Pseudolikelihood Decimation Algorithm Improving the Inference of the Interaction Network in a General Class of Ising Models

Aurélien Decelle[1] and Federico Ricci-Tersenghi[1,2]

[1]*Dipartimento di Fisica, Università La Sapienza, Piazzale Aldo Moro 5, I-00185 Roma, Italy*
[2]*INFN-Sezione di Roma 1 and CNR-IPCF, UOS di Roma, Piazzale Aldo Moro 5, I-00185 Roma, Italy*

In this Letter we propose a new method to infer the topology of the interaction network in pairwise models with Ising variables. By using the pseudolikelihood method (PLM) at high temperature, it is generally possible to distinguish between zero and nonzero couplings because a clear gap separate the two groups. However at lower temperatures the PLM is much less effective and the result depends on subjective choices, such as the value of the ℓ_1 regularizer and that of the threshold to separate nonzero couplings from null ones. We introduce a decimation procedure based on the PLM that recursively sets to zero the less significant couplings, until the variation of the pseudolikelihood signals that relevant couplings are being removed. The new method is fully automated and does not require any subjective choice by the user. Numerical tests have been performed on a wide class of Ising models, having different topologies (from random graphs to finite dimensional lattices) and different couplings (both diluted ferromagnets in a field and spin glasses). These numerical results show that the new algorithm performs better than standard PLM.

Recent years have seen a growing interest of the statistical physics community in the inverse Ising problem [1–7]. Although the problem has been known for a long time as Boltzmann machine learning [8], the renewed interest is linked to the large number of data sets coming from many different fields—e.g., biology, physics, neuroscience—that require a quantitative description. Buiglding reliable models for these data sets has become therefore a fundamental problem [9–12]. Given a data set and a model, the inverse problem aims to find the parameters of the model that best fit the data. Among all possible models, the Ising one, although very simple—it involves only pairwise interactions between discrete variables—can take into account a wide range of phenomena. It is thus natural to develop inference methods for this statistical model, as initially done by Boltzmann machine learning [8].

A common approach in Bayesian inference is to infer the model couplings by maximizing the likelihood function. Unfortunately, the likelihood depends on the partition function which is extremely difficult to compute in general. Therefore we use the pseudolikelihood approximation [13,14] to perform inference on Ising models of reasonable size. This method has at least three advantages. (i) The pseudolikelihood function (PLF) can be maximized in polynomial time, (ii) the method is known to be exact in the case of infinite sampling [2], and (iii) as we will see in this Letter, the PLF can be used as an indicator of how good the reconstruction is, allowing one to decrease accurately the number of model parameters.

We begin by introducing the pseudolikelihood method (PLM) that has been already tested in inferring finite dimensional models and found to be very effective [2].

We also recall the ℓ_1-regularization extension and the thresholding procedure for discriminating nonzero couplings. We discuss the problem of very large coupling appearing in the PLM and how we solve this issue without using the ℓ_1 regularization. Finally, we describe the new decimation procedure to select accurately the model parameters and we show that the PLF can be used as a reliable likelihood. We conclude by showing that our new symmetrized PLM with decimation provides a very good inference of model couplings, better than standard PLM, in a wide class of Ising models.

Let us consider an Ising model with probability measure $P(\underline{s}) = Z^{-1} \exp(\beta \sum_{i<j} J_{ij} s_i s_j + \beta \sum_i h_i s_i)$, whose parameters $\{J_{ij}, h_i\}$ we want to infer, given a set of M configurations $\{\underline{s}^{(k)}\}_{k=1...,M}$ independently sampled from $P(\underline{s})$. Bayesian inference prescribes to maximize with respect to the model parameters the likelihood of the data $\mathcal{L} = \beta \sum_{i<j} J_{ij} \langle s_i s_j \rangle_D + \beta \sum_i h_i \langle s_i \rangle_D - \log(Z)$, where $\langle \cdot \rangle_D$ is the average over the data. The maximum of $\mathcal{L}$ corresponds to parameters $\{J_{ij}^*, h_i^*\}$ matching the magnetizations and the correlations of the inferred model with those from the data. There are two general methods to find the parameters $\{J_{ij}^*, h_i^*\}$ that fulfill these conditions. The first one is by finding the maximum of $\mathcal{L}$ using a gradient descent approach, but this requires to compute the partition function many times, which is impossible in practice. The second one, named Boltzmann machine learning [8], computes magnetizations and correlations by Monte Carlo methods and updates the parameters $\{J_{ij}, h_i\}$ accordingly. Again this requires to run several long Monte Carlo simulations evaluating the average values with a precision good enough.

An important aspect in the inverse problem is therefore to find an approximation that provides a reasonably good estimate for the model parameters in a quick time (usually polynomial in the system size). A review of known approximations can be found in [15], to which we should add the more recent methods, such as those based on the Bethe approximation [3,4], the adaptive cluster expansion [1], and the probabilistic flow method [16]. In this Letter we compare our new decimation technique based on the PLM against the ℓ_1 regularization, which is often used to find the interaction topology [2,14,17]. The standard mean-field method will not be considered, since in general it performs weakly on finite dimensional systems. Neither will the adaptive cluster expansion or the probabilistic flow method be considered, since they strongly depend on several subjective choices (thresholds, dynamics, etc…).

Instead of using the likelihood function, which is very hard to compute, we use the PLF, $\mathcal{PL} = \sum_r \mathcal{L}_r$, where the index r runs over all variables and the "local" likelihood functions are $\mathcal{L}_r = M^{-1} \sum_{k=1}^{M} \log p(s_r^{(k)}|\underline{s}_{\backslash r}^{(k)})$, with $p(s_r|\underline{s}_{\backslash r}) \equiv [1 + \exp(-2\beta s_r(h_r + \sum_{j \neq r} J_{rj}s_j))]^{-1}$ being the conditional probability of variable s_r given the rest of the system.

The standard implementation of the PLM consists in maximizing each of the N local likelihood functions $\mathcal{L}_r$ separately, thus getting two different estimates for each coupling J_{ij}: a first one J_{ij}^{*i} from the maximum of $\mathcal{L}_i$ and a second one J_{ij}^{*j} from the maximum of $\mathcal{L}_j$ (hereafter we assume $h_i = 0$ and $J_{ij} \in \{0, 1\}$ for the ease of presentation). Since the Ising model has symmetric couplings, the final estimate for the coupling J_{ij} is then obtained by taking the average, $J_{ij}^* = (J_{ij}^{*i} + J_{ij}^{*j})/2$.

This method has been confronted against mean-field methods for the SK model in [2] and it clearly estimates the couplings much better at low temperatures. For sparse models, its ability to infer correctly the interaction network (i.e., which couplings are nonzero) is largely improved by the use of the ℓ_1 regularization [14], thus maximizing the local functions $\mathcal{L}_r^{\ell_1} = \mathcal{L}_r + \lambda \sum_{j \neq r} |J_{rj}|$, with a suitably chosen (and not too large) λ regularizer. A further improvement in inferring the model topology has been achieved [2] by setting to zero all couplings whose estimate is below a threshold, $|J_{ij}^*| < \delta$ (but the choice of δ is delicate, as we discuss below).

Using the standard PLM we observe that, in difficult situations (e.g., when M is not large enough and temperature is low) some couplings are largely overestimated, $|J_{ij}^*| \gg 1$. In those cases, the inferred couplings are not symmetric ($J_{ij}^{*i} \neq J_{ij}^{*j}$) and only one of the two estimates is very large. The origin of this problem is to be found in the data information content, that is sometimes very poor around some variable s_i: in that case the estimates J_{ij}^{*i} are strongly unreliable (think, e.g., to what happen if in the M samples s_i and its neighbors were almost always

perfectly aligned). This problem is often partially solved by using the ℓ_1 regularization: the regularization parameter λ adds a penalty on nonzero couplings and therefore prevents a coupling from being too large. A drawback of this approach is the tendency to underestimate globally the couplings and is therefore not completely satisfying. In this Letter, we choose a different solution: by maximizing the PLF $\mathcal{PL}$ (rather than the N functions $\mathcal{L}_r$ separately) we look for a compromise, where the estimate J_{ij}^* must be such that both $\mathcal{L}_i$ and $\mathcal{L}_j$ are reasonably large. The advantage of this maximization is that it provides a *unique* estimate for each coupling, which is in general of the right order of magnitude (unless the information content of the data is poor in a wide region). The PLF can be maximized by a standard Newton method, while for the ℓ_1-regularized functions we use the one of Ref. [18].

In inference problems, the "model selection" is the ability to reduce the number of model parameters in order to keep only the essential ones (by Occam's razor rule the simplest model fitting a data set is the one to be preferred). This model selection is both crucial to identify the structure underlying a data set (e.g., the interaction network or the model topology) and to improve the quality of the coupling estimates (relevant couplings are better inferred after excluding the insignificant couplings from the model).

A common approach in the inverse Ising problem is to choose which are the null couplings by putting a cut-off on their estimated values. This is a reasonable choice as long as a clear gap separates the estimates of null couplings ($\boldsymbol{J}_0$) from those of nonzero couplings ($\boldsymbol{J}_1$), as it happens in the left panel of Fig. 1. However at a lower temperature the gap is completely absent and a good δ thresholding to split $\boldsymbol{J}_0$ and $\boldsymbol{J}_1$ is impossible (central panel of Fig. 1). For sparse models, the use of the ℓ_1 regularization may induce a new gap (see right panel of Fig. 1), but this ℓ_1-induced gap does not always separates correctly $\boldsymbol{J}_1$ from $\boldsymbol{J}_0$ couplings and the
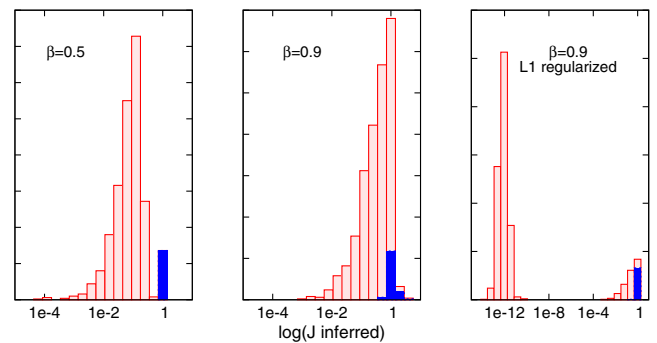

FIG. 1 (color online).   Histograms of couplings inferred by PLM in a typical sample of 2D ferromagnetic Ising model with 30% dilution ($M = 4500$). Left: at $\beta = 0.5$ a clear gap identifies couplings in $\boldsymbol{J}_1$ (drawn in dark blue). Center: at $\beta = 0.9$ there is no gap and couplings in $\boldsymbol{J}_1$ can not be identified. Right: with ℓ_1 regularization ($\lambda = 0.01$) a gap is present, but does not allow us to recover the topology perfectly.

topology can not be correctly recovered. Moreover the estimates of J_1 are systematically smaller than the true values, and this is unavoidable when using the ℓ_1 regularization. Finally, there is no consensus on how to choose the λ value.

We propose a new method for inferring nonzero couplings J_1 that solves all the above problems. Our idea is to recursively set to zero couplings which are estimated very small by the PLM: we always maximize $\mathcal{PL}$ so as to avoid too large couplings and the bias due to the ℓ_1 regularization. At each decimation step we set to zero a finite fraction ρ of the remaining couplings; as such, the total number of steps is $O(\log N)$ and the PLM + decimation algorithm is competitively fast. This fraction ρ is actually the only choice left to the user and the results are largely independent on it (in our tests we have used $\rho \leq 0.05$). Setting couplings to zero gradually is equivalent to using an adaptive threshold with a very small δ value: so our new method should perform better than any standard thresholding procedure, especially because it does not require the existence of a gap in the inferred couplings (thus avoiding the use of ℓ_1 regularization that produces biased estimates and a strong λ dependence).

The stopping criterion for the decimation procedure is based on the behavior of the PLF. Indeed, we expect that as long as the decimation procedure sets to zero couplings in J_0 which are unnecessary to fit the data, the PLF should not change significantly. On the contrary, the pruning of a coupling in J_1 should produce a drastic decrease in the PLF value. This expected behavior is confirmed by the numerical simulations. In practice we would like to stop the decimation at the point where the PLF variation, $\Delta\mathcal{PL}/\Delta n$ with Δn being the number decimated couplings in the last step, goes from "small" to "large" values. To make these two adjectives quantitative, we can compute the overall mean PLF variation during the decimation, that is the change in PLF between the fully connected model [where $\mathcal{PL}$ is maximized over all the $N(N-1)/2$ possible couplings and takes value $\mathcal{PL}_{\max}$] and the model of independent variables [no couplings left by the decimation and PLF equal to $-N \log(2)$]. The mean PLF variation is thus equal to $(\mathcal{PL}_{\max} + N \log(2))2/(N(N-1))$, and we propose to stop the decimation where the PLF variation reaches this mean value, that should separate "small" from "large" PLF variations. In practice it is more convenient to define the stopping point as the maximum of the *tilted PLF* (tPLF) $\mathcal{PL}^{\mathrm{tilted}} \equiv \mathcal{PL} - x\mathcal{PL}_{\max} + (1-x)N \log(2)$, where x is the fraction of nondecimated coupling. It is easy to check that $\mathcal{PL}^{\mathrm{tilted}} = 0$ both before starting the decimation ($x = 1$) and on a model with no coupling ($x = 0$). In the interval [0,1] a maximum appears if correlations are present in the data set.

In Fig. 2 we show, for a case where inference is difficult, the tPLF as a function of the number of nondecimated couplings and the corresponding error in inferring
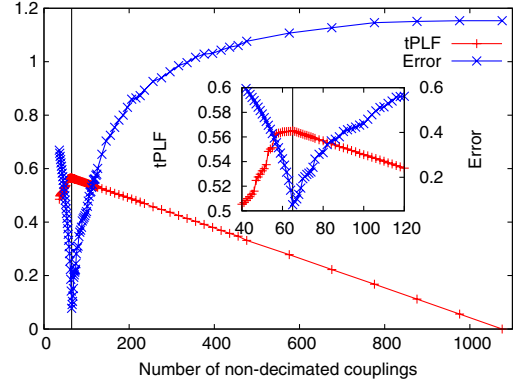

FIG. 2 (color online). Same model as Fig. 1 with $\beta = 0.9$ and $M = 4500$. The tPLF increases as the number of nondecimated couplings is reduced until it exhibits a maximum. In this case the maximum corresponds to a complete recovery of the graph topology and therefore to a small reconstruction error. The inset is a zoom on the maximum region.

couplings, defined as $\epsilon = \sqrt{\sum_{i<j}(J_{ij} - J^*_{ij})^2 / \sum_{i<j} J^2_{ij}}$. The zoom in the inset clearly shows that the maximum in the tPLF does corresponds to the minimum in the inference error. Please notice also as the density of data points along the curves changes, because we have decreased the value of ρ during the decimation in order to spend less time in the initial part (which is easy) and have denser points close to maximum (and thus improve its location).

Our new inference algorithm (PLM + decimation) can be very well compared to the standard PLM with ℓ_1 regularization and δ thresholding by plotting the corresponding ROC curves (see Fig. 3). Each ROC curve is
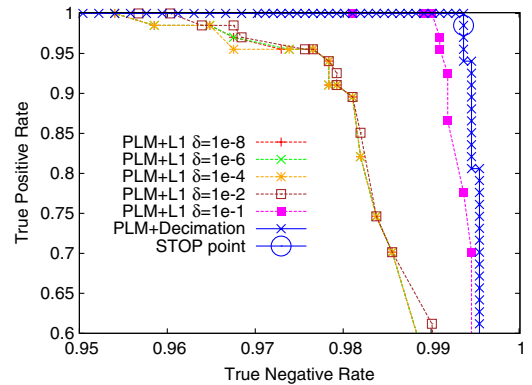

FIG. 3 (color online). ROC curves for a typical sample of the 2D ferromagnetic Ising model with 30% of dilution at $\beta = 1.0$ and $M = 4500$ (difficult case). The upper right corner corresponds to the perfect reconstruction. For PLM $+ \ell_1$ curves are drawn by varying λ and they almost coincide if the threshold δ is correctly chosen in the gap, $\delta \in [10^{-8}, 10^{-2}]$. The better value $\delta = 0.1$ can not be chosen without previous knowledge about the topology. The PLM + decimation is clearly inferring topology better, even if PLM $+ \ell_1$ is finely tuned over λ and δ.

obtained by plotting parametrically the true positive rate (i.e., couplings in J_1 inferred as nonzero divided by the total number of couplings in J_1) versus the true negative rate (i.e., couplings in J_0 inferred as null divided by the total number of couplings in J_0). The ROC curve for an exact inference method run on noiseless data would pass through the upper right corner. In general an inference method is better the larger the area below the ROC curve. In the present case, the ROC curves for the standard PLM have been drawn by varying λ at fixed δ, while for the PLM + decimation the ROC curve has been drawn by varying the fraction of decimated couplings. Clearly the new method is outperforming standard PLM, even if the latter were optimized over λ and δ. To this respect, it is worth noticing that δ values in the range $[10^{-8}, 10^{-2}]$ do actually fall in the gap (see lower panel of Fig. 1) and lead to very similar ROC curves, while $\delta = 0.1$ is outside the gap and would be impossible to choose that value without knowing in advance the topology we are looking for. So the better ROC curve with $\delta = 0.1$ is practically unachievable, and in general we believe that optimizing over δ should not produce any sensible improvement as long as δ is correctly chosen in the gap (to this respect the improvement obtained in Ref. [2] makes us suspect that a too-large δ value was chosen based on the previous knowledge of the topology). So a fair comparison of our new method with the standard PLM should be made by choosing $\delta \leq 10^{-2}$ and the improvement is then very large. The stopping point selected by maximizing the tPLF is shown by a large dot in Fig. 3, and is indeed the closest to the upper right corner (full topology recovery).

We applied the standard PLM + ℓ_1 and our new decimation algorithm to other models to compare their ability in recovering the topology of the underlying graph. We studied two diluted ferromagnetic models (in 3D with 55% dilution and in 2D with 30% dilution and random fields) and two spin glass models ($J_{ij} = \pm 1$) on Erdös-Renyi random graphs with average degrees $c = 3$ and $c = 4$. To measure the performance of the inference methods we show in Fig. 4 the fraction of successfully reconstructed neighborhoods as a function of the number of samples M. It is clear that the new decimation algorithm outperforms standard PLM + ℓ_1 in reconstructing any model topology. Please consider that for PLM + ℓ_1 method λ is chosen as to minimize the distance on the ROC curve from perfect reconstruction (the point closer to the upper right corner) and therefore this choice is made knowing the topology. The threshold δ was always set in the gap (typically $\delta \sim 10^{-6}$) and we observe that this method does not manage to separate J_0 from J_1, even at high sampling. In order to keep the figure more readable, the uncertainties on data points are drawn only for one data set in the upper left panel (the others errors being similar or smaller). These uncertainties have been computed by resampling the configurations passed as input to the inference algorithm.
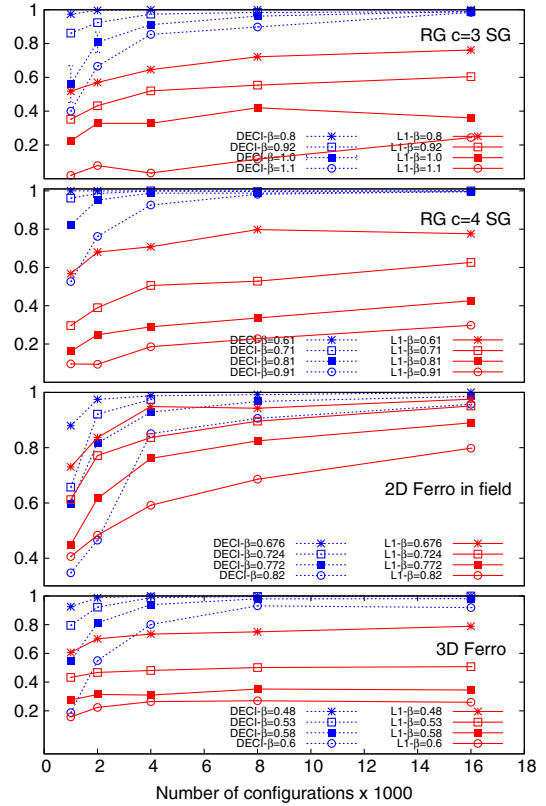

FIG. 4 (color online). Neighborhood reconstruction for four different models as a function of the number of samples. Dashed blue lines are results by the new decimation algorithm, while full red lines are by PLM + ℓ_1. Each point is the average over 10 samples. Top to bottom: spin glass models, $J_{ij} = \pm 1$, on random graph, $N = 50$, with average degree $c = 3$ and $c = 4$; diluted ferromagnetic models, in 2D, $N = 49$, with 30% of dilution and a random Gaussian field with zero mean and variance 0.3, and in 3D, $N = 64$, with 55% of dilution. The decimation algorithm is always better and performs particularly well on random graphs.

Compared to the methods presented in [19,20], our algorithm performs similarly and in a quicker time, since it does not need to compute the inverse of the Hessian.

In conclusion, we have presented a new method for inferring the interaction topology of Ising models which is based on the PLM and a decimation procedure that recursively sets to zero couplings which are inferred as the weakest. The procedure is fully automated (apart from the choice of ρ which is mostly irrelevant for the results) and provides a unique answer to the inverse Ising problem and the corresponding model selection problem. Execution times are comparable to those of standard PLM (i.e., polynomial in system size), apart from an extra $O(\log N)$ multiplicative factor (but remind that maximization without the ℓ_1 regularizer is easier). As the standard PLM, also our new method is exact in the limit of very large number of samples M. For finite (and small) values of M we have run extensive numerical tests on a wide class of Ising models, with different topologies (from random

graphs to finite dimensional lattices) and different couplings (both diluted ferromagnets in a field and spin glasses). The results show that the new algorithm performs much better than standard PLM with ℓ_1 regularization and δ thresholding, which was considered among the best inference techniques available.

———————

[1] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011).
[2] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).
[3] F. Ricci-Tersenghi, J. Stat. Mech. (2012) P08015.
[4] H. C. Nguyen and J. Berg, J. Stat. Mech. (2012) P03004; H. C. Nguyen and J. Berg, Phys. Rev. Lett. **109**, 050602 (2012).
[5] S. Cocco and R. Monasson, J. Stat. Phys. **147**, 252 (2012).
[6] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[7] J. Raymond and F. Ricci-Tersenghi, Phys. Rev. E **87**, 052111 (2013); J. Raymond and F. Ricci-Tersenghi, *Proceedings of IEEE International Conference on Communications Workshops (ICC)* (2013), p. 1429.
[8] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cogn. Sci. **9**, 147 (1985).
[9] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).
[10] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2009).
[11] S. Cocco, S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. U.S.A. **106**, 14058 (2009).
[12] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Proc. Natl. Acad. Sci. U.S.A. **107**, 5405 (2010).
[13] J. Besag, The Statistician **24**, 179 (1975).
[14] B. Ravikumar, M. Wainwright, and J. Lafferty, Ann. Stat. **38**, 1287 (2010).
[15] Y. Roudi, E. Aurell, and J. A. Hertz, Front. Comput. Neurosci. **3**, 22 (2009).
[16] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, Phys. Rev. Lett. **107**, 220601 (2011).
[17] J. Bento and A. Montanari, in *Proceeding of NIPS 22* (2009), p. 1303.
[18] K. Koh, S.-J. Kim, and S. Boyd, J. Mach. Learn. Res. **8**, 1519 (2007).
[19] Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel, NIPS'89, 598 (1990).
[20] B. Hassibi and D. G. Stork, NIPS'92, 164 (1993).