\delta_{ij}$; these are the Gaussian subspace. The remaining five sources have very low reproducibility that fluctuates very little; these sources are spurious sources resulting from overextraction.

https://doi.org/10.1371/journal.pone.0175775.g004

row dimension of the data matrix, and then use source reproducibility to estimate the total source content.

We performed a test to determine how MIPReSt performs for overextraction. Specifically, we wanted to know if spurious sources were clearly distinguishable in $R$, $\delta_{ij}$ space from both Gaussian and nongaussian sources. We therefore generated five input sources; two supergaussians (both Inverse Cosh) and three Gaussian sources, each of which had $5 \times 10^5$ samples. (Using two subgaussians or one subgaussian and one supergaussian yielded identical results.) These were then mixed with a $10 \times 5$ Gram-Schmidt orthogonalized mixing matrix to obtain a ten dimensional data matrix which consists of only five real sources, Gaussian or otherwise. We applied MIPReSt to this mixture; in each case we used an ensemble of fifteen random 2-fold subsampled data matrices.

Fig 4 shows the results of this experiment. Based on the $R$ and $\delta_{ij}$ values the extracted sources sort themselves into three categories: (*i*) signals with near unit reproducibility and zero delta, (*ii*) signals with high reproducibility but also high subsample-to-subsample fluctuations, and (*iii*) signals with very low reproducibiliy that does not fluctuate very much from subsample to subsample. Comparison of these three sets of sources to the known input sources by solving the assignment problem shows that the nongaussian sources are contained in the first group, all the Gaussian sources are in the second, and all spurious sources are completely unreproducible. In some subsamples, FastICA exhausts the variance in this data with fewer than five spurious sources; these missing spurious sources are assigned a reproducibility of zero. When we project out the recovered sparse sources and estimate the dimension of the Gaussian subspace, all PCA stopping rules arrive at the correct dimension of three (see Table 2). Based on this analysis, MIPReSt can recover the true sparse sources and the correct basis dimension for the Gaussian subspace even if the extraction dimension is larger than the true data dimension.

**Table 2. Results for estimated dimension of Gaussian subspaces.**

| N (matrix/Gaussian sources) | Overextraction | Speech | Fisher |
|---|---|---|---|
| | 10/3 | 15/5 | 4/? |
| KG | 3 | 5 | 3 |
| Joliffe KG | 3 | 5 | 3 |
| Broken Stick | 3 | 5 | 3 |
| Inf. Dim. | 2.93 | 4.98 | 2.01 |
| PA | 3 | 5 | 2 |
| Random Lambda | 3 | 5 | 1 |

https://doi.org/10.1371/journal.pone.0175775.t002

## Separation of complex, real-world signals

Next, we wanted to see if MIPReSt was still successful in recovering the dimension of the sparse subspace when that subspace consisted of sources with more realistic structure. We therefore continued to use simulated mixtures, but the nongaussian subspace was constructed from random sections of the public domain audiobooks described in Methods. Each nongaussian source consisted of $2 \times 10^4$ contiguous audio samples, starting at a random location. At 2.75 kHz (see Methods) this consists of 7.3 seconds of audio. Each speech source was standardized. The simulated data matrices we constructed consisted of five such speech sources and five Gaussian sources. These were overmixed using a Gram-Schmidt orthogonalized random mixing matrix to a fifteen dimensional data matrix. We used fifty two-fold subsampled data matrices for MIPReSt calculations of $R$ and $\delta_{ij}$.

Fig 5 shows the results of running MIPReSt on the overmixed speech examples. Again, as in Fig 4, one can see three distinct categories of sources, corresponding to the speech signals (high $R$, low $\delta_{ij}$), the five-dimensional Gaussian subspace (variable $R$ but high $\delta_{ij}$), and the five spurious sources resulting from overmixing (low or zero $R$ and low $\delta_{ij}$). This is the same pattern we saw in our previous experiments using simulated sparse sources. The sparse sources are near-perfectly reproducible and highly stable. The Gaussian components are occasionally

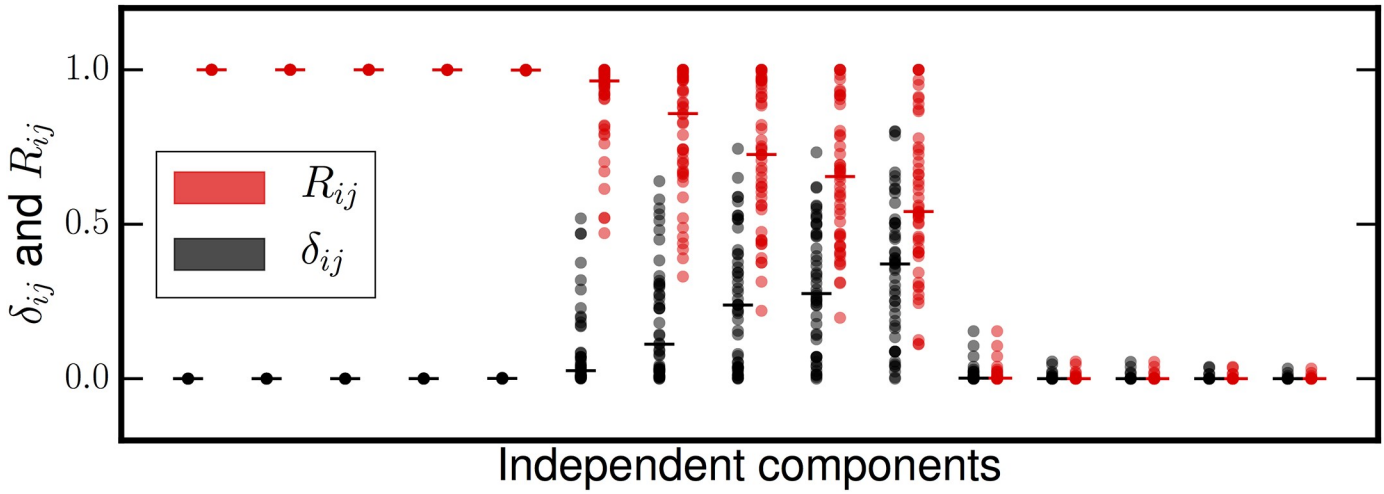


**Fig 5. Reproducibility ($R$) and reproducibility fluctuations ($\delta_{ij}$) for speech signals mixed with Gaussian sources.** For each of the fifteen extracted sources, $R$ is shown in red and $\delta_{ij}$ in black. For both quantities, values for each of the fifty subsampled data matrices are shown as points and the median value as a horizontal bar. The sources clearly group into three categories: high $R$ with low $\delta_{ij}$ (true sparse sources), variable $R$ with high $\delta_{ij}$ (Gaussian sources), and low $R$ and $\delta_{ij}$ (spurious sources).

https://doi.org/10.1371/journal.pone.0175775.g005

very reproducible, but as before quite unstable to decimation. Finally any spurious sources related to overextraction have almost no reproducibility whatsoever. We performed multiple iterations of this experiment and the results were consistent every time. Contaminating each speech signal with Gaussian noise of varying signal-to-noise ratio had no effect on estimation of any of the subspace dimensions. This is expected; the added noise had identical characteristics to the signals making up the Gaussian subspace and hence caused no difficulties in extraction. As before, all PCA stopping rules agreed that the Gaussian subspace had a dimension of five.

## Fisher's Iris data

Finally, we examined the performance of MIPReSt on Fisher's famous Iris data set [20], originally introduced in the manuscript in which Fisher developed the linear discriminant. Fisher's Iris data is probably the most famous classification dataset in existence; a partial list of papers which cite the Iris data, maintained on the UCI Machine learning Repository [30], contains over 200 papers. The Iris data has 150 samples measured on four dimensions (features): petal length, petal width, sepal length, and sepal width. In order to use MIPReSt on the Iris data, we subjected it to eighty random two-fold decimations. We emphasize that this data could be quite challenging for our method, as it has far fewer samples (by two or more orders of magnitude) than the test data we have considered previously.

Fig 6 shows that there is one and only one sparse component present in the data. When we examine the RAICAR sources from the parent data in Fig 7 the nongaussianity of the high $R_{ij}$, low $\delta_{ij}$ source is obvious (far left panel in the figure). We note that our estimate of a single sparse source in the Iris data agrees with that obtained by the algorithm of Woods et al. [15]. The fourth coulumn of Table 2 shows the estimates for dimension of the residual (after

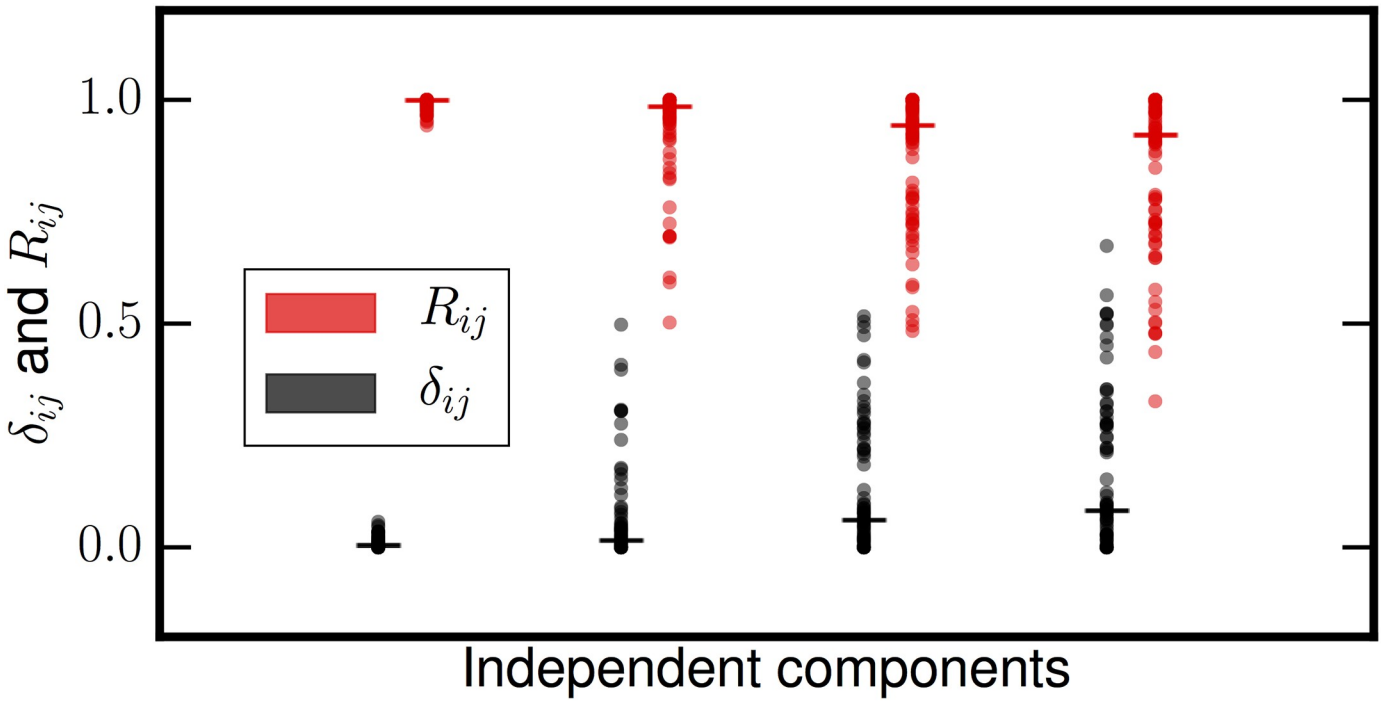


**Fig 6. Reproducibility plot for the Iris data.** The format and color scheme for this figure is identical to that used in Figs 4 and 5. Based on this information and related discussion in the text, it appears that there is one (and likely only one) sparse source present in the iris data.

https://doi.org/10.1371/journal.pone.0175775.g006

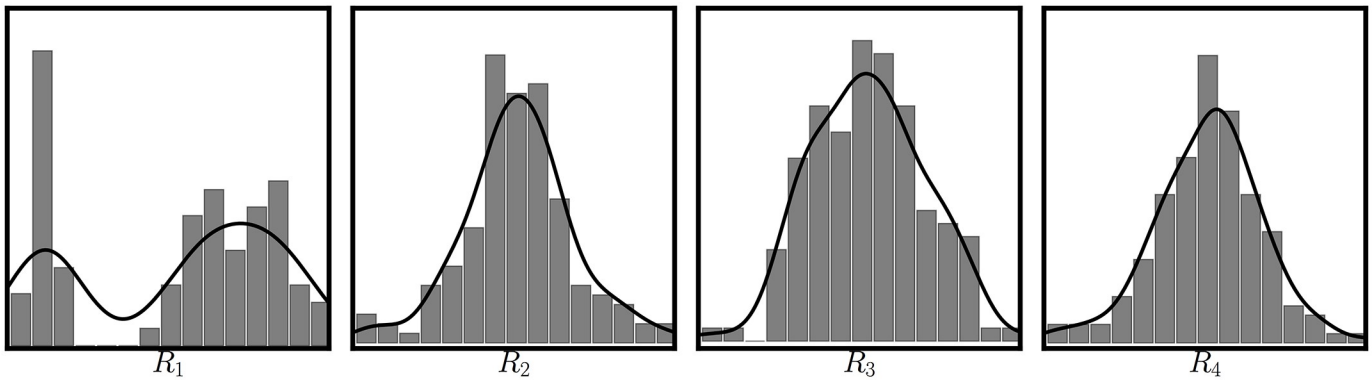


**Fig 7. Histograms of extracted sources from the Iris data.** Each panel shows a histogram (bars) and kernel density estimate (Gaussian kernel, solid line) for one of the four RAICAR sources extracted from the iris data. The nongaussianity of the most reproducible source (upper left) is clearly evident.

https://doi.org/10.1371/journal.pone.0175775.g007

projection) Gaussian subspace. Here, there is less of a consensus than in the simulated mixture cases. The stopping rules indicate there are anywhere between 1 and 3 Gaussian sources present. Our reasons for including the Kaiser-Guttman criterion (and its modification by Joliffe) are its simplicity, speed, and wide use by practitioners. It has, however, been roundly criticized [24]. In a detailed simulation study, Peres-Neto and colleagues give high marks to PA and Random Lambda (along with four other rules we did not consider) [22]; these methods tend to produce the correct number of relevant components for the widest variety of correlation structures in the data. In addition, in a study of several stopping rules applied to microarray data [21] find similar disagreement and recommend a consensus approach based on multiple rules. If we exclude KG and Joliffe's KG and simply average the results from the remainder of the rules, we obtain a Gaussian subspace dimension of two.

We therefore find that the Iris data, despite having a potentially problematic number of samples, did not pose a significant challenge to MIPReSt. We are able to unambiguously find only a single sparse source, a result that agrees with a previous mixed ICA/PCA method that requires a full likelihood model for the data [15]. These results, along with our decompositions of the audiobook speech signals above, strongly indicate MIPReSt will be a valuable algorithm for a variety of real-world data.

## Discussion

We have presented MIPReSt, a new algorithm for mixed ICA/PCA and demonstrated its utility for both simulated mixtures and empirical data. MIPReSt performs many repeated ICA realizations on both the original, parent data matrix $X$ as well as a number of derived data matrices obtained from $X$ via randomly dropping some fraction of the samples in $X$. Using a combination of component reproducibility and a measure of subsample-to-subsample fluctuations in reproducibility, we are clearly able to separate a complex mixture into sparse and Gaussian subspaces, as well as flag potentially spurious sources resulting from extraction of more sources than are actually present in the data. Even on data matrices with an extremely limited number of samples (150 for the Iris data), MIPReSt still obtains results consistent with other algorithms which are more heavily parameterized and much more computationally expensive [15]. In addition, MIPReSt's use of FastICA allows it to recover both supergaussian and subgaussian sources without any need to specify the relative numbers of each.

As currently stands, MIPReSt uses a very basic version of FastICA. A single nonlinearity (logcosh) was used for all the data matrices we decomposed, and for every extracted source

within those data matrices. Given that the sparse subspaces of real mixtures may be composed of sources with a variety of shapes, the use of a single nonlinearity may be questionable. Precisely these concerns have led others to develop a "reloaded" deflationary FastICA method that tries to adaptively find the optimal nonlinearity for each extracted source as the algorithm progresses [33]. This method showed improved performance over traditional FastICA when sources with varying distributional shapes were mixed together. It would be interesting and valuable to compare the performance of MIPReSt with this adaptive FastICA method to what we have used here. However, we should note that while gains could be made, they are not likely to be dramatic, given the performance of MIPReSt in this study. Using multiple estimations, we were able to recover non-Gaussian subspaces to high accuracy even when they consisted of a mixture of both super-Gaussian and sub-Gaussian components.

When dealing with datasets of much larger dimension, say those typical in ICA analyses of electroencephalographic [9] or functional magnetic resonance imaging [10] data, some sparse components may have reproducibility further from unity than we see here. This could cause a potential problem, since then the RAICAR averaged mixing matrix columns corresponding to these sources deviate from orthogonality. In these cases, it should be possible to either correct the mixing matrix to orthogonality via Gram Schmidt, or more simply use a single ICA realization on the parent data $X$ in which we identify the true sparse sources and corresponding mixing matrix columns solving the assignment problem between the RAICAR/MIPReSt sources and the single-run ICA sources. As above, the cross-correlation matrix between the two sets of sources gives the cost matrix for the assignment problem.

Our method for estimation of the dimension of the Gaussian subspace relies on using one or more PCA stopping rules, an area in which there is guidance but not very much consensus [21, 22, 24]. We find very consistent results for dimension estimation in simulated mixtures, even when those mixtures contain real-world sources (speech). On Fisher's Iris data, the results are less clear. The best approach would be to obtain a dimension estimate from a combination of stopping rules [21] that have proven to work well under a variety of correlation structures in the data [22]. However, we should point out here that for the applications considered here, obtaining the exact dimension for the Gaussian subspace is not really a concern. All we can recover is a basis for the Gaussian subspace, not the individual sources which compose it (which is impossible). In this case, it may actually be desirable to underestimate the dimension of the Gaussian subspace in order to obtain some amount of data compression.

In other cases, more careful evaluation of eigenvalue selection criteria will be warranted. If the Gaussian subspace were to consist of signals with non-identical power spectra—for example AR/ARMA models with nonidentical coefficients—then by using algorithms like SOBI [34] or AMUSE [35] we should be able to recover the constituent Gaussian processes and not just a basis for the subspace. In this case, estimation of the dimension of the residual data matrix $\tilde{X}$ becomes much more important, and a comprehensive study of the performance of eigenvalue selection algorithms for simulated mixtures of sparse and Gaussian sources will be necessary. We are currently working on a version of MIPReSt tailored to unmixing of time series and investigating this issue.

It has recently become much easier to collect EEG data from within an MRI scanner [36, 37], leading to the possibility of combining EEG and fMRI data during a cognitive task to obtain a single view of human brain activity with simultaneously high spatial and temporal resolution. Many methods have been proposed for this problem [38–46], and use of ICA as some part of the analysis or processing pipeline is a feature of many of these methods [18, 19, 36, 47–51]. However, none of these methods deal with the problem of nongaussian components in the data and the possible contamination of sparse sources during the ICA steps. We are

currently working to integrate MIPReSt into BICAR, an ICA-based method for producing reproducible joint components from concurrent EEG-fMRI data [18, 19]. This should produce fewer spurious joint maps, and enhance the interpretability of the real ones.

## Acknowledgments

We thank members of the Brown research group for helpful discussions.

## Author Contributions

**Conceptualization:** AA KSB.

**Data curation:** AA KSB.

**Formal analysis:** AA KSB.

**Funding acquisition:** KSB.

**Investigation:** AA KSB.

**Methodology:** AA KSB.

**Project administration:** KSB.

**Software:** AA KSB.

**Supervision:** KSB.

**Validation:** AA KSB.

**Visualization:** AA KSB.

**Writing – original draft:** AA KSB.

**Writing – review & editing:** AA KSB.

## References

1. Jutten C, Herault J. Blind separation of sources, Part 1: an adaptive algorithm based on neuromimetic architecture. Signal Process. 1991; 24:1–10. https://doi.org/10.1016/0165-1684(91)90079-X
2. Comon P, Jutten C, editors. Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press;.
3. Joliffe IT. Principal Component Analysis. Springer; 2002. https://doi.org/10.1002/0470013192.bsa501
4. Lorenz EN. Empirical orthogonal functions and statistical weather prediction. Statistical Forecasting Project, Massachusetts Institute of Technology Department of Meterology; 1956. 1.
5. Gerbrands JJ. On the relationships between SVD, KLT and PCA. Pattern Recognit. 1981; 14:375–381. https://doi.org/10.1016/0031-3203(81)90082-0
6. Aubry N. On the hidden beauty of the proper orthogonal decomposition. Theor Comput Fluid Dyn. 1991; 2:339–352. https://doi.org/10.1007/BF00271473
7. Aires F, Rossow WB, Chédin A. Rotation of EOFs by the Independent Component Analysis: Toward a Solution of the Mixing Problem in the Decomposition of Geophysical Time Series. J Atmos Sci. 2002; 59:111–123. https://doi.org/10.1175/1520-0469(2002)059%3C0111:ROEBTI%3E2.0.CO;2
8. Baccigalupi C, Bedini L, Burigana C, Zotti GD, Farusi A, Maino D, et al. Neural networks and the separation of cosmic microwave background and astrophysical signals in sky maps. Mon Not R Astron Soc. 2000 Nov; 318(3):769–780. https://doi.org/10.1046/j.1365-8711.2000.03751.x
9. Makeig S, Bell A, Jung T, Sejnowski T. Independent component analysis of electroencephalographic data. In: Advances in neural information processing systems 8; 1996. p. 7.
10. McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, et al. Analysis of fMRI Data by Blind Separation Into Independent Spatial Components. Hum Brain Mapp. 1998; 6:160–188. https://doi.org/10.1002/(SICI)1097-0193(1998)6:3%3C160::AID-HBM5%3E3.3.CO;2-R PMID: 9673671

11. Pham DT, Garrat P. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans Signal Process. 1997; 45:1712–1725.
12. Bell A, Sejnowski T. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 1995; 7:1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129 PMID: 7584893
13. Nadal JP, Parga N. Non-linear neurons in the low noise limit: a factorial code maximizes information transfer. Network. 1994; 5:565–581.
14. Cardoso JF. Infomax and maximum likelihood for source separation. IEEE Lett Signal Process. 1997; 4:112–114. https://doi.org/10.1109/97.566704
15. Woods R, Hansen L, Strother S. How many separable sources? Model selection in Independent Components Analysis. PLoS One. 2015; 10:e0118877. https://doi.org/10.1371/journal.pone.0118877 PMID: 25811988
16. Hyvärinen A, Oja E. A fast fixed-point algorithm for independent component analysis. Neural Comput. 1997; 9(7):1483–1492.
17. Yang Z, Laconte S, Weng X, Hu X. Ranking and averaging independent component analysis by reproducibility (RAICAR). Hum Brain Mapp. 2008; 29(6):711–725. https://doi.org/10.1002/hbm.20432 PMID: 17598162
18. Brown K, Grafton S, Carlson J. BICAR: A new algorithm for multiresolution spatiotemporal data fusion. PLoS One. 2012; 7:e50268. https://doi.org/10.1371/journal.pone.0050268 PMID: 23209693
19. Brown KS, Kasper R, Giesbrecht B, Carlson JM, Grafton ST. Reproducible paired components from concurrent EEG-fMRI data using BICAR. J Neurosci Meth. 2013; 219:205–219. https://doi.org/10.1016/j.jneumeth.2013.07.012 PMID: 23933055
20. Fisher R. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7:178–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
21. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biol Direct. 2007; 2:1–21. https://doi.org/10.1186/1745-6150-2-2 PMID: 17229320
22. Peres-Neto P, Jackson D, Somers K. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal. 2005; 49:974–977. https://doi.org/10.1016/j.csda.2004.06.015
23. Guttman L. Some necessary conditions for common factor analysis. Psychometrika. 1954; 19:149–161. https://doi.org/10.1007/BF02289162
24. Jackson D. Stopping rules in principal component analysis: a comparison of heuristic and statistical approaches. Ecology. 1993; 74:2204–2214. https://doi.org/10.2307/1939574
25. MacArthur R. On the relative abundance of bird species. Proc Natl Acad Sci USA. 1957; 43:293–295. https://doi.org/10.1073/pnas.43.3.293 PMID: 16590018
26. Fontier S. Étude de la décroissance des values propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. Biol Ecol. 1976; 25:67–75.
27. Legendre P, Legendre L. Numerical Ecology. Elsevier Science BV; 1998.
28. Horn J. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;p. 178–185. https://doi.org/10.1007/BF02289447 PMID: 14306381
29. ter Braak C. CANOCO—a Fortan program fr canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis, and redundancy analysis (version 2.1). Agricultural Mathematic Group, Wageningen; 1988. LWA-88-02.
30. Lichman M. UCI Machine Learning Repository; 2013. Available from: http://archive.ics.uci.edu/ml.
31. Burkard R, Dell'Amico M, Martello S. Assignment Problems. SIAM; 2009. https://doi.org/10.1137/1.9780898717754
32. Munkres J. Algorithms for the assignment and transportation problems. SIAM J Appl Math. 1957; 5:32–38. https://doi.org/10.1137/0105003
33. Miettinen J, Nordhausen K, Oja H, Taskinen S. Deflation-based FastICA with adaptive choices of nonlinearities. IEEE Trans Signal Process. 2014; 62(21):5716–5724. https://doi.org/10.1109/TSP.2014.2356442
34. Belouchrani A, Abed-Meraim K, Cardoso JF. A Blind Source Separation Technique Using Second-Order Statistics. IEEE Trans Signal Process. 1997; 45:434–444. https://doi.org/10.1109/78.554307
35. Tong L, Soon V, Huang Y, Liu R. AMUSE: A New Blind Identification Algorithm. In: Circuits and Systems, 1990., IEEE International Symposium on; 1990. p. 1784–1787 vol. 3.
36. Debener S, Ullsperger M, Siegel M, Engel AK. Single-trial EEG-fMRI reveals the dynamics of cognitive function. Trends Cogn Sci. 2006; 10(12):558–563. https://doi.org/10.1016/j.tics.2006.09.010 PMID: 17074530

37. Rosenkranz K, Lemieux L. Present and future of simultaneous EEG-fMRI. MAGMA. 2010; 23:309–316. https://doi.org/10.1007/s10334-009-0196-9 PMID: 20101434

38. Grouiller F, Thornton RC, Groening K, Spinelli L, Duncan JS, Schaller K, et al. With or without spikes: localization of focal epileptic activity by simultaneous electroencephalography and functional magnetic resonance imaging. Brain. 2011; 134:2867–2886. https://doi.org/10.1093/brain/awr156 PMID: 21752790

39. Yuah H, Zotev V, Phillips R, Drevets WC, Bodurka J. Spatiotemporal dynamics of the brain at rest—exploring EEG microstates as electrophysiological signatures of BOLD resting state networks. Neuroimage. 2012; 60:2062–2072. https://doi.org/10.1016/j.neuroimage.2012.02.031 PMID: 22381593

40. Babiloni F, Babiloni C, Carducci F, Del Gratta C, Rossini P, Cincotti F. Cortical source estimate of combined high resolution EEG and fMRI data related to voluntary movements. Methods Inf Med. 2002; 41:443–450. PMID: 12501818

41. Liu Z, Kecman F, He B. Effects of fMRI-EEG mismatches in cortical current density estimation. Clin Neurophysiol. 2006; 117:1610–1622. https://doi.org/10.1016/j.clinph.2006.03.031 PMID: 16765085

42. Bojak I, Oostendorp TF, Reid AT, Kötter R. Connecting mean field models of neural activity to EEG and fMRI data. Brain Topogr. 2010; 23:139–149. https://doi.org/10.1007/s10548-010-0140-3 PMID: 20364434

43. Schultze-Kraft M, Becker R, Breakspear M, Ritter P. Exploiting the potential of three dimensional spatial wavelet analysis to explore the nesting of temporal oscillations and spatial variance in simultaneous EEG-fMRI data. Prog Biophys Mol Bio. 2011; 105:67–79. https://doi.org/10.1016/j.pbiomolbio.2010.11.003 PMID: 21094179

44. Sato JR, Rondioni C, Sturzbecher M, de Araujo DB, Amaro E. From EEG to BOLD: brain mapping and estimating transfer functions in simultaneous EEG-fMRI acquisitions. Neuroimage. 2010; 50:1416–1426. https://doi.org/10.1016/j.neuroimage.2010.01.075 PMID: 20116435

45. Ostwald D, Porcaro C, Bagshaw AP. An information theoretic approach to EEG-fMRI integration of visually evoked responses. Neuroimage. 2010; 49:498–516. https://doi.org/10.1016/j.neuroimage.2009.07.038 PMID: 19632339

46. Daunizeau J, Grova C, Marrelec G, Mattout J, Jbabdi S, Pélégrini-Issac M, et al. Symmetrical event-related EEG/fMRI information fusion in a variational Bayesian framework. Neuroimage. 2007; 36:69–87. https://doi.org/10.1016/j.neuroimage.2007.01.044 PMID: 17408972

47. Eichele T, Calhoun V, Moosmann M, Specht K. Unmixing concurrent EEG-fMRI with parallel independent component analysis. Int J Pyschophysiol. 2008; 67:222–234. https://doi.org/10.1016/j.ijpsycho.2007.04.010 PMID: 17688963

48. Eichele T, Calhoun VD, Debener S. Mining EEG–fMRI using independent component analysis. Int J Pyschophysiol. 2009;.

49. Moosmann M, Eichele T, Nordby H, Hugdahl K. Joint independent component analysis for simultaneous EEG–fMRI: Principle and simulation. Int J Pyschophysiol. 2008; 67:212–221.

50. Brown KS, Ortigue S, Grafton ST, Carlson JM. Improving human brain mapping via joint inversion of brain electrodynamics and the BOLD signal. Neuroimage. 2010 Feb; 49(3):2401–2415. https://doi.org/10.1016/j.neuroimage.2009.10.011 PMID: 19833215

51. Brookings T, Ortigue S, Grafton S, Carlson J. Using ICA and realistic BOLD models to obtain joint EEG/fMRI solutions to the problem of source localization. Neuroimage. 2009; 44:411–420. https://doi.org/10.1016/j.neuroimage.2008.08.043 PMID: 18845263