Comparing Measures of Sparsity

Niall Hurley and Scott Rickard

Abstract—Sparsity of representations of signals has been shown to be a key concept of fundamental importance in fields such as blind source separation, compression, sampling and signal analysis. The aim of this paper is to compare several commonly-used sparsity measures based on intuitive attributes. Intuitively, a sparse representation is one in which a small number of coefficients contain a large proportion of the energy. In this paper six properties are discussed: (Robin Hood, Scaling, Rising Tide, Cloning, Bill Gates and Babies), each of which a sparsity measure should have. The main contributions of this paper are the proofs and the associated summary table which classify commonly-used sparsity measures based on whether or not they satisfy these six propositions. Only one of these measures satisfies all six: the Gini Index.

I. INTRODUCTION

Whether with sparsity constraints or with sparsity assumptions, the concept of sparsity is readily used in diverse areas such as oceanic engineering [1], antennas and propagation [2], face recognition [3], image processing [4], [5] and medical imaging [6]. Sparsity has also played a central role in the success of many machine learning algorithms and techniques such as matrix factorization [7], signal recovery/extraction [8], denoising [9], [10], compressed sensing [11], dictionary learning [12], signal representation [13], [14], support vector machines [15], sampling theory [16], [17] and source separation/localization [18], [19]. For example, one method of source separation is to transform the signal to a domain in which it is sparse (e.g. time-frequency or wavelet) where the separation can be performed by a partition of the transformed signal space due to the sparsity of the representation [20], [21]. There has also been research in the uniqueness of sparse solutions in overcomplete representations [22], [23].

There are many measures of sparsity. Intuitively, a sparse representation is one in which a small number of coefficients contain a large proportion of the energy. This interpretation leads to further possible alternative measures. Indeed, there are dozens of measures of sparsity used in the literature. Which of the sparsity measures is the best? In this paper we suggest six desirable characteristics of measures of sparsity and use them to compare fifteen popular sparsity measures in the literature. We elaborate on one of these measures, the Gini Index, as it has many desirable characteristics including the ability to measure the sparsity of a distribution. We also show graphically how some measures treat components of different magnitude. In Sec. IV we present the main result of this work, namely, the comparison of the fifteen commonly-used sparsity measures using the six criteria. We show that the only measure to satisfy all six is the Gini Index. Proofs of the table are attached in Appendices A and B. A preliminary report on these results (without proofs) appeared in [24]. We then compare the fifteen measures graphically on data drawn from two sets of parameterized distributions. We select distributions for which we can control the ‘sparsity’. This allows us to visualize the behavior of the sparsity measures in view of the sparse criteria. In Sec. V we present some conclusions. The main conclusion is that from the fifteen measures, only the Gini Index satisfies all six criteria, and, as such, we encourage its use and study.

II. THE SIX CRITERIA

The following are six desirable attributes of a measure of sparsity. The first four, $D_1$ through $D_4$, were originally applied in a financial setting to measure the inequity of wealth distribution in [25]. The last two, $P_1$ and $P_2$, were proposed in [26]. Distribution of wealth can be used interchangeably with distribution of energy of coefficients and where convenient in this paper, we will keep the financial interpretation in the explanations. Inequity of distribution is the same as sparsity. An equitable distribution is one with all coefficients having the same amount of energy, the least sparse distribution.

$D_1$ Robin Hood - Robin Hood decreases sparsity (Dalton’s 1st Law). Stealing from the rich and giving to the poor decreases the inequity of wealth distribution (assuming we do not make the rich poor and the poor rich). This comes directly from the definition of a sparse distribution being one for which most of the energy is contained in only a few of the coefficients.

$D_2$ Scaling - Sparsity is scale invariant (Dalton’s modified 2nd Law [24]). Multiplying wealth by a constant factor does not alter the effective wealth distribution. This means that relative wealth is important, not absolute wealth. Making everyone ten times more wealthy does not affect the effective distribution of wealth. The rich are still just as rich and the poor are still just as poor.

$D_3$ Rising Tide - Adding a constant to each coefficient decreases sparsity (Dalton’s 3rd Law). Give everyone a trillion dollars and the small differences in overall wealth are then negligible so everyone will have effectively the same wealth. This is intuitive as adding a constant energy to each coefficient reduces the relative difference of energy between large and small coefficients. This law
assumes that the original distribution contains at least two individuals with different wealth. If all individuals have identical wealth, then by D2 there should be no change to the sparsity for multiplicative or additive constants.

D4 Cloning - Sparsity is invariant under cloning (Dalton’s 4th Law). If there is a twin population with identical wealth distribution, the sparsity of wealth in one population is the same for the combination of the two.

P1 Bill Gates - Bill Gates increases sparsity. As one individual becomes infinitely wealthy, the wealth distribution becomes as sparse as possible.

P2 Babies - Babies increase sparsity. In populations with non-zero total wealth, adding individuals with zero wealth to a population increases the sparseness of the distribution of wealth.

These criteria give rise to the sparsest distribution being one with one individual owning all the wealth and the least sparse being one with everyone having equal wealth.

Dalton [25] proposed that multiplication by a constant should decrease inequality. This was revised to the more desirable property of scale invariance. Dalton’s fourth principle, D4, is somewhat controversial. However, if we have a distribution from which we draw coefficients and measure the sparsity of the coefficients which we have drawn, as we draw more and more coefficients we would expect our measure of sparsity to converge. D4 captures this concept.

‘Mathematically this [D4] requires that the measure of inequality of the population should be a function of the sample distribution function of the population. Most common measures of inequality satisfy this last principle.’ [27]

Interestingly, most measures of sparsity do not satisfy this principle, as we shall see.

We define a sparse measure $S$ as the a function with the following mapping

$$S : \left( \bigcup_{n \geq 1} \mathbb{C}^n \right) \rightarrow \mathbb{R}$$

where $n \in \mathbb{N}$ is the number of coefficients. Thus $S$ maps complex vectors to a real number.

There are two crucial, core, underlying attributes which our sparsity measures must satisfy. As all measures satisfy these two conditions trivially we will not comment on them further except to define them.

A1 $S(\tilde{c}) = S(\Pi \tilde{c})$ where $\Pi$ denotes permutation, that is, the sparsity of any permutation of the coefficients is the same. This means that the ordering of the coefficients is not important.

A2 The sparsity of the coefficients is calculated using the magnitudes of the coefficients. This means we can assume we are operating in the positive orthant, without loss of generality.

By A2 we can assume we are operating in the positive orthant, and as such we can rewrite (1) as

$$S : \left( \bigcup_{n \geq 1} \mathbb{R}_+^n \right) \rightarrow \mathbb{R},$$

which is more consistent with the wealth interpretation.

We will use the convention that $S(\tilde{c})$ increases with increasing sparsity where $\tilde{c} = [c_1 \ c_2 \ \cdots \ c_N]$ are the coefficient strengths. Given vectors

$$\tilde{c} = [c_1 \ c_2 \ \cdots \ c_N]$$

$$\tilde{d} = [d_1 \ d_2 \ \cdots \ d_M]$$

we define concatenation, which we use $||$ to denote, as

$$\|	ilde{c}||\tilde{d} = [c_1 \ c_2 \ \cdots \ c_N \ d_1 \ d_2 \ \cdots \ d_M].$$

We also define the addition of adding a constant to a vector as the addition of that constant to each element of the vector, that is, for $\alpha \in \mathbb{R}$,

$$\tilde{c} + \alpha = [c_1 + \alpha \ c_2 + \alpha \ \cdots \ c_N + \alpha].$$

The six sparse criteria can be formally defined as follows:

D1 Robin Hood:

$$S([c_1 \ \cdots \ c_i - \alpha \ \cdots \ c_j + \alpha \ \cdots]) < S(\tilde{c})$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

D2 Scaling:

$$S(\alpha \tilde{c}) = S(\tilde{c}), \forall \alpha \in \mathbb{R}, \alpha > 0.$$

D3 Rising Tide:

$$S(\tilde{c} + \alpha) < S(\tilde{c}), \alpha \in \mathbb{R}, \alpha > 0$$

(We exclude the case $c_1 = c_2 = c_3 = \cdots = c_i = \cdots \forall i$ as this is equivalent to scaling.).

D4 Cloning:

$$S(\tilde{c}) = S(\tilde{c}||\tilde{c}||\tilde{c}||\tilde{c}) = S(\tilde{c}||\tilde{c})||\tilde{c}).$$

P1 Bill Gates:

$$\forall \exists \beta_i > 0, \text{ such that } \forall \alpha > 0 :$$

$$S([c_1 \ \cdots \ c_i + \beta + \alpha \ \cdots]) > S([c_1 \ \cdots \ c_i + \beta \ \cdots]).$$

P2 Babies:

$$S(\tilde{c}||0) > S(\tilde{c}).$$

As stated above, when proving Rising Tide we exclude the scenario where all coefficients are equal. In this case, adding a constant is actually a form of scaling. Another interpretation is that the case with all coefficients equal is, in fact, the minimally sparse scenario and hence adding a constant cannot decrease the sparsity.

A. Two Proofs

As one would surmise there is some overlap between the criteria. We present and prove two theorems which demonstrate this overlap. Theorem 2.1 states that if a measure satisfies both criteria D1 and D2, the sparsity measure also satisfies P1 by default. Theorem 2.2 states that a measure satisfying D1, D2 and D4 necessarily satisfies P2.

Theorem 2.1: D1 & D2 $\Rightarrow$ P1, that is, if both D1 and D2 are satisfied, P1 is also satisfied.

Proof: Without loss of generality, we begin with the vector $\tilde{c}$ sorted in ascending order

$$\tilde{c} = [c_1 \ c_2 \ \cdots \ c_N]$$

with $c_1 \leq c_2 \leq \cdots \leq c_N$. We then perform a series of inverse Robin Hood steps to get a vector $\tilde{d}$, that is, we take from smaller coefficients and give to the largest coefficient

$$d_i = c_i - \Delta c_i$$

$$d_N = c_N + \Delta c_i$$

where

$$\Delta c_i = \min(c_{i+1}, \cdots, c_N) - c_i.$$
with condition $\Delta < 1$. As these are inverse Robin Hood steps (inverse $D_1$), they increase sparsity and result in the vector

$$\tilde{d} = [(c_1 - \Delta c_1) (c_2 - \Delta c_2) \cdots (c_{N-1} - \Delta c_{N-1}) (\Delta c_1 + \cdots + \Delta c_{N-1} + c_N)]$$

Without affecting the sparsity we can then scale (D2) $\tilde{d}$ by $\frac{1}{1-\Delta}$ to get

$$\bar{c} = \frac{1}{1-\Delta} (c_1 \ c_2 \ \cdots \ c_{N-1} \alpha + c_N),$$

where

$$\alpha = \frac{1}{1-\Delta} (c_1 + c_2 + \cdots + c_N).$$

It is clear that

$$S(\bar{c}) = S(\tilde{d}) > S(\bar{c}),$$

which is equivalent to $P1$ with the given $\alpha$ and $\beta = 0$. If we wish to operate on $c_i$ (instead of $c_N$ as above), $\beta$ can be chosen sufficiently large to make the desired coefficient the largest, that is, we set

$$\beta > c_N - c_i \quad \blacksquare$$

**Theorem 2.2:** $D_1$ & $D_2$ & $D_4 \Rightarrow P2$, that is, if $D_1$, $D_2$ and $D_4$ are satisfied, $P2$ is also satisfied.

**Proof:** We begin with vector $\bar{c}$

$$\tilde{c} = [c_1 \ c_2 \ \cdots \ c_N].$$

We then clone (D4) this $N+1$ times to get

$$\bar{c} = \left[\begin{array}{cccc} \tilde{c} & \tilde{c} & \cdots & \tilde{c} \end{array}\right].$$

We then take one of the $\tilde{c}$ from $\bar{c}$, which we shall refer to as $\tilde{c}$ and by a series of inverse Robin Hood operations (D1) we distribute this $\tilde{c}$ in accordance with the size of each element to form new vector $\tilde{D}$. That is to say, each $c_i$ of each $\tilde{c}$ (excluding $\tilde{c}$) becomes $c_i + \frac{\tilde{c}}{N}$ by $N$ consecutive inverse Robin Hood operations which increase sparsity. The result is

$$\tilde{D} = \left[\begin{array}{cccc} \frac{\tilde{c}}{N} & \frac{\tilde{c}}{N} & \cdots & \frac{\tilde{c}}{N} & 0 & 0 & \cdots & 0 \end{array}\right].$$

We can then scale (D2) $\tilde{D}$ by a factor of $\frac{N}{1+N}$ without affecting the sparsity to get

$$\bar{D} = \left[\begin{array}{cccc} \frac{\tilde{c}}{N} & \frac{\tilde{c}}{N} & \cdots & \frac{\tilde{c}}{N} & 0 & 0 & \cdots & 0 \end{array}\right].$$

which by cloning (D4) we know is equivalent to

$$\bar{F} = \left[\begin{array}{cc} \tilde{c} & 0 \end{array}\right].$$

In summation, we have shown that

$$S(\bar{c}) = S(\bar{D}) = S(\bar{F}) = S(\tilde{c}),$$

that is,

$$S(\bar{c}) < S(\bar{c})|0,$$

which is also known as $P2$. 

### III. The Measures of Sparsity

In this section we discuss a number of popular sparsity measures. These measures are used to calculate a number which describes the sparsity of a vector $\bar{c} = [c_1 \ c_2 \ \cdots \ c_N]$. The measures’ monikers and their definitions are listed in Table I. Some measures in Table I have been manipulated (in general negated) to ensure that the an increase in sparsity results in a (positive) increase in the sparse measure.

#### Table I

| Measure | Definition |
|---------|------------|
| $\ell^0$ | $\# \{j : c_j = 0\}$ |
| $\ell_0^c$ | $\# \{j : c_j < \epsilon\}$ |
| $-\ell^1$ | $-\sum_j c_j$ |
| $-\ell^p$ | $-\left(\sum_j c_j^p\right)^{1/p}$, $0 < p < 1$ |
| $\ell_2^p$ | $\sqrt{\sum_j c_j^2}$ |
| $-\tan\theta_{ab}$ | $-\sum_j \tan\left((ac_j)b\right)$ |
| $-\log$ | $-\sum_j \log (1 + c_j^2)$ |
| $\kappa_4$ | $\sum_j c_j^4$ |
| $u_0$ | $1 - \min_{i=1,2,\ldots,N}[|\theta|\gamma_i + 1 - \frac{\kappa_i(|\theta|\gamma_i - c_i)}{c_i}]$, s.t. $|\theta| \neq N$ for ordered data, $c_1 \leq c_2 \leq \cdots \leq c_N$ |
| $-\mu_0$ | $-\sum_j \log c_j^{-2}$, $p < 0$ |
| $H_G$ | $-\sum_j \log c_j^{-2}$ |
| $H_S$ | $-\sum_j c_j \log c_j^{-2}$ |
| Hoyer | $\left((\sqrt{N} - \sum_j \gamma_j) / (\sqrt{N} - 1)\right)^{-1}$ |
| Gini | $1 - 2 \sum_{k=1}^N \frac{e(k)}{N!} \left(\frac{N-k+1}{N}\right)$ for ordered data, $c_1 \leq c_2 \leq \cdots \leq c_N$ |

In [28] the $\ell^0$, $\ell^0_c$, $\ell^1$, $\ell^p$, $\tan\theta_{ab}$, log and $\kappa_4$ were compared. The most commonly used and studied sparsity measures are the $\ell^p$ norm-like measures,

$$||\bar{c}||_p = \left(\sum_j c_j^p\right)^{1/p}$$

for $0 \leq p \leq 1$.

The $\ell^0$ measure simply calculates the number of non-zero coefficients in $\bar{c}$,

$$||\bar{c}||_0 = \# \{c_j \neq 0, j = 1,\ldots,N\}.$$
are interested in the number of coefficients, \( c_j \) that are greater than a threshold \( \epsilon \). Clearly, the value of \( \epsilon \) is crucial for \( \ell_0 \) to be meaningful. This is undesirable. As optimization using \( \ell_0 \) is difficult because the gradient yields no information, \( \ell^p \) with \( 0 < p < 1 \) is often used in its place. In many optimization problems, as linear programming offers a fast, computationally efficient solution.

In several alternative measures of sparsity are noted which approximate the \( \ell^0 \) measure but emphasize different properties. \( \tanh_{a,b} \) is sometimes used in place of \( \ell^p \), \( 0 < p < 1 \), as it is limited to the range \( (0, 1) \) and better models \( \ell^0 \) and \( \ell^1 \) in this respect. A representation is more sparse if it has one large component, rather than dividing up the large component into two smaller ones. \( \tanh_{a,b} \) and \( \ell^p \) preserve this. In it is shown that the \( \log \) measure enforces sparsity outside some range, but for distributions with low energy coefficients the opposite is achieved by effectively spreading the energy of the small components. \( k_4 \) is the kurtosis which measures the peakedness of a distribution. \( \theta \) measures the smallest range which contains a certain percentage of the data. This is achieved by sorting the data and determining the minimum difference between the largest and smallest sample in a range containing the specified percentage \( \theta \) of data points as a fraction of the total range of the data. The reason that a continuous parameter \( \theta \) is used in the model is to maintain compatibility with pre-existing literature.

For measuring ‘diversity’, some different measures. Three of these are entropy measures: the Shannon entropy diversity measure \( H_S \), a modified version of the Shannon entropy diversity measure \( H_{S^t} \) and the Gaussian entropy diversity measure \( H_G \). They also extend the \( \ell^p \) measure to negative exponents, that is, \( -1 < p < 0 \). We call this measure \( \ell_p \) to avoid confusion.

Some of the measures can be normalized to satisfy more of the constraints, although in general for the measures, forcing satisfaction of one constraint means breaking another. The exception to this is the Hoyer measure which is a normalized version of the \( \ell_1 \) measure as is obvious from its definition. In Fig. we can get an insight into how component magnitude affect certain measures. In general, the smaller the magnitude the less it impinges on the sparsity of the measure. We can see how many of the measures approximate the \( \ell^0 \) measure but as they are not flat like the \( \ell^0 \) measure, they have a gradient that can be used in optimization problems. The \( \ell^0, \ell^1, \tanh, \log \), \( \ell^p \) \( (0 < p < 1) \), \( \ell^1 \) measures all prefer components to be zero or near zero. Oddly, the Shannon entropy based measures \( H_S \) and \( H_{S^t} \) prefer components to be at a non-zero value less than 1.

A. The Gini Index

Having perused the measures thus far, some desirable aspects of a sparsity measure emerge. Like \( \tanh \) and Hoyer, a measure should be some kind of weighted sum of the coefficients. This means that unlike \( \ell^0 \) when a coefficient changes slightly we have a weighted effect on the corresponding change in the value of the sparsity measure based on how ‘important’ that particular coefficient is to the overall sparsity. Large coefficients should have a smaller weight than the small coefficients so that they do not overwhelm them to the point that smaller coefficients have a negligible (or no) effect on the measure of sparsity. If even one of the smaller coefficients is changed, that change should be reflected by a change in the value of the sparsity measure. A weighted sum achieves this. In other words, we have a gradient which we can use in optimization problems. Another important aspect of a sparsity measure is normalization. A set of coefficients should not be rated more or less sparse simply because it has more coefficients than another set, nor should it be deemed more or less sparse simply due to having louder or quieter coefficients.

In short, there should be two forms of normalization. Firstly, the measure of sparsity should be dependent on the relative values of coefficients as a fraction of the total value. Secondly, the measure of sparsity should be independent of the number of coefficients so that sets of different size can be compared. Lastly, it would be useful if the measure was 0 for the least sparse case and 1 for the most sparse case. All these qualities are embodied by the Gini Index, which we now define.

Given a vector, \( \bar{c} = [c_1, c_2, c_3, \ldots] \), we order from smallest to largest, \( c_1(1) \leq c_2(2) \leq \cdots \leq c_i(N) \) where \( (1), (2), \ldots, (N) \) are the new indices after the sorting operation. The Gini Index is given by

\[
S(\bar{c}) = 1 - 2 \sum_{k=1}^{N} \frac{c_{(k)}}{|{\bar{c}}|} \left( \frac{N - k + \frac{1}{2}}{N} \right).
\]

The Gini Index also has an interesting graphical interpretation which we see in Fig. If percentage of coefficients versus percentage of total coefficient value is plotted for the sorted coefficients we can define the Gini Index as twice the area between this line and the 45° line. The 45° line represents the least sparse distribution, that with all the coefficients being equal.

If we have a distribution from which we draw coefficients and measure the sparsity of the coefficients which we have drawn, as we draw more and more coefficients we would expect our measure of sparsity to converge. The Gini Index meets these expectations. The Gini Index of a distribution with probability density function \( f(x) \) (which satisfies \( f(x) = 0, x < 0 \) and cumulative distribution function \( F(x) \) is given by

\[
G = 1 - 2 \int_0^1 \frac{1}{\int_0^\infty t f(t) dt} dt dF(x).
\]

As a side note, the Gini Index was originally proposed in economics as measure of the inequality of wealth and is still studied in relation to wealth distribution as well as other areas. ‘Inequality in wealth’ in signal processing language is ‘efficiency of representation’ or ‘sparsity’. The utility of the Gini Index as a measure of sparsity has been demonstrated in.
IV. COMPARISON OF SPARSITY MEASURES

In this section we present the main result of the paper, the comparison of the measures using the criteria. Many of the measures fail for simple test cases which prove non-compliance. For example, \([0, 1, 3, 5]\) is more sparse than \([0, 2, 3, 4]\) because a Robin Hood operation maps one sequence to the other. Six of the measures do not correctly handle this case. Others fail on similar examples. Seven of the measures, however, satisfy \(D_1\). An example for each sparse criterion is given in Table II along with the desired outcome when the sparsity of the examples are measured with sparsity measure \(S(\cdot)\). Table III details which of the six sparse criteria hold for each of the fifteen measures. The information is based on proofs and counter-examples which are contained in their entirety in Appendices A and B. There are essentially two types of proof, Type A and Type B. Type A is the standard form of proof which uses inequalities, an example of which is the following:

**Theorem 4.1:** \(\ell^2_2/\ell^1_1\) satisfies

\[
S\left(\frac{c_1 \cdots c_i - \alpha \cdots c_j + \alpha \cdots}{\ell^2_2/\ell^1_1}\right) < S(\cdot),
\]

![Fig. 1. Component contribution to sparsity measure vs component amplitude.](image1)

**TABLE II**

MOST COMMON COUNTER-EXAMPLE FOR A GIVEN PROPERTY WITH MEASURE OF SPARSITY AND DESIRED OUTCOME WITH SPARSITY MEASURE \(S(\cdot)\).

| Property | Most common counter-example | Desired outcome |
|----------|-----------------------------|-----------------|
| \(D_1\) | [0, 1, 3, 5] vs [0, 2, 3, 4] | \(S([0, 1, 3, 5]) > S([0, 2, 3, 4])\) |
| \(D_2\) | [0, 1, 3, 5] vs [0, 2, 6, 10] | \(S([0, 1, 3, 5]) = S([0, 2, 6, 10])\) |
| \(D_3\) | [1, 3, 5] vs [1.5, 3.5, 5.5] | \(S([1, 3, 5]) < S([1.5, 3.5, 5.5])\) |
| \(D_4\) | [0, 1, 3, 5] vs [0, 0, 1, 1, 3, 5] | \(S([0, 1, 3, 5]) = S([0, 0, 1, 1, 3, 5])\) |
| \(P_1\) | [0, 1, 3, 5] vs [0, 1, 3, 20] | \(S([0, 1, 3, 5]) < S([0, 1, 3, 20])\) |
| \(P_2\) | [0, 1, 3, 5] vs [0, 0, 0, 1, 3, 5] | \(S([0, 1, 3, 5]) < S([0, 0, 0, 1, 3, 5])\) |

![Fig. 2. Percentage of coefficients versus percentage of total coefficient value is plotted for the sorted coefficients for [0 0 0 0 1] (top) and [1 1 2 3 10] (bottom). The Gini Index is twice the shaded area.](image2)
for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:** As $\frac{c^2}{\pi^2} = \frac{\sum_{i} c_i^2}{\sum_{j} c_j}$, we can restate the above as

$$\frac{\sum_{k \neq \ell, j} c_k^2 (c_i - \alpha)^2 + (c_j + \alpha)^2}{\sum_{k} c_k + \alpha - \alpha} < \frac{\sum_{k} c_k^2}{\sum_{k} c_k}.$$ 

This simplifies to

$$(c_i - \alpha)^2 + (c_j + \alpha)^2 < c_i^2 + c_j^2.$$ 

Expand this to get

$$c_i^2 - 2c_i\alpha + \alpha^2 + c_j^2 + 2c_j\alpha + \alpha^2 < c_i^2 + c_j^2$$

which we know is true as $0 < \alpha < \frac{c_i - c_j}{2}$.

A type B proof on the other hand uses derivatives, for example:

**Theorem 4.2:** $-\ell^p$ satisfies

$$S(\left[ c_1 \quad \cdots \quad c_i - \alpha \quad \cdots \quad c_j + \alpha \quad \cdots \right]) < S(\bar{c})_i,$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:**

$$-\ell^p = -\left(\sum_{k} c_k^p\right)^{1/p}, \quad 0 < p < 1.$$ 

We wish to show that the following holds true for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

$$\frac{\partial}{\partial \alpha} \left[ -\left(\sum_{n \neq i,j} c_n^p + (c_i - \alpha)^p + (c_j + \alpha)^p\right)^{1/p}\right] < 0.$$ 

Expand this to get

$$\frac{-1}{p} \left(\sum_{k \neq i,j} c_k^p + (c_i - \alpha)^p + (c_j + \alpha)^p\right)^{\frac{1}{p} - 1} \left(-p(c_i - \alpha)^{p-1} + p(c_j + \alpha)^{p-1}\right) < 0.$$ 

Which holds true if

$$(c_j + \alpha)^{p-1} - (c_i - \alpha)^{p-1} > 0.$$ 

As $p - 1 < 0$ we can rewrite the above as

$$\frac{1}{(c_j + \alpha)^{1-p}} - \frac{1}{(c_i - \alpha)^{1-p}} > 0$$

$$\frac{1}{(c_j + \alpha)} > \frac{1}{(c_i - \alpha)}$$

$$\frac{c_i - \alpha}{c_i - c_j} > \frac{c_j + \alpha}{\alpha},$$

which is necessarily true as it is one of the constraints upon $\alpha$.

From Table III we can see that $D3$ (Rising Tide) is satisfied by most measures. This shows that relative size of coefficients is of the utmost importance when desiring sparsity. As previously mentioned, most measures do not satisfy $D4$ (Cloning). Each of the other criteria is satisfied by a varying number of the fifteen measures of sparsity. This demonstrates the variety of attributes to which measures of sparsity attach importance. $\kappa_4$ and the Hoyer measure satisfy most of the criteria. The Gini Index alone satisfies all six criteria.

| Measure | $D1$ | $D2$ | $D3$ | $D4$ | $P1$ | $P2$ |
|---------|------|------|------|------|------|------|
| $p^r$  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $p^e$  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $-\ell^r$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $-\ell^e$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $\frac{\pi}{\pi^2}$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $-\tan(\alpha, \beta)$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $-\log$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $\kappa_4$ | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $H_C$  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $H_G$  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| $H_s$  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| Hoyer | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |
| Gini  | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    |

### A. Numerical Sparse Analysis

In this section we present the results of using the fifteen sparse measures to measure the sparsity of data drawn from a set of parameterized distributions. We select data sets and distributions for which we can change the ‘sparsity’ by altering a parameter. By applying the fifteen measures to data drawn from these distributions as a function of the parameter, we can visualize the criteria. The examples are based on the premise that all coefficients being equal is the least sparse scenario and all coefficients being zero except one is the most sparse scenario.

In the first experiment we draw a variable number of coefficients from a probability distribution and measure their sparsity. We expect sets of coefficients from the same distribution to have a similar sparsity. As we increase the number of coefficients we expect the measure of sparsity to converge.

In this experiment we examine the sparsity of sets of coefficients from a Poisson distribution (Fig.[3]) with parameter $\lambda = 5$ as a function of set size. From the normalized version of the sparsity plot in Fig.[4] we can see that three measures converge. They are $\kappa_4$, the Hoyer measure and the Gini Index. As this is similar in nature to $D4$ we expect the Gini Index to converge. The convergence of Hoyer measure is unsurprising as this measure almost satisfies $D4$ especially for large $N$.

The results are also normalized for clearer visualization in that they are modified so that the sparsity falls between 0 and 1.

In the second experiment we take coefficients from a Bernoulli distribution where coefficients are either 0 with probability $p$ or 1 with probability $1 - p$. For this experiment the set size remains constant and the probability $p$ varies from 0 to 1. With a low $p$ most coefficients will be 1 and very few zero. The energy distribution of such a set is not sparse and accordingly has a low value (see Fig.[5]). As $p$ increases so should the sparsity measure. We can see this is the case in some form for all of the measures except $H_{ST}$. We note that $\kappa_4$ does not rise steadily with increasing $p$ but rises dramatically as the set approaches its sparsest. This is of some concern if optimizing sparsity using $\kappa_4$ as there is not much indication that the distribution is getting more sparse until its already
Fig. 3. Sample Poisson distribution probability density functions for $\lambda = 5, 10, 15, 30$. We expect the distributions with a ‘narrower’ peak (small $\lambda$) to have a higher sparsity than those with a ‘wider’ peak (large $\lambda$).

Fig. 4. Sparsity of sets of coefficients drawn from a Poisson distribution ($\lambda = 5$) vs the length of the vector of coefficients. The erratically ascending measures are $\ell_0^\epsilon$ and $\ell_0^\rho$. The measures $\ell_1^1$, $\ell_1^p$, $\text{log}$, $\text{Tanh}_{ab}$, $\text{log}$, $\text{H}_G$, $\text{HS}$, $\kappa_4$ and $\ell_p^-$ are grouped in an almost-straight decreasing line. The measures are scaled to be between 0 and 1.
quite sparse.

V. CONCLUSIONS

In this paper we have presented six intuitive attributes of a sparsity measure. Having defined these attributes mathematically, we then compared commonly-used measures of sparsity. The goal of this paper is to provide motivation for selecting a sparsity measure. Having defined these attributes mathematically, we then compared commonly-used measures of sparsity. The measures are scaled to fit between a sparsity range of 0 to 1.

We use these measures to calculate a number which describes the sparsity of a set of coefficients \( \hat{c} = [c_1, c_2, \ldots, c_N] \).

APPENDIX

We use these measures to calculate a number which describes the sparsity of a set of coefficients \( \hat{c} = [c_1, c_2, \ldots, c_N] \).

Note - ignore the trivial cases, for example, \( D2 \) with \( \alpha = 1 \).

**D1 Robin Hood:**

\[
S([c_1, \ldots, c_i - \alpha, \ldots, c_j + \alpha, \ldots]) < S(\hat{c}) \quad \text{for all } \alpha, c_i, c_j \text{ such that } c_i > c_j \text{ and } 0 < \alpha < \frac{c_i - c_j}{2}. 
\]

**D2 Scaling:**

\[
S(\alpha \hat{c}) = S(\hat{c}), \forall \alpha \in \mathbb{R}, \alpha > 0.
\]

**D3 Rising Tide:**

\[
S(\alpha + \hat{c}) < S(\hat{c}), \alpha \in \mathbb{R}, \alpha > 0 \quad \text{(We exclude the case } c_1 = c_2 = c_3 = \cdots = c_i = \cdots \forall i \text{ as this is equivalent to scaling.).}
\]

**D4 Cloning:**

\[
S(\alpha \hat{c}) = S(\hat{c} \parallel \hat{c}) = S(\hat{c} \parallel \hat{c} \parallel \hat{c}) = S(\hat{c} \parallel \hat{c} \parallel \hat{c} \parallel \hat{c}).
\]

**P1 Bill Gates:**

\[
\forall i \exists \beta > 0, \text{ such that } \forall \alpha > 0 : S([c_1, \ldots, c_i + \beta + \alpha, \ldots]) > S([c_1, \ldots, c_i + \beta, \ldots]).
\]

**P2 Babies:**

\[
S(\hat{c} \parallel 0) > S(\hat{c}).
\]

A. Counter-Examples

The most parsimonious method of showing non-compliance with the sparse criteria is through the following simple counter-examples. As an sample we take the \( \ell^1 \) measure and \( D1 \). \( D1 \) states that the \( \ell^1 \) measure of \([0, 1, 3, 5]\) should be greater than the \( \ell^1 \) measure of \([0, 2, 3, 4]\). Using counter example we see that

\[
S([0, 1, 3, 5]) = -9, \quad S([0, 2, 3, 4]) = -9.
\]

Fig. 5. Sparsity vs \( p \) for a Bernoulli distribution with coefficients being 0 with probability \( p \) and 1 otherwise. The measures are scaled to fit between a sparsity range of 0 to 1.


**TABLE IV**

**GUIDE TO COUNTER-EXAMPLES AND PROOFS EACH FOLLOWED BY REFERENCE NUMBER.** A ✓ indicates compliance of the measure with the relevant criterion. 'obv' means that the proof is obvious and as such is not included.

| Measure       | D1     | D2     | D3     | D4     | P1     | P2     |
|---------------|--------|--------|--------|--------|--------|--------|
| $\theta$     | C.Ex A.1 | ✓ obv  | C.Ex A.3 | C.Ex A.4 | C.Ex A.5 | ✓ obv  |
| $\phi_0$     | C.Ex A.1 | C.Ex A.2 | ✓ obv  | C.Ex A.4 | C.Ex A.5 | ✓ obv  |
| $-\ell^1$    | C.Ex A.1 | C.Ex A.2 | ✓ obv  | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $-\ell^p$    | ✓ Proof B1 | C.Ex A.2 | ✓ Proof B2 | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $\tau$       | ✓ Proof B3 | ✓ obv  | Proof B4 | C.Ex A.4 | ✓ Proof B5 | C.Ex A.6 |
| $-\tanh_{a,b}$ | ✓ Proof B6 | C.Ex A.3 | ✓ Proof B7 | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $-\log$      | C.Ex A.1 (*) | ✓ Proof B9 | ✓ Proof B10 | C.Ex A.4 | ✓ Proof B11 | C.Ex A.6 |
| $\kappa_4$   | C.Ex A.1 (*) | ✓ Proof B13 | ✓ Proof B14 | C.Ex A.4 | ✓ Proof B15 | C.Ex A.6 |
| $u_0$        | Proof B18 | ✓ obv  | Proof B19 | ✓ Proof B20 | ✓ Proof B21 | ✓ obv  |
| $-\ell^p_-$  | C.Ex A.1 | C.Ex A.2 | ✓ obv  | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $H_G$        | ✓ Proof B18 | C.Ex A.2 | ✓ obv  | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $H_S$        | C.Ex A.1 | C.Ex A.2 | ✓ C.Ex A.3 (*) | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| $H'_S$       | C.Ex A.1 | C.Ex A.2 | ✓ C.Ex A.3 (*) | C.Ex A.4 | C.Ex A.5 | C.Ex A.6 |
| Hoyer        | ✓ Proof B19 | ✓ obv  | ✓ Proof B20 | C.Ex A.4 | ✓ Proof B21 | ✓ obv  |
| Gini         | ✓ Proof B22 | ✓ Proof B23 | ✓ Proof B24 | ✓ Proof B25 | ✓ Proof B26 | ✓ Proof B27 |
As the Robin Hood operation had no effect on the sparsity of the vectors as measured by the $\ell^p$ measure the measure does not satisfy $D1$. In the case of $-\ell^p$ the zeros in the counter-examples are omitted.

**Counter Example A.1:**

$$[0, 1, 3, 5] \text{ vs } [0, 2, 3, 4]$$

**Counter Example A.1 (**)**

$$[.3, 1, 2] \text{ vs } [.31, .99, 2]$$

**Counter Example A.2:**

$$[0, 1, 3, 5] \text{ vs } [0, 2, 6, 10]$$

**Counter Example A.3:**

$$[1, 3, 5] \text{ vs } [1.5, 3.5, 5.5]$$

**Counter Example A.3 (**)**

$$[.1, 3, 5] \text{ vs } [.15, .35, .55]$$

**Counter Example A.4:**

$$[0, 1, 3, 5] \text{ vs } [0, 0, 1, 3, 5]$$

**Counter Example A.5:**

$$[0, 1, 3, 5] \text{ vs } [0, 1, 3, 20]$$

**Counter Example A.6:**

$$[0, 1, 3, 5] \text{ vs } [0, 0, 0, 1, 3, 5]$$

### B. Proofs

This section contains the proofs that were longer than Table IV permitted. The obvious method of proving that the measures satisfy the criteria, is to plug the formulae for the measures into the mathematical definitions of the six criteria. Another method used below is to differentiate the modified sparse measure with respect to the parameter that modifies it and observe the result. For example if we show that $\frac{\partial S(\alpha+c\ell)}{\partial \alpha} < 0$ for $\alpha > 0$ this proves $D3$ as any change in $\alpha$ causes the measure to drop.

1) $-\ell^p$ and $D1$:

**Theorem A.1:** $-\ell^p$ satisfies

$$S([c_1 \cdots c_i - \alpha \cdots c_j + \alpha \cdots ]) < S(\bar{c})$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:**

$$-\ell^p = -\left(\sum_k c_k^p\right)^{1/p}, \quad 0 < p < 1.$$ 

We wish to show that the following holds true for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$

$$\frac{\partial}{\partial \alpha} \left[-\left(\sum_{n \neq i,j} c_n^p + (c_i - \alpha)^p + (c_j + \alpha)^p\right)\right] < 0.$$ 

$$-\frac{1}{p} \left(\sum_{k \neq i, j} c_k^p + (c_i - \alpha)^p + (c_j + \alpha)^p\right)^{1/p - 1} (p(c_i - \alpha)^{p-1} + p(c_j + \alpha)^{p-1}) < 0.$$ 

Which holds true if

$$(c_j + \alpha)^{p-1} - (c_i - \alpha)^{p-1} > 0.$$ 

As $p - 1 < 0$ we can rewrite the above as

$$\frac{1}{(c_j + \alpha)^{1-p}} - \frac{1}{(c_i - \alpha)^{1-p}} > 0.$$ 

$$\frac{1}{(c_j + \alpha)} > \frac{1}{(c_i - \alpha)}.$$ 

$$\frac{c_i - c_j}{2} > \alpha,$$

which is necessarily true as it is one of the constraints upon $\alpha$. $\blacksquare$

2) $-\ell^p$ and $D3$:

**Theorem A.2:** $-\ell^p$ satisfies

$$S(\alpha + \bar{c}) < S(\bar{c}), \quad \alpha \in \mathbb{R}, \quad \alpha > 0.$$ 

**Proof:**

$$-\left(\sum_{k=1}^N (\alpha + c_k)^p\right)^{1/p} < -\left(N\alpha^p + \sum_{k=1}^N c_k^p\right)^{1/p}$$

$$< -\left(\sum_{k=1}^N c_k^p\right)^{1/p}.$$ 

$\blacksquare$

3) $\ell^2 \forall p$ and $D1$:

**Theorem A.3:** $\ell^2 \forall p$ satisfies

$$S([c_1 \cdots c_i - \alpha \cdots c_j + \alpha \cdots ]) < S(\bar{c})$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:** As $\ell^2 = \sqrt{\sum_{k=1}^N c_k^2}$ we can restate the above as

$$\sqrt{\sum_{k \neq i,j} c_k^2 + (c_i - \alpha)^2 + (c_j + \alpha)^2} < \sqrt{\sum_{k=1}^N c_k^2}.$$ 

This simplifies to

$$\sum_{k \neq i,j} c_k^2 + (c_i - \alpha)^2 + (c_j + \alpha)^2 < \sum_k c_k^2$$

$$(c_i - \alpha)^2 + (c_j + \alpha)^2 < c_i^2 + c_j^2.$$ 

$$c_i^2 - 2c_i\alpha + \alpha^2 + c_j^2 + 2c_j\alpha + \alpha^2 < c_i^2 + c_j^2$$

$$c_j - c_i + \alpha < 0,$$

which we know is true as $0 < \alpha < \frac{c_i - c_j}{2}$. $\blacksquare$
4) \( \frac{\alpha}{\tau} \) and D3:

**Theorem A.4:** \( \frac{\alpha}{\tau} \) does not satisfy

\[
S(\alpha + \vec{c}) < S(\vec{c}), \alpha \in \mathbb{R}, \alpha > 0.
\]

**Proof:**

\[
\frac{\sqrt{\sum_j(\alpha + c_j)^2}}{\sum_j(\alpha + c_j)} = \frac{\sqrt{\sum_j(\alpha^2 + 2c_j\alpha + c_j^2)}}{\sum_j(\alpha + c_j)}.
\]

To simplify matters we make the following substitutions

\[
s_1 = \sum_j c_j
\]

\[
s_2 = \sum_j c_j^2
\]

and note that \( s_1^2 > s_2 \). We now have

\[
\frac{\sqrt{s_2 + 2\alpha s_1 + N\alpha^2}}{s_1 + N\alpha} < \frac{\sqrt{s_2}}{s_1},
\]

\[
s_1^2(s_2 + 2\alpha s_1 + N\alpha^2) < s_2(s_1^2 + 2s_1 N\alpha + N^2\alpha^2)
\]

\[
\alpha < \frac{N}{2s_1} \left( \frac{s_2 - s_1^2}{N s_1^3 - s_2} \right),
\]

which is false as \( \left( \frac{s_2 - s_1^2}{N s_1^3 - s_2} \right) < 0 \) which violates the condition \( \alpha > 0 \).

5) \( \frac{\alpha}{\tau} \) and P1:

**Theorem A.5:** \( \frac{\alpha}{\tau} \) satisfies \( \forall \exists \beta = \beta_1 > 0, \forall \alpha > 0 : S(\alpha) > S(c) \).

**Proof:** We make the following substitutions

\[
s_1 = \sum_j c_j
\]

\[
s_2 = \sum_j c_j^2
\]

and wish to show that

\[
\sqrt{s_2 + \alpha^2 + \beta^2 + 2(\alpha\beta + \alpha c_j + \beta c_j)} < \frac{s_2 + \beta^2 + 2c_j\beta}{s_1 + \beta}.
\]

Squaring both sides and cross-multiplying gives

\[
\alpha > \frac{2 s_1 s_2 + 2 \beta s_1^2 c_j - 2 \beta s_1^2 - 2 c_j s_1^2}{s_1^2 + 2 s_1 \beta - s_2 - 2 \beta c_j}.
\]

We want RHS \( < 0 \) and therefore want a \( \beta \) such that

\[
\frac{2 s_1 s_2 + 2 \beta s_1^2 c_j - 2 \beta s_1^2 - 2 c_j s_1^2}{s_1^2 + 2 s_1 \beta - s_2 - 2 \beta c_j} \leq 0.
\]

As the denominator is always positive, we are only interested in the numerator, that is, finding a \( \beta \) such that

\[
s_1 s_2 + \beta^2 c_j - \beta s_1^2 - c_j s_1^2 \leq 0.
\]

This is satisfied for \( \beta = s_1 \)

\[
s_1 s_2 + s_1^2 c_j - s_1^3 - c_j s_1^2 \leq 0,
\]

which is clearly true.

6) \( -\tanh_{a,b} \) and D1:

**Theorem A.6:** \( -\tanh_{a,b} \) satisfies

\[
S\left( \begin{array}{c} c_1 \\ \ldots \\ c_i - \alpha \\ \ldots \\ c_j + \alpha \\ \ldots \end{array} \right) < S(\vec{c}),
\]

for all \( \alpha, c_i, c_j \) such that \( c_i > c_j \) and \( 0 < \alpha < \frac{c_i - c_j}{2} \).

**Proof:** Need to show that

\[
-\tanh (ac_i - \alpha a)^b - \tanh (ac_j + \alpha a)^b < - \tanh (ac_i)^b - \tanh (ac_j)^b
\]

Making the substitutions \( x = ac_i, y = ac_j \) and \( z = \alpha a \) we get

\[
\tanh (x - y)^b + \tanh (y + z)^b > \tanh (x)^b + \tanh (y)^b
\]

with \( x > y > 0 \) and \( 0 < z < \frac{x - y}{2} \). Setting

\[
f(z) = (\tanh(x - y)^b - \tanh(x)^b) + (\tanh(y + z)^b - \tanh(y)^b),
\]

we use the mean value theorem of differential calculus to prove that

\[
\tanh(x - y)^b - \tanh(x)^b = -zb \left( 1 - \tanh^2(\theta_1)^b \right)
\]

\[
\tanh(y + z)^b - \tanh(y)^b = zb \left( 1 - \tanh^2(\theta_2)^b \right)
\]

where \( x - z < \theta_1 < x \) and \( y < \theta_2 < y + z \). However, because \( 1 - \tanh^2(x)^b \) is strictly decreasing for \( x > 0 \) and \( b > 0 \) because \( z < \frac{x - y}{2} \) \( \iff y + z < x - z \), it follows that

\[
f(z) = zb \left( \left[ 1 - \tanh^2(\theta_2)^b \right] - \left( 1 - \tanh^2(\theta_1)^b \right) \right) > 0.
\]

7) \( -\tanh_{a,b} \) and D3:

**Theorem A.7:** \( -\tanh_{a,b} \) satisfies

\[
S(\alpha + \vec{c}) < S(\vec{c}), \alpha \in \mathbb{R}, \alpha > 0.
\]

**Proof:** It is enough to show that \( \frac{\partial S(\alpha + \vec{c})}{\partial(\alpha + \vec{c})} < 0 \) as if the derivative of the measure with respect to the parameter \( \alpha \) is negative then any \( \alpha \) causes the measure to drop.

\[
\frac{\partial}{\partial \alpha} \left[ -\sum_j \tanh (a \alpha + ac_j)^b \right]
\]

\[
= -\sum_j \left[ 1 - \tanh^2((a \alpha + c_j \alpha)^b) \right] b \left( a \alpha + ac_j \right)^{b-1} \alpha < 0,
\]

which is true as \( a, b > 0 \) and \( \tanh^2 \theta < 1 \).

8) \( -\log \) and D3:

**Proof:** \( -\log \) satisfies

\[
S(\alpha + \vec{c}) < S(\vec{c}), \alpha \in \mathbb{R}, \alpha > 0.
\]

as

\[
-\sum_j \log \left( \frac{1 + (a \alpha + c_j)^2}{1 + c_j^2} \right) > 0
\]

Which is true because

\[
\iff \frac{1 + (a \alpha + c_j)^2}{1 + c_j^2} > 1, \alpha > 0.
\]
9) \(κ_4\) and \(D2\):
Theorem A.8: \(κ_4\) satisfies

\[ S(αc) = S(\hat{c}), \quad ∀\ α ∈ \mathbb{R}, \ α > 0 \]

Proof:

\[ (∑_j (αc)_j)^4 \over (∑_j (αc)_j)^2 ) = α^4 (∑_j c_j^4) \over (∑_j c_j^2)^2 \]

We can ignore the denominator as it is clearly positive. We claim that \(\frac{∂f}{∂α} < 0\) for \(α > 0\). This is because, for positive \(x_i\), it is always true that

\[ ∑_i x_i^2 ∑_i x_i^3 < ∑_i x_i^4 ∑_i x_i \]

as

\[ ∑_i x_i^2 ∑_i x_i^3 - ∑_i x_i^4 ∑_i x_i = ∑_{i≠j} (x_i^2 x_j^3 + x_i^3 x_j^2 - x_i^4 x_j - x_i x_j^4) \]

\[ = ∑_{i≠j} x_i x_j (x_i^2 x_j^3 + x_i^3 x_j^2 - x_i^4 x_j - x_i x_j^4) \]

\[ = - ∑_{i≠j} x_i x_j (x_i - x_j)^2 (x_i + x_j) < 0. \]

10) \(κ_4\) and \(D3\):
Theorem A.9: \(κ_4\) satisfies

\[ S(α + \hat{c}) < S(\hat{c}), \quad α ∈ \mathbb{R}, \ α > 0. \]

Proof: Set

\[ f(α) = \frac{∑_j (c_i + α)^4}{(∑_j (c_i + α)^2)^2} \]

It follows that

\[ \frac{∂f}{∂α} = 4 \frac{][∑_j (c_i + α)^3 ∑_j (c_i + α)^2 - ∑_j (c_i + α)^4 ∑_j (c_i + α)]}{(∑_j (c_i + α)^2)^3} \]

Multiplying out and substituting back in for \(c_i\) this becomes

\[ c_i + α + β > \sqrt{∑_{j≠i} c_j^4 \over ∑_{j≠i} c_j^2} . \]

Clearly there exists a \(β\) such that the above expression holds true for all \(α > 0\).

12) \(-P_2\) and \(P_1\):

Theorem A.11: \(-P_2\) satisfies ∀\(i∃β = β_i > 0\), such that ∀\(α > 0\):

\[ S( [ c_1 \ldots c_i + β + α \ldots ] ) > S( [ c_1 \ldots c_i + β \ldots ] ) . \]

Proof: Without loss of generality we can change the conditions slightly by replacing \(p\) (\(p < 0\)) with \(-p\) and correspondingly update the constraint to \(p > 0\).

\[ - ∑_{j≠i,c_j≠0} c_j − (c_i + β + α)^p > - ∑_{j≠i,c_j≠0} c_j − (c_i + β)^p \]

\[ (c_i + β + α)^p < (c_i + β)^p \]

\[ 1 \over (c_i + β + α)^p < 1 \over (c_i + β)^p , \]

which is true if \(β > 0\).

13) \(u_θ\) and \(D1\):

Theorem A.12: \(u_θ\) does not satisfy

\[ S([ c_1 \ldots c_i - α \ldots c_j + α \ldots ]) < S(\hat{c}) . \]

for all \(α, c_i, c_j\) such that \(c_i > c_j\) and \(0 < α < \frac{c_i - c_j}{2} . \)

Proof: For \(θ = .5 , \)

\[ S([1,2,4,9]) = .6667 \]

\[ S([1,1.9,4,9]) = .7333 \].

The Robin Hood operation increased sparsity and hence does not satisfy \(D1\).

14) \(u_θ\) and \(D3\):

Theorem A.13: \(u_θ\) does not satisfy

\[ S(α + \hat{c}) < S(\hat{c}), \quad α ∈ \mathbb{R}, \ α > 0. \]

Proof: The support of \(c_i\) is \([c_{i(1)}, c_{i(N)}]\). Assume the support of the \([θN]\) points that correspond to the minimum is \([c_{i(k)}, c_{i(j)}]\). By adding a constant, \(α\), to each coefficient in the distribution we shift the distribution to \(c_i + α\). Clearly, neither of the two supports mentioned above changes: \((c_{i(j)} - α) - (c_{i(k)} - α) = c_{i(j)} - c_{i(k)}\). Hence \(u_θ\) does not satisfy \(D3\).}

15) \(u_θ\) and \(D4\):

Theorem A.14: \(u_θ\) satisfies

\[ S(\hat{c}) = S(\hat{c}) | | \hat{c} = S(\hat{c}) | | \hat{c} = S(\hat{c}) | | \hat{c} = S(\hat{c}) | | \hat{c} . \]

Proof: The support of \(c_i\) is \([c_{i(1)}, c_{i(N)}]\). Assume the support of the \([θN]\) points that correspond to the minimum is \([c_{i(k)}, c_{i(j)}]\). The new set \{\(c_i | | \hat{c}\}\} has \(2|Nθ|\) points lying between values \(c_{i(j)}\) and \(c_{i(k)}\), that is, neither of the previously mentioned two supports has changed. This reasoning holds for cloning the data more than once. Hence \(u_θ\) satisfies \(D4\).
16) $u_0$ and $P1$: 

**Theorem A.15:** $u_0$ satisfies $\forall i \exists \beta = \beta_i > 0$, such that $\forall \alpha > 0$:

$$S(\begin{bmatrix} c_1 & \ldots & c_i + \beta + \alpha & \ldots \end{bmatrix}) > S(\begin{bmatrix} c_1 & \ldots & c_i + \beta & \ldots \end{bmatrix}).$$

**Proof:** The support of $\hat{c}$ is $\{c_1, \ldots, c_N\}$. Without loss of generality we focus on $\alpha(c_N)$ as the effect of adding sufficiently large $\beta$ to any other coefficient will result in this coefficient becoming the largest. We choose $\beta$ sufficiently large so that $c(N) + \beta$ is set sufficiently far apart from the other coefficients for the support of $[\hat{\theta}N]$ points that correspond to the minimum not to contain $c(N)$. Consequently, the numerator of the minimization term is a constant $K$ not depending on $\beta$ or $\alpha$. We can rewrite

$$S(\begin{bmatrix} c_1 & \ldots & c_i + \beta + \alpha & \ldots \end{bmatrix}) > S(\begin{bmatrix} c_1 & \ldots & c_i + \beta & \ldots \end{bmatrix})$$

as

$$1 - \frac{K}{c(N) - c(1) + \alpha + \beta} < 1 - \frac{K}{c(N) - c(1) + \alpha}$$

which is clearly true and the proof is complete. 

17) $u_0$ and $P2$: 

**Theorem A.16:** $u_0$ does not satisfy $S(\hat{c})[0] > S(\hat{c})$.

**Proof:** Assume $\hat{c}$ has total support $c(N) - c(1)$ and the support of $[\theta N]$ points lying between values $c(j) - c(k)$. If $0$ lies within the range $c(j) - c(k)$ adding a $0$ will decrease the range to $c(j-1) - c(k)$ without increasing the total support. 

18) $H_G$ and $D1$: 

**Theorem A.17:** $H_G$ satisfies

$$S(\begin{bmatrix} c_1 & \ldots & c_i - \alpha & \ldots & c_j + \alpha & \ldots \end{bmatrix}) < S(\hat{c}),$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:**

$$-\sum_{k \neq i,j} \ln c_k^2 - \ln (c_i - \alpha)^2 - \ln (c_j + \alpha)^2 < -\sum_k \ln c_k^2 - 2 \ln (c_i - \alpha) - 2 \ln (c_j + \alpha) < -2 \ln c_i - 2 \ln c_j (c_i - \alpha)(c_j + \alpha) + c_i c_j$$

$$a < c_i - c_j,$$

which is clearly true.

19) Hoyer and $D1$: 

**Theorem A.18:** Hoyer satisfies

$$S(\begin{bmatrix} c_1 & \ldots & c_i - \alpha & \ldots & c_j + \alpha & \ldots \end{bmatrix}) < S(\hat{c}),$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:**

$$\frac{\partial}{\partial \alpha} \left( \frac{\sqrt{N} - c_i^2}{\sqrt{N} - 1} \right) \equiv \frac{\partial}{\partial \alpha} \left[ \frac{-1}{\sqrt{N} - 1} \left( \sum_{j \neq i} c_j \right) \left( c_i + \alpha + \beta \right) \right],$$

which is

$$\frac{\sum_{j \neq i} c_j}{\sqrt{N} - 1} \left( \frac{\sum_{j \neq i} c_j}{\sum_{j \neq i} \left( c_i + \alpha + \beta \right)^2} \right)^\frac{3}{2}.$$ 

Clearly for sufficiently large $\beta$ the above quantity is $> 0$. 

20) Hoyer and $D3$: 

**Theorem A.19:** Hoyer satisfies $S(\hat{c} + \alpha) < S(\hat{c})$, $\alpha \in \mathbb{R}$, $\alpha > 0$.

**Proof:**

$$\frac{\partial}{\partial \alpha} \left( \frac{\sqrt{N} - \sum_{i=1}^{N} (c_i + \alpha)^2}{\sqrt{N} - 1} \right) \equiv \frac{\partial}{\partial \alpha} \left[ \frac{-1}{\sqrt{N} - 1} \left( \sum_{i=1}^{N} (c_i + \alpha) \left( \sum_{i=1}^{N} (c_i + \alpha)^2 \right)^{-\frac{1}{2}} \right) \right].$$

With the substitution

$$s_1 = \sum_{i=1}^{N} c_i,$n

$$s_2 = \sum_{i=1}^{N} c_i^2,$n

this becomes

$$(s_1 + N \alpha)^2 (s_2 + 2 \alpha s_1 + N \alpha^2)^{-\frac{3}{2}} - N(s_2 + 2 \alpha s_1 + N \alpha^2) < 0,$n

which simplifies to

$$N > \frac{s_1^2}{s_2}.$n

We rewrite this as

$$N s_2 = \sum_{i=1}^{N} \sum_{i=1}^{N} c_i^2 > \left( \sum_{i=1}^{N} c_i \right)^2 = s_1,$$n

which is true by Cauchy-Schwarz. 

21) Hoyer and $P1$: 

**Theorem A.20:** Hoyer satisfies $\forall i \exists \beta = \beta_i > 0$, such that $\forall \alpha > 0$:

$$S(\begin{bmatrix} c_1 & \ldots & c_i + \beta + \alpha & \ldots \end{bmatrix}) > S(\begin{bmatrix} c_1 & \ldots & c_i + \beta & \ldots \end{bmatrix}).$$

**Proof:**

$$\frac{\partial}{\partial \alpha} \left( \frac{\sqrt{N} - c_i^2}{\sqrt{N} - 1} \right) \equiv \frac{\partial}{\partial \alpha} \left[ \frac{-1}{\sqrt{N} - 1} \left( \sum_{j \neq i} c_j \left( c_i + \alpha + \beta \right) - \sum_{j \neq i} c_j^2 \right) \right],$$

which is

$$\frac{\sum_{j \neq i} c_j}{\sqrt{N} - 1} \left( \left( \sum_{j \neq i} c_j \right) (c_i + \alpha + \beta) - \sum_{j \neq i} c_j^2 \right)^\frac{3}{2}.$$ 

Clearly for sufficiently large $\beta$ the above quantity is $> 0$. 

22) Gini and $D1$: 

**Theorem A.21:** The Gini Index satisfies

$$S(c_1, \ldots, c_i - \alpha, \ldots, c_j + \alpha, \ldots) < S(c),$$

for all $\alpha, c_i, c_j$ such that $c_i > c_j$ and $0 < \alpha < \frac{c_i - c_j}{2}$.

**Proof:** The Gini Index of $\hat{c} = \left[ c_1 \ c_2 \ c_3 \ \ldots \right]$ is given by

$$S(\hat{c}) = 1 - 2 \sum_{k=1}^{N} \frac{c(k)}{\|\hat{c}\|_1} \left( \frac{N - k + \frac{1}{2}}{N} \right),$$

where $(k)$ denotes the new index after sorting from lowest to highest, that is, $c(1) \leq c(2) \leq \cdots \leq c(N)$. 

Without loss of generality we can assume that the two coefficients involved in the Robin Hood operation are \(c(i)\) and \(c(j)\). After a Robin Hood operation is performed on \(\vec{c}\) we label the resulting set of coefficients \(\vec{d}\) which are sorted using an index which we denote \([\cdot]\), that is, \(d[1] \leq d[2] \leq \cdots \leq d[N]\). Let us assume that the Robin Hood operation alters the sorted ordering in that the new coefficient obtained by the subtraction of \(\alpha\) from \(c(i)\) has the new rank \(i-n\), that is,

\[
d[i-n] = c(i) - \alpha
\]

and the new coefficient obtained by the addition of \(\alpha\) to \(c(j)\) has the new rank \(j+m\), that is,

\[
d[j+m] = c(j) + \alpha.
\]

The correspondence between the coefficients of \(\vec{c}\) and \(\vec{d}\) is shown in Fig. 6 and in mathematical terms is

\[
\begin{align*}
d[k] &= c(k) & \text{for} & & 1 \leq k \leq j - 1 \\
d[k] &= c(k+1) & \text{for} & & j \leq k \leq j + m - 1 \\
d[k] &= c(j) + \alpha & \text{for} & & k = j + m \\
d[k] &= c(k) & \text{for} & & j + m + 1 \leq k \leq i - n - 1 \\
d[k] &= c(i) - \alpha & \text{for} & & k = i - n \\
d[k] &= c(k-1) & \text{for} & & i - n + 1 \leq k \leq i \\
d[k] &= c(k) & \text{for} & & i + 1 \leq k \leq N.
\end{align*}
\]

We wish to show

\[
S(\vec{d}) > S(\vec{c}).
\]

Removing common terms and noting that \(||\vec{c}||_1 = ||\vec{d}||_1\) we can simplify this to

\[
\sum_{k \in \Delta} c(k) \left( N - k + \frac{1}{2} \right) < \sum_{k \in \Delta} d[k] \left( N - k + \frac{1}{2} \right),
\]

where \(\Delta = \{j, j+1, \ldots, j+m, i-n, i-n+1, \ldots, i\}\). Using the correspondence above we can express the coefficients of \(\vec{d}\) in terms of the coefficients of \(\vec{c}\). We then get

\[
\sum_{k=1}^{N} c(k) \left( N - k + \frac{1}{2} \right) \leq \sum_{k=1}^{N} d[k] \left( N - k + \frac{1}{2} \right)
\]

which becomes

\[
\sum_{k=1}^{m} (c(j+k) - c(j)) + \sum_{k=1}^{n} (c(i) - c(i-k)) + \alpha \left( (i-n) - (j+m) \right) > 0.
\]

This is true as the two summations are positive as the negative component has a lower sorted index than the positive and is hence smaller and the last term is positive due to the condition on \(\alpha\).

23) **Gini and D2:**

**Theorem A.22:** The Gini Index satisfies

\[
S(\alpha \vec{c}) = S(\vec{c}), \forall \alpha \in \mathbb{R}, \alpha > 0.
\]

**Proof:**

\[
S(\alpha \vec{c}) = 1 - 2 \sum_{k=1}^{N} \frac{\alpha c(k)}{||\alpha \vec{c}||_1} \left( \frac{N - k + \frac{1}{2}}{N} \right)
\]

\[
= 1 - 2 \sum_{k=1}^{N} \frac{c(k)}{||\vec{c}||_1} \left( \frac{N - k + \frac{1}{2}}{N} \right) = S(\vec{c}).
\]

\[\blacksquare\]

24) **Gini and D3:**

**Theorem A.23:** The Gini Index satisfies

\[
S(\alpha + \vec{c}) < S(\vec{c}), \alpha \in \mathbb{R}, \alpha > 0.
\]

**Proof:** Rewriting \(S(\alpha + \vec{c}) < S(\vec{c})\) and making the substitution

\[
f(k) = \left( \frac{N - k + \frac{1}{2}}{N} \right),
\]

we get the following:

\[
\sum_{k=1}^{N} \frac{c(k)}{||\vec{c}+\alpha||_1} f(k) + \frac{N \alpha}{||\vec{c}+\alpha||_1} \sum_{k=1}^{N} f(k) - \sum_{k=1}^{N} \frac{c(k)}{||\vec{c}||_1} f(k) > 0
\]

\[
\sum_{k=1}^{N} c(k) f(k) \left( \frac{1}{||\vec{c}+\alpha||_1} - \frac{1}{||\vec{c}||_1} \right) + \frac{N \alpha}{||\vec{c}+\alpha||_1} \sum_{k=1}^{N} f(k) > 0
\]

\[
\sum_{k=1}^{N} c(k) f(k) \left( \frac{-N \alpha}{||\vec{c}+\alpha||_1} \right) + \frac{N \alpha}{||\vec{c}||_1} \sum_{k=1}^{N} f(k) > 0
\]

\[
\sum_{k=1}^{N} f(k) \left( 1 - \frac{c(k)}{||\vec{c}||_1} \right) > 0.
\]

This is clearly true for \(N > 1\).

\[\blacksquare\]

25) **Gini and D4:**

**Theorem A.24:** The Gini Index satisfies

\[
S(\vec{c}) = S(\vec{c}) = S(\vec{c}||\vec{c}||\vec{c}) = S(\vec{c})||\vec{c}||\cdots||\vec{c}).
\]

\[\blacksquare\]
Without loss of generality we have chosen to perform the operation on $\vec{c}_N$ as $\beta$ can absorb the additive value needed to change any of the $c_{(k)}$ to $c_{(N)}$.

We wish to show that $S(\vec{c}) > S(\vec{d})$.

Proof: We use the following notation,

$$\vec{c} = \{c_{(1)}, c_{(2)}, \ldots, c_{(N)} + \beta\}.$$

We can simplify the above to

$$\sum_{i=1}^{N} c_{(i)} \left( \frac{N-i+\frac{1}{2}}{N} \right) > \frac{\beta}{2N(\|\vec{c}\|_1 + \beta)}$$

Hence, the Gini Index satisfies $P1$.

27) Gini and $P2$:

Theorem A.26: The Gini Index satisfies satisfy

$$S(\vec{c}) > S(\vec{d}).$$

Proof: Let us define

$$\vec{d} = \vec{c} - 0 [c_{1} c_{2} c_{3} \cdots c_{N}]$$

and we note that $\|\vec{d}\|_1 = \|\vec{c}\|_1$. Without loss of generality we assign the lowest rank to the added coefficient $0$, that is, $d_{N+1} = d_{(0)}$. We can now make the assertion $d_{(i+1)} = c_{(i)}$, yielding

$$S(\vec{d}) = 1 - 2 \sum_{k=0}^{N} \frac{d_{(k)}}{d} \left( \frac{N+1-k+\frac{1}{2}}{N+1} \right)$$

Making the substitution $i = k - 1$ we get

$$S(\vec{d}) = 1 - 2 \sum_{i=1}^{N} \frac{d_{(i+1)}}{|\vec{d}|} \left( \frac{N+1-i+\frac{1}{2}}{N+1} \right).$$

References

[1] W. Li and J. Preisig, “Estimation of rapidly time-varying sparse channels,” Oceanic Engineering, IEEE Journal of, vol. 32, no. 4, pp. 927–939, Oct. 2007.
[2] R. Adams, Y. Xu, and F. Canning, “Sparse pseudo inverse of the discrete plane wave transform,” Antennas and Propagation, IEEE Transactions on, vol. 56, no. 2, pp. 475–484, Feb. 2008.
[3] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. accepted for publication, 2008.
[4] J. Mairal, M. Elad, and G. Sapiro, “Sparse representation for color image restoration,” Image Processing, IEEE Transactions on, vol. 17, no. 1, pp. 53–69, Jan. 2008.
