Maximum-entropy Monte Carlo method for the inversion of the structure factor in simple classical systems

Marco D'Alessandro
Institute for Complex Systems, National Research Council (CNR), Via del Fosso del Cavaliere 100, IT-00133 Rome, Italy
(Received 21 July 2011; published 20 October 2011)

We present a method for the evaluation of the interaction potential of an equilibrium classical system starting from the (partial) knowledge of its structure factor. The procedure is divided into two phases, both of which are based on the maximum entropy principle of information theory. First we determine the maximum entropy estimate of the radial distribution function constrained by the information contained in the structure factor. Next we invert the pair function and extract the interaction potential. The method is tested on a Lennard-Jones fluid at high density and the reliability of its results with respect to the missing information in the structure factor data are discussed. Finally, it is applied to the experimental data of liquid sodium at 100 °C.

DOI: 10.1103/PhysRevE.84.041130 PACS number(s): 05.20.Jj, 05.10.Ln

I. INTRODUCTION

The radial distribution function (RDF) of an equilibrium statistical system contains useful information concerning its physical properties. Indeed, at least for systems governed by pairwise additive interactions, its knowledge allows one to compute the ensemble average for observable quantities such as internal energy and pressure. Furthermore, if the same hypotheses are satisfied, the RDF is in one-to-one correspondence with the microscopic interaction potential [1,2] and represents the starting point for the solution of the so-called "inverse problem" of statistical mechanics [3–8].

Despite its central role in the analysis of a statistical system, the RDF is not directly accessible from the experiments and its estimation passes through the measurement of the structure factor. This former quantity is formally related to the RDF by an inverse Fourier transform, which for a homogeneous and isotropic system reads

$$g(r) = 1 + \frac{1}{2\pi^2\rho}\int_0^\infty dk \frac{\sin(kr)}{kr} k^2 [S(k) - 1], \qquad (1)$$

where $g(r)$ and $S(k)$ are the RDF and the structure factor, respectively, and ρ is the density. So the measurement of the RDF is reduced to the evaluation of the integral appearing in Eq. (1). Unfortunately this procedure, although conceptually correct, cannot be directly applied due to some typical limitations in the measurement of the $S(k)$. Indeed, the experimental information is obtained by an analysis of the x-ray and/or neutron diffraction data. These techniques provide results over a finite k range and a number of nontrivial corrections on measured data are needed. So the resulting experimental structure factor turns out to be incomplete and typically spoiled by systematic and statistical errors. As a consequence, the RDF obtained by means of Eq. (1) may present nonphysical features and spurious structures could emerge.

In order to overcome these difficulties different approaches have been pursued. A promising class of solutions is provided by simulation assisted methods in which a molecular dynamics or a Monte Carlo (MC) simulation is driven with the aim to minimize the differences between the simulated structure factor and the experimental data. Among the results belonging to this class we cite the reverse Monte Carlo technique, proposed by McGreevy and Pusztai in Ref. [9], which implements the transition probability on the basis of the χ^2 function between the reference and the simulated structure factors; this procedure, however, does allow one to determine the pair interaction potential. Further approaches are provided by the method proposed by Tóth [10] and based on the previous work of Lyubartsev and Laaksonen [5], and by the solution due to Almarza, Lomba, and Molina [11]. These methods consist in an iterative procedure for the evaluation of an effective pair potential compatible with the experimental data, so they attempt to provide a true solution of the inverse problem starting from the structure factor. A comprehensive review of the simulation assisted methods is given in Ref. [12].

Since we are dealing with the reconstruction of the RDF starting from the partial knowledge of the experimental $S(k)$ one question concerning the uniqueness of the solution naturally arises; at the same time it is desirable that no information besides that contained in the structure factor is transferred to the RDF during the reconstruction. Both of these issues can be addressed using the maximum entropy principle (ME) [13] as the guideline for the definition of the reconstruction procedure. Indeed ME has the remarkable feature of producing the highest entropy solution compatible with the given constraints, so the corresponding estimate for the RDF is "maximally noncommittal with regard to the missing information" [13]. ME-based algorithms for the inversion of the structure factor were first developed by Root, Egelstaff, and Nickel [14] and by Soper [15]. In these papers it has been shown that the adoption of ME improves the Fourier transform of the structure factor data and reduces the spurious structure in the RDF. ME has been introduced for the first time in contest of the inverse problem by Cilloco in Ref. [16]; the method described in this paper used ME inside a Monte Carlo simulation scheme. It has been shown that a maximum entropy ensemble of configurations compatible with a given reference RDF can be built adopting a suitable definition of the transition probability between neighbor states. This approach has been recovered and extended in Ref. [8]; the transition probability has been reinterpreted as an information-based feedback controller and an "integral term" has been added. The authors evidenced that this quantity converges to the

interaction potential thus providing a ME-based solution of the inverse problem.

The purpose of this paper is to present a ME Monte Carlo method for the inversion of the experimental structure factor. The procedure is mainly based on the statistical properties of the pair distribution functions both in the r and in the k space. We will show that, thanks to the above mentioned features of ME, the algorithm provides a reliable reconstruction of the RDF starting from a limited knowledge of the experimental structure factor. Once the $S(k)$ has been inverted, we can apply the technique described in Ref. [8] to the resulting RDF and extract the (pair) interaction potential.

The paper is organized as follows. Section II contains a detailed theoretical description of our procedure. In Sec. III we test the method for a Lennard-Jones fluid assuming different cutting points of the input $S(k)$ and we invert the experimental data of liquid sodium at 100 °C. Finally, in Sec. IV we discuss our results and present some concluding remarks.

II. THEORY

We present a statistical description of a simple monoatomic fluid, in an analogous way of what has been done in Ref. [8], and extend this analysis to the Fourier transform of the RDF. Then we describe a procedure for the construction of a ME ensemble of configurations constrained by the (partial) knowledge of the structure factor.

A. Preliminaries

We define the notion of a *model* system, which is a homogeneous and isotropic collection of pointlike elements with average density ρ. Given an arbitrary configuration $\mathbf{x}$ of the model we define two quantities that will be relevant for the subsequent analysis: the local sampling of the elements pair function (PF) n and its Fourier transform $\bar{n}$. The former quantity provides the number of particles n_i inside the ith spherical shell of width δr centered on a reference element; the sampling is performed up to the maximum value r_M and consequently the index i runs from 1 to $N = r_M/\delta r$. The Fourier transform (FT) of the local PF is defined through the equation

$$\bar{n}_j = \mathcal{F}_j(n) = \sum_{i=1}^{N} \frac{\sin(k_j r_i)}{k_j r_i} n_i, \quad j = 1, \ldots, N, \tag{2}$$

where r_i is the value of the radius associated to the ith shell: $r_i = i\delta r$. Given the local pair function n, its FT $\bar{n}$ represents an N element vector in the k space. The component $\bar{n}_j$ contains the k-space value at $k_j = j\delta k$; the sampling is performed with a uniform step of width δk, chosen according to the relation

$$\delta r \delta k = \frac{\pi}{N}. \tag{3}$$

We also introduce the notion of inverse Fourier transform (IFT). Given a k-space vector $\bar{n}$ we define its IFT through the equation

$$n_i = \mathcal{F}_i^{(-1)}(\bar{n}) = \frac{2}{N} \sum_{j=1}^{N} r_i^2 k_j^2 \frac{\sin(k_j r_i)}{k_j r_i} \bar{n}_j. \tag{4}$$

Equation (3) ensures the orthonormality of the discrete basis of functions adopted in Eqs. (2) and (4), namely,

$$\sum_{j=1}^{N} \sin(k_j r_i) \sin(k_j r_l) = \frac{N}{2} \delta_{il}, \tag{5}$$

so the transformation of a PF from the r space to the k space and back again will reproduce the initial function [17]. It is worth mentioning that, according to Eq. (3), a sampling of width δr in the r space produces a k-space vector with a maximum wave number given by $k_M = \pi/\delta r$, in agreement with the Shannon-Nyquist sampling theorem [18].

Since we are interested in the construction of the average pair functions (both in the r and k space) we have to extend the notion of local PF and of its FT to a large number of configurations. So we introduce a probability function $p(\mathbf{x})$ defined upon the model configuration space and we collect an ensemble of s elements extracted according to p. The global samplings over the ensemble are defined as the average values of the local ones:

$$\begin{aligned} m_i &= \frac{1}{s} \sum_{\alpha=1}^{s} n_i^{(\alpha)}, \\ \bar{m}_j &= \mathcal{F}_j(m) = \frac{1}{s} \sum_{\alpha=1}^{s} \bar{n}_j^{(\alpha)}, \end{aligned} \tag{6}$$

where the index α labels the elements of the ensemble and the last equality holds due to the linearity of the FT. The radial distribution function and the structure factor of the model system are defined starting from the global quantities (6) in the limit $s \to \infty$. The RDF is obtained by normalizing the global PF built on the model ensemble with the pair function of a uniform reference system (perfect gas) with the same density of the model one:

$$g(r_i) = \frac{m_i}{m_i^{(pg)}}, \tag{7}$$

where $m_i^{(pg)} = 4\pi\rho r_i^2 \delta r$ is the perfect gas pair function. The structure factor is defined in terms of the FT of the global pair function via the relation

$$S(k_j) = 1 + \bar{m}_j - \bar{m}_j^{(pg)}. \tag{8}$$

This definition provides the usual notion of $S(k)$ for an isotropic statistical system, indeed making use of Eqs. (2) and (7), and performing the continuum limit gives

$$S(k) = 1 + 4\pi\rho \int_0^{r_M} dr \frac{\sin(kr)}{kr} r^2 [g(r) - 1], \tag{9}$$

which is the formal definition of structure factor adopted in statistical mechanics. We observe that, due to the finite size of the model system, the integral in Eq. (9) extends up to the maximum sampled value of the model RDF. Consequently, according to the Shannon-Nyquist sampling theorem, the maximum allowed k resolution is given by π/r_M.

We conclude this preliminary section by introducing a useful notation for dealing with the Fourier transforms. Since

the FT and its inverse are realized as linear combinations among the $\bar{n}$ and the n variables, respectively, we can write

$$\bar{n}_j = \sum_{i=1}^{N-1} c_{ij} n_i, \quad n_i = \sum_{j=1}^{N-1} c_{ij}^{-1} \bar{n}_j, \tag{10}$$

where according to Eqs. (2), (3), and (4) both c_{ij} and its inverse are symmetric matrices given by

$$c_{ij} = \frac{N}{\pi} \frac{\sin\left[\left(\frac{\pi}{N}\right) ij\right]}{ij}, \quad c_{ij}^{-1} = 2\pi \frac{\sin\left[\left(\frac{\pi}{N}\right) ij\right] ij}{N^2}. \tag{11}$$

The sum in Eqs. (10) has been restricted to the first $N-1$ elements since the last one gives a zero contribution for algebraic reason related to the definition of the matrices (11). We point out that Eq. (5) ensures that the matrix product of these quantities gives the identity matrix as expected.

B. Analysis of the model distribution function

We are interested in computing the probability distribution of the FT of the global pair function built on the model ensemble. So we suppose that an ensemble of configurations has been extracted according to a given probability distribution $p(\mathbf{x})$ and we compute the probability associated to a particular sampling $\bar{m}$ as a function of the parameters of the underlying ensemble distribution.

To achieve this task we start from the probability of the local sampling of the pair function n. The ith shell of the PF follows a Poisson distribution with expectation value μ_i [8]; we assume that the system exhibits a hard core (HC) structure with radius r_0 so that the expected number of particles μ_i is zero for i lower than the threshold value $N_0 = r_0/\delta r$ and is strictly positive otherwise. Since there is no correlation among different shells the complete distribution is obtained as the product of the single shell values and reads

$$\mathcal{P}_\mu(n) = \prod_{i=N_0}^{N} e^{-\mu_i} \frac{(\mu_i)^{n_i}}{n_i!}. \tag{12}$$

The HC structure of the reference distribution imposes that $\mathcal{P}_\mu$ is zero if there is some $n_i > 0$ for $i < N_0$. The FT of the local sampling of the pair function $\bar{n}$ is defined through a linear combination of the n variables (10), so its expectation value and its covariance can be expressed in terms of the μ_i parameters:

$$\begin{aligned} E(\bar{n}_j) = \bar{\mu}_j &= \sum_i c_{ij} \mu_i, \\ \mathrm{Cov}(\bar{n}_j, \bar{n}_k) = \xi_{jk} &= \sum_i c_{ij} c_{ik} \mu_i. \end{aligned} \tag{13}$$

We observe that, due to the linear combination (10), the covariance matrix of the $\bar{n}$ variables is not diagonal even if the original variables n are uncorrelated. The variable $\bar{m}$ is defined as the average of the $\bar{n}^{(\alpha)}$ (6), so it has the same expectation value $\bar{\mu}$ and a covariance given by ξ/s. Its asymptotic distribution can be computed using a multivariate central limit theorem; this theorem states that the distribution function of the reduced variable $\sqrt{s}(\bar{m}_j - \bar{\mu}_j)$ converges, in the limit $s \to \infty$, to a multivariate Gaussian with zero mean and covariance given by ξ. So we have

$$\sqrt{s}(\bar{m} - \bar{\mu}) \underset{s \gg 1}{\sim} \mathcal{N}_0, \tag{14}$$

where

$$\mathcal{N}_0(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} |\xi|^{1/2}} e^{-(1/2)\mathbf{x}^T (\xi)^{-1} \mathbf{x}} \tag{15}$$

is the multivariate distribution function with zero mean and $|\xi|$ represents the determinant of the covariance matrix. It turns out that the elements of $\bar{m}$ are linear dependent and consequently the covariance matrix is singular. This is due to the fact that only the nonzero components of the local PF contribute to the linear combination (10), so the number of independent elements of $\bar{m}$ is $N - N_0$. In order to avoid a singular covariance matrix we have to restrict our analysis to a set of independent elements of $\bar{m}$; in this domain the covariance matrix can be inverted and its inverse reads

$$\xi_{jk}^{-1} = \sum_{i=N_0}^{N-1} \tilde{c}_{ij}^{-1} \tilde{c}_{ik}^{-1} \frac{1}{\mu_i}, \tag{16}$$

where the indices j,k run from 1 to $N - N_0$ and the tilde indicates that the matrices are restricted to the subset of independent variables.

This analysis shows that the asymptotic distribution for the independent subset of the $\bar{m}$ components is described by a multivariate Gaussian distribution $\mathcal{N}_{\bar{\mu}}$ with mean $\bar{\mu}$ and inverse covariance $s\xi^{-1}$. We observe that both the expectation value (13) and the inverse covariance matrix (16) are functions of the parameters of the original distribution function (12).

C. Maximum entropy approach to the inverse problem

We consider a monoatomic system and assume that for a given density ρ and temperature T the structure factor $S_t(k)$ of the system is known up to the maximum value k_M. We will refer to this system as the *target*.

The aim of this section is to define a procedure for the evaluation of an *equilibrium* model distribution function $p(\mathbf{x})$ compatible with the information contained in the target structure factor. The method is based on the maximum entropy principle and is realized inside a Monte Carlo simulation scheme. MC represents an effective tool to pursue this approach: the maximization of configurational entropy is produced by the MC random movements for the construction of the trial configurations (source of entropy) while the transition probability among neighbor states selects the configurations and acts as a source of information. At equilibrium these two mechanisms are in balance, the net amount of information loss is zero, and the system approaches a state of maximum entropy consistently with the given constraints.

The main advantage of this kind of procedure is that the ME solution is sought in terms of a "real" physical system which possesses a true configuration space beyond the two-body pair function; so its equilibrium distribution implicitly defines the correlation functions of any order. Inside this scheme, the implementation of the ME algorithm realizes the maximization of the whole configurational entropy and not only of the two-body contribution. This method provides the maximum

entropy estimate of the complete equilibrium distribution of the model system and the ensemble of configurations built according to it can be used to compute the average value of any quantity of interest.

Since, under this perspective, the transition probability is the natural object in which the knowledge on the system is codified, we seek this quantity with the aim of building a model distribution function that produces an expectation value of $\bar{m}$ consistent with the target reference value $\bar{\mu}_t$ (the proper definition of this parameter on the basis of the target $S(k)$ will be discussed in the next section). To achieve this task we make use of the method developed in Refs. [8,16] and we maximize the log-likelihood function between the model pair function and the target reference value. This choice is based on statistical reasons: in the limit of a large number of configurations the average $\bar{m}$ computed over the model ensemble converges to the expectation value $\bar{\mu}$ of the model distribution function and the log-likelihood can be related to the relative entropy D (Kullback-Leibler divergence [19]) between the model and the target distributions:

$$\ln N_{\bar{\mu}_t}(\bar{m}) = -D(\mathcal{N}_{\bar{\mu}} || \mathcal{N}_{\bar{\mu}_t}), \tag{17}$$

so the maximization of the log-likelihood function is asymptotically equivalent to the minimization of the relative entropy (17). Given two distributions p and q, the relative entropy $D(p||q)$ is positive definite and vanishing only if $p = q$, so a complete maximization of the likelihood function implies the equality of the distributions.

So our general strategy is the following: we perform a MC simulation using a transition probability which maximizes the log-likelihood function defined above. This procedure builds a maximum entropy ensemble of configurations constrained by the target $S(k)$ and the radial distribution function computed over this ensemble is the maximum entropy estimate of the inverse Fourier transform of the target structure factor. Since the maximum entropy principle has the feature of providing reliable estimates on the basis of a partial input of information, we expect that this procedure should be able to produce a correct reconstruction of the radial distribution function starting from a limited knowledge of the structure factor.

In the next section we will describe some details of the implementation of this procedure. The applications of the method and some checks of its reliability and sensitivity to the amount of missing information are discussed in Sec. III.

D. Maximization of the log-likelihood function

Assume that the $S_t(k)$ has been measured with a resolution δk up to the value k_M; so the target input is given by $N_t = k_M/\delta k$ samplings of the structure factor.

The first step consists in a proper definition of the model system (see Sec. II A): the value of the model density is chosen equal to the target one and the model pair function is sampled up to the maximum value $r_M = \pi/\delta k$. This choice ensures that the model structure factor is sampled with the same resolution as the one of the target system. The spatial resolution δr in the model configuration space is chosen with the double task of producing an accurate sampling of the model RDF and to ensure that the maximum sampling value of the model structure factor (given by $\pi/\delta r$) is greater than the target value k_M.

We define the procedure for the construction of the model ensemble based on the maximization of the log-likelihood function described in the previous section. First we build the reference distribution on the basis of available information concerning the target structure factor [see Eq. (8)]:

$$\bar{\mu}_{tj} = \bar{m}_j^{(pg)} + S_t(k_j) - 1, \tag{18}$$

where j extends over all the shells in the model system (from 1 to $N = r_M/\delta r$) and we impose that $S_t(k_j)$ is equal to 1 for all $j > N_t$. Given the $\bar{\mu}_t$ we compute its inverse Fourier transform μ_t, which represents the expectation value of the target pair function. Due to the lacking information in the target structure factor the μ_t provides a *biased* reconstruction of the true target pair function; typically this function exhibits nonphysical behavior such as, for instance, strong oscillations inside the hard core radius.

Next we analyze the construction of the log-likelihood ratio. Assume that we have performed s MC iterations. For each point of the path we compute a local sampling of the PF $n^{(\alpha)}$ and its FT $\bar{n}^{(\alpha)}$ and we construct the global pair functions m and $\bar{m}$. Then we select a reference particle and compute a local sampling of the PF $n^{(1)}$; at the same time the particle is randomly moved and the new local sampling of the PF is stored in the array $n^{(2)}$. In this way we obtain two different samplings of $\bar{m}$ at the level $s + 1$, namely, $\bar{m}^{(1)}$ and $\bar{m}^{(2)}$. Then we perform a *cut* in the model system consistently with the missing information in the target reference function: so we substitute the perfect gas value in both the $\bar{m}$ samplings for $j > N_t$. This procedure provides $\bar{m}^{\mathrm{cut}\,(1)}$ and $\bar{m}^{\mathrm{cut}\,(2)}$ which are the natural quantities to be compared with $\bar{\mu}_t$. Finally we define the log-likelihood ratio via the relation

$$\delta\lambda = \ln \frac{\mathcal{N}_{\bar{\mu}_t}(\bar{m}^{\mathrm{cut}\,(1)})}{\mathcal{N}_{\bar{\mu}_t}(\bar{m}^{\mathrm{cut}\,(2)})}. \tag{19}$$

The transition probability selects all the trial configurations which maximize the likelihood function ($\delta\lambda < 0$) [23] and consequently the distribution of the $\bar{m}^{\mathrm{cut}}$ converges to a multivariate Gaussian defined by the target parameters $\bar{\mu}_t$ and μ_t [see Eqs. (13) and (16)]. At the same time the model global sampling m converges to the *unbiased* reconstruction of the target RDF and its FT $\bar{m}$ builds the complete target structure factor. So, thanks to the ME approach, we build a complete estimate of the target distribution function on the basis of a limited amount of information.

We conclude this section by analyzing the expression of the log-likelihood ratio. In the limit of a large number of configurations we can expand Eq. (19) in power of s. The leading order contribution reads

$$\delta\lambda = \sum_{i,j=1}^{N_t} \left(\bar{n}_i^{\mathrm{cut}\,(2)} - \bar{n}_i^{\mathrm{cut}\,(1)}\right) \xi_{ij}^{-1} \left(\bar{m}_j^{\mathrm{cut}} - \bar{\mu}_{tj}\right). \tag{20}$$

Equation (20) evaluates the log-likelihood ratio as the weighted sum of the differences between the actual and the trial local sampling $\bar{n}$. The weights are proportional to the discrepancy between the global $\bar{m}^{\mathrm{cut}}$ and the target reference function, and due to the nondiagonal correlation matrix, each term in the sum depends on the whole difference $(\bar{m}^{\mathrm{cut}} - \bar{\mu}_t)$. It is interesting

to recast this equation in terms of the local PFs in the r space; to obtain this result we make use of Eqs. (10) and (16); this provides

$$\delta\lambda = \sum_{i=N_0}^{N} \left(n_i^{(2)} - n_i^{(1)}\right) \frac{m_i^{(\mathrm{bias})} - \mu_{ti}}{\mu_{ti}}, \tag{21}$$

where $m^{(\mathrm{bias})}$ represents the inverse Fourier transform of $\bar{m}^{\mathrm{cut}}$. We see that, once reformulated in the configuration space, the log-likelihood becomes *diagonal*: the i th term in the sum (21) is weighted by the ith shell value of the discrepancy between the model and the target global (biased) PF. This behavior is a consequence of the independence between the local pair functions related to different shells [see Eq. (12)]. It is worth mentioning that Eq. (21) is strongly reminiscent of the log-likelihood ratio computed in [8,16] starting from the knowledge of the target RDF. Indeed we recognize that the weighted difference between $m^{(\mathrm{bias})}$ and μ_t is the first order expansion [24] of $\ln(m^{(\mathrm{bias})}/\mu_t)$. Furthermore, if the complete target structure factor is provided, then the input of information content becomes equivalent to the knowledge of the target RDF; in this case μ_t represents the true value of the target PF and the $m^{(\mathrm{bias})}$ coincides with the model global PF m thus providing an identical expression of the log-likelihood ratio.

E. A remark on the transition probability

In Sec. II D we stated that the transition probability is defined in a way to accept all the trial configurations with a log-likelihood ratio lower than zero ($\delta\lambda < 0$). In order to better comprehend the reasons behind this choice it is useful to briefly recall the main results concerning the analysis of the transition probability described in Ref. [8]. Following the approach outlined in this reference we can interpret Eq. (21) has a *proportional feedback controller* which selects the model system configurations on the basis of the "error" $e = (m^{(\mathrm{bias})} - \mu_t)/\mu_t$ between the target and the model global pair function. This interpretation suggests the formulation of an improved expression for $\delta\lambda$, based on the theory feedback systems, that also includes an *integral term* apart from the proportional one; this latter quantity keeps into account all the errors in the steps preceding the actual one. In this way we realize a proportional-integral controller, schematically defined as

$$\delta\lambda = \sum_i \left(n_i^{(2)} - n_i^{(1)}\right) \left(k_p e_i + k_I \sum_\alpha e_i^{(\alpha)}\right),$$

where k_p and k_I are the coefficients of the proportional and integral terms, respectively. The transition probability is defined as $\min\{1, \exp(-\delta\lambda)\}$ and the proportional and integral coefficients are fixed with the aim to ensure the correct fluctuation of the model PF around its average value. Results reported in [8] evidence that this approach allows one to build the correct equilibrium distribution of the target system; furthermore, the interaction potential emerges as the asymptotic limit of the integral term.

The procedure delineated above can be applied in the present case and would allow one to directly extract the interaction potential from the knowledge of the structure factor. Instead, we have adopted a different implementation of the feedback controller in which the integral term is absent and the proportional coefficient is virtually infinite: so only the trial configurations with a log-likelihood ratio lower than zero are accepted. The main advantage of this choice is that a pure proportional controller ensures a straightforward and stable convergence of the inversion procedure, so the RDF is obtained without the need of setting any parameters. In this way we can perform an intermediate check of the inversion procedure. Obviously, the extraction of the interaction potential requires the subsequent inversion of the RDF using the method described in [8].

III. APPLICATIONS

The technique previously described has been applied to a simple Lennard-Jones fluid and to the experimental structure factor data of the liquid sodium at 100 °C measured by Greenfield, Wellendorf, and Wiser in Ref. [22].

The inverse MC simulation is realized in the NVT ensemble: the model configuration space is a cubic volume of linear length L with N_p pointlike particles and the periodic boundary conditions together with the minimum image convention have been adopted.

The transition probability between neighbor states has been evaluated by using Eq. (21) for the computation of the log-likelihood ratio. This formula requires knowledge of the HC radius which is *a priori* unknown; a brief estimate of its value can be obtained (as suggested by Reatto in Ref. [4]) by computing the inverse FT of the structure factor and by taking a fraction of the r position of its first peak. ME will allow this estimate to be corrected to its optimal value during the simulation.

A. Results for the Lennard-Jones system

We consider a system described by the Lennard-Jones potential with argonlike parameters $\sigma = 3.405$ Å and $\epsilon/k_B = 119.76$ K. The target structure factor is evaluated by performing a metropolis MC simulation on a system of 864 particles at reduced density $\rho^* = \rho\sigma^3 = 0.84$ and reduced temperature $T^* = k_B T/\epsilon = 0.75$, near the triple point. The simulation runs for 2×10^4 cycles after equilibration. The $g(r)$ has been evaluated up to $r^* = r/=7.05$ (24 Å), the width of the shells for the measure of the $g(r_i)$ was $\delta r = 2.4 \times 10^{-2}$ Å, and the number of measured points was 10^3. The structure factor has been evaluated using the procedure described in Sec. II A; the k resolution is given by Eq. (3) and is equal to $\delta k = 0.13$ Å^{-1}.

Once the target $S(k)$ was computed we performed the inversion procedure for the reconstruction of the RDF. In order to check the sensitivity of this approach we truncated the target $S(k)$ at different values of k and we proceeded to the reconstruction for each of the truncated functions. So we built three structure factors, namely, $S_{t1}(k)$ (truncated at $k_M = 13$ Å^{-1}), $S_{t2}(k)$ (truncated at $k_M = 6.5$ Å^{-1}), and $S_{t3}(k)$ (truncated at $k_M = 3.2$ Å^{-1}). Then the reconstruction procedure started for 2×10^4 cycles after equilibration. In this way we produced three radial distribution functions, namely, $g_1(r)$, $g_2(r)$, $g_3(r)$, and the corresponding structure factors $S_1(k)$, $S_2(k)$, and $S_3(k)$.

Results are reported in Fig. 1. The first line contains the outcomes of the inversion starting from $S_{t1}(k)$. The maximum difference between the target and the model structure factor for

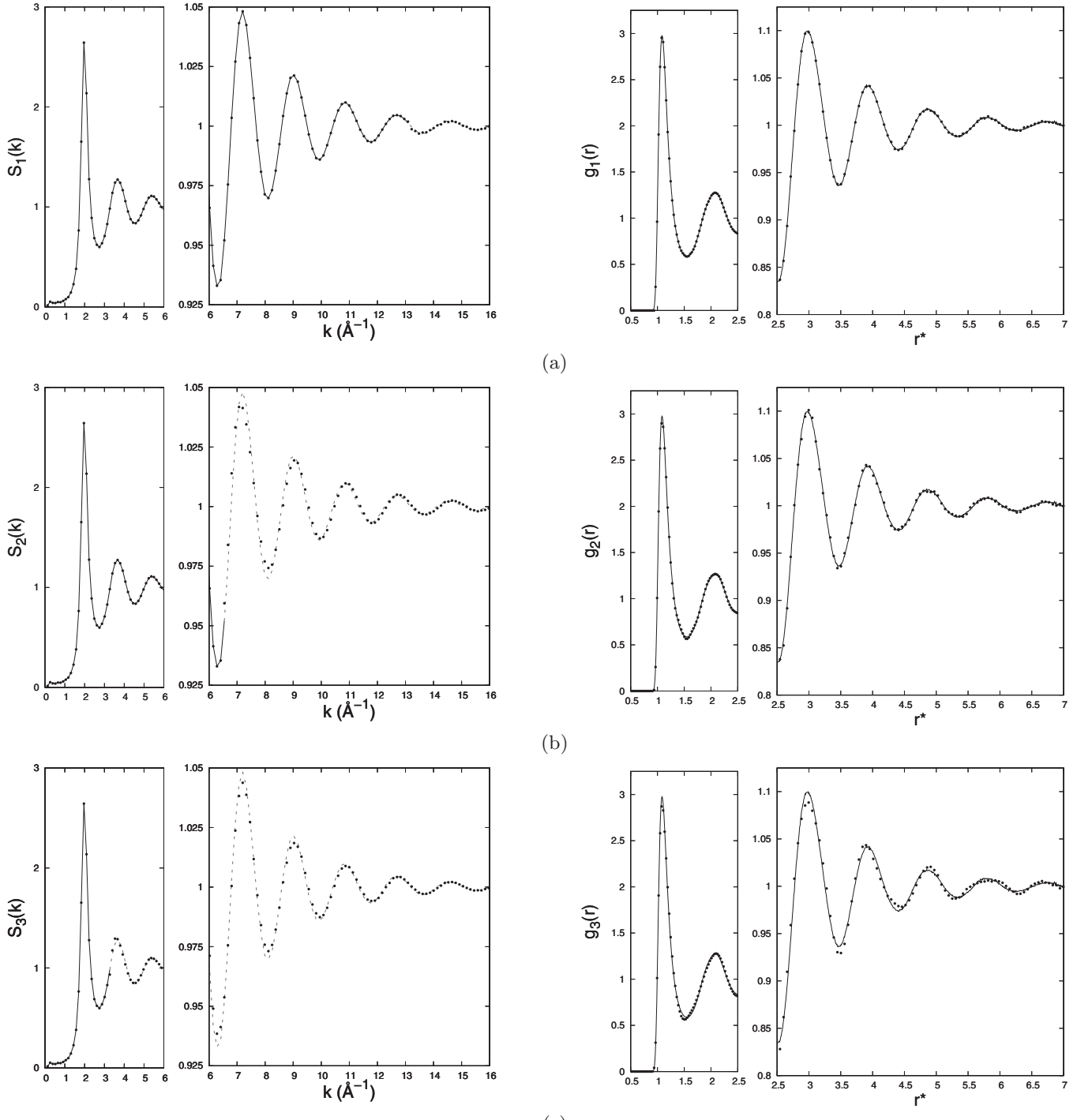

FIG. 1. (Color online) Results of the inversion procedure for a Lennard-Jones system. The left column contains the plots of the structure factor: continuous line for the target $S(k)$ used as model input, dotted line for the target $S(k)$ in the k region beyond k_M, and filled circles for the model $S(k)$. The right column contains the radial distribution functions: continuous line for the target RDF and filled circles for the model RDF. (a) Target $S(k)$ truncated at 13 Å^{-1}; (b) target $S(k)$ truncated at 6.5 Å^{-1}; and (c) target $S(k)$ truncated at 3.2 Å^{-1}.

k up to k_M is about 4×10^{-4} and the procedure reconstructed the target $S(k)$ for $k > k_M$ with an error lower than 1×10^{-3}; the model RDF reproduces the target values with a maximum difference of about 2×10^{-2}. The second line reports results obtained using the information content of $S_{t_2}(k)$. Even in this case the maximum discrepancy up to k_M (6.5 Å^{-1}) is about 4×10^{-4} and the procedure reconstructed the target structure for $k > k_M$ with an error lower than 1×10^{-2}; the maximum difference between the target and the model RDF is about 5×10^{-2}. Finally, in the last line of Fig. 1 we present the results of the inversion of $S_{t_3}(k)$. The maximum difference between the target and the model structure factor up to k_M (3.2 Å^{-1}) is

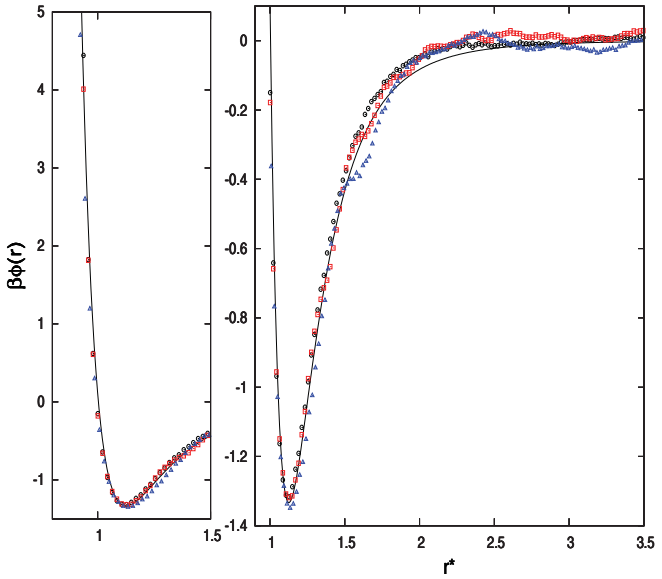

FIG. 2. (Color online) Results for the Lennard-Jones system. The target potential (continuous line) and the model potential [circles for $g_1(r)$, squares for $g_2(r)$, and triangles for $g_3(r)$] are plotted.

about 4×10^{-4} as in the previous cases, and despite the modest information content of the target structure factor, the model $S(k)$ reproduces the target one for $k > k_M$ with an error lower than 4×10^{-2}. The corresponding RDF reconstructs the target function with a maximum discrepancy of about 6×10^{-2}. This analysis evidences the effectiveness of the maximum entropy principle to provide accurate reconstructions on the basis of a limited amount of information.

In order to complete the inversion procedure we have evaluated the interaction potential associated to the three RDFs computed above. These results have been obtained by using the method described in [8] and are presented in Fig. 2. The analysis of this figure shows that the potentials extracted from $g_1(r)$ and $g_2(r)$ are essentially equivalent and provide a good estimate of the target one (with a maximum discrepancy of about 5×10^{-2} distributed over the r axis). The potential extracted from $g_3(r)$ is less accurate with respect to the previous ones. In this case the main features of the target potential (such as the amplitude and location of the absolute minimum) are reproduced correctly but some spurious oscillations are present. This behavior is a consequence of the presence of small oscillations in the RDF $g_3(r)$ which are hardly visible at the scale of Fig. 1. This fact indicates that a very precise reconstruction of the target RDF is needed in order to obtain a correct solution of the inverse problem.

B. Inversion of the Na data

We present the result of the inversion of the structure factor of the liquid Na at 100 °C [22]. Since we are dealing with a real fluid at high density we expect that the many-body contributions in the interaction potential cannot be neglected, so our solution of the inverse problem will produce an effective pair potential.

Experimental data have been measured with a variable k resolution up to 8.9 Å^{-1}. In order to apply the inversion technique defined in Sec. II we need a target $S(k)$ sampled uniformly with a δk value compatible with the size of the simulation box; so a preliminary operation on experimental data is needed. Our prescription is the following: we perform the inverse Fourier transform of the experimental $S(k)$ and compute the (biased) RDF; then we "cut" this function at a value r_M consistent with the linear dimension of the model system in which we will perform the inversion procedure. Finally, we transform back to the k space and compute the new structure factor which is ready to be used as the target input function. The reliability of this method has been tested for a Lennard-Jones system in which the evaluation of the target $S(k)$ and the inversion procedure for the reconstruction of the RDF have been performed in boxes of different linear length. In all the cases we have obtained a correct reconstruction of the target RDF. For the present case of the Na data we have chosen $r_M = 22$ Å which corresponds to the maximum sampled value for a model system made of 864 particles.

The inverse simulation procedure for the reconstruction of the Na radial distribution function took 2×10^4 cycles after equilibration. The result for the RDF is reported in the left panel of Fig. 3. This function evidences a HC radius of 2.65 Å and the first peak is located at $r = 3.72$ Å and is equal to 2.32. It is interesting to compare our result with the RDF obtained in [4] using an iterative method for the inversion of the Na structure factor. The two functions are in substantial agreement: the RDF of [4] has a HC radius of 2.7 Å, whereas the first peak is located at $r = 3.66$ Å and is equal to 2.43; furthermore even the relative positions of the other minima and maxima differ less than the 2%.

The Na pair effective potential has been extracted from the RDF computed above by using the method described in Ref. [8]. The result is presented in the right panel of Fig. 3. The potential reported in the figure evidences a highly repulsive part in the low r region; then there is an attractive zone with the minimum located in $r = 4.05$ Å and equal to -0.91 and a further weak repulsive part with a local maximum at $r = 5.45$ Å. Finally, the potential approaches zero with some smooth oscillations. Again we compare our solution with the one obtained in Ref. [4]. We observe that the shapes of the two potentials are in qualitative agreement but a quantitative comparison reveals some differences: in particular, the locations of the absolute minima coincide but the depth of the potential wells differ by about 15%. This fact has to be interpreted on the basis of the high sensitivity of the inverse problem on the input RDF, so minor differences among the RDFs could produce a sensible discrepancy at the level of the interaction potentials.

IV. DISCUSSION AND CONCLUSIONS

We have presented a method, based on the maximum entropy principle of information theory, for the reconstruction of the radial distribution function of an equilibrium statistical system starting from the partial knowledge of its structure factor. The procedure is realized inside a Monte Carlo simulation scheme which is revealed to be an effective tool for the implementation of the ME; indeed the maximization of the configuration entropy is realized by the MC random

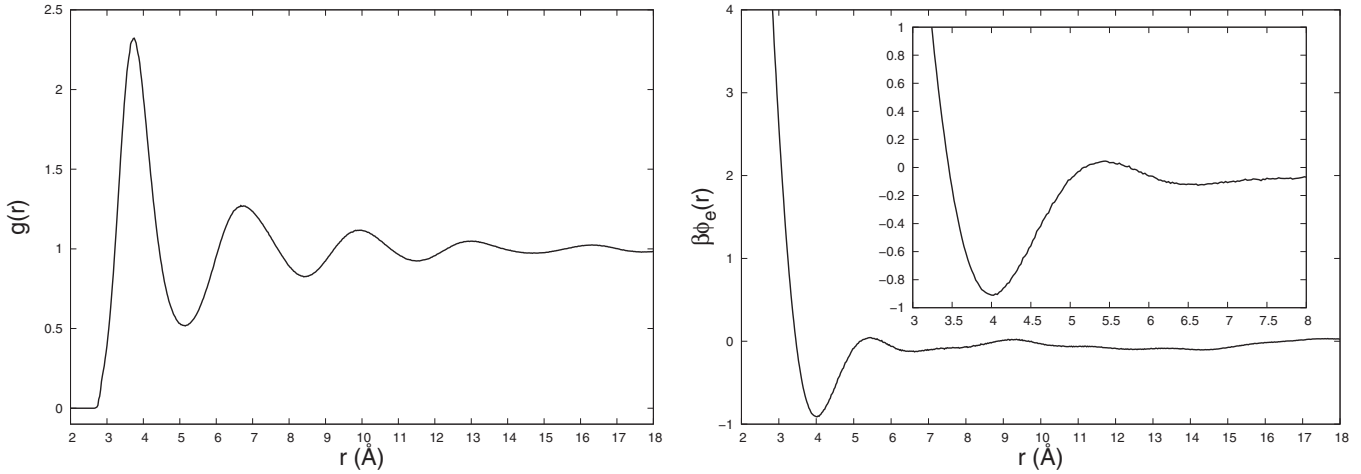

FIG. 3. Results of the inversion procedure for Na at 100 °C. The left panel contains the plot of the radial distribution function. The right panel contains the two-body effective potential.

displacements, whereas the transition probability between neighbor states is defined consistently with the information input codified in the target $S(k)$. Once the RDF has been computed we can derive the two-body effective potential by using the method defined in Ref. [8], thus providing a true ME-based solution of the inverse problem.

As stated in Sec. II C, the realization of the ME approach inside a MC-based procedure presents some interesting features. Indeed, this method realizes a complete maximization of the model system configurational entropy (beyond the two-body term) and provides the maximum entropy estimate of the complete equilibrium distribution of the model system. So within this approach it is possible to extract information concerning the physical system under inspection that goes beyond the simple improvement of the Fourier transform of the structure factor. Furthermore, since the correlators are obtained through the ensemble average over the model system configuration space, any nonphysical feature (such as, for instance, negative values for the RDF inside the hard core region) is automatically avoided.

The applications of the method are presented in Sec. III. Results analyzed in the first part of this section are designed to test our approach with respect to the missing information in the input structure factor. ME has the feature of being "maximally noncommittal with regard to the missing information" [13], and indeed, the results discussed in Sec. III A demonstrate a reliable reconstruction of the system RDF even for very limited knowledge of the $S(k)$. Finally, Sec. III B contains the analysis of the real experimental data of the liquid sodium at 100 °C. We evaluated the Na RDF and then we extracted the effective pair interaction potential; both of the procedures converged to a stable result. The solution of the inverse problem for this system has been compared with the one presented in Ref. [4]. The discrepancies between the two potentials have been motivated in terms of the (small) differences among the RDFs. It is well known that the solution of the inverse problem is highly sensible to the details of the pair function used as the input of the reconstruction procedure. So, under this perspective, the adoption of the maximum entropy principle as a general and solid guideline for the definition inversion procedure could guarantee the correctness of the results.

A last comment concerns the possible extensions of the technique described in the present paper. ME principle holds for any system at equilibrium, so the main idea at the basis of this approach can be extended to systems other than the simple monoatomic fluid discussed in the present paper. For instance, polyatomic fluids are often characterized by strong directional interactions and an effective description of their physical properties in terms of the model system defined in this paper could be revealed as very crude. In these cases, however, it is possible to define an improved model system with new degrees of freedom which provide a better match with the ones of the experimental system under inspection. The statistical analysis presented in Sec. II has to be extended in order to include these new degrees of freedom and the same kind of procedure based on the maximization of the log-likelihood ratio can be performed. Obviously, the feasibility of this strategy requires a higher involvement of information concerning the target system and further experimental data, beyond the two-body pair function, has to be provided.

ACKNOWLEDGMENTS

The author is grateful to Francesco Cilloco for many stimulating discussions and suggestions. A further acknowledgment goes to Luciana Silvestri for assistance.

[1] R. L. Henderson, Phys. Lett. A **49**, 197 (1974).
[2] J. T. Chayes and L. Chayes, J. Stat. Phys. **36**, 471 (1984).
[3] W. Schommers, Phys. Rev. A **28**, 3599 (1983).
[4] L. Reatto, D. Levesque, and J. J. Weis, Phys. Rev. A **33**, 3451 (1986).
[5] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).

[6] A. K. Soper, J. Chem. Phys. **202**, 295 (1996).
[7] N. G. Almarza and E. Lomba, Phys. Rev. E **68**, 011202 (2003).
[8] M. D'Alessandro and F. Cilloco, Phys. Rev. E **82**, 021128 (2010).
[9] R. L. McGreevy and L. Pusztai, Mol. Simul. **1**, 359 (1988).
[10] G. Tóth, J. Chem. Phys. **115**, 4770 (2001).
[11] N. G. Almarza, E. Lomba, and D. Molina, Phys. Rev. E **70**, 021203 (2004).
[12] G. Tóth, J. Phys.: Condens. Matter **19**, 335220 (2007).
[13] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[14] J. Root, P. Egelstaff, and B. Nickel, *Neutron Scattering Data Analysis*, Institute of Physics Conf. Series (Institute of Physics, Bristol, 1986).
[15] A. K. Soper, Chem. Phys. **107**, 61 (1986).
[16] F. Cilloco, J. Mol. Struct. **296**, 253 (1993).
[17] F. Lado, J. Comput. Phys. **8**, 417 (1971).
[18] C. E. Shannon, Proc. IRE **37**, 10 (1949).
[19] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[20] In this paper we follow the method introduced in Ref. [16] for the maximization of the likelihood function.
[21] The expansion of the logarithm is due to the use of a central limit theorem, performed in Sec. II B, for the determination of the probability distribution of $\bar{m}$.
[22] A. J. Greenfield, J. Wellendorf, and N. Wiser, Phys. Rev. A **4**, 1607 (1971).
[23] In this paper we follow the method introduced in Ref. [16] for the maximization of the likelihood function.
[24] The expansion of the logarithm is due to the use of a central limit theorem, performed in Sec. II B, for the determination of the probability distribution of $\bar{m}$.