INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 6375–6380 PII: S0305-4470(01)25737-0

A measure of data collapse for scaling

Somendra M Bhattacharjee[1,2] **and Flavio Seno**[2]

[1] Institute of Physics, Bhubaneswar 751005, India
[2] INFM and Dipartimento di Fisica, Università di Padova, Via Marzolo 8, 35131 Padova, Italy

E-mail: somen@iopb.res.in and flavio.seno@pd.infn.it

Received 6 June 2001
Published 10 August 2001
Online at stacks.iop.org/JPhysA/34/6375

Abstract

Data collapse is a way of establishing scaling and extracting associated exponents in problems showing self-similar or self-affine characteristics as, for example, in equilibrium or non-equilibrium phase transitions, in critical phases, in dynamics of complex systems and many others. We propose a measure to quantify the nature of data collapse. Via a minimization of this measure, the exponents and their error-bars can be obtained. The procedure is illustrated by considering finite-size-scaling near phase transitions and quite strikingly recovering the exact exponents.

PACS numbers: 05.10.-a, 05.70.Jk, 03.65.-w, 64.60.-i

(Some figures in this article are in colour only in the electronic version)

Scaling, especially finite size scaling (FSS), has emerged as an important framework for understanding and analysing problems involving diverging length scales. Such problems abound in condensed matter, high-energy and nuclear physics, equilibrium and non-equilibrium situations, thermal and non-thermal problems and many more. The operational definition of scaling is this: a quantity $m(t, L)$ depending on two variables, t and L, is considered to have scaling if it can be expressed as

$$m(t, L) = L^d f(t/L^c). \tag{1}$$

Depending on the nature of the problem of interest, m may refer to magnetization, specific heat, size or some other characteristic of a polymer, width of a growing or fluctuating surface and so on. Equation (1) is the FSS form if L is a linear dimension of the system and t is any other variable, could even be time in dynamics. In the thermodynamic limit of infinite-sized systems, such a scaling would have t and L representing two thermodynamic parameters like the magnetic field, pressure, chemical potential etc or one could be time. If L is a length scale, then d would look like the dimension of this quantity m, and c of variable t. In fluctuation-dominated cases, it is generally a rule, rather than an exception, that d and c

assume nontrivial values, different from what one expects from a dimensional analysis. The exponents and the scaling function $f(x)$ then characterize the behaviour of the system. The fact that two completely independent variables (both conceptually and as controlled in experiments) combine in a nontrivial way to form a single one leads to an enormous simplification in the description of the phenomenon. This underlies the importance of scaling.

A quantitative way of showing scaling is data collapse (also called scaling plot) that goes back to the original observation of Rushbrooke that the coexistence curves for many simple systems could be made to fall on a single curve [1]. For example, the values of $m(t, L)$ (equation (1)) for various t and L can be made to collapse on a single curve if mL^{-d} is plotted against tL^{-c}. The method of data collapse therefore comes as a powerful means of establishing scaling. It is in fact now used extensively to analyse and extract exponents especially from numerical simulations. Given the importance of scaling in wide varieties of problems, it is imperative to have an appropriate measure to determine the 'goodness of collapse'—not to be left to the eyes of the beholder.

It might be mentioned here that there are situations where pure power laws are expected, or the exponents are known with sufficient accuracy but data collapse suffers from correction-to-scaling terms. In such situations, several methods are available to extract the exponents [2] or to take care of the correction terms [3]. Our method finds relevance in situations where no such prior knowledge is available especially as many phenomena in sciences, social sciences, economics, etc are being analysed for possible scaling behaviour. In such cases the data collapse remains a viable approach.

In this paper, we propose *a measure* that can be used to quantify 'collapse'. This measure can be used, via a minimization principle, for an automatic search for the exponents thereby removing the subjectiveness of the approach. To show the power of the method and the measure, we use it for two exactly known cases, namely the FSS of the specific heat for: (1) the one-dimensional ferro-electric six-vertex model [4] showing a first-order transition [5], and (2) the Kasteleyn dimer model [6] exhibiting the continuous anisotropic Pokrovsky–Talapov transition [7, 8]. In addition, to show the usefulness of the method in case of noisy data as expected in any numerical simulation, we consider the one-dimensional case with extra Gaussian noise added (by hand). It is worth emphasizing that the proposed procedure, without any bias, recovered the exactly known exponents from the specific heat data for finite systems.

If the scaling function $f(x)$ of equation (1) is known, then the sum of residuals

$$R = \frac{1}{N} \sum \mid L^{-d}\, m - f(t/L^{c}) \mid \tag{2}$$

where the sum is over all the data points, is minimum for the right choice of (d, c). In absence of any statistical or systematic error, the minimum value is zero.

However, in most situations the function itself is not known but is generally an analytic function. In the case of a perfect collapse, any one of the sets (say set p) can be used for $f(x)$. An interpolation scheme can then be used to estimate the residuals. For this estimate, only the *overlapping regions* are considered, i.e. the regions where points from both the sets are present (see below). Since this can be done for any set as the basis, we repeat the procedure for all sets. Let the tabulated values of m and t be denoted by m_{ij}, t_{ij} (ith value of t for the jth set of L (i.e. $L = L_j$ for set j)). We now define a quantity P_b:

$$P_b = \left[\frac{1}{\mathcal{N}_{\text{over}}} \sum_{p} \sum_{j \neq p} \sum_{i,\text{over}} \mid L_j^{-d}\, m_{i,j} - \mathcal{E}_p(L_j^{-c}\, t_{ij}) \mid^{q} \right]^{1/q} \tag{3}$$

where $\mathcal{E}_p(x)$ is the interpolating function based on the values of set p bracketing the argument in question (of set j). The innermost sum over i is done only over the overlapping regions

(denoted by the 'i, over'), i.e. only those points of set j are considered for which $x_{ij} = L_j^{-d} t_{ij}$ belongs to the interval spanned by the corresponding x-values of set p. Let $\mathcal{N}_{\text{over}}$ being the total number of such pairs. Though defined with a general q, we use $q = 1$. For $\mathcal{E}_p(x)$, a four-point polynomial interpolation can be used and if any complex singularity is suspected a rational approximation may be used. Extrapolations are avoided. The minimum of this function P_b is zero[3] and is achievable in the ideal case of perfect collapse with correct values of (d, c), i.e.

$$P_b \geqslant P_b|_{\text{abs min}} = 0. \tag{4}$$

This inequality can then be exploited and a minimization of P_b over (d, c) can be used to extract the optimal values of the parameters.

In addition to the values of the exponents, estimates of errors can be obtained from the width of the minimum. This can be obtained by diagonalizing the inverse of the Hessian matrix for P_b. A simpler approach (at least for illustration) is to take the quadratic part in the individual directions along the (d, c) plane. From an expansion of $\ln P_b$ around the minimum at (d_0, c_0), the width is estimated as

$$\Delta d = \eta d_0 \left[2 \ln \frac{P_b(d_0 \pm \eta d_0, c_0)}{P_b(d_0, c_0)} \right]^{-1/2} \tag{5a}$$

and

$$\Delta c = \eta c_0 \left[2 \ln \frac{P_b(d_0, c_0 \pm \eta c_0)}{P_b(d_0, c_0)} \right]^{-1/2} \tag{5b}$$

for a given η. Choosing $\eta = 1\%$, the final estimate for the exponents would be $d_0 \pm \Delta d$, $c_0 \pm \Delta c$ with the error bar reflecting the width of the minimum at the 1% level.

We now use the proposed method for different test cases. In order to implement the program[4], we have used the routines of numerical recipes [9]. To calculate P_b, POLINT or RATINT has been used for interpolation with HUNT to place a point in the table. This procedure treats the data points as 'exact'. In case of data points with known error bars (as, e.g., from Monte Carlo simulations) one may use piecewise continuous fits to obtain the interpolated values. For minimization, AMOEBA has been used thrice to locate the minimum, each time using the current estimates to generate a new triangle enclosing the minimum. In the examples given below, there was no need for more sophisticated minimization routines, which could be needed in case of subtle crossover behaviours or with nearby minima.

Let us first consider the one-dimensional six-vertex model which shows a first-order transition [5]. With the partition function $Z = 2 + (2x)^N$ for N sites with $x = \exp(-\epsilon/k_B T)$ as the Boltzmann factor, ϵ being the energy of the high-energy vertices, T the temperature and k_B the Boltzmann constant, the specific heat can be computed exactly. The first-order transition is at $x = 1/2$, for $N \to \infty$, with a δ-function jump in specific heat. The N-dependent specific heat (per site), c_N, is given by [5]

$$c_N = k_B (\ln x)^2 2N \frac{(2x)^N}{[2 + (2x)^N]^2} \tag{6}$$

which for large N and small $t = 2x - 1$ has the scaling form of equation (1) with $d = 1$, $c = -1$ and

$$f(z) = k_B (\ln 2)^2 2 \frac{e^z}{(2 + e^z)^2}. \tag{7}$$

[3] Pathological cases of no overlap are avoided.

[4] This program is available on request from the authors. It is flexible enough that the Numerical Recipes [9] routines could be replaced by any other suitable package.

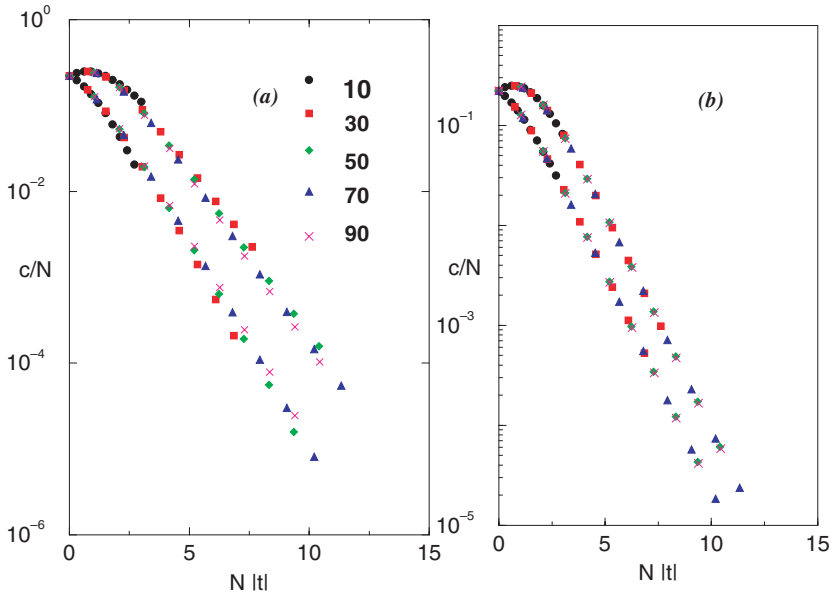

Figure 1. The collapse of specific heat for the $1-d$ vertex model as calculated from equation (6) (see (a)) and from equation (7) (see (b)). The upper (lower) branch is for $t > 0$ ($t < 0$).

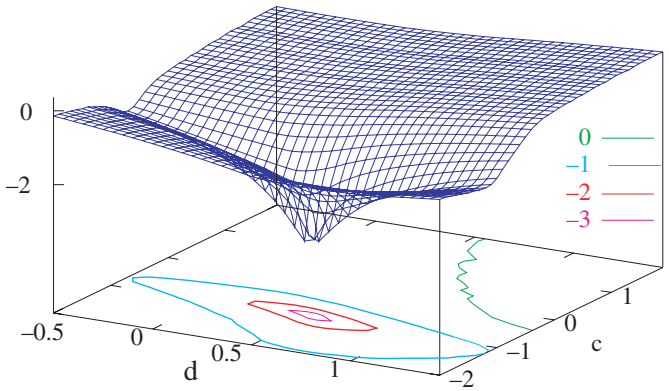

Figure 2. The residue P_b with $q = 1$ is shown over the (d, c) plane. The z-axis is in log scale. A few contours of constant $\ln(P_b)$ are shown by projecting the curves on the (d, c) plane.

From the exact formula, equation (6), data were generated for $N = 10, 30, 50, 70$ and 90, for various values of temperatures. A minimization of P_b gave us the estimate $d = 0.997 \pm 0.04$, $c = -0.98 \pm 0.06$, with $P_b = 0.56881\text{E-}01$. The exponents are very close to the exact ones. The error-bars or the width of the minimum is to be interpreted as an indication of the presence of non-scaling corrections. To test this, we have generated data from the exact scaling function of equation (7). An unbiased minimization of P_b then gave $d = 1 \pm 0.004$, $c = -1 \pm 0.004$ with $P_b = 0.34876\text{E-}03$. The smallness of the residue and of the errors (or the width of the minimum) represents a good data collapse. The nature of the data collapse for both the cases is shown in figure 1.

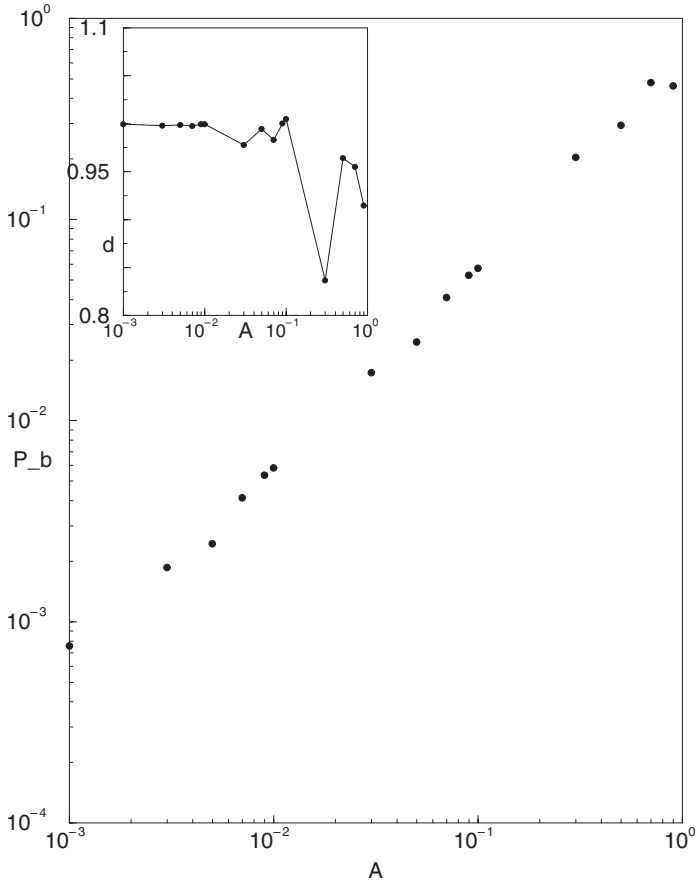

Figure 3. Noisy data. Plots of P_b against A. Inset shows the estimated value of d as a function of A.

A similar minimization of P_b was carried out for the two-dimensional Kasteleyn dimer model (also isomorphic to a two-dimensional five-vertex model). This is an exactly solvable lattice model of the continuous anisotropic Pokrovsky–Talapov transition for surfaces, and shows a square-root singularity for specific heat with different correlation lengths in the two directions [8][5]. The specific heat for lattices of size M along the direction of the 'walls' and infinite in the transverse direction is known exactly and its FSS form has been discussed in [8]. Using the following formula [8] for the specific heat per site c_M:

$$\frac{c_M}{k_B a} = M \int_0^{2\pi} \frac{(2x\cos\phi)^M}{[1+(2x\cos\phi)^M]^2}\,\mathrm{d}\phi \tag{8}$$

specific heats data were generated for $M = 10, 30, 50, 70$ and 90. In this formula, a few unimportant factors are put under a and not explicitly shown. The critical point is at $x = 1/2$. A minimization of P_b gave $d = 0.5 \pm 0.03, c = -0.945 \pm 0.02$ to be compared with the exact values $d = 0.5, c = -1$. The residue factor is $P_b = 0.12424\text{E-}01$. The importance of correction terms are clear from figure 5 of [8], and in our approach it gets reflected in the not too small value of P_b.

[5] Along the direction of the 'walls' the bulk length-scale exponent is $\nu_y = 1$ while in the transverse direction $\nu_x = 1/2$.

The function P_b for $q = 1$ for the above two-dimensional problem is shown as a surface plot over the (d, c) plane in figure 2. The sharpness of the minimum is noteworthy. In both the examples considered, the performance of the method is remarkable[6].

The last example we consider is a case of noisy data [10] where c is calculated from equation (7) and a Gaussian noise was added to it so that $c_{n,\eta} = |c_n\ (1 + A\eta)|$, where η is a Gaussian deviate and A is the amplitude of the noise added, and the absolute value is taken to keep c_n positive. The values of the exponents are found to be insensitive to A for $A < 0.1$ and starts changing for higher values of A. In figure 3 we show P_b against A. The larger values of P_b for larger A is a sign of poor collapse, as one finds by direct plotting with the estimated values or exact values.

To summarize, we have proposed a measure to quantify the nature of data collapse in any scaling analysis of the form given by equation (1). This measure, equation (3), can be used for an automated search for the exponents. The method is quite general and even though we formulated it in terms of power-laws as in equation (1), it can very easily be adopted to other forms of scaling[7]. We conclude that the subjectiveness of data collapse can be removed and P_b could be used as a quantitative measure to test or compare 'goodness of collapse' in a scaling analysis. This is an invaluable advantage of our method in contrast to the current usage of the 'best-by-eye' data-collapse method.

Acknowledgment

We acknowledge financial support from MURST(COFIN-99).

References

[1] Stanley H E 1999 *Rev. Mod. Phys.* **71** S358
[2] Schmittbuhl J, Vilotte J P and Roux S 1995 *Phys. Rev.* E **51** 131
[3] Bruce A D and Wilding N B 1999 *Phys. Rev.* E **60** 3748
[4] Lieb E and Wu F Y 1971 *Phase Transitions and Critical Phenomena* vol 1, ed C Domb and M Green (New York: Academic)
[5] Bhattacharjee S M 2001 *Field Theories in Condensed Matter Systems* ed S Rao (New Delhi: Hindustan) (Bhattacharjee S M 2000 *Preprint* cond-mat/0011011)
[6] Kasteleyn P W 1963 *J. Math. Phys.* **4** 287
[7] Pokrovksy V L and Talapov A L1979 *Phys. Rev. Lett.* **42** 65
Nagle J F 1975 *Phys. Rev. Lett.* **34** 1150
[8] Bhattacharjee S M and Nagle J F 1985 *Phys. Rev.* A **31** 3199
[9] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes in FORTRAN 77* (Cambridge: Cambridge University Press)
[10] For analysis of noisy series, see, e.g., Dekeyser R, Iglói F, Mallezie F and Seno F 1990 *Phys. Rev.* A **42** 1923

[6] In the case of perfect scaling as in equation (7), the choice of q is not crucial. Since larger values of q amplifies larger deviations more, there will be a sensitivity to q if P_b is large. This is seen for the other two cases, equations (6) and (8).

[7] For example, when tested on a log-type singularity as one finds for the specific heat in the two-dimensional Ising model, the method does not give a good collapse with power laws, but yields the correct results when the power laws in equation (1) are changed to log's.