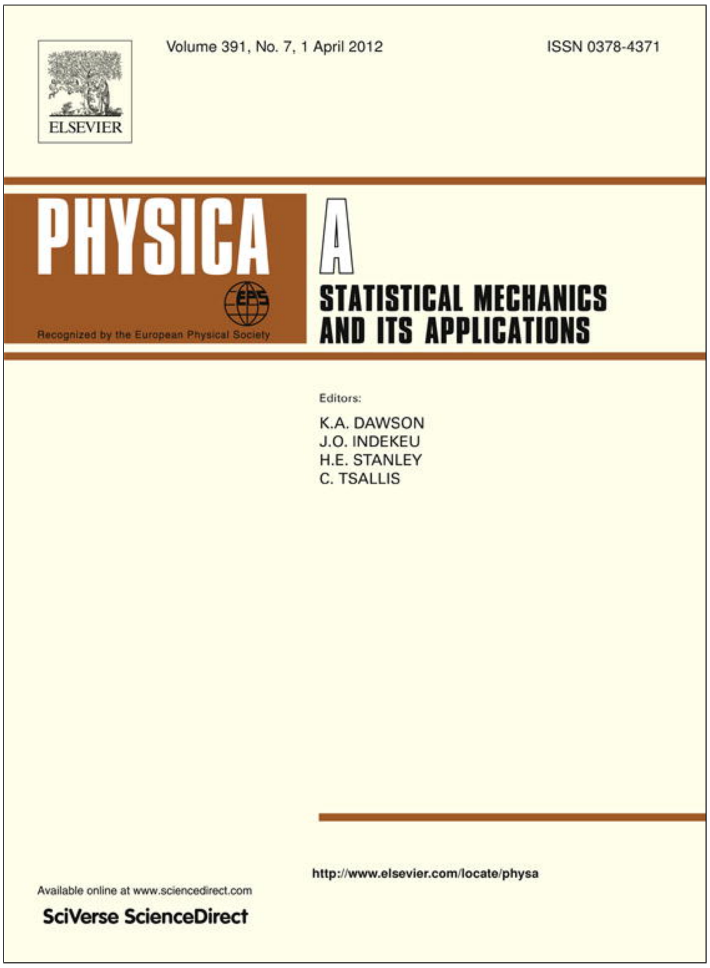

Volume 391, No. 7, 1 April 2012

ISSN 0378-4371

ELSEVIER

PHYSICA A

Recognized by the European Physical Society

STATISTICAL MECHANICS
AND ITS APPLICATIONS

http://www.elsevier.com/locate/physa

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physica A 391 (2012) 2401–2407

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Large-scale data analysis using the Wigner function

R.A. Earnshaw [a], C. Lei [a], J. Li [b], S. Mugassabi [a], A. Vourdas [a,*]

[a] School of Computing, Informatics and Media, University of Bradford, Bradford BD7 1DP, United Kingdom

[b] School of Management, University of Bradford, Bradford BD9 4JL, United Kingdom

ARTICLE INFO

Article history:
Received 6 October 2011
Received in revised form 28 November 2011
Available online 14 December 2011

Keywords:
Wigner function
Data analysis

ABSTRACT

Large-scale data are analysed using the Wigner function. It is shown that the 'frequency variable' provides important information, which is lost with other techniques. The method is applied to 'sentiment analysis' in data from social networks and also to financial data.

1. Introduction

The Wigner function and more generally time–frequency methods have been studied in a mathematical, physical and engineering (especially signal processing) context for a long time (e.g., Refs. [1–4]). In this paper we apply these techniques for fast analysis and compression of the information in large-scale data. There exists much work (e.g., Refs. [5,6]) on the compression of information in large-scale data. In time–frequency analysis to each variable corresponds a 'frequency variable' which provides very important information, which is lost with other techniques.

Our first example is a sentiment study of information in the media and the internet. Fast analysis of information from social networks (Twitter) about current affairs is important for the government, for political commentators; etc. Here we present a quantitative analysis of information in the internet about the BP oil spill in the USA. Our second example is in the area of accounting and finance discipline. We study different disclosure practices across companies, using the data in Refs. [7,8].

2. Methodology

The first step is to describe the data with a real and positive function $F(x)$ of a variable x which takes values $x = 1, \ldots, N$. In the first example x describes the location of the information (each integer value of x represents an internet address) and $F(x)$ the number of comments in the address x about the BP oil spill. A criterion is required here for the ordering of the addresses (e.g., the geographical location). In the second example x is the size of the company (see below) and $F(x)$ the disclosure of information.

The time t can be included in the function, but for simplicity we have not studied how the information changes as a function of time. Also we consider functions of one variable, but the use of more variables can give a more accurate description of the information (e.g., another variable y can quantify the strength of the sentiment about the BP oil spill).

[*] Corresponding author. Tel.: +44 1274 233950.
E-mail address: A.Vourdas@Bradford.ac.uk (A. Vourdas).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.11.060

We first study whether the issue that we consider is a local or a global one. We quantify this with the entropy

$$I = -\sum_{x=1}^{N} p(x) \log p(x); \quad p(x) = \frac{F(x)}{\sum_x F(x)}. \tag{1}$$

In the calculations we used logarithms with base 2 and the results are in bits.

The maximum value of the entropy is $\log N$. The quantity

$$G = \frac{I}{\log N}; \quad 0 \le G \le 1 \tag{2}$$

can be used as a measure of the local or global nature of $p(x)$. Values of G close to 0 indicate that the issue that we study is a local one, and values of G close to 1 indicate that it is a global one.

The real and positive function $F(x)$ is defined above for integer values of x, within a finite interval of the real line. With interpolation and with the assumption that outside this finite interval $F(x)$ goes very fast to zero (or alternatively that $F(x)$ is a periodic function of x), it can be defined for all real values of x. Also $F(x)$ is a real function, but it will be convenient to use the well known Hilbert transform (in signal analysis it is called 'analytic signal') to convert it into a complex function $f(x)$:

$$f(x) = \frac{1}{\pi} \mathrm{PV} \int_{-\infty}^{\infty} \frac{F(x')}{x - x'} \mathrm{d}x'. \tag{3}$$

Here 'PV' indicates the principal value. Both $f(x)$ and $F(x)$ contain exactly the same information, but it is easier to apply our method to the complex function $f(x)$. The Hilbert transform is available in all computer libraries (e.g., in MATLAB).

The Fourier transform of $f(x)$ is the function

$$\tilde{f}(\nu_x) = \int \mathrm{d}x f(x) \exp(-ix\nu_x) \tag{4}$$

where ν_x is a 'spatial frequency' corresponding to the variable x. This frequency plays a central role in our analysis. If the function $\tilde{f}(\nu_x)$ is significant only at low (high) values of ν_x, this means that the function $f(x)$ varies slowly (quickly) as a function of x.

We study the Wigner function

$$W(x, \nu_x) = \int_{-\infty}^{\infty} \mathrm{d}a f(x - a) f^*(x + a) \exp(i2a\nu_x) \tag{5}$$

corresponding to the function $f(x)$. Areas of the $x - \nu_x$ plane where the Wigner function is large, indicate strong activity related to the information we are studying in the corresponding addresses and with the corresponding frequencies.

The marginal properties of the Wigner function are

$$\int \mathrm{d}x W(x, \nu_x) = |\tilde{f}(\nu_x)|^2$$
$$\int \mathrm{d}\nu_x W(x, \nu_x) = \pi |f(x)|^2. \tag{6}$$

Below we plot the $|f(x)|$ and $|\tilde{f}(\nu_x)|$ for the examples that we consider, and we use them to check that our Wigner functions obey these marginal properties.

3. Application to data from social networks

For the first example we searched for 'BP oil spill in USA' in Google news. We have used 19 741 articles on this topic in 37 different websites (which we have ordered alphabetically with $x = 1, 2, \ldots, 37$). These articles appeared in the two month period (December 2010–January 2011). For simplicity we do not consider how the data change as a function of time. For these data we calculated that $G = 0.71$.

From these data we constructed the real function $F(x)$ for $x = 1, 2, \ldots, 37$. With interpolation we constructed the function $F(x)$ for all real values in the interval $(1, 37)$. As we have already mentioned earlier, for the remainder values of x, we assume that the function is equal to zero. We then used the Hilbert transform to obtain the complex function $f(x)$. Results for $|f(x)|$ and $|\tilde{f}(\nu_x)|$ are given in Figs. 1, 2, correspondingly.

Using $f(x)$ we calculated the Wigner function using Eq. (2). A contour diagram of the Wigner function is shown in Fig. 3 (the continuous lines correspond to $W(x, \nu_x) = 80$ and the broken lines to $W(x, \nu_x) = -8$). The results indicate that there were many articles about the BP oil spill in the addresses labelled with $0 \le x \le 17$ and also in the addresses labelled with $27 \le x \le 37$. The Wigner function takes high values at the high frequencies $12 \le \nu_x \le 22$. This indicates a non-uniform

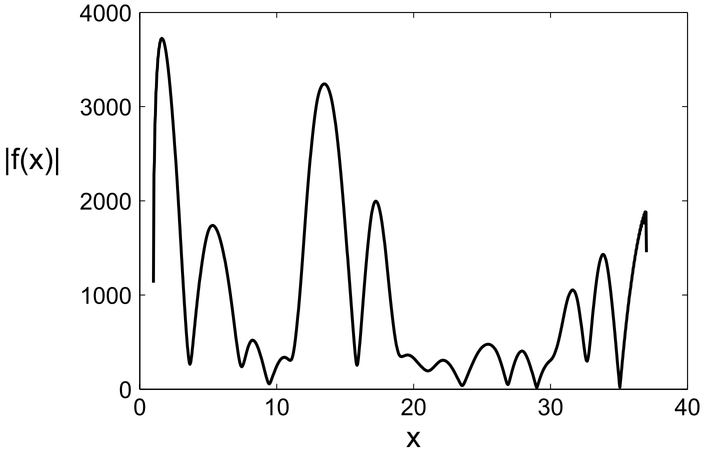

Fig. 1. $|f(x)|$ for the first example.

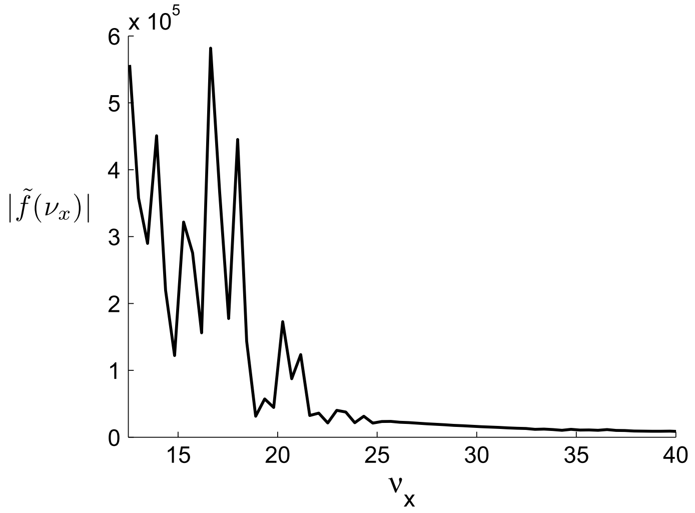

Fig. 2. $|\tilde{f}(\nu_x)|$ for the first example.

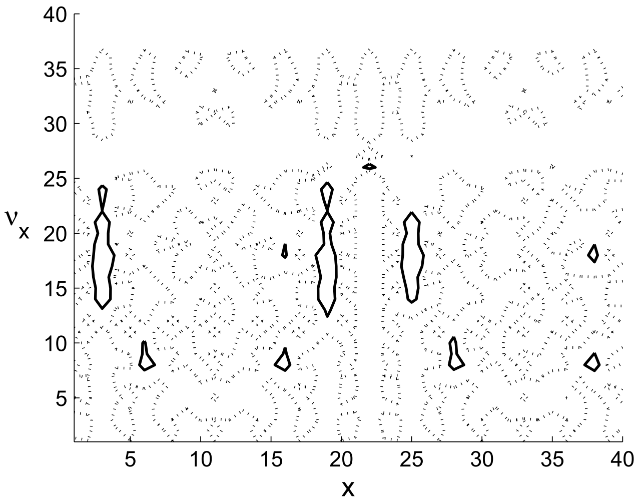

Fig. 3. A contour plot of the Wigner function $W(x, \nu_x)$ for the first example, showing the levels $W(x, \nu_x) = 80$ (continuous line) and $W(x, \nu_x) = -8$ (broken line).

(oscillatory) distribution of the articles in the various addresses. The merit of our approach is precisely the fact that the frequency can detect uniform or non-uniform behaviour in a neighbourhood of *x* (which here are websites).

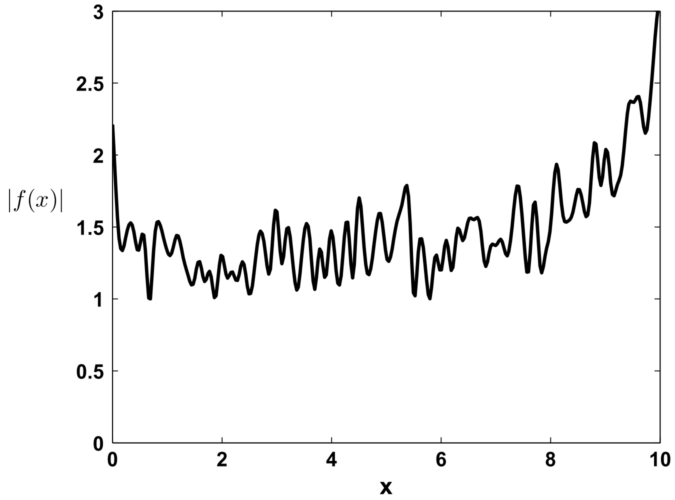

Fig. 4. $|f(x)|$ for the second example.

The regions with negative values are a well known feature of the Wigner function related to interference. Interference is related to many peaks in the function $F(x)$, which in the present context indicates a non-uniform (oscillatory) distribution of the articles in the various addresses.

4. Application to financial data

The Wigner function can also be usefully applied to the accounting and finance discipline. Many studies have examined the information provided in the annual report and various other communication channels using content analysis, with an aim to understand what and how companies communicate with its information users, such as shareholders, potential investors, fund managers and analysts. Findings from previous studies that examine the relationship between explanatory factors and various types of disclosure help shareholders and other groups of information users as well as the regulatory bodies to identify factors that may encourage disclosure in the annual report.

One of the limitations of the existing disclosure studies is that the small sample size restricts our ability to run the regression analysis by splitting the sample into smaller groups, such as by size and industry type. Splitting companies into smaller groups allows better analysis of the variations in disclosure practice within each group of companies. Whilst this is difficult to achieve using the regression analysis, given the limitation in sample size, the Wigner function allows variations in disclosure practice within different groups of companies to be identified. Findings from such analysis have policy implications. They could help information users and regulatory bodies identify companies that require greater regulation in order to protect the interest of various groups of investors, to determine resource allocation by identifying the type of companies that require greater assistance.

In this study, we use intellectual capital (IC) disclosure as an example to demonstrate how the Wigner function can be applied to identify the type of companies that have greater variation in their IC disclosure practice in the annual report. Refs. [6,7] examined the extent of IC disclosure in the annual report of 100 London Stock Exchange (LSE) listed UK companies and found a significant positive association between the extent of IC disclosure and company size. The finding is consistent with those of previous studies. However, the association is a general trend among the 100 companies. It is unclear whether the variations are greater or smaller among the larger or smaller companies.

We used the same 100 sample companies as those in Refs. [6,7] and the IC disclosure index scores obtained there. The 100 companies are ranked by its size, using market capitalisation (we refer to this as second example) and company sales (we refer to this as third example). The parameter G takes the values 0.99 and 0.99 in the second and third examples, correspondingly. This indicates the global nature of the problem.

In Figs. 4–6 we present the $|f(x)|$, the $|\tilde{f}(\nu_x)|$ and the Wigner function, for the second example. In Figs. 7–9 we present the $|f(x)|$, the $|\tilde{f}(\nu_x)|$ and the Wigner function for the third example. In Figs. 6, 9 the continuous lines correspond to $W(x, \nu_x) = 3.8$ and the broken lines to $W(x, \nu_x) = -2$. The IC disclosure index scores are compressed into a complex function $f(x)$ (using the methodology described earlier). The variable x describes the 100 companies ranked from small to large, according to two proxies. Each value represents a company. As is shown in Figs. 6, 9, IC disclosure is found to vary more significantly for companies that fall within the size range of 45–55 (i.e. market capitalisation of £339–615 million, and sales of £272–519 million) and also 80–100 (i.e. market capitalisation of £3287–98 258 million, and sales of £3174–39 792 million) compared to the other companies. It is interesting that although the same companies have been ordered in a different way in Figs. 6, 9 (according to market capitalisation and sales) the results are consistent with each other. The findings suggest that attention is needed for large companies (and also some medium size companies) to help achieve a more standardised IC reporting practice.

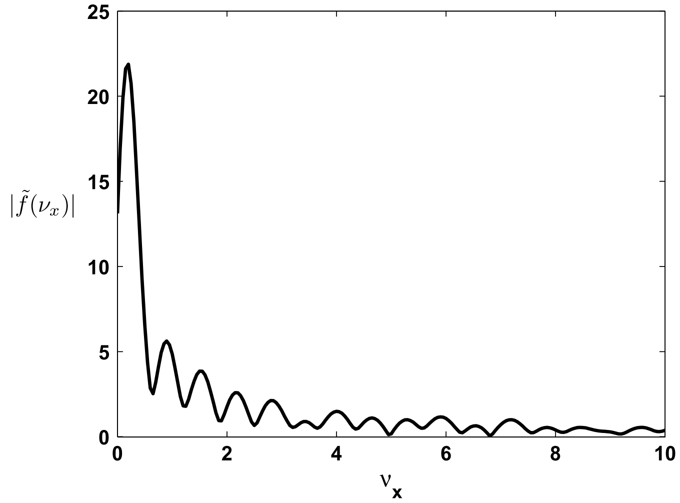

Fig. 5. $|\tilde{f}(\nu_x)|$ for the second example.

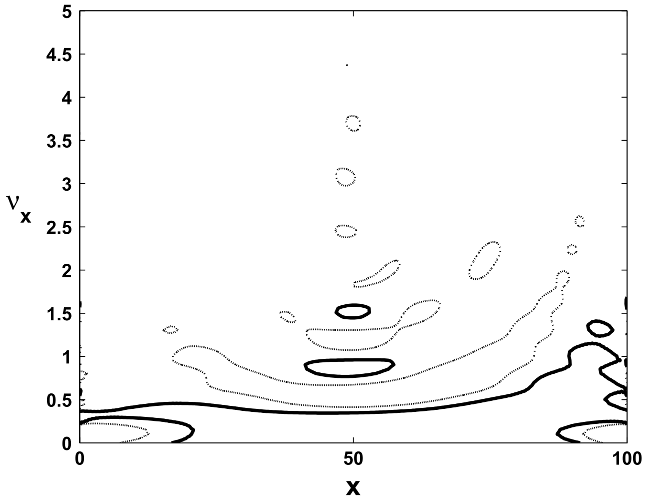

Fig. 6. A contour plot of the Wigner function $W(x, \nu_x)$ for the second example, showing the levels $W(x, \nu_x) = 3.8$ (continuous line) and $W(x, \nu_x) = -2$ (broken line).

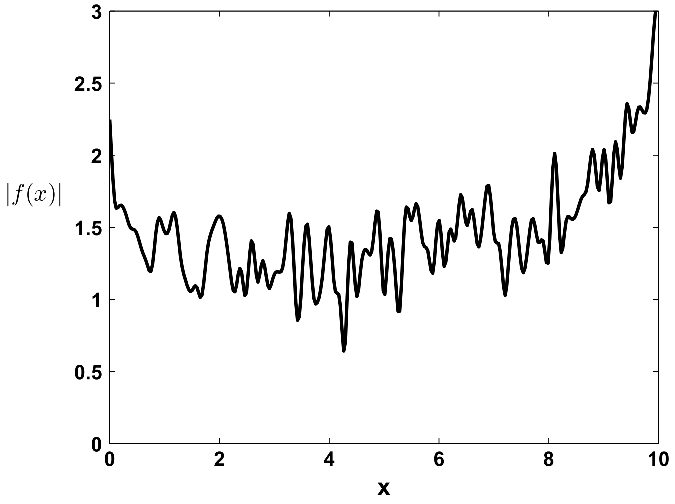

Fig. 7. $|f(x)|$ for the third example.

5. Discussion

Given the rapidly increasing size and complexity of data sets produced by scientific experiments or the observations of natural, social, and financial phenomena, the volume of data to be analysed in the future will present major challenges.

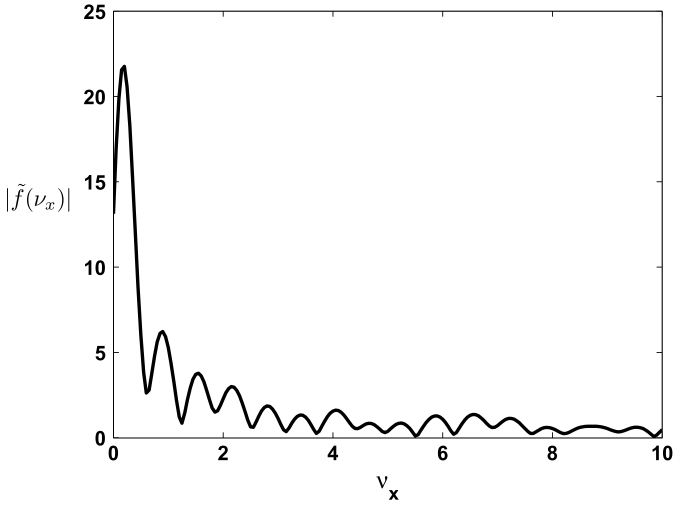

Fig. 8. $|\tilde{f}(\nu_x)|$ for the third example.

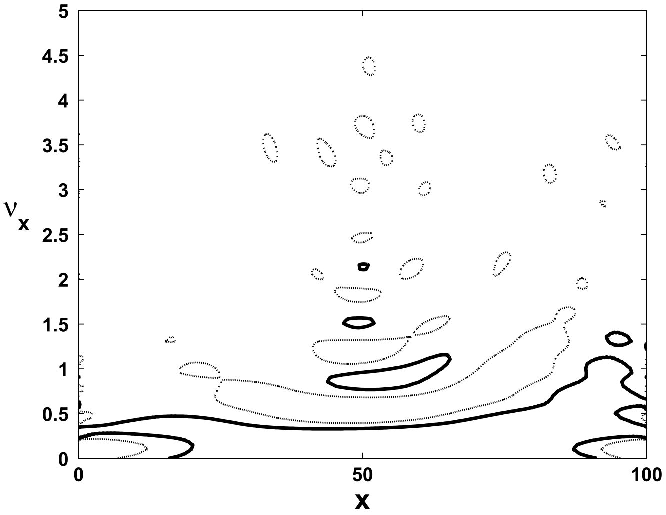

Fig. 9. A contour plot of the Wigner function $W(x, \nu_x)$ for the third example, showing the levels $W(x, \nu_x) = 3.8$ (continuous line) and $W(x, \nu_x) = -2$ (broken line).

The Wigner function has been used extensively in mathematics, physics and engineering. Use of the Wigner function in the present context enables meaningful information to be extracted from large-scale data in near real time. In particular, the frequency variable describes how uniform the behaviour is in the neighbourhood of a variable.

We have considered data as a function of one variable x (and the corresponding frequency ν_x). More variables can be used for a more accurate description of the information. For example, in the BP oil spill case, a second variable y which takes the values $1, \ldots, M$ can quantify the degree of opposition against the event. In this case the Wigner function will be $W(x, y, \nu_x, \nu_y)$. In addition to that the time variable can also be included, indicating how sentiment changes as a function of time.

There is a plethora of potential applications of our approach. In addition to the examples that we discussed, another important application is commercial (or national) security. For example we can count key-words in emails from employees in a company, and look for anomalies outside the normal pattern. The frequency variable does precisely that, by detecting non-uniform behaviour among similar employees. Therefore the Wigner function approach can be very useful in this context.

There are other distributions related to the Wigner function, which can also be used for the extraction of information from large data. For example the ambiguity or characteristic or Weyl function (e.g., Refs. [1–3]) is related through a two-dimensional Fourier transform to the Wigner function. Both of these functions are related to the moments of the distribution of the data. The first two moments (which give the average and standard deviation) and also the other higher moments, have been used extensively for data analysis (especially for financial data). The aim of this paper is to show that in the context of data analysis, the Wigner function through the frequency variable, reveals information which is not easily seen through the moments.

In summary, the Wigner function approach is a powerful technique for extracting information from large-scale data.

References

[1] L. Cohen, Time Frequency Analysis, Prentice Hall, London, 1995.
[2] R. Tolimieri, M. An, Time Frequency Representations, Birkhäuser, Boston, 1998.
[3] K. Grohening, Foundations of Time–Frequency Analysis, Birkhäuser, Boston, 2001.
[4] S. Chountasis, A. Vourdas, Weyl functions and their use in the study of quantum interference, Physical Review A 58 (1998) 848.
[5] K.A. Cook, R.A. Earnshaw, J. Stasko, Discovering the unexpected, IEEE Computer Graphics and Applications 27 (2007) 15–19.
[6] R.A. Earnshaw, R.A. Guedj, A. van Dam, J.A. Vince (Eds.), Frontiers of Human-Centered Computing, Online Communities and Virtual Environments, Springer Verlag, 2001.
[7] J. Li, An investigation of intellectual capital disclosure in annual reports of UK firms: practices and determinants, Ph.D. Thesis, University of Bradford School of Management, 2009.
[8] J. Li, R. Pike, R. Haniffa, Intellectual capital and corporate governance structure in UK firms, Accounting and Business Research 38 (2) (2008) 139–157.