14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SPARSE SOURCES ARE SEPARATED SOURCES

Scott Rickard

Sparse Signal Processing Group
University College Dublin
Dublin, Ireland
scott.rickard@ucd.ie

ABSTRACT

Sparse representations are being used to solve problems previously thought insolvable. For example, we can separate more sources than sensors using an appropriate transformation of the mixtures into a domain where the sources are sparse. But what do we mean by sparse? What attributes should a sparse measure have? And how can we use this sparsity to separate sources? We investigate these questions and, as a result, conclude that sparse sources are separated sources, as long as you use the correct measure.

1. INTRODUCTION

Sparse signal representations lead to efficient and robust methods for compression, detection, sensing, denoising, and signal separation [1–3]. However, there is no standard practical measure of sparsity that is universally accepted. In a strict sense, sparsity means that most signal components are zero. In a practical sense, sparsity means that most signal components are relatively small. Indeed, it is probably appropriate that the definition of sparsity be application specific. In this paper we examine sparsity measures for source separation algorithms.

Sparsity has garnered much interest in the blind source separation community. In this domain, the goal is, given M-by-L matrix $\mathbf{Y}$ of the form

$$\mathbf{Y} = \mathbf{AX} + \mathbf{N} \qquad (1)$$

determine M-by-N matrix $\mathbf{A}$ and N-by-L matrix $\mathbf{X}$ which minimize

$$\|\mathbf{Y} - \mathbf{AX}\|_F + \lambda\|\mathbf{X}\|_G + \mu\|\mathbf{A}\|_H \qquad (2)$$

for matrix cost functions ($\|\cdot\|_F$, $\|\cdot\|_G$, $\|\cdot\|_H$) and regularization parameters (λ, μ). The interpretation for source separation problems is that $\mathbf{Y}$ is L observations of M mixtures, $\mathbf{A}$ is a mixing matrix, $\mathbf{X}$ is L observations of N sources and M-by-L matrix $\mathbf{N}$ is noise. Thus the goal is given the mixtures, to determine the sources. When $M < N$, the system of equations is underdetermined and the purpose of the matrix cost functions and regularization parameters is to select one solution from the infinite number of possible solutions. The sources (and/or mixing matrix) have certain desirable properties that we know a priori and this helps us to select one of the solutions. For example, we often prefer solutions with sparse representations because the original signals themselves have sparse representations.

For many signal processing applications, we do not care about the sparsity of $\mathbf{A}$ and thus $\mu = 0$. We do care about the sparsity of $\mathbf{X}$, however, and setting $\lambda > 0$, $\|\mathbf{X}\|_G$ is used to force the solution to be sparse. Often, $\|\mathbf{X}\|_G = \sum_{i=1}^{L} G(\mathbf{x}_i)$ where $\mathbf{x}_i$ is the ith column of $\mathbf{X}$ and $G(\mathbf{x})$ measures the sparseness of vector $\mathbf{x}$. Typically, $G(\mathbf{x})$ is of the form,

$$G(\mathbf{x}) = \sum_{j=1}^{N} g(\mathbf{x}(j)) \qquad (3)$$

where $\mathbf{x}(j),\ j = 1, \ldots, N$ are the N components of vector $\mathbf{x}$. So rather than enforce sparsity for each measurement index across all sources, the entire source matrix is treated as an ensemble of coefficients without regard to their ordering and the ensemble is forced to be sparse. Thus, for notational simplicity in order to avoid double summations, we will consider the vector $\mathbf{x} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(NL)\}^\top$ where

$$\mathbf{x}(k) = \mathbf{X}_{i,j} \text{ for } i = 1+((k-1)\text{mod } L) \text{ and } j = 1+\left\lfloor\frac{k-1}{L}\right\rfloor \qquad (4)$$

so that $\mathbf{x}$ is simply the stacking of the columns of $\mathbf{X}$ into one column vector. The most commonly studied sparse measures is the ℓ^p norm-like measure,

$$\|\mathbf{X}\|_p = (\sum_{i,j} |\mathbf{x}(j)|^p)^{1/p} \qquad (5)$$

where $0 \leq p \leq 1$. When $p = 1$, ℓ^p is a norm. When $0 < p < 1$, ℓ^p is a quasi-norm. When $p = 0$,

$$\|\mathbf{X}\|_0 = \#\{j, \mathbf{x}(j) \neq 0\} \qquad (6)$$

which simply counts the number of non-zero components and ℓ^0 is not even a quasi-norm as it is not linear with respect to scalar multiplication. There are a myriad of other possible measures of sparsity introduced in the literature, these are discussed and compared in Section 3.

The seminal paper [4] introduced to a wider community the concept of non-negative matrix factorization for finding component 'parts' (or sources) given their mixtures. The technique is often applied to repetitive speech or music mixtures by transforming each mixture into the time-frequency domain and then taking the magnitude of each component. If only one mixture is available, the columns of the mixing matrix correspond to frequencies which are co-active in the mixtures and M is the number of frequencies, L is the number of time windows in the time-frequency transform, and N is the number of objects. The co-active frequency vectors make up the atoms/objects present in the source, and the source matrix $\mathbf{X}$ is then a train of pulses representing the times at which each atom occurs in the mixture. It is assumed that the objects occur only occasionally and thus the sources activations are sparse [5].

One rarely cited caveat with this approach is that the mixing model (1) assumes additive mixing and in this case is only valid if the original source components do not overlap. [6] asked the question 'when does NMF work?' and obtained a similar result, in a sense. Our problem is that the model NMF is based on is not accurate because the phase terms do not necessarily align. When $\mathbf{Y}$ is actually based on a complex-valued matrix and the magnitude is used in order to apply NMF, the components of $\mathbf{Y}$ are made up of linear combinations of a complex-valued mixing matrix. So each component y is made up of a sum of components $a_1 + \ldots + a_L$, but $|y| = |a_1 + \ldots + a_L|$ only if the phases of all the a_i are identical, or, if at most one a_i is non-zero. So

This material is based upon work supported by the Science Foundation Ireland under the PIYRA Programme.

NMF is appropriate in this setting only if the sources have disjoint support.

The non-overlapping requirement of the time-frequency representations of speech/music required by NMF is the basis for several source separation techniques [7–9]. The DUET technique [9] relies on *W-Disjoint Orthogonality* which is the requirement that at most one source is active at any given time-frequency point. Due to the time-frequency uncertainty principle, this requirement is not satisfied for interesting signals such as speech and music, although it could be satisfied for time-disjoint or frequency-disjoint signals, provided the appropriate choice of window length and window overlap. W-disjoint orthogonality which was based on time-frequency representations can easily be generalized to any arbitrary representation and the concept is a power weighted measure of how little the supports of a group of signals overlap. Note that disjointness is what is required by NMF and by the DUET-like techniques. The reason why there is so much focus on sparsity is that, sparsity coupled with independence of occurrence of components leads to a low probability of overlap and results in near disjointness.

The paper organization is as follows. In Section 2 a generalization of the W-disjoint orthogonality measure is discussed and we feel that this measure, *disjoint orthogonality* captures the essence of what is required of signal transformation such that demixing via binary masking is possible. In this section, also, the disjoint orthogonality of speech mixtures is measured and the optimal time-frequency window length from a disjoint orthogonality perspective is determined. In Section 3 we list and compare a number of sparsity measures that are common in the literature. We then measure the sparseness of time-frequency transformations of speech for various window lengths and compare the results to those from the disjoint orthogonality tests to see if any of the sparsity measures can be used to indicate the level of disjoint orthogonality for speech. For these tests, it is shown that the Kurtosis and Gini Index sparsity measures are well matched to disjoint orthogonality.

2. DISJOINT ORTHOGONALITY

In this section we trivially generalize the concept of W-disjoint orthogonality so that it can be applied to signals in any representation. The presentation follows that in [9]. We call two vectors $\mathbf{x}$ and $\mathbf{y}$ **disjoint orthogonal** if their supports are disjoint,

$$\mathbf{x}(j)\mathbf{y}(j) = 0, \ \forall j = 1, \ldots, L. \tag{7}$$

Here the signals can be in any arbitrary domain (e.g., time domain, time-frequency domain, or frequency domain). This condition is a more stringent requirement than simple orthogonality which would require that the (expected value of the) inner product of the two signals to be zero. For example, two independent white noise signals are orthogonal, but will not be disjoint orthogonal. This condition is a mathematical idealization of the condition that usually at most one source has significant energy at a given index and allows for the perfect separation of sources from one mixture. Consider the mixture $\mathbf{y}$ of N sources $\mathbf{x}_i$,

$$\mathbf{y} = \sum_{i=1}^{N} \mathbf{x}_i \tag{8}$$

and the masking vector $\mathbf{m}_i$,

$$\mathbf{m}_i(j) := \begin{cases} 1 & \mathbf{x}_i(j) \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

for $j = 1, \ldots, L$ so that $\mathbf{m}_i$ is the indicator function for the support of $\mathbf{x}_i$. Knowledge of the masking functions $\mathbf{m}_i$ is knowledge of the original sources if the sources are disjoint orthogonal as,

$$\mathbf{x}_i = \mathbf{m}_i \otimes \mathbf{y} \ \forall i = 1, \ldots, N, \tag{10}$$

where $\otimes$ is element-wise multiplication (the Hadamard product). The disjoint orthogonality assumption is not strictly satisfied for our signals of interest. In order to measure to what degree the above condition is approximately satisfied, we consider the following which generalizes the approximate W-DO measure discussed in [9, 10]. In order to measure approximate disjoint orthogonality for a given mask, we combine two important performance criteria: (1) how well the mask preserves the source of interest, and (2) how well the mask suppresses the interfering sources. These two criteria, the preserved-signal ratio (PSR) and the signal-to-interference ratio (SIR), are defined below.

First, given a mask $\mathbf{m}$ such that $0 \leq \mathbf{m}(j) \leq 1$ for all j, we define $\mathrm{PSR}_{\mathbf{m}}$, the PSR of the mask $\mathbf{m}$, as

$$\mathrm{PSR}_{\mathbf{m}} := \frac{\|\mathbf{m} \otimes \mathbf{x}_i\|_2^2}{\|\mathbf{x}_i\|_2^2} \tag{11}$$

which is the portion of energy of the ith source remaining after demixing using the mask. Note that $\mathrm{PSR}_{\mathbf{m}} \leq 1$ with $\mathrm{PSR}_{\mathbf{m}} = 1$ only if $\mathbf{m}_i(j) = 1 \Rightarrow \mathbf{m}(j) = 1, \forall j$. Now, we define

$$\mathbf{z}_i := \sum_{\substack{k=1 \\ k \neq i}}^{N} \mathbf{x}_k \tag{12}$$

so that $\mathbf{z}_i$ is the summation of the sources interfering with the ith source. Then, we define the signal-to-interference ratio of mask $\mathbf{m}$

$$\mathrm{SIR}_{\mathbf{m}} := \frac{\|\mathbf{m} \otimes \mathbf{x}_i\|_2^2}{\|\mathbf{m} \otimes \mathbf{z}_i\|_2^2} \tag{13}$$

which is the output signal-to-interference ratio after using the mask to demix to the ith source.

We now combine the $\mathrm{PSR}_{\mathbf{m}}$ and $\mathrm{SIR}_{\mathbf{m}}$ into one measure of approximate disjoint orthogonality. We propose the normalized difference between the signal energy maintained in masking and the interference energy maintained in masking as a measure of the approximate disjoint orthogonality associated with a particular mask:

$$\begin{aligned} D_{\mathbf{m}} &:= \frac{\|\mathbf{m} \otimes \mathbf{x}_i\|_2^2 - \|\mathbf{m} \otimes \mathbf{z}_i\|_2^2}{\|\mathbf{x}_i\|_2^2} && (14) \\ &= \mathrm{PSR}_{\mathbf{m}} - \mathrm{PSR}_{\mathbf{m}}/\mathrm{SIR}_{\mathbf{m}}. && (15) \end{aligned}$$

For signals which are disjoint orthogonal, using the mask $\mathbf{m}_i$ defined in (9), we note that $\mathrm{PSR}_{\mathbf{m}_i} = 1$, $\mathrm{SIR}_{\mathbf{m}_i} = \infty$, and $D_{\mathbf{m}_i} = 1$. This is the maximum obtainable disjoint orthogonality value because $D_{\mathbf{m}} \leq 1$ for all $\mathbf{m}$ such that $0 \leq \mathbf{m}(j) \leq 1$. Moreover, for any $\mathbf{m}$, $D_{\mathbf{m}} = 1$ implies that $\mathrm{PSR}_{\mathbf{m}} = 1$, $\mathrm{SIR}_{\mathbf{m}} = \infty$, and that (7) is satisfied. That is, $D_{\mathbf{m}} = 1$ implies that the signals are disjoint orthogonal and that mask $\mathbf{m}$ perfectly separates the ith source from the mixture. In order for a mask to have $D_{\mathbf{m}} \approx 1$, i.e., good demixing performance, it must simultaneously preserve the energy of the signal of interest while suppressing the energy of the interference. The failure of a mask to accomplish either of the goals can result in a small, even negative, value of D. For example, $D_{\mathbf{m}} = 0$ implies either that $\mathrm{PSR}_{\mathbf{m}} = 0$ (the mask kills all the energy of the source of interest) or that $\mathrm{SIR}_{\mathbf{m}} = 1$ (the mask results in equal energy for source and interference). Masks with $\mathrm{SIR}_{\mathbf{m}} < 1$ have associated $\mathrm{D}_{\mathbf{m}} < 0$.

If it is desirable that the disjoint orthogonal measure be bounded between 0 and 1, then we suggest the following mapping,

$$d_{\mathbf{m}} = 2^{D_{\mathbf{m}} - 1} \tag{16}$$

which has the desirable properties that:

1. $d_{\mathbf{m}} = 1$ implies that x_i is disjoint orthogonal with all interfering signals,

2. $d_{\mathbf{m}} = 1/2$ implies that application of mask $\mathbf{m}$ results in a demixture with equal source of interest and interference energies, and
3. $d_{\mathbf{m}} = 0$ implies that the mask $\mathbf{m}$ results in a demixture with $\mathrm{SIR}_{\mathbf{m}} \to 0$.

Which mask should we use when (7) is not satisfied? From (15), it follows that

$$\mathbf{m}_i^*(j) = \begin{cases} 1, & |\mathbf{x}_i(j)| > |\mathbf{z}_i(j)| \\ 0, & |\mathbf{x}_i(j)| \le |\mathbf{z}_i(j)| \end{cases} \qquad (17)$$

maximizes $D_{\mathbf{m}}$ as it turns 'on' signal coefficients where the source of interest dominates the interference and turns 'off' the remaining coefficients. The terms of equal magnitude in (17) we have arbitrarily turned 'off', but including them or excluding them makes no difference to the disjoint orthogonal measure as the terms cancel. The mask $\mathbf{m}_i^*$ is the optimal mask for demixing from a disjoint orthogonal performance standpoint.

2.1 Disjoint Orthogonality of Speech

We present results in Figure 1 which measure the disjoint orthogonality for pairwise (and 3-way and 4-way) mixing as a function of window size. For the tests, two (or three or four) 16kHz sampled speech files were selected at random from the TIMIT database and each file transformed into the time-frequency domain using a Hamming window of size $\{2^0, 2^1, \dots, 2^{15}\}$. The magnitude of the coefficients of a target source were compared to the sum of the remaining sources to generate the the mask $\mathbf{m}_i^*$. Using the mask, $d_{\mathbf{m}_i^*}$ was calculated. Over 150 mixtures were generated and the results averaged to form each data point shown in the figure. In all three cases the Hamming window size of size 1024 produced the representation that was the most disjoint orthogonal. A similar conclusion regarding the optimal time-frequency resolution of a window for speech separation was arrived at in [7]. Note that even when the window size is 1 (i.e., time domain), the mixtures still exhibit a high level of disjoint orthogonality. This fact was exploited by those methods that used the time-disjoint nature of speech [11–14]. Figure 1 clearly shows the advantage of moving from the time domain to the time-frequency domain: the speech signals are more disjoint in the time-frequency domain provided the window size is sufficiently large. Choosing the window size too large, however, results in reduced disjoint orthogonality.

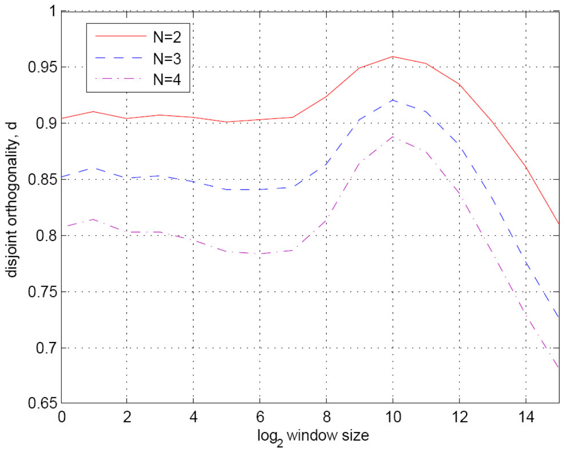

Figure 1: Disjoint orthogonality for time-frequency representations of $N = 2, 3, 4$ speech source mixing as a function of window size used in the time-frequency transformation. Speech is most disjoint orthogonal when a window of 1024 samples is used, corresponding to 64 ms length.

3. SPARSE MEASURES

In this section we examine and compare the common sparse measures found in the literature and see which of them yield desirable results when measuring the sparsity of time-frequency representations of speech. The desirable attribute in this case is that the sparseness of the signal has a similar shape to the disjoint orthogonality curve in Figure 1. It is reasonable to desire that our sparse measure indicates correctly which window size results in the best demixing performance.

Many commonly used sparse measures of the form $G(\mathbf{x})$ discussed in the Introduction are investigated and compared in [2]. These measures, and others, are listed and defined in Table 1. As discussed before, the ℓ^0 cost is a 'direct' measure of sparsity as it counts the number of non-zero elements of $\mathbf{x}$. When noise is present, ℓ^0_ϵ is often used as the noise results in very few components being truly zero, despite the fact the representation is still sparse in an intuitive sense. As optimization using ℓ^0_ϵ is difficult because the gradient yields no information, ℓ^p is often used in its place, with $p < 1$. $\tanh_{a,b}$ is sometimes used in place of ℓ^p, $p < 1$, because it is limited to the range $(0, 1)$ and better models ℓ^0 and ℓ^0_ϵ in this respect. The fact that ℓ^p, $p < 1$ and $\tanh_{a,b}$ are concave enforces sparsity. That is, a representation is more sparse if we have one large component, rather than dividing up the large component into two smaller ones. The log measure is concave outside some range, but convex near the origin, which in effect spreads the small components. κ_4 is the kurtosis which measures the peakedness of a distribution. The last measure considered in [2] is u_θ which measures the smallest range which contains are certain percentage of the data. For many of these measures, sometimes the data vector $\mathbf{x}$ is replaced with a whitened version $\mathbf{x}_{\text{white}}$. Plots of the individual measures of sparseness that have a functional form are shown in Figure 2.

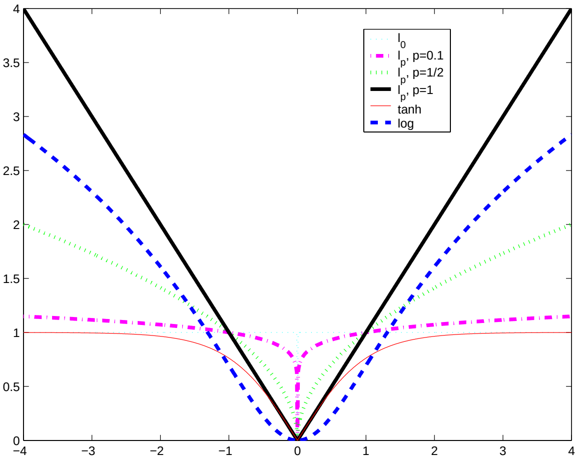

Figure 2: Component sparsity contributions as a function of component amplitude.

[15] uses several additional measures of sparsity for optimal basis selection. They generalize the ℓ^p measure to include negative values of p, which we will label ℓ^p_-. [15] also considers the Gaussian entropy diversity measure, $H_G(\mathbf{x})$, and the Shannon entropy diversity measure, $H_S(\mathbf{x})$. These three sparsity measures are also listed in Table 1. [16] considers the additional choices for $\tilde{\mathbf{x}}(j)$, $\tilde{\mathbf{x}}(j) = |\mathbf{x}(j)|$, which we label H'_S.

Our favorite measure of sparsity is the Gini index as we feel it captures several desirable characteristics that a sparsity measure should have. These characteristics are described here with regard to inequity of wealth distribution as that was the original application of the Gini index

ℓ^0	$\#\{j, \mathbf{x}(j) \neq 0\}$
ℓ^0_ϵ	$\#\{j, \lvert\mathbf{x}(j)\rvert \geq \epsilon\}$
ℓ^1	$\left(\sum_j \lvert\mathbf{x}(j)\rvert\right)$
ℓ^p	$\left(\sum_j \lvert\mathbf{x}(j)\rvert^p\right)^{1/p}, \ \ 0 < p < 1$
$\tanh_{a,b}$	$\sum_j \tanh\left(\lvert a\mathbf{x}(j)\rvert^b\right)$
log	$\sum_j \log\left(1 + \lvert\mathbf{x}(j)\rvert^2\right)$
κ_4	$\frac{\sum_j \lvert\mathbf{x}(j)\rvert^4}{\left(\sum_j \lvert\mathbf{x}(j)\rvert^2\right)^2}$
u_θ	$\min_{i,j}(\mathbf{x}(\{i\}) - \mathbf{x}(\{j\}))$ s.t. $\frac{i-j}{L} \geq \theta$ for ordered data, $\mathbf{x}(\{1\}) \leq \mathbf{x}(\{2\}) \leq \cdots \leq \mathbf{x}(\{L\})$
ℓ^p_-	$\sum_{j,\mathbf{x}(j)\neq 0} \lvert\mathbf{x}(j)\rvert^p, \ \ p < 0$
H_G	$-\sum_j \ln\lvert\mathbf{x}(j)\rvert^2$
H_S	$-\sum_j \tilde{\mathbf{x}}(j)\ln\lvert\tilde{\mathbf{x}}(j)\rvert^2$ where $\tilde{\mathbf{x}}(j) = \frac{\lvert\mathbf{x}(j)\rvert^2}{\lVert\mathbf{x}\rVert_2^2}$
H'_S	$-\sum_j \tilde{\mathbf{x}}(j)\ln\lvert\tilde{\mathbf{x}}(j)\rvert^2$ where $\tilde{\mathbf{x}}(j) = \lvert\mathbf{x}(j)\rvert$
Gini	see text

Table 1: Measuring sparsity.

[17, 22].

- (Dalton's 1st Law) Robin Hood decreases sparsity. Stealing from the rich and giving to the poor, decreases the inequity of wealth distribution (assuming you don't make the rich poor and the poor rich).
- (Dalton's modified 2nd Law) Sparsity is scale invariant. Multiplying wealth by a constant factor does not alter the effective wealth distribution.
- (Dalton's 3rd Law) Adding a constant decreases sparsity. Give everyone a trillion dollars and the small differences in overall wealth are then negligible.
- (Dalton's 4th Law) Sparsity is invariant under cloning. If you have a twin population with identical wealth distribution, the sparsity of wealth in one population is the same for the combination of the two.
- (Proposed in [22]) Bill Gates increases sparsity. As one individual becomes infinitely wealthy, the wealth distribution becomes as sparse as possible.
- (Proposed in [22]) Babies increase sparsity. Adding individuals with zero wealth to a population increases the sparseness of the distribution of wealth.

With these in mind, we define the Gini index. We order the coefficient data from smallest to largest, $|\mathbf{x}(\{1\})| \leq |\mathbf{x}(\{2\})| \leq \cdots \leq |\mathbf{x}(\{L\})|$, where $\{1\}, \{2\}, \ldots, \{L\}$ are the indices of the sorting operation. The Lorenz curve is used to measure wealth distribution in society and was originally defined in [18]. We parameterize this curve with parameter p and introduce here the parameterized-Lorenz curve Λ_p which is the function with support $(0, 1)$, that is piecewise linear with $L + 1$ points defined,

$$\Lambda_p\left(\frac{i}{L}\right) = \frac{1}{\|\mathbf{x}\|_p^p}\sum_{j=1}^{i}|\mathbf{x}(\{j\})|^p, \quad \text{for } i = 0, \ldots, L. \qquad (18)$$

Note, $\Lambda_p(0) = 0$ and $\Lambda_p(1) = 1$. With $p = 2$, each point on the Lorenz curve $(x = a_0, y = b_0)$ has the interpretation that $100 \times a_0$ percent of the sorted signal coefficients captures $100 \times b_0$ percent of the total signal power. Thus, the slower the curve rises to 1, the fewer coefficients are needed to accurately represent the signal. If all coefficients were equal, which we could argue is the least sparse scenario, the curve would rise at a 45 degree angle. Thus, the area between the Lorenz curve and the 45 degree line will increase as the sparsity of the signal increases. Indeed, twice the area of this region was originally proposed (in English) in 1921 in [19] as a measure of the inequality of wealth distribution; 'Inequity in distribution' is another way of describing sparsity. The area beneath the Lorenz curve is,

$$A_p(\mathbf{x}) = \frac{1}{2N}\sum_{n=1}^{N}\left(\Lambda_p\left(\frac{n-1}{N}\right) + \Lambda_p\left(\frac{n}{N}\right)\right) \qquad (19)$$

and twice the area between the Lorenz curve and the 45 degree, which is known as the Gini index, is then simply,

$$\text{Gini}_p(\mathbf{x}) = 1 - 2A_p(\mathbf{x}). \qquad (20)$$

The Lorenz curve and its potential use in sparse basis selection is discussed in [16, 20] but the Gini index is not mentioned as a potential sparse measure. In [22] the Gini index was used a a measure of sparsity to determine if speech was more sparse in time-scale or time-frequency.

Figure 3 shows the Lorenz curve and Gini index for four simple vectors. Note that the distribution in which all individuals have equal wealth is the least sparse and the distribution in which all the wealth is concentrated in one individual is the most sparse.

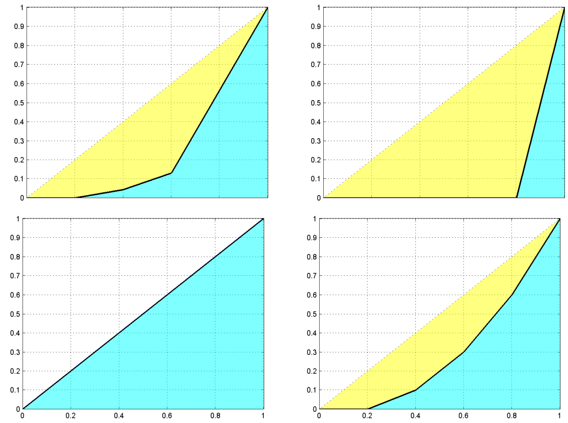

Figure 3: Lorenz curve for [0 1 2 10 10] (top left), [0 0 0 0 1] (top right), [1 1 1 1 1] (bottom left), and [0 1 2 3 4] (bottom right). The Gini index is twice the lightly shaded/yellow area. The Gini indexes are 0.5043, 0.8, 0.0, and 0.4, respectively.

The Gini index has many nice properties:

- A representation with equal wealth distribution has $\text{Gini}_p(\mathbf{x}) = 0$, no sparsity.
- (Dalton's 1st & 2nd Law) $\text{Gini}_p(\mathbf{x})$ satisfies the Robin Hood Principle and is scale invariant.
- (Dalton's 3rd Law) $\text{Gini}_p(\mathbf{x} + k) \rightarrow 0$ as scaler $k \rightarrow \infty$.
- (Dalton's 4th Law) $\text{Gini}_p(\{x_1, x_2, \ldots x_N\})$ is identical to $\text{Gini}_p(\{x_1, x_1, x_2, x_2, \ldots, x_N, x_N\})$.
- (Proposal 1) As one component of a representation goes to infinity, $\text{Gini}_p(\mathbf{x}) \rightarrow 1$.
- (Proposal 2) If an infinite number of zero components are added to a vector, $\text{Gini}_p(\mathbf{x}) \rightarrow 1$.

3.1 The Sparseness of Speech

Our goal is to determine which of the above discussed sparsity measures can be used as a potential indicator for the level of disjoint orthogonality. In order to gain some insight into this question, we measured the sparseness of speech in the time-frequency domain using the ℓ^0, ℓ^1, $\ell^{0.1}$, $\tanh_{1,1}$, log, κ_4, ℓ^1_-, H_G, H_S, H'_S, and Gini_2 measures for various window lengths, as before. Examination of the sparseness as a function of window size revealed that the ℓ^0, ℓ^1_-, H_G, and H_S measures all indicated that speech was sparser for smaller windows with a clear peak in the sparseness for window sizes of 2^1 or 2^2. On the other hand, the ℓ^1, $\ell^{0.1}$, $\tanh_{1,1}$, log, and H'_S measures all indicated a clear peak for larger windows with a clear peak in the sparseness for window sizes of 2^{12} to 2^{14}. Plots of the sparseness normalized so that the maximum average sparseness is unity for the H_G and ℓ^1 are shown in Figure 4. The two measures that did match well with the disjoint orthogonality results, the Kurtosis (κ_4) and Gini Index with parameter $p = 2$ (Gini_2), are also shown in Figure 4. Both have broad peaks for a window size of 2^{10}.

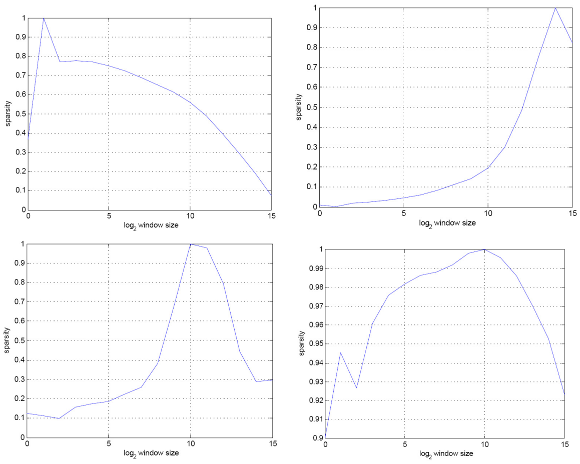

Figure 4: The H_G (upper left) and ℓ^1 (upper right) sparsity measures applied to speech signals in the time-frequency domain as a function of window size. These graphs, which are indicative of the majority of the measures tested, do not match well to the results for the disjoint orthogonality tests. The Kurtosis (κ_4) (lower left) and Gini Index with parameter $p = 2$ (Gini_2) (lower right) sparsity measures applied to speech signals in the time-frequency domain as a function of window size. Both of this measures behave similarly to the disjoint orthogonality results.

4. CONCLUSIONS

Sparsity is important for source separation of speech mixtures in the application of NMF and DUET-like techniques only because what is really required is disjointness. Sparsity coupled with independence of occurrence of the coefficients results in a low probability that coefficients from different sources/object overlap, and this leads to approximate disjointness. We argued in this paper that what we should be interested in for source separation applications such as speech and music is not sparse representations, but rather disjoint representations. We defined a measure, disjoint orthogonality, of how approximately disjoint a class of signals are for the purposes of measuring potential demixing performance via binary masking. We also examined common sparsity measures and evaluated which ones have the potential to indicate when a representation is optimally disjoint orthogonal. The results indicate that both the Kurtosis and the Gini Index are reasonable indicators for when sources are disjoint, and thus separable.

References

[1] D. L. Donoho. Sparse components analysis and optimal atomic decompositions. *Constructive Approximation*, 17:353–382, 2001.

[2] J. Karvanen and A. Cichocki. Measuring sparseness of noisy signals. In *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 125–130, Nara, Japan, April 2003.

[3] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[4] D. D. Lee and H. S. Seung. Learning of the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[5] P. Smaragdis. Non-negative matrix factor deconvolution; Extraction of multiple sound sources from monophonic inputs. In *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 494–499, September 2004.

[6] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[7] M Aoki, M Okamoto, S Aoki, H Matsui, T Sakurai, and Y Kaneda. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science and Technology*, 22(2):149–157, 2001.

[8] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

[9] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.

[10] S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 529–532, Orlando, Florida, USA, May 2002.

[11] J.-K. Lin, D. G. Grier, and J. D. Cowan. Feature extraction approach to blind source separation. In *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 398–405, Amelia Island Plantation, Florida, September 24–26 1997.

[12] T-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Proc. Letters*, 6(4):87–90, April 1999.

[13] M. Van Hulle. Clustering approach to square and non-square blind source separation. In *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 315–323, Madison, Wisconsin, August 23–25 1999.

[14] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda, and J. C. Principe. Underdetermined blind source separation in a time-varying environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 3049–3052, Orlando, Florida, USA, May 13–17 2002.

[15] B. D. Rao and K. Kreutz-Delgado. An affine scaling methodology for best basis selection. *IEEE Transactions on Signal Processing*, 47(1):187–200, January 1999.

[16] K. Kreutz-Delgado and B. D. Rao. Measures and algorithms for best basis selection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1881–1884, Seattle, Washington, USA, May 1998.

[17] B. C. Arnold. *Majorization and the Lorenz Order: A Brief Introduction.* Springer-Verlag, 1986.

[18] M. O. Lorenz. Methods of measuring concentrations of wealth. *J. Amer. Stat. Assoc.*, 1905.

[19] C. Gini. Measurement of inequality of incomes. *Economic Journal*, 31:124–126, 1921.

[20] K. Kreutz-Delgado and B. D. Rao. Sparse basis selection, ica, and majorization: Towards a unified perspective. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1081–1084, Phoenix, Arizona, USA, March 1999.

[21] H. Dalton. The measurement of the inequity of incomes. *Economic Journal*, 30:348–361, 1920.

[22] S. Rickard and M. Fallon. The Gini index of speech. In *Conference on Information Sciences and Systems*, Princeton, NJ, USA, March 2004.