Nonparametric estimation of Fisher information from real data

Har-Shemesh, O.; Quax, R.; Miñano, B.; Hoekstra, A.G.; Sloot, P.M.A.

DOI
10.48550/arXiv.1507.00964
10.1103/PhysRevE.93.023301

Publication date
2016

Document Version
Final published version

Published in
Physical Review E

Nonparametric estimation of Fisher information from real data

Omri Har-Shemesh,[1,*] Rick Quax,[1,†] Borja Miñano,[2,‡] Alfons G. Hoekstra,[1,3,§] and Peter M. A. Sloot[1,3,4,‖]
[1]*Computational Science Lab, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands*
[2]*IAC³UIB, Mateu Orfila, Carretera de Valldemossa km 7.5, 07122 Palma, Spain*
[3]*ITMO University, Saint Petersburg, Russia*
[4]*Complexity Institute, Nanyang Technological University, 60 Nanyang View, Singapore 639673, Republic of Singapore*
(Received 7 July 2015; revised manuscript received 18 January 2016; published 8 February 2016)

The Fisher information matrix (FIM) is a widely used measure for applications including statistical inference, information geometry, experiment design, and the study of criticality in biological systems. The FIM is defined for a parametric family of probability distributions and its estimation from data follows one of two paths: either the distribution is assumed to be known and the parameters are estimated from the data or the parameters are known and the distribution is estimated from the data. We consider the latter case which is applicable, for example, to experiments where the parameters are controlled by the experimenter and a complicated relation exists between the input parameters and the resulting distribution of the data. Since we assume that the distribution is unknown, we use a nonparametric density estimation on the data and then compute the FIM directly from that estimate using a finite-difference approximation to estimate the derivatives in its definition. The accuracy of the estimate depends on both the method of nonparametric estimation and the difference $\Delta\theta$ between the densities used in the finite-difference formula. We develop an approach for choosing the optimal parameter difference $\Delta\theta$ based on large deviations theory and compare two nonparametric density estimation methods, the Gaussian kernel density estimator and a novel density estimation using field theory method. We also compare these two methods to a recently published approach that circumvents the need for density estimation by estimating a nonparametric f divergence and using it to approximate the FIM. We use the Fisher information of the normal distribution to validate our method and as a more involved example we compute the temperature component of the FIM in the two-dimensional Ising model and show that it obeys the expected relation to the heat capacity and therefore peaks at the phase transition at the correct critical temperature.

DOI: 10.1103/PhysRevE.93.023301

I. INTRODUCTION

The Fisher information matrix (FIM) is a measure of the sensitivity of a probability distribution function (PDF) to the value of the parameters θ on which it depends. We designate the PDF as $p(x;\theta)$, where x is either discrete or continuous (and possibly a vector) and θ is a vector of continuous parameters whose members we designate with Greek letter indices. Then the FIM is given by

$$g_{\mu\nu}(\theta) = \langle(\partial_\mu \ln p)(\partial_\nu \ln p)\rangle. \tag{1}$$

Here $\partial_\mu \equiv \partial/\partial\theta^\mu$ and $\langle\cdot\rangle$ is an average with respect to $p(\cdot;\theta)$. It is a positive semidefinite matrix (in the case of multiple parameters) or a positive number if only one parameter is taken under consideration. In the theory of statistical inference it quantifies the difficulty of estimating the value of θ from a set of samples $\{x_i\}_{i=1,\ldots,N}$ through the Cramér-Rao lower bound (CRLB) [1]. It is widely used in many domains of science ranging from optimal experimental design [2] to its interpretation as a Riemmanian metric on the statistical manifold [3] and its relation to theories of phase transitions [4–13] and complex systems [14–21].

It is often the case that the distribution of the data is assumed to have a certain form. Then it is possible, at least in principle, to compute the FIM analytically from Eq. (1). However, even this might turn out to be a formidable task and one must then resort to numerical methods, such as the Monte Carlo based method described in Ref. [22]. When an analytic expression of the FIM is available, we estimate the FIM by estimating the parameters of the PDF and plugging them into the known expression. This we term "parametric estimation" of the FIM, since it is estimated through estimation of its parameters from the data.

We are interested in estimating the FIM in cases where the parameters θ can be set precisely but the response of the system to the parameters is complicated. This is the inverse of the parameter estimation problem, since our interest lies in the response of the system rather than in the determination of the parameters. In these cases the FIM can be seen as a generalized susceptibility measure [20]. In these settings there seem to be two main paths to compute the FIM: either directly from Eq. (1) by first estimating the density $p(x;\theta)$ nonparametrically (hence "nonparametric estimation") and using a finite-difference approximation to the derivatives in the definition (see Sec. III) or by estimating it indirectly through its relation to a class of information divergences called f divergences [23,24]:

$$D_f[p(x;\theta), p(x;\theta+d\theta)] = \tfrac{1}{2} g_{\mu\nu}(\theta) d\theta^\mu d\theta^\nu + O(d\theta^3). \tag{2}$$

Here summation is implied over repeated Greek indices. Berisha and Hero [24] have very recently shown that it is

*O.HarShemesh@uva.nl
†R.Quax@uva.nl
‡bminyano@mail.iac3.eu
§A.G.Hoekstra@uva.nl
‖P.M.A.Sloot@uva.nl

possible to estimate the FIM by use of a statistic called the Friedman-Rafsky two-sample multivariate statistic [25] that converges to a type of f divergence known as an α divergence in the limit of large numbers of samples. Their method requires obtaining data at various parameter values around θ, computing a Euclidean minimal spanning tree (EMST) of the samples, and solving Eq. (2) for the FIM.

Other approaches to nonparametric estimation of FI primarily deal with PDFs with locationlike parameters, i.e., $p(x;\theta) = p(x-\theta)$. There Huber [26] found a unique density with minimal FI given a set of $k \geqslant 2$ samples from the cumulative distribution function. Kostal and Pokora [27] adapted the maximized penalized likelihood method of Good and Gaskins [28] to compute the FI. Kostal and Pokora rejected the use of a kernel density estimation (KDE) for the direct computation of the FI because no appropriate bandwidth parameter to control of the p'/p term in Eq. (1) is known [27].

In this paper we focus on two aspects of the problem of nonparametric estimation of Fisher information (FI). We first present a theoretical argument based on Sanov's theorem [29] for the optimal selection of the parameter $d\theta$ that is applicable for both the density estimation approach and the f-divergence approach. Second we show that a new Bayesian approach to nonparametric density estimation called "density estimation using field theory" (DEFT) accurately estimates the FI for low data dimensions (so far DEFT is implemented for one and two dimensions), compared with a more standard density estimation method called Gaussian KDE and compared with the f-divergence method of Berisha and Hero [24] (which we refer to here as EMST).

The paper is organized as follows. In Sec. II we describe the general problem of density estimation and introduce DEFT. In Sec. III we define the finite difference approximation we use. In Sec. IV we present a theoretical argument that guides the selection of $d\theta$. In Sec. V we show the results of numerical experiments performed for the estimation of the univariate normal distribution, whose FIM is analytically known and therefore can form a benchmark for our method. We present the results of estimating the FIM for the two-dimensional Ising model as a more involved example where the sensitivity of the distribution of energies to the value of the temperature can be used to locate the critical temperature of the Ising model. Finally in Sec. VI we give some final remarks about our results.

II. DENSITY ESTIMATION

The general density estimation problem [30] aims to obtain the best estimate Q_{est} of the distribution Q_{true} given N independently drawn samples. We distinguish between parametric and nonparametric estimation. Parametric estimates constrain Q_{est} to depend on a few parameters that are estimated from the data [31]. By the Cramér-Rao inequality [1] the inverse of the FI is a lower bound on the variance of the estimated parameters. FI is therefore often computed in the parametric setting. In these cases the FI is computed analytically from the assumed function.

When we do not assume a specific form for the PDF, we estimate the density nonparametrically. Thus, the data determine the shape of the distribution. Areas with higher probability density will contain more data points than areas with lower probability density. The main problem of nonparametric methods is how to balance the goodness of fit to the data and the smoothness of the estimated curve [30]. For example, kernel density estimators (KDEs) are a sum over kernel functions with width h, positioned at each data point, i.e., $Q_{\text{est}}(x) = (hN)^{-1}\sum_{i=1}^{N} K[(x-x_i)/h]$, where x_i is a data point and K is a kernel function. The bandwidth h controls the smoothness of the estimate. In the limit $h \to 0$ the estimate is a sum of delta functions at each data point; in the limit $h \to \infty$ it is uniform. Choosing the correct bandwidth is therefore important. Taken too large, the estimate will hide crucial features. If it is too small it will cause spurious peaks in the estimate, especially for long-tailed distributions [30]. Important to this study, the amount of smoothing directly affects the value of the FI. This can be seen from the definition of the FIM Eq. (1) which depends on the derivatives of the PDF. If, e.g., the estimated PDF Q_{est} is smoother than the true PDF Q_{true}, the estimate for the FI will be smaller than the true FI.

One elegant approach that derives the smoothness from the data itself was proposed in Ref. [32]. The authors used field theory to formulate the notion of a smoothness scale as an high-frequency cutoff, treating the smoothness length scale ℓ as a parameter in a Bayesian inference procedure. They showed that, in the large N limit, the data select an appropriate length scale. Recently, this method was developed into a fast and accurate algorithm called DEFT [33]. The algorithm was only implemented in one and two dimensions, since it suffers from the "curse of dimensionality" [33].

III. FINITE-DIFFERENCE APPROXIMATION

Our finite difference approximation is obtained by replacing the derivatives in Eq. (1) with a centered derivative:

$$g_{\mu\nu}(\theta) \approx \int \frac{p(x;\theta+\Delta\theta^{\mu}) - p(x;\theta-\Delta\theta^{\mu})}{2\Delta\theta^{\mu}} \times \frac{p(x;\theta+\Delta\theta^{\nu}) - p(x;\theta-\Delta\theta^{\nu})}{2\Delta\theta^{\nu}} \frac{dx}{p(x;\theta)} \tag{3a}$$

$$\approx \int \frac{\ln p(x;\theta+\Delta\theta^{\mu}) - \ln p(x;\theta-\Delta\theta^{\mu})}{2\Delta\theta^{\mu}} \times \frac{\ln p(x;\theta+\Delta\theta^{\nu}) - \ln p(x;\theta-\Delta\theta^{\nu})}{2\Delta\theta^{\nu}} p(x;\theta)dx. \tag{3b}$$

Here $\Delta\theta^{\mu}$ indicates a change in the value of only one parameter, θ^{μ}, keeping all other parameters fixed, i.e., $\theta + \Delta\theta^{\mu} \equiv (\theta^1,\ldots,\theta^{\mu}+\Delta\theta^{\mu},\ldots,\theta^d)$. The error introduced by this replacement is proportional to $O(\Delta\theta^2/6)$ (for each derivative) as can be verified by performing a Taylor expansion. Higher-order finite-difference schemes can be used but not, in our experience, a lower-order one-sided derivative because the estimate does not converge to the true value (data not shown).

To obtain all entries of the FIM we proceed in the following way: first, obtain N samples at all parameter positions required by Eq. (3). For the diagonal elements three parameter positions are required (at θ and at $\theta \pm \Delta\theta^{\mu}$). For the off-diagonal elements five positions are necessary. Obtain a nonparametric

estimate for each of those positions (using DEFT) and integrate these estimates numerically.

There are several sources for errors in the estimation resulting from this method. First of all, the replacement of the derivatives with centered finite-difference derivatives introduces an error that scales as $\Delta\theta^2$, as mentioned earlier. The second comes from the accuracy of the estimation of p and is related both to the method of estimation and to the number N of samples at each point. A third source of error occurs when the densities are too close together to be distinguishable. We discuss the balance between this error and the finite difference error in the next section where we show that an optimum in the selection of $\Delta\theta$ exists.

IV. CHOICE OF $\Delta\theta^\mu$

The value of $\Delta\theta^\mu$ strongly influences the accuracy of the computation. Two sources of error determine the optimal $\Delta\theta^\mu$: the aforementioned numerical derivative error and the finite sample size N. The first error, which scales like $O[(\Delta\theta^\mu)^2]$, becomes smaller with smaller $\Delta\theta^\mu$. The second source, however, becomes smaller when increasing $\Delta\theta^\mu$. This happens because a density estimate from a finite number of samples is always underdetermined. Any estimate is one curve from a group of curves that are close, but not equal, to the true density. The larger the number of samples is, the smaller the size of the group is. If $\Delta\theta^\mu$ is too small, the groups of the densities in the numerical derivatives will overlap and the difference $p(x;\theta+\Delta\theta^\mu)-p(x;\theta-\Delta\theta^\mu)$ will be ill-defined. This leads to one of our main results: since $\Delta\theta^\mu$ cannot be too small or too large, there is an optimal value with minimal error between the two extremes.

To estimate the curve group size and avoid overlaps, we use large deviations theory. According to Sanov's theorem [29] the appropriate distance measure is the Kullback-Leibler (KL) divergence:

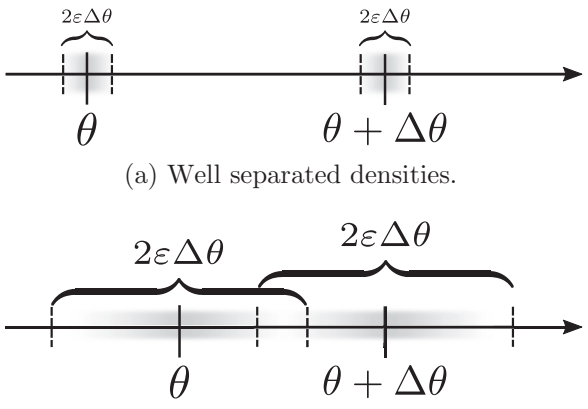
$$\mathcal{D}_{\mathrm{KL}}[Q||P] \equiv \int_{x\in\mathcal{X}} Q(x)\ln\frac{Q(x)}{P(x)}dx, \tag{4}$$

which is defined for two densities $P(x)$ and $Q(x)$ where the support of P and Q overlap. The probability that a set of N samples independently drawn from P appears to be drawn from Q is proportional to

$$\exp(-N\mathcal{D}_{\mathrm{KL}}[Q||P]). \tag{5}$$

In the limit of infinite sample size this tends to zero. For finite N the set of distributions whose KL divergence with P is small enough, such that this probability is finite, forms the curve group. This can be interpreted as a hypersphere in parameter space centered at θ with an N- and θ-dependent radius. To minimize error, the radius at θ and $\theta+\Delta\theta^\mu$ should be small compared to $\Delta\theta^\mu$. The ideal case is drawn schematically in Fig. 1(a) with well-separated densities and in Fig. 1(b) where $\Delta\theta^\mu$ is too small.

To compute the hypersphere radius we take $P=p(x;\theta)$ and $Q=p(x;\theta+\varepsilon\Delta\theta^\mu)$ in Eq. (5). We thus seek the density Q at the edge of the hypersphere and parametrize it with ε, the hypersphere radius in units of $\Delta\theta^\mu$. The KL divergence of



(a) Well separated densities.

(b) Overlapping densities.

FIG. 1. Schematic drawing in one dimension with points of estimation θ and $\theta+\Delta\theta$. The gray area is the hypersphere. ε is the radius of the hypersphere in units of $\Delta\theta$.

two neighboring distributions is approximately [34]

$$\mathcal{D}_{\mathrm{KL}}[P(\theta)||P(\theta+\varepsilon\Delta\theta^\mu)] \approx \frac{\varepsilon^2}{2}g_{\mu\nu}(\theta)\Delta\theta^\mu\Delta\theta^\nu = O(\Delta\theta^2). \tag{6}$$

Inserting Eq. (6) into Eq. (5) we get

$$\exp\left[-\frac{N\varepsilon^2}{2}g_{\mu\nu}(\theta)\Delta\theta^\mu\Delta\theta^\nu\right]. \tag{7}$$

If we define the boundary of the hypersphere as the point where the probability is equal to e^{-1}, we obtain the radius ε:

$$\varepsilon^2 = \frac{2}{Ng_{\mu\nu}(\theta)\Delta\theta^\mu\Delta\theta^\nu}. \tag{8}$$

The radius depends on the number of samples N, θ, and $\Delta\theta^\mu$. At a given N and θ, *increasing* $\Delta\theta^\mu$ will *decrease* the radius and thus increase accuracy.

As an analytically solvable example, we take the univariate normal distribution $\mathcal{N}(\mu,\sigma)$. Its FI is as follows:

$$g_{\mu\mu}=\frac{1}{\sigma^2};\quad g_{\sigma\sigma}=\frac{2}{\sigma^2};\quad g_{\mu\sigma}=g_{\sigma\mu}=0. \tag{9}$$

We focus on the FI of σ, which is not a location parameter. Inserting this in Eq. (8) yields

$$\Delta\sigma=\sqrt{\frac{2}{\varepsilon^2 N g_{\sigma\sigma}}}=\frac{\sigma}{\varepsilon\sqrt{N}}, \tag{10}$$

with $\Delta\sigma\equiv\Delta\theta^\sigma$. This guides the choice of $\Delta\sigma$ for a given N, σ, and desired radius ε. We can get the same result using the Cramér-Rao inequality. The minimal variance of an unbiased estimator for σ is $1/g_{\sigma\sigma}$. Given N samples this equals $\sigma^2/2N$. Demand that the variance of σ is equal to $\frac{1}{2}(\varepsilon\Delta\sigma)^2$ (the factor $\frac{1}{2}$ ensures a consistent definition of ε). This variance is equivalent to a hypersphere radius of $\varepsilon\Delta\sigma$. We then have $\sigma^2/2N=(\varepsilon\Delta\sigma)^2/2$. Solving for $\Delta\sigma$ yields Eq. (10).

For real data the FI is unknown and can be estimated iteratively. First compute the FI with $\Delta\theta^\mu$ that ensure a good approximation of the numerical derivatives. Then use the FI to compute ε. If it is too large (based on our simulations, up

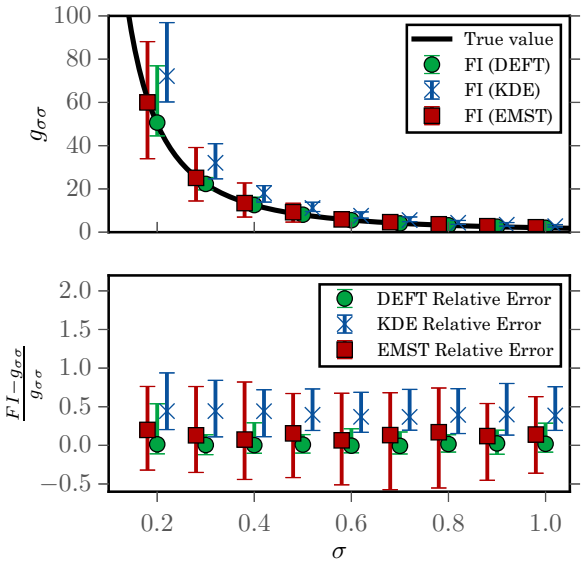


FIG. 2. A comparison between Gaussian KDE, DEFT, and EMST for FI estimation. The top figure shows median FI estimates using the different methods. Error bars represent 5 and 95 percentiles. The bottom figure shows the relative errors. The values were computed with $N = 10^4$, $\varepsilon = 0.05$, and 100 repetitions at each σ. The same samples were used by all three methods. The KDE and EMST estimates were shifted by ± 0.02 along the σ axis for clearer presentation.

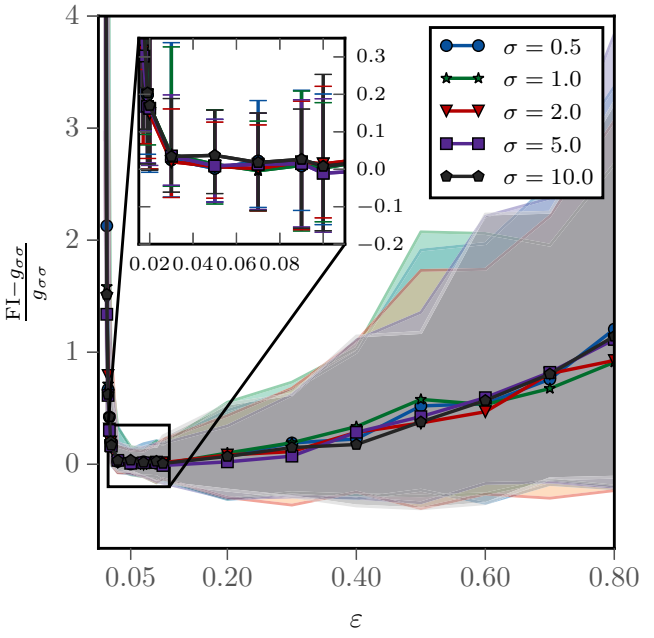


FIG. 3. The median relative error as a function of ε for different values of σ. FI stands for the computed value and $g_{\sigma\sigma}$ the analytic value. The shaded areas and error bars in the inset indicate the 5 and 95 percentiles computed over 100 repetitions of the computation with $N = 2 \times 10^4$.

to about $\varepsilon \approx 0.1$ seems reasonable, see Fig. 4), increase $\Delta\theta^\mu$ or N.

V. RESULTS

A. Fisher information of the Gaussian distribution

We demonstrate our main results by computing $g_{\sigma\sigma}$ from independently drawn normally distributed samples. We first compare DEFT (with number of grid points $G = 100$, smoothness parameter $\alpha = 3$, and a bounding box twice the interval between the smallest and the largest sample [33]), KDE (using Scott's rule for the bandwidth), and the EMST method of Ref. [24]. We use the same samples with all three methods and compute the FI. In the top plot of Fig. 2 the FI estimate is shown. The black curve is the analytic value, the green dots, blue $\times$'s, and red squares are the median estimates after 100 repetitions (error bars are 5 and 95 percentiles) for DEFT, KDE, and EMST, respectively. We use $N = 10^4$ for each density estimate and a value of $\varepsilon = 0.05$ since this yields the best results (see Fig. 3). The KDE and EMST plots are slightly shifted along the σ axis by ± 0.02 for clearer presentation but were computed at the same value of σ as DEFT.

All three methods follow the analytic curve; however from the relative errors it is clear that KDE consistently overestimates the FI by about 40% and the distance between 5 and 95 percentiles is about 100% of the original value. DEFT has zero bias and a spread of 30%−40%. EMST does not suffer from the same bias as KDE but has larger error bars. We conclude that in this example DEFT provides an improvement over KDE and EMST both in the estimated value and in the error margins. In the above computations we used Eq. (3a) for computation with DEFT and Eq. (3b) for KDE, because KDE was extremely unstable when computed using Eq. (3a) while DEFT performed slightly better with Eq. (3a). A few notes about the implementation of the EMST method for our example are in order. In the one-dimensional case, the Friedman-Rafsky statistic is simply the number of times samples from two different distributions are adjacent when arranged along the real line. Unlike Ref. [24] we only use one perturbation $\Delta\theta$ in our computation and solve the FI by inverting Eq. (2), i.e., $g_{\sigma\sigma} = 2D_\alpha/(d\sigma)^2$. This emulates the situation where obtaining the samples is relatively expensive and therefore a one-shot estimate of the FIM is desirable. In the Appendix we study the behavior of the EMST method at different values of ε. The computation, at least in this way, appears to be less stable than the use of DEFT and might hint that at lower dimensions DEFT outperforms the EMST method.

In the following we use DEFT exclusively for the density estimation. To see how the error depends on ε we varied it at a fixed $N = 2 \times 10^4$ and plotted the relative error. We computed the FI for $\sigma = 0.5$, 1, 2, 5, and 10. Each computation was repeated 100 times at different ε and the median and 5 and 95 percentiles of the relative error ($[g_{\sigma\sigma} - FI]/g_{\sigma\sigma}$, where FI is the estimated FI) were computed. All curves have the same functional dependence on ε and, as we predicted, there is an optimal value for $\Delta\sigma$, at $\varepsilon \approx 0.05$. Thus the absolute errors depend on σ through the combination in Eq. (8), as shown in Fig. 3. All the curves have a minimum in the range of $\varepsilon \in [0.04, 0.1]$. At small ε they grow due to errors in the numerical derivative ($\Delta\sigma$ too large). At large ε they grow due to overlapping densities. The spread (the 90%

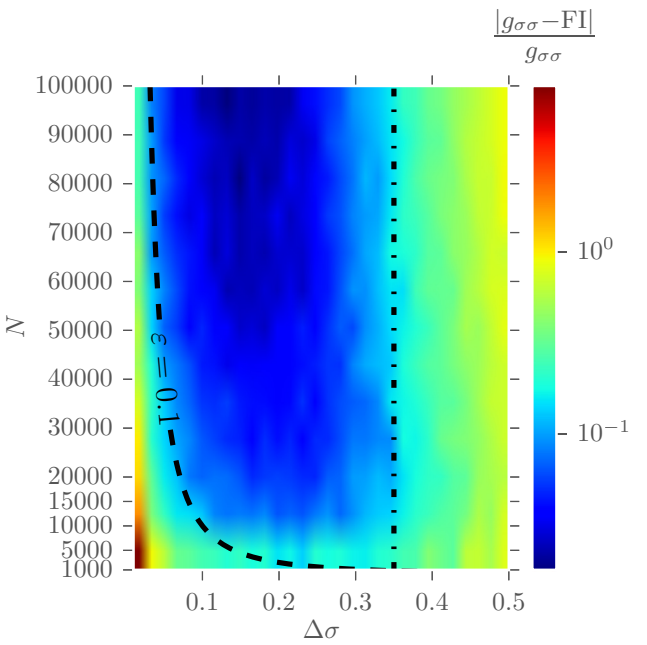


FIG. 4. Relative error in the computation of the FI for $\sigma = 1.0$ as a function of both $\Delta\sigma$ and N. Computed using DEFT with 100 repetitions per point. Dashed line represents the $\varepsilon = 0.1$ line and the dash-dotted line is the $\Delta\sigma = 0.35$ line. Unlike the previous plots, here we compute the absolute-value relative error to avoid problems with the logarithmic color-bar scale.

interpercentile range) is minimal at $\varepsilon = 0.05$ as well. The shaded regions in the plot represent the interpercentile range of the various σ curves.

To verify the N and ε dependence of the errors we varied both and computed $g_{\sigma\sigma}$. The result is presented as a heat map in Fig. 4. The color represents the absolute-value relative estimation error in logarithmic scale. The dashed line indicates the $\varepsilon = 0.1$ line, which represents the highest value of ε where good results are still obtained. The dash-dotted line represents the $\Delta\sigma = 0.35$ line. All computations were done with $\sigma = 1.0$ and 100 repetitions. The errors due to small $\Delta\sigma$ seem to follow the $\varepsilon = 0.1$ curve, showing again the dependence of this type of error on ε. Above $\Delta\sigma = 0.35$ we see increasing errors due to the large value of $\Delta\sigma$. The best area for the estimation is between the two lines.

B. Fisher information for the two-dimensional Ising model

One of the applications of the computation of FI from samples is in detecting phase transitions [18]. As a further validation we took the two-dimensional Ising model, which is the prototypical model of a continuous phase transition. It is a model of binary spins, s_i, on a square lattice with nearest-neighbors interaction. Its Hamiltonian is

$$\mathcal{H} = -\sum_{\langle i,j\rangle} J_{ij} s_i s_j - h \sum_i s_i, \tag{11}$$

where $\langle i,j\rangle$ indicates the sum is on nearest neighbors, $s_i = \pm 1$ is the value of a spin at site i, J_{ij} is the interaction energy, and h is an external applied magnetic field. In more than one dimension there is a critical order-disorder phase transition at a finite temperature. Onsager [35] solved the model exactly in two dimensions in the thermodynamic limit (infinite number of spins) and at zero applied external field. The critical temperature in the isotropic case ($J_{ij} \equiv J$) is

$$T_c = \frac{2J}{\ln(1+\sqrt{2})} \approx 2.269J. \tag{12}$$

For simplicity we set $J \equiv 1$ and Boltzmann's constant $k_B \equiv 1$.

Prokopenko *et al.* [18] computed both the TT and hh components of the FIM (computed for the Gibbs distribution with $\theta^1 = h$ and $\theta^2 = T$) in terms of the susceptibility χ_T and the specific heat C_h and showed that

$$g_{TT} = \frac{C_h}{T^2}; \quad g_{hh} = \frac{\chi_T}{T}. \tag{13}$$

We therefore expect both to diverge as the system approaches the critical temperature. In a finite system this means that the FI peaks at the critical temperature.

To validate this result we simulate the Ising model and compute the FI. We use the Metropolis-Hastings Monte Carlo algorithm to obtain samples of the configuration energy with the Gibbs distribution (at zero external field):

$$p(S;T) = \frac{1}{Z(T)} \exp\left[-\beta\mathcal{H}(S,T,h=0)\right]. \tag{14}$$

Here $\beta = 1/T$ is the inverse temperature, $S = \{s_i\}_{i=1,\ldots,L^2}$ is a configuration of the spins on a $L \times L$ square lattice, and Z is the partition function. Since the Gibbs distribution in our case is L^2 dimensional we cannot use DEFT to estimate it directly. Instead we compute the temperature component of the FI of the distribution of energies:

$$p(E) = \frac{1}{Z(T)} g(E) \exp[-\beta E], \tag{15}$$

where $g(E)$ is the density of states. Since $g(E)$ is independent of the temperature it drops from the calculation of the FI and we therefore expect Eq. (13) to still hold. We then estimate the TT component of the FI using Eq. (3) with densities estimated from the sampled energies. We also compute the specific heat:

$$C_h(T) = \frac{1}{L^2T^2}(\langle E^2\rangle - \langle E\rangle^2), \tag{16}$$

where L^2 is the total number of spins, E is the energy of the configuration, and the average is performed over different configurations at the same temperature.

We plot the result of both the FI and the specific heat C_h computation in Fig. 5. The simulation is run on a 25×25 lattice of spins with periodic boundary conditions in the temperature range [0.5,4.0] which we divide into 200 segments, leading to a parameter difference of $dT \simeq 0.17$. We repeat the simulation five times and compute the median and 5 and 95 percentiles. We use a warm-up period of 5×10^6 time steps and take $N = 15\,000$ samples of the configuration energy. We use DEFT (with $G = 200$, $\alpha = 3$, and a bounding box of $[-4,1]$) for the density estimation. Because the FI depends on T, ε is not constant. Its median is $\varepsilon = 0.12^{+0.12}_{-0.07}$ for the values of ε which were not infinite. To verify that Eq. (13) holds, we plot the ratio of the two sides of the equation. This is presented in the inset in Fig. 5.

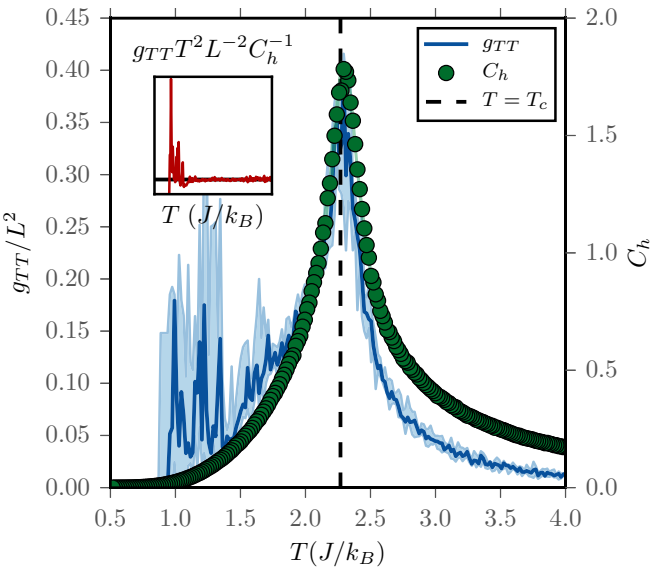


FIG. 5. Blue continuous curve is the TT component of the FIM and the green dots are the heat capacity in the two-dimensional Ising model on a 25×25 grid. Shaded blue and green regions indicate the 5 and 95 percentiles computed from five simulations. The inset shows the ratio of FI to heat capacity ($g_{TT}T^2C_h^{-1}L^{-2}$) which according to Eq. (13) is equal to 1 (black horizontal line in the inset).

VI. DISCUSSION

There are several technical points we wish to mention about the implementation of the method. First, we performed the same Ising computation with a smaller grid spacing ($dT = 0.007$). This led to a much worse signal-to-noise ratio because the very close densities caused large peaks to occur, especially in the low-temperature range. Second, it is important to find the most suitable parameters for DEFT. If the bounding box is too small or the number of grid points too small or too large, the estimated density will have multiple peaks which are not apparent in the data. Thus we recommend plotting the result of DEFT together with a histogram for several data points to make sure the convergence is good. Third, in the computation of Eq. (3a) the term $1/p$ may contribute large values at very small p. Equivalently with Eq. (3b), when $p(x|\theta \pm \Delta\theta)$ are small, their logarithm will again be large. This requires the introduction of a numerical cutoff. It is common practice to set the contribution of a term where $p(x) = 0$ to zero [19]. We thus introduced a cutoff such that if any of the estimates at a particular point are less than the cutoff, the contribution of this point to the integral will be zero. We investigated the effect of this cutoff for a range of values between 10^{-20} and 10^{-2}. The value of the cutoff had very little effect. In the Ising model, the only effect was to change the size of the low-temperature region where the FI is exactly zero (the lower the cutoff was, the smaller the region was). In producing Fig. 5 we used a value of 10^{-10}. Last, we mention that the plots in Fig. 5 are obtained by the use of Eq. (3b).

In summary, the algorithm to compute the FI from samples is the following. First obtain a good nonparametric estimate of the density at each parameter point. When using DEFT, make sure to adjust G, α, and the bounding box for proper convergence. Second, find the appropriate parameter distance $\Delta\theta$. This can be aided by computing the ε parameter. Third, if necessary, use a cutoff for very low values of the probability density. When using DEFT to perform the density estimation, the procedure is limited by the limitations of DEFT. It is especially important to note that so far DEFT has been implemented in one and two dimensions. This is because the number of grid points necessary to evaluate the density using DEFT increases exponentially with the number of dimensions. It is important to note, however, that this limits only the dimensionality of the data, not the number of parameters or the number of samples, which scales well to higher dimensions. A software implementation in Python of the method is available (see Ref. [36]).

ACKNOWLEDGMENTS

O.H.S. would like to thank Joan Massó and Antoni Arbona from the University of the Balearic Islands and Fredrik Jansson and Lisa Jenny Krieg from the University of Amsterdam for enlightening discussions. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grants No. 317534 and No. 318121 (Sophocles and TopDrim projects). A.G.H. wishes to acknowledge partial funding by the Russian Scientific Foundation, under Grant No. 14-11-00826. P.M.A.S. wishes to acknowledge partial funding by the Russian Scientific Foundation, under Grant No. 14-21-00137. We would like to thank an anonymous reviewer for pointing us to the EMST method and to all three anonymous reviewers for their constructive comments.

APPENDIX: COMPARISON OF KDE, DEFT, AND EMST FOR DIFFERENT VALUES OF ε

Here we add additional comparison plots between KDE, DEFT, and EMST for values of ε which are not the "optimal"

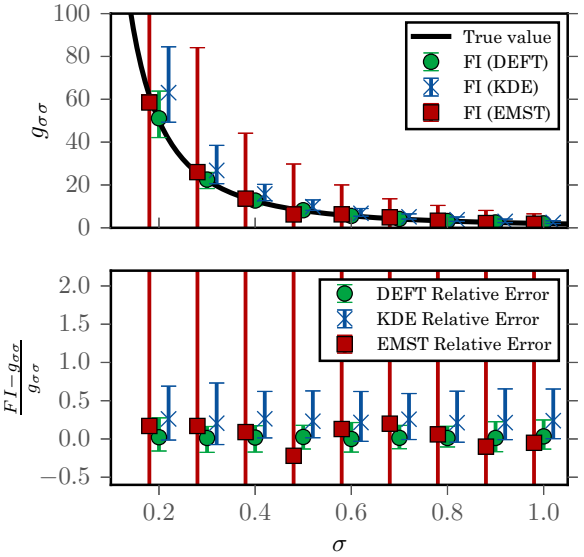


FIG. 6. Comparison between the three methods reviewed in the main text for $\varepsilon = 0.1$. All other parameters are equal to those in Fig. 2.

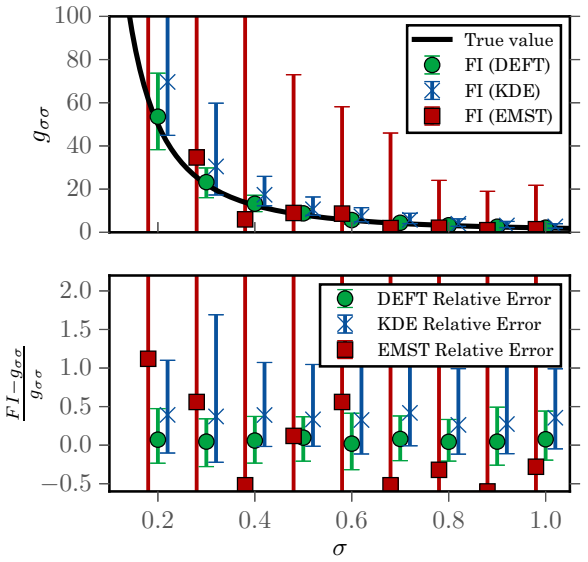


FIG. 7. Comparison between the three methods reviewed in the main text for $\varepsilon = 0.2$. All other parameters are equal to those in Fig. 2.

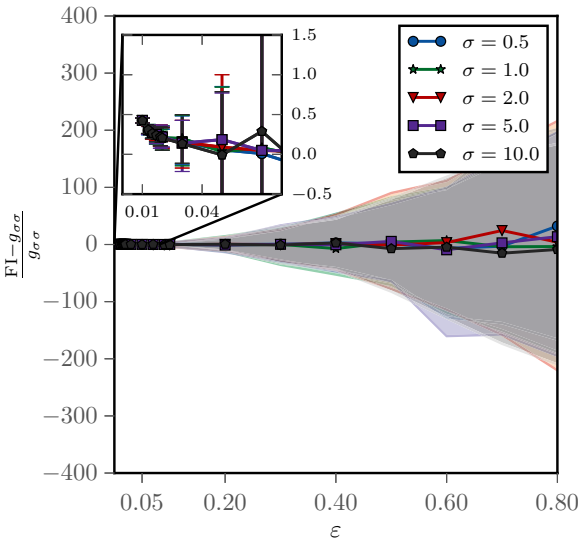


FIG. 8. A plot of the accuracy of EMST as a function of ε with the same parameters as in Fig. 3.

0.05. As we can clearly see, when ε is increased from 0.05 to 0.1 (in Fig. 6) the error bars for the EMST method increase dramatically while DEFT remains quite accurate. This becomes even more pronounced in Fig. 7 for $\varepsilon = 0.2$. In Fig. 8 we plot the error of the EMST method for various values of σ as a function of ε. The errors increase dramatically in comparison with DEFT as can be seen by the scale on the y axis. One also sees in the inset the convergence of the method for low values of ε (high $\Delta\sigma$) to a constant (about 0.57). We believe this to be related to the convergence of the Friedman-Rafsky statistic for large separations of the densities but leave the exact study of this for future research.

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley & Sons, New York, 2006), p. 640.
[2] V. V. Fedorov, *Theory of Optimal Experiments* (Academic Press, New York, 1972).
[3] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry: Translations of Mathematical Monographs*, Vol. 191 (American Mathematical Society, Providence, 2000).
[4] G. Ruppeiner, Phys. Rev. A **20**, 1608 (1979).
[5] G. Ruppeiner and C. Davis, Phys. Rev. A **41**, 2200 (1990).
[6] G. Ruppeiner, Rev. Mod. Phys. **67**, 605 (1995).
[7] G. Ruppeiner, A. Sahay, T. Sarkar, and G. Sengupta, Phys. Rev. E **86**, 052103 (2012).
[8] R. S. Ingarden, H. Janyszek, A. Kossakowski, and T. Kawaguchi, Tensor (NS) **37**, 105 (1982).
[9] H. Janyszek and R. Mrugala, Phys. Rev. A **39**, 6515 (1989).
[10] H. Janyszek, J. Phys. A.: Math. Gen. **23**, 477 (1990).
[11] D. Brody and N. Rivier, Phys. Rev. E **51**, 1006 (1995).
[12] D. C. Brody and D. W. Hook, J. Phys. A: Math. Theor. **42**, 023001 (2009).
[13] P. Kumar, S. Mahapatra, P. Phukon, and T. Sarkar, Phys. Rev. E **86**, 051117 (2012).
[14] T. Obata, H. Hara, and K. Endo, Phys. Rev. A **45**, 6997 (1992).
[15] T. Obata, H. Oshima, and H. Hara, Phys. Rev. E **56**, 213 (1997).
[16] A. L. Mayer, C. W. Pawlowski, and H. Cabezas, Ecol. Modell. **195**, 72 (2006).
[17] S. A. Frank, J. Evol. Biol. **22**, 231 (2009).
[18] M. Prokopenko, J. T. Lizier, O. Obst, and X. R. Wang, Phys. Rev. E **84**, 041116 (2011).
[19] X. R. Wang, J. T. Lizier, and M. Prokopenko, Artif. Life **17**, 315 (2011).
[20] J. Hidalgo, J. Grilli, S. Suweis, M. A. Muñoz, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. USA **111**, 10095 (2014).
[21] J. Hidalgo, J. Grilli, S. Suweis, A. Maritan, and M. A. Munoz, arXiv:1510.05941.
[22] J. C. Spall, J. Comput. Graph. Stat. **14**, 889 (2005).
[23] S. Amari and A. Cichocki, Bull. Polish Acad. Sci. Tech. Sci. **58**, 183 (2010).
[24] V. Berisha and A. O. Hero, IEEE Signal Process. Lett. **22**, 988 (2015).
[25] J. H. Friedman and L. C. Rafsky, Ann. Stat. **7**, 697 (1979).
[26] P. J. Huber, Ann. Stat. **2**, 1029 (1974).
[27] L. Kostal and O. Pokora, Entropy **14**, 1221 (2012).
[28] I. Good and R. Gaskins, Biometrika **58**, 255 (1971).
[29] I. N. Sanov, Mat. Sb. **42**, 11 (1957).
[30] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (CRC Press, Boca Raton, FL, 1986), Vol. 26.
[31] E. Walter and L. Pronzato, *Communications and Control Engineering* (Springer Verlag, New York, 1997).
[32] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996).
[33] J. B. Kinney, Phys. Rev. E **90**, 011301 (2014).
[34] This well-known result is a special case of the more general expression, Eq. (2), since the KL divergence is also an f divergence. It can be derived using a Taylor expansion and the definition of the FI.
[35] L. Onsager, Phys. Rev. **65**, 117 (1944).
[36] http://uva.computationalscience.nl/research/software/npfi.