Sparsity in social robotics experiments: an abstract view

João Silva Sequeira[1]

Instituto Superior Técnico, University of Lisbon, 1049-001 Lisbon, Portugal,
joao.silva.sequeira@tecnico.ulisboa.pt

Abstract. The paper addresses the use of several performance indexes in the analysis of data acquired in contexts such as social robotics (SR) experiments. The main claim is that the analysis using several correlation/distance indexes can be misleading and data sparsity can reveal the presence of abnormal data (and hence have the potential to influence and conclusions). In general, SR experiments embed multiple sensors and tend to produce huge amounts of data both during short and long time intervals. Furthermore, the social nature often means that observations must be spread in time, making them prone to errors, e.g., perception errors caused by memory fails. In particular, the paper reviews the use of the Gini sparsity index to detect bias, e.g., potentially induced by human factors and causing perception problems. By analyzing sparsity variations one can obtain, for example, models for the degradation of human perception over time, which can be relevant information for SR programming. These ideas are illustrated with simulation experiments, with synthetic data, which (i) allow a simple construction of the experiment and, (ii) to avoid any potential bias that real experiments often, inadvertently, induce in the data. Moreover, despite the social robotics flavour the ideas are applicable to other domains.
Keywords: Sparse Data, Sparsity Indexes, Inequality Indexes, Human-Robot Interaction, Social Robotics, Intelligent Robots

1. Introduction

Social robotics experiments tend to generate relevant data according to the dynamics of the environments they are in. In a typical experiment, data will be collected during one or more time intervals, e.g., by observing robots in a first moment and by recollecting observations/measurements in a second moment, or observing robots in two different moments, or comparing the real observed events with delayed observations. Differences between the observations at different periods will often happen. The observation at a first period may differ from that at a second period, and the correspondence may not even be known, e.g., because they are acquired from sensors of unknown dynamics, thus invalidating the possibility of direct comparison. Following generally accepted models of human memory, perception degradation tends to increase with time and hence it is expectable that variations of the sparsity will occur during experiments spanning large durations (i.e., memory does not get any better with time and/or boredom/tiredness).

Other examples arise when a social robot interacts from time to time with humans, as in a hosting service in a hotel reception, or moving along in wide areas making the robot break interaction for some periods.

In any of these scenarios the dynamics of the observing agent, e.g., a human, is likely to affect the observations, as when delayed answers to questionnaires get affected by memory issues. Similarly, when a social robot emits social clues, e.g., open/close eyes, or issues nonverbal utterances, or moves arms, with abnormal temporal sparsity, this may induce an unwanted perception in human observers.

Analysis of data collected in such experiments often relies in estimating and comparing distributions or measuring correlations, of which the literature describes numerous variants.

Data sparsity has predominantly been studied in the context of communications, [10, 22], medical experiments, [12, 13], and financial markets/transactions, [1, 19]. Sparsity may be given a temporal interpretation (looking to the distribution of data along some timeline), or a spatial interpretation (looking for the variety of data along some range of values).

Processing data acquired in experiments may include filtering, e.g., to remove outliers, to weight the relative importance of the measurements along time (as in ARMAX filters), and/or using sliding temporal windows and constraining any analysis to the data inside the window. Abnormal measurements that are not ruled out as outliers and discarded may affect any indicators used to derive conclusions. However, in the specific case of SR experiments, unfrequent measurements may be ruled out as abnormal and their effect attenuated/removed, e.g., as in averaging processes, may be inconvenient/incorrect and bias conclusions.

The goal of this paper is to compare performance indexes that are common in the SR domain with other indexes, e.g., sparsity indexes, to detect potential potential biasing situations, using simplified simulation scenarios.

The organization is as follows. Section 2 reviews generic aspects related to social robotics experiments that determine the type of data generated. Section 3 describes two simulation experiments, with synthetic data, that can mimic conditions often encountered in real SR experiments. Section 4 discusses the results obtained.

2. Distinctive aspects of SR experiments

The coexistence of humans and robots in the same area and their interdependencies while sharing information naturally lead to the introduction of noise, uncertainties and biases. SR experiments are prone to generate data that can easily resemble noise or outliers and hence filtering approaches may not be adequate. A typical example would be a social robot in a noisy environment that needs to detect and respond to voice calls that happen unfrequently and should not be considered as outlier stimulae because they may be important.

The goal of SR experiments is often related to trend analysis and identification of patterns, e.g., to test the acceptance of a robot by assessing if a population exhibits any evidence of accepting/rejecting the robot. Statistical analysis and model identification, among other techniques, play a relevant role, e.g., analyzing time series, estimating models that can be used for forecast. Trends have been recognized to be, potentially, statistically important (see, for instance, [2] for a study on the quality of time trends). In addition, SR experiments tend to produce data exhibiting a wide range of both spatial and temporal sparsities. Multivariate models and "flexible distributions" (that can be parameterized to mimic the shape of multiple different distributions), [15], and combinations, e.g., linear, of distributions, [21], can be used to model data. In this paper, we argue that in many situations, namely in decision scenarios arising in SR experiments, it is not necessary to have explicit knowledge of the distributions involved. Instead, one can rely on indexes that can discriminate relevant properties of the data.

Furthermore, in SR experiments, if people are reporting results from an experiment, it is also often inadequate to impose (or assume) that they report their data at the specific instants of time, i.e., as soon as they observe it or at regular time instants. Instead, people are given a reasonable time window on which reporting of observations will be accepted. Also, asking people to state the times of the observations may also lead to strong inaccuracies due to potentially incorrect perceptions of time. SR experiments will often produce questionnaire data, hence mostly untimed.

The dynamics of the report timings can affect both the short term and the long term view of the events. Observations with a high concentration of values (many values reported in a short interval) may induce an incorrect perception (a bias) of a dynamic event (that is, an event occurring frequently).

3. Experiments

This section covers hypothetical experiments using synthetic data analyzed according correlation and sparsity indexes, namely L_2-norm, Bhattacharyya coefficient, [3], Spearman and Chronback-α correlations, [5, 14], and Gini index, [6, 18]. There are numerous methods to compare data in the literature. These indexes in this paper cover methodologies involved correlation among datasets, comparison among distributions, and sparsity.

To illustrate the ideas in the paper let us first consider data generated according to different ranges of values observed. The following data structures are considered: (i) a vector r, of size n_r, containing the true events generated by the experiment, and (ii) vectors o_i of size n_{o_i}, containing the corresponding observations. In the following, it is assumed that $n_r = n_{o_i}$, meaning that each event occurrence will have a corresponding value in the observations, not necessarily identical. Also, it is assumed that the meaning of the data is compatible with the indexes, i.e., when comparing r and o_i using the Bhattacharyya coefficient the meaning of the values is that of a relative frequency in a histogram. The goal is to discuss the merits of the indexes agnostic to the meaning of the data.

Furthermore, assume that the events generate both nominal and ordinal data, e.g., as representing the presence/absence of a given movement or quantifying the amplitude of that movement (the former can be in a binary space whereas the latter can be in an ordered set). For ordinal events, the values observed may differ from the reference ones, thus accounting for possible uncertainties. For nominal events, numeric labelling of the outcomes also simplifies the assessment of differences.

Statistics based on the type of events are likely to yield a good picture (assuming that observation fails are small and the observation window is large enough to ensure that any observation fails are not due to the window being too small). If the observation window is too small (which is a possibility if the

data is sparsely generated along time) most statistics will have a significant error and any correlations may be misleading.

Besides the temporal sparsity, spatial sparsity may also affect perception. This is related with the concentration of values associated to an event in the admissible range (many values in a short range may indicate that, overall, there are regions of the full range which have no values observed and hence the spatial sparsity may be high.

Multiple sparsity measures have been used in literature for the analysis of datasets. The survey in [11] selects Gini's index as the most adequate (the only that verifies six, allegedly intuitive, criteria described in their paper). Similarly, [17] also concludes by the superior consistency of the Gini index (with consistency being defined as being scale-invariant and independency of signal energy). In the case of datasets having negative values, data must be transformed to a positive domain in order to preserve the usual interpretation (see for instance [8, 16]), or an extension of the usual definition used, [18].

3.1. Experiment 1

This experiment aims at verifying the effect in multiple indicators of errors in the range of values, i.e., as when the events issued observed are in a wrong range. This mimics SR scenarios in which people report events that did not occurred, i.e., report values that were not generated, or fail to report events that occurred, i.e., the values reported do not include the complete range issued. The methodology is to generate random data and apply the performance indexes.

Figure 1a shows the results of several correlation measures (L_2-norm, Bhattacharyya coefficients, Spearman correlation, Chronback-α correlation, and Gini index), for collections of observations/datasets with 100 samples each, obtained from uniform distributions, $r = U(0,1), o_1 = U(0,1.1), o_{1b} = U(0,1.1), o_{1c} = U(0,1.1), o_2 = U(0,8), o_3 = U(0,12)$, with r representing multiple occurrences of a single event which can have values in $[0,1]$ which is to be observed by 3 independent sensors/observers. The three columns in the figure, o_1, o_2, o_3, stand for the corresponding independent observations. The columns o_{1b} and o_{1c} stand for datasets composed by stacking two and three datasets of similar statistical characteristics of o_1. o_{123} is formed by stacking o_1, o_2, and o_3 in a single dataset (as in a data fusion process).

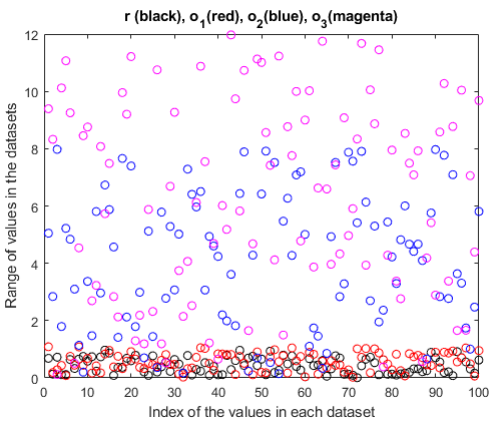
The Spearman rank correlation is used as alternative to the Pearson correlation when (i) there is no missing data, i.e., r and o_i have the same size, (ii) the relation between r and o is not necessarily linear, (iii) their distributions may be non-normal, (iv) data is of ordinal type, and (v) outliers' effects are to be avoided, [4, 7]. As yet another alternative to the Spearman's rank, the Cronhback-α also compute a correlation between elements of two datasets, [14], and provides a reliability/consistency measure between them.

This synthetic scenario mimics, in an abstract way, a SR experiment in which the events characterizing the interaction, i.e., r, are known to the people supervising/analyzing the experiment, and the observations, i.e., $o_1, o_{1b}, o_{1c}, o_2, o_3$ may correspond to the impressions of human observers (e.g., answers to a survey on the interaction with the robot).

No assumption is made on the timings of the observations, i.e., each sensor has its own timing to acquire the data and the order of the values at issue time may not be the same at observation time (though, L_2-norm comparisons implicitly assume that the order of issue-observation is the same, which in SR scenarios may not be adequate). This mimics, for example, situations in which the conscient perception of the observer can have delays leading to a change in the ordering of the observed events.

The L_2-norm shows effectively the differences in the range of each dataset. L_2 and Spearman indicators agree that o_1 is the observer closest to the real events issued, r, though the Spearman correlation means a very weak correlation. The Spearman correlation shows significant differences between o_1 and o_2, o_3 (the interpretation of the numerical value in terms of the correlation strength can be found in papers from multiple areas, see for instance, [20]). Bhattacharyya coeficients suggest that all three datasets have similar statistical properties, and are not significantly different from the "true event" dataset, r, as their values are close to 1. For this particular sample, the Cronbach-α concludes that o_2, o_3 are too close to r, as the values are close to 1, which is clearly not the case, and clearly illustrates how this indicator can be misleading. In what concerns o_1, the negative Chronbach-α correlation leads to the conclusion of abnormalities in the sampling/generation of data, [5]. However, this is clearly an erroneous conclusion and is an example of the limitations of this indicator.

The Gini index shows similar/close values for $r, o_1, o_2, o_3, o_{1b}, o_{1c}$ and was able to capture the difference for o_{123} (o_{1b} and o_{1c} provide a comparison with datasets where each sensor has similar statistical characteristics). Given the relatively small values of the index, the interpretation is that sparsity is not



(a) Raw data

	r, o_1	r, o_2	r, o_3
L_2	4.2285	43.0422	66.9019
Bhattacharyya	0.88117	0.88551	0.88368
Spearman	0.10644	-0.015278	-0.05991
Cronbach-α	-4.6287	0.99566	0.9964

Gini index

r	o_1	o_2	o_3
0.33331	0.3769	0.32668	0.31282

o_{123}	o_{1b}	o_{1c}
0.51367	0.3683	0.34974

(b) Correlations and Gini index

Fig. 1. A random uniform sequence r, of 1D data, with observation patterns of similar sparsities.

high, i.e., each set of observations tends to be evenly distributed along its range. Moreover, the values of the index can be considered close to the reference, with the exception of that of o_{123}.

When several observations are stacked, e.g., as with the stacking o_{123}, of o_1, o_2, and o_3, as to emulate observations reported at different times by different sensors, the Gini sparsity index in Table 1 shows the effect of observations with uncertainty. This result can be expected as o_{123} contains more values than any of o_1, o_2, o_3 but only a small fraction of these will be far from o_1, hence increasing sparsity of o_{123}. Whereas the Gini sparsity index shows a similarity among the observations and with the reference dataset, with the augmented dataset o_{123} shows the presence of dissimilar values (as the Gini index is higher than for isolated datasets). Stacking observations with similar characteristics, e.g., o_{1b} and o_{1c}, amounts to fusing observations from different sensors (possibly for redundancy). In this scenario, the Gini index for the stacking of data with similar statistical nature continues to be close to that of the original dataset, r. Interestingly, testing different stackings using the Gini index may be useful to decide on faulty sensors. Observations of such different quality would have been entirely possible in a real scenario and the Gini index highlights such difference in quality, i.e., as expected, the Gini index detected the bias introduced by the abnormal data from o_2 and o_3. An advantage relative to other indexes is that it is not necessary to use r; by comparing with the indexes (or the actual data) from other observers, it is possible to conclude on the potential existence of a bias.

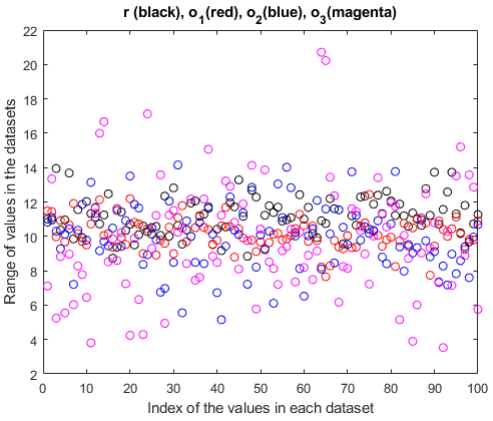
This simple experiment already shows the differences in performance for common indexes. A similar experiment using datasets obtained with normal distributions, also of size 100 samples each, with $r = N(11, 1), o_1 = N(10, 1.1), o_1b = N(10, 1.1), o_1c = N(10, 1.1), o_2 = N(10, 2), o_3 = N(10, 3)$, is reported in Figure 2.

The Bhattacharyya coefficients indicate similarities between the real values and the observed ones. The Gini index clearly points to differences in the datasets, namely those corresponding to o_2 and o_3 and the stacking o_{123}. The negative Spearman correlation indicates opposite directions of growth. The values close to 0 indicate weak correlations, which is a questionable conclusion. The Cronbach-α shows differences between o_3 and the reference, r (as the value is negative), or, at least, some inconsistency in the values.

The Gini index clearly shows differences between o_1 and any other of the o_2, o_3, o_{123}, specially the last ones As in the previous example, this index can be used to detect anomalies/differences among them. Moreover, if the true index is known, i.e., that of r, then Gini index provides a quality measure for the observations relative to real value (instead of only relative quality measurements). In the case reported in Figure 2, it is clear that o_1 is closer to r than the other observations.

This first series of experiments can be identified with multiple meaningful SR scenarios. As possible examples, the values in each o_i may represent the times between observations of a collection of events, or a set of distances obtained by a Lidar, or a set of temperatures measured in different points in space, or pixel values in an image.

Note that the ordering of the values in each observation does not affect the corresponding Gini index, i.e., the index is invariant to random permutations (see [9]), meaning that the Gini index can be used as



(a) Raw data

	r, o_1	r, o_2	r, o_3
L_2	14.044	20.9112	30.297
Bhattac.	0.99756	0.99462	0.98651
Spearman	0.1324	-0.0397	0.1624
Cronbach-α	0.48307	0.54032	-0.35514

Gini index

r	o_1	o_2	o_3
0.0562	0.0602	0.0991	0.158

o_{123}	o_{1b}	o_{1c}
0.109	0.062	0.0620

(b) Correlations and Gini index

Fig. 2. A random normal sequence r of 1D data, with observation patterns with varying sparsities.

indicator for missing/excess data (exact index in case of zero observation error). This can also be used to identify a rough memory model for the observers, i.e., as when an observer fails to report all data.

3.2. Experiment 2

The previous experiment shows how differences in the values of synthetic data, hypothetically resulting from an SR experiment, affect the indexes considered. This experiment aims at assessing effects of temporal sparsity, namely, as resulting from changes in the order of observations and of the temporal horizon used to analyze the data.

Changes in the order of observations amount to a permutation of the reference and hence Gini index is useless to assess the range sparsity. However, if the analysis of data is based on a temporal window, hence not using all the information, the Gini index may still be relevant.

In terms of SR experiments, human perception has its own dynamics, i.e., memories from some observation fade as time evolves, and hence as people may take longer to report an observation that may fall outside the observation temporal window (thus making the observations sparser). Therefore, a variation in sparsity in the observations of a given time window relative to the original sequence of events provides information about the behaviors/dynamics of the observers.

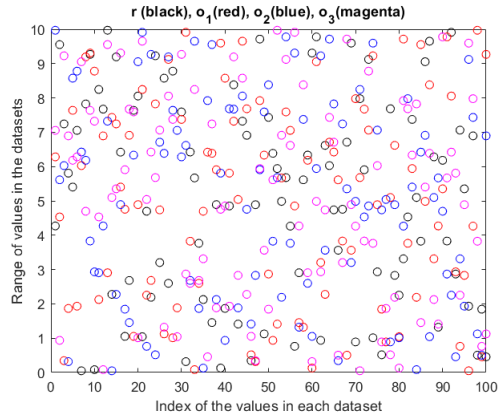
Figure 3 illustrates a scenario in which only temporal sparsity changes relative to the reference, i.e., the observations have no errors relative to the reference but they are reported at arbitrarily different times or by an arbitrarily different order and a sequence of windows is used to analyze the data. Three observers, plotted in red, magenta, and green, are shown, with the reference plotted in blue. The indices for each observation are computed with the windows moving over the whole dataset and concatenated for display as single plots.

The last window (the righthand side of the plot) amounts to the whole dataset. As the windows enlarge, the corresponding indexes approach the values of whole dataset. This clearly shows (i) a potential effect of an inadequate window size, and (ii) different permutations of the original/reference dataset may exhibit ripples (substantially different from the true values). In particular, for small size windows, some indexes, namely Spearman correlation, may thus lead to incorrect decisions. Bhattacharyya coefficients and Gini index can be considerably less affected.

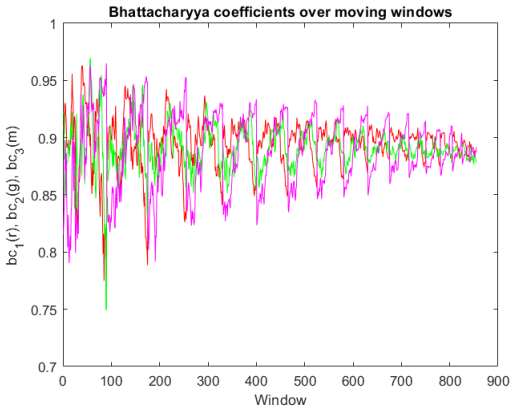
4. Conclusions and future work

Using social robotics as motivational domain, the paper showed, through simple experiments, using synthetic data agnostic to the data types, that it is easy to obtain flawed conclusions when using common correlation/distance indexes. Besides social robotics experiments, e.g., involving duration of the interactions, assessing the results of Likert questionnaires, and the dynamics of specific events identified from sensors onboard the robots, the ideas discussed in the paper are applicable to other domains, e.g., generic sensor data processing and fusion.

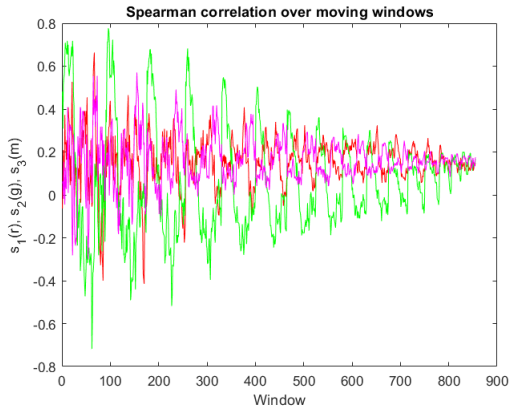
SR experiments have specific characteristics, that may rend analysis challenging. For example, the observers may be assumed independent of each other, though it is entirely possible that word-of-mouth phenomena exists, e.g., with one observer confirming his/her own beliefs about some specific observations



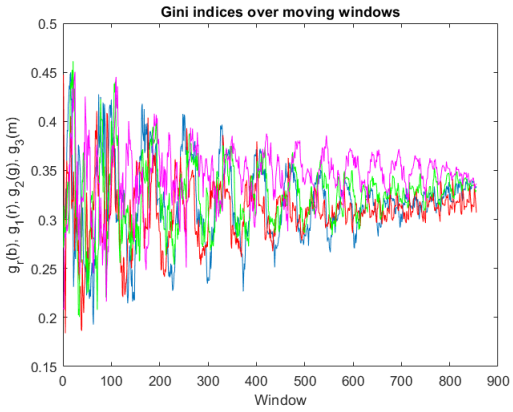
(a) Raw data



(b) Bhattacharyya coefficient



(c) Spearman correlation



(d) Gini index

Fig. 3. A random uniform sequence r of 1D data, with observations differing by a random permutation – moving windows of size between 10 and 100 samples in increments of 5 instants.

with other observer and hence, possibly, introducing biases (this is the so called "confirmation bias"). The numerous statistical analysis techniques available may also make the selection of an adequate one a difficult problem (with the literature advising the use of multiple techniques in the same problem). In SR experiments it is common not to know in advance the distributions of the events generated/observed and hence using analysis via distances between distributions may require additional estimation steps, which this paper tried to avoid. The paper focus on the Gini index, Bhattacharyya coefficients, Spearman correlation, L_2 norm, and Cronbach-α, these last ones being more common in analysis of SR experiments can serve as comparison to the Gini and Bhattacharyya.

In the experiments, sparsity indexes, namely the Gini index, and Bhattacharyya coefficients showed a consistent behavior (in the sense that the ranges spawned do not disturb assessments, e.g., Bhattacharyya coefficients stays in a range to which a common assessment can be assigned, in the case of Figure 3b the observations assigned some degree of similarity.

Future work will include other inequality/sparsity indexes, e.g., Hoyer, pq-means (though the literature reports a superior performance by the Gini index, [11]).

Acknowledgements
This work was supported by LARSyS FCT funding (DOI: 10.54499/LA/P/0083/2020, 10.54499/UIDP/50009/2020, and 10.54499/UIDB/50009/2020).

References

1. Mert Akyuz, Ghislain Nono Gueye, and Cagin Karul. Long-run dynamics between trade liberalization and income inequality in the European Union: a second generation approach. *Empirica*, 49(5), 2022.
2. Andreas C. Bryhn and Peter H. Dimberg. An Operational Definition of a Statistically Meaningful Trend. *PLoS ONE*, 6(4:e19241), 2011.
3. Sung-Hyuk Cha and Sargur N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35:1355–1370, 2002.
4. Trisha Chandra. How to measure the relationship between variables – An attempt to explain covariance and correlation in the simplestform, 2021.
5. Eunseong Cho and Seonghoon Kim. Cronbach's Coefficient Alpha: Well Known but Poorly Understood. *Journal of Organizational Research Methods*, 18(2):207–230, April 2015.
6. Martin Čihák and Ratna Sahay. *Finance and Inequality*. International Monetary Fund, January 2020. In collaboration with Adolfo Barajas, Shiyuan Chen, Armand Fouejieu, and Peichu Xie; IMF Discussion Note.
7. Gina M. D'Angelo, Jingqin Luo, and Chengjie Xiong. Missing Data Methods for Partial Correlations. *Journal Biometrics and Biostatistics*, 3(8), 2013.
8. Francesca De Battisti, Francesco Porro, and Achille Vernizzi. The Gini coefficient and the case of negative values. *Electronic Journal of Applied Statistical Analysis*, 12(1):85–107, 2019.
9. Etienne Billette de Villemeur and Justin Leroux. Assessing Inequality Assessments: A General Representation of Inequality Indices, 2021.
10. Swati Goswami, C.A. Murthy, and Asit Kumar Das. Sparsity Measure of a Network Graph: Gini Index. *Information Sciences*, 462, December 2016.
11. Niall Hurley and Scott Rickard. Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, 55(10):3723–4741, November 2009.
12. Gustav Kjellsson and Ulf-G. Gerdtham. Measuring Health Inequalities Using the Concentration Index Approach. In Anthony J. Culyer, editor, *Encyclopedia of Health Economics*. Elsevier, December 2014.
13. Jonathan Levy, Susan M. Chemerynski, and Jessica Leibler. Incorporating concepts of inequality into health benefits analysis. *International Journal for Equity in Health*, 5(1), February 2006.
14. Mike Lopez. Estimation of Cronbach's alpha for sparse datasets. In Samuel Mann and Noel Bridgeman, editors, *20th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ 2007)*. 2007. Nelson, New Zealand.
15. Mark McDonald, Kais Zaman, and Sankaran Mahadevan. Probabilistic Analysis with Sparse Data. *American Institute of Aeronautics and Astronautics Journal*, 51(2), February 2013.
16. Katarzyna Ostasiewicz and Achille Vernizzi. Decomposition of the Gini Index in the Presence of Observations With Negative Values. *Mathematical Economics*, 14(21), 2018.
17. Y. Parkale and S. Nalbalwar. Investigation on 1-D and 2-D Signal Sparsity Using the Gini Index, L1-Norm and L2-Norm for the Best Sparsity Basis Selection. In B. Iyer, S. Nalbalwar, and R.Pawade, editors, *ICCASP/ICMMD-2016. Advances in Intelligent Systems Research*, volume 137, pages 642–651. Atlantis Press, 2016.
18. Emanuela Raffinetti, Elena Siletti, and Achille Vernizzi. On the Gini coefficient normalization when attributes with negative values are considered. *Statistical Methods & Applications*, 24(3), September 2015.
19. Ewa Weychert. Financial development and income inequality. *Central European Economic Journal*, 7(54):84–100, October 2020.

20. Zhihong Yan, Shuqian Wang, Ding Ma, Bin Liu, Hong Lin, and Su Li. Meteorological Factors Affecting Pan Evaporation in the Haihe River Basin and China. *Water*, 11(2):317–335, February 2019.
21. Xiaowei Yang, Huiming Zhang, Haoyu Wei, and Shouzheng Zhang. Sparse Density Estimation with Measurement Errors. *Entropy*, 24, 2022.
22. Dornoosh Zonoobi, Ashraf Kassim, and Yedatore V. Venkatesh. Gini Index as Sparsity Measure for Signal Reconstruction from Compressive Samples. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):927–932, October 2011.