ELSEVIER

9 July 2001

PHYSICS LETTERS A

Physics Letters A 285 (2001) 279–285

www.elsevier.com/locate/pla

Cost functions for pairwise data clustering

L. Angelini [a,c], L. Nitti [b,c], M. Pellicoro [a,c], S. Stramaglia [a,c,*]

[a] *Universitá degli Studi, Dipartimento di Fisica, via Amendola 173, 70126 Bari, Italy*
[b] *Universitá degli Studi, DETO Sezione Fisica Medica, Piazza Giulio Cesare 11, 70124 Bari, Italy*
[c] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, via Amendola 173, 70126 Bari, Italy*

Received 20 March 2001; accepted 7 May 2001
Communicated by A.R. Bishop

Abstract

Cost functions for non-hierarchical pairwise clustering are introduced, in the probabilistic autoencoder framework, by the request of maximal average similarity between input and the output of the autoencoder. Clustering is thus formulated as the problem of finding the ground state of Potts spins Hamiltonians. The partition, provided by this procedure, identifies clusters with dense connected regions in the data space.

Clustering methods aim at partitioning a set of data-points in classes such that points that belong to the same class are more alike than points that belong to different classes [1]. These classes are called clusters and their number may be preassigned or can be a parameter to be determined by the algorithm. There exist applications of clustering in such diverse fields as pattern recognition [2], astrophysics [3], communications [4], biology [5], business [6] and many others. Two main approaches to clustering can be identified: parametric and non-parametric clustering.

Non-parametric approaches make few assumptions about about the data structure and, typically, follow some local criterion for the construction of clusters. Typical examples of the non-parametric approach are the agglomerative and divisive algorithms that produce dendrograms. In the last years non-parametric clustering algorithms have been introduced employing the statistical properties of physical systems. The super-paramagnetic approach by Domany and coworkers [7] exploits the analogy to a model granular magnet: the spin–spin correlation of a Potts model, living on the data-points lattice and with pair couplings decreasing with the distance, is used to partition points in clusters. The synchronization properties of a system of coupled chaotic maps are used in [8] to produce hierarchical clustering.

Parametric methods make some assumptions about the underlying data structure. Generative mixture models [9] treat clustering as a problem of density estimation: data are viewed as coming from a mixture of probability distributions, each representing a different cluster, and the parameters of these distributions are adjusted to achieve a good match with the distribution of the input data. This can be obtained by maximizing the data likelihood (ML) or the posterior (MAP) if additional prior information on the parameters is available [10].

Many parametric clustering methods are based on a cost function: the best partition of points in clusters is assumed to be the one with minimum cost. Often cost functions incorporate the loss of information in-

* Corresponding author.
E-mail address: stramaglia@ba.infin.it (S. Stramaglia).

0375-9601/01/$ – see front matter
PII: S0375-9601(01)00373-5

curred by the clustering procedure when trying to reconstruct the original data from the compressed cluster representation: the most popular algorithm to optimize a cost function is K-means [9]. Starting from a statistical *ansatz* and invoking maximum likelihood leads to a cost function which has been observed to work for clustering financial time series [11].

It is important to stress the difference between *central* clustering, where it is assumed that each cluster can be represented by a prototype [12], and *pairwise* clustering where data are indirectly characterized by pairwise comparison instead of explicit coordinates [13]; pairwise algorithms require as input only the matrix of dissimilarities. Obviously the choice of the measure of dissimilarity is not unique and it is crucial for the performance of any pairwise clustering method. It is worth remarking that it often happens that the dissimilarity matrix violates the requirements of a distance measure, i.e., the triangular inequality does not necessarily holds.

Folded Markov chains are used in the probabilistic autoencoder framework to derive cost functions for clustering [14]. Some examples of two-stage folded Markov chains, and the corresponding algorithms for clustering and topographic mapping [15], are thoroughly analyzed in [16], where it is also shown that the cost function for pairwise clustering, introduced in [13], may be seen as a consequence of Bayes' theorem and the requirement of minimal average distorsion in a probabilistic autoencoder.

It is the purpose of this Letter to introduce a new class of cost functions for pairwise clustering which can be obtained, in the autoencoder frame, by requiring *maximal similarity* instead of minimal distorsion. We show that the cost functions here introduced provide a non-hierarchical clustering of points where dense connected regions of points in the data space are recognized as clusters.

Let us now discuss autoencoders described by one-stage folded Markov chains. Let us consider a point x, in a data space, sampled with probability distribution $P_0(x)$; a code index $\alpha \in \{1,\dots,q\}$ is assigned to x according to conditional probabilities $P(\alpha|x)$. A reconstructed version of the input, x', is then obtained by use of the Bayesian decoder:

$$P(x'|\alpha) = \frac{P(\alpha|x')P_0(x')}{P(\alpha)}. \tag{1}$$

The joint distribution of x, x' and α, describing this encoding–decoding process, is

$$P(x,x',\alpha) = P_0(x)P(\alpha|x)P(x'|\alpha); \tag{2}$$

owing to (1), the joint distribution reads

$$P(x,x',\alpha) = \frac{P_0(x)P_0(x')P(\alpha|x)P(\alpha|x')}{P(\alpha)}. \tag{3}$$

The conditional probabilities $\{P(\alpha|x)\}$ are the free parameters that must be adjusted to force the autoencoder to emulate the identity map on the data space.

Let $d(x,x')$ be a measure of the distorsion between input and output of the autoencoder. The average distorsion is then given by

$$\mathcal{D} = \sum_{\alpha=1}^{q} \int dx \int dx' \, \frac{P_0(x)P_0(x')P(\alpha|x)P(\alpha|x')}{P(\alpha)} \times d(x,x'). \tag{4}$$

Moreover, let $s(x,x')$ be a measure of the similarity between input and output; the average similarity is then given by

$$\mathcal{S} = \sum_{\alpha=1}^{q} \int dx \int dx' \, \frac{P_0(x)P_0(x')P(\alpha|x)P(\alpha|x')}{P(\alpha)} \times s(x,x'). \tag{5}$$

It is natural to postulate a one-to-one mapping between values of distorsion and similarity, $s = F(d)$, with F a strictly decreasing function. A good autoencoder is obviously characterized by a low value of $\mathcal{D}$ and high value of $\mathcal{S}$. However, we remark that the two requirements $\mathrm{Min}(\mathcal{D})$ and $\mathrm{Max}(\mathcal{S})$, for reasonable choices of F, are not generically equivalent.

Now we turn back to the clustering problem. Given a data-set $\{x_i\}$ of cardinality N, partitioning these points in q classes corresponds, in this frame, to design an autoencoder, with q code indexes, acting on data space. We choose the encoder to be deterministic:

$$P(\alpha|x) = \delta_{\alpha\,\sigma(x)}, \tag{6}$$

$\sigma(x) \in \{1,\dots,q\}$ being the code index associated to x. The estimate for the average distorsion (4), based on the data-set at hand, is given by $\hat{\mathcal{D}} = N H_d[\sigma]$, where we introduce the Hamiltonian H_d for the Potts variables $\{\sigma_i\}$:

$$H_d[\sigma] = \sum_{\alpha=1}^{q} \frac{\sum_{i,j=1}^{N} \delta_{\alpha\sigma_i}\delta_{\alpha\sigma_j} d_{ij}}{\sum_{k=1}^{N} \delta_{\alpha\sigma_k}}, \tag{7}$$

where $\sigma_i = \sigma(x_i)$, $d_{ij} = d(x_i, x_j)$. It turns out that H_d is equivalent to the cost function for pairwise clustering, influential in the clustering literature, introduced in [13].

The estimate for the average similarity is, similarly, given by $\hat{\mathcal{S}} = -NH_s[\sigma]$, where we introduce the Hamiltonian H_s:

$$H_s[\sigma] = -\sum_{\alpha=1}^{q} \frac{\sum_{i,j=1}^{N} \delta_{\alpha\sigma_i}\delta_{\alpha\sigma_j} s_{ij}}{\sum_{k=1}^{N} \delta_{\alpha\sigma_k}}. \tag{8}$$

If we choose the autoencoder by minimizing the average distorsion, then the best partition of the data-set in q classes corresponds to the ground state of H_d. If we choose it by maximizing the average similarity, then the ground state of H_s must be sought for, instead. Since both $\{d_{ij}\}$ and $\{s_{ij}\}$ may be taken positive, it follows that H_d is characterized by antiferromagnetic couplings between the Potts variables, while H_s is made of ferromagnetic couplings. Denominators in both H_d and H_s serve to enforce the coherence among the q clusters. In particular, without the denominator the ground state of H_s would correspond to a single big cluster.

The form of the function F, determining the relation between s and d, has to be specified. In what follows we consider two forms of this relation. A scale-free relation

$$s_{ij} = F_\gamma(d_{ij}) = \left(\frac{d_{ij}}{\langle d\rangle}\right)^{-\gamma}, \tag{9}$$

depending on the exponent γ, and a scale-dependent relation

$$s_{ij} = F_a(d_{ij}) = \exp\left(-\frac{1}{2a^2}\left(\frac{d_{ij}}{\langle d\rangle}\right)^2\right), \tag{10}$$

depending on the scale a. In the formulas above, $\langle d\rangle$ is the average dissimilarity over all the pairs of data-set points. The exponent γ will be restricted to assume small values so as to characterize the corresponding Potts model by long-range ferromagnetic couplings; the scale parameter a will be bounded in $[0, 1]$.

At this point it is worth stressing that minimization of the distorsion and maximization of the similarity yield, in the autoencoder frame, different cost functions. The Hamiltonian H_d embodies the requirement that pairs of distant points (large d_{ij}) should belong to different clusters. On the other hand, the Hamiltonian H_s, for reasonable choices of F, concentrates on pairs of close points (small d) and forces them to belong to the same cluster. In other words, H_s may be seen to implement the idea that clusters should be searched for as dense connected regions in the data space.

We describe now the application of the variational criterions for clustering, described above, to some artificial and real data-sets. We consider two optimization algorithms to find the configuration of minimum cost: simulated annealing [17] and mean-field annealing [18]. Both approaches associate a Gibbs probability distribution to the functional to be optimized. Simulated annealing is a Monte Carlo technique which samples the Gibbs distribution as the temperature is reduced to zero, while mean-field annealing attempts to track an approximation, to the mean of the distribution, known as *mean field* approximation [19]. We remark that an efficient mean-field annealing algorithm for cost function (7), based on the EM scheme [20], is described in [13]: the generalization of that algorithm to (8) is straightforward.

In many cases cost functions H_d and H_s have very close global minima. For example, in Fig. 1(a) we depict an artificial data-set generated by two overlapping isotropic Gaussian distributions. In this case the natural measure of dissimilarity is Euclidean metrics, and we use $q = 2$. In Fig. 1(b) the corresponding ground state of H_d [21] is depicted: it is very close to the Bayesian solution, i.e., the solution obtained drawing the symmetry plane for the centers of the two Gaussians. A similar partition is obtained minimizing, by simulated annealing, H_s. As a measure of the difference between two partitions $\{\sigma_i\}$ and $\{\eta_i\}$, we evaluate the following quantity:

$$\epsilon = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1,\, j\neq i}^{N} (\delta_{\sigma_i\sigma_j} - \delta_{\eta_i\eta_j})^2 \tag{11}$$

which counts the number of pairs of points upon which the two partitions disagree. Using the scale-dependent F_a, we find the ground state of H_s to differ from those of H_d by $\epsilon < 0.01$ varying a in $[0.05, 1]$. Analogously, using the scale-free F_γ, with $\gamma \in [0.1, 1.5]$, we find $\epsilon < 0.02$ when we compare the ground state of H_s with those of H_d. Hence, on this data-set, the cost functions introduced above work similarly within wide ranges of γ and a values.

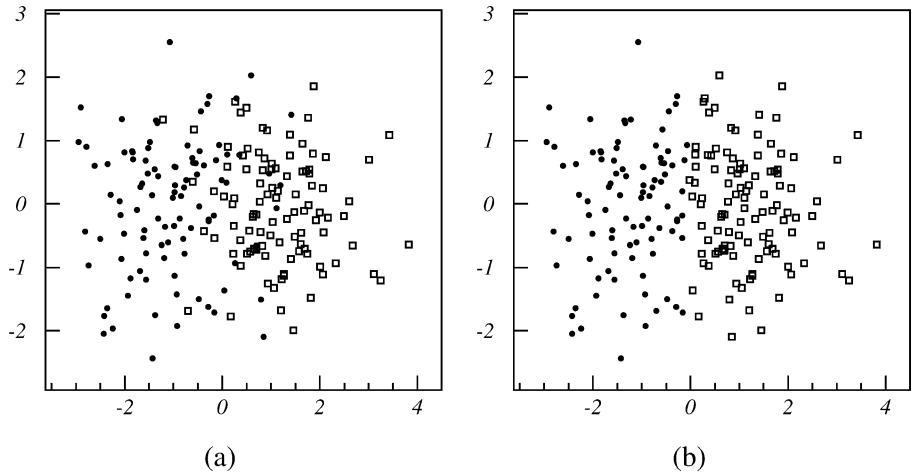

Fig. 1. (a) An artificial data-set made of two Gaussian clusters, each consisting of 100 points. Empty squares and black circles refer to the two different clusters. (b) Clustering result obtained by minimization of H_d (see the text).

We find a similar behaviour with respect to the famous IRIS data of Anderson [22]. This data-set has often been used as a standard for testing clustering algorithms: it consists of three clusters (Virginica, Versicolor and Setosa) and there are 50 objects in $\mathbf{R}^4$ per cluster. Two clusters (Virginica, Versicolor) are very overlapping. The clustering result, with $q = 3$ and minimizing H_d, consists of three clusters of 61, 39 and 50 points, respectively, with 90% of correct classification percentage. We obtain exactly the same partition by minimizing H_s using a scale-free F (with $\gamma \in [0.15, 1.45]$), and using a scale-dependent F (with $a \in [0.25, 1]$). For $a \in [0.1, 0.25]$ we obtain, in the scale-dependent case, a slightly different partition with clusters' sizes 58, 42, 50 and correct classification percentage 93.3%. These results show that also in the IRIS case the pairwise clustering procedures by distorsion minimization and similarity maximization are almost equivalent.

A typical situation resulting in different answers from H_d and H_s is depicted in Fig. 2(a). This two-dimensional data-set is made of an elongated cluster and a Gaussian distributed circular one. It is evident that two dense connected regions are present, and that the farthest pairs of points belong to the same connected region. This is the type of data-set such that minimizing the distorsion is not equivalent to maximizing the similarity. In Fig. 2(b) the partition we obtain minimizing H_d is depicted: it fails to recognize the structure in the data-set. Let us now consider the ground state of H_s with the scale-dependent F. For $a < 0.7$ the ground state, depicted in Fig. 2(c), recognizes with 99% accuracy the data structure. At $a \sim 0.7$ a transition phenomenon occurs: the configuration depicted in Fig. 2(c) ceases to be the global minimum, the new ground state (Fig. 2(d)) being very close to the solution by H_d. In Fig. 3(a) we depict the efficiency of the classification versus the resolution parameter a, for the scale dependent F, while in Fig. 3(b) we consider a sequence of a-values and we plot the ϵ between partitions corresponding to adjacent values of a. The peak at $a = 0.7$ is the indicator of the transition between global minima. Finally, in Fig. 3(c) the size of the two clusters, versus a, is depicted. Concerning the scale-free F, in Fig. 4 the same plots as in Fig. 3 are depicted, showing that the *good* minimum is stable for a wide range of γ.

The choice of the optimization algorithm deserves a comment. All the results described above are obtained by simulated annealing; we also apply the mean-field annealing scheme, described in [13], and we always find a configuration very close to the one from simulated annealing, while spending less computational time. This confirms that optimization algorithms rooted on mean-field theory yield quickly a good solution on these problems [18].

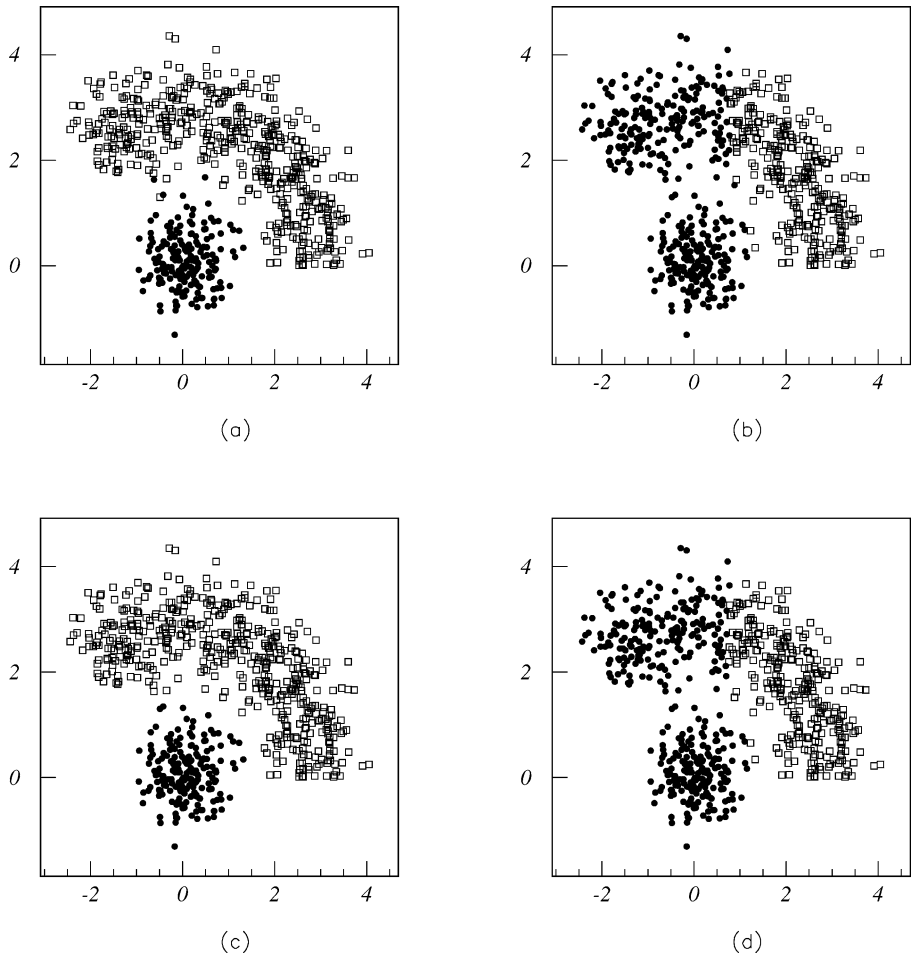

Fig. 2. (a) An artificial data-set made of an elongated cluster of 500 points (empty circles) and a circular cluster of 200 points (black circles). (b) Partition by minimizing H_d. (c) Partition by minimizing scale dependent H_s with $a < 0.7$. (d) Partition by minimizing scale dependent H_s with $a > 0.7$.

In summary, we address non-hierarchical pairwise clustering and, working in the probabilistic autoencoder frame, we introduce a class of cost functions arising from the request of maximal average similarity between the input and the output of the autoencoder. Our simulations show that the partition provided by these new cost functions corresponds to extract dense connected regions in data space, and that a relevant discrepancy with the partition provided by the cost function introduced in [13] is to be expected in case of non-trivial geometry of clusters. We note that the approach to clustering here described has some similarities with the method in [7]: indeed, in both cases clustering is mapped onto a ferromagnetic Potts model with couplings decreasing with the distance. In the superparamagnetic approach, however, q is not related to the number of classes present in the data-set and one obtains hierarchical clustering as the temperature of the Potts model is varied. In the present case q is the number of classes, which is supposed to be known (non-hierarchical clustering), and the denominators in the Hamiltonian, ensuring clusters's coherence, leads to a non-trivial ground state which reflects data structure. We consider two classes of cost func-

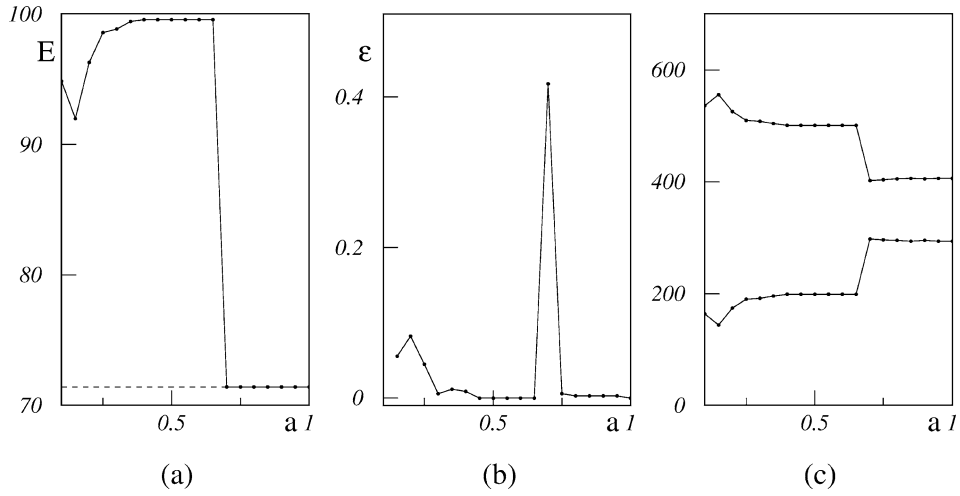

Fig. 3. (a) The efficiency (percentage of correctly classified points) versus a, obtained on the data-set depicted in Fig. 2 by minimizing H_s with scale-dependent F. The dashed line is the efficiency obtained by minimization of H_d. (b) The ϵ parameter (see the text), between partitions corresponding to adjacent values of a, is plotted versus a. (c) The size of the two output clusters versus a.

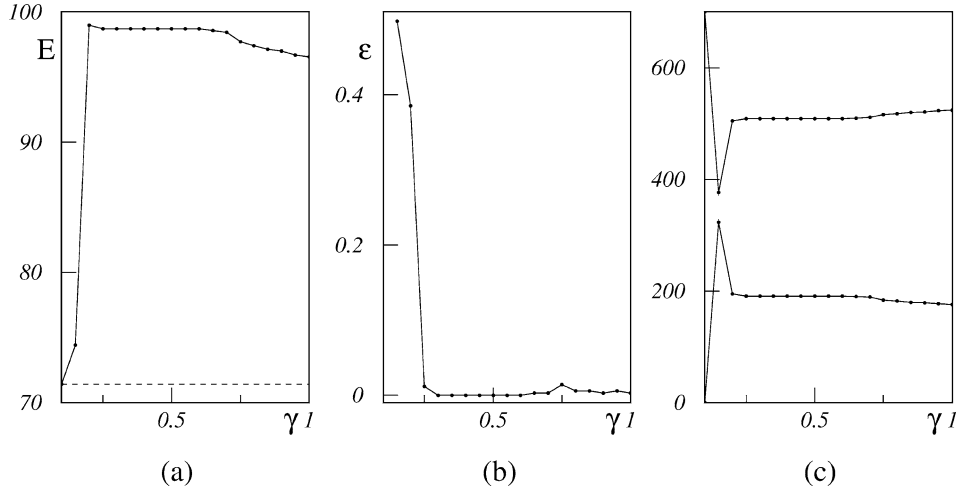

Fig. 4. (a) The efficiency (percentage of correctly classified points) versus γ, obtained on the data-set depicted in Fig. 2 by minimizing H_s with scale-independent F. The dashed line is the efficiency obtained by minimization of H_d. (b) The ϵ parameter (see the text), between partitions corresponding to adjacent values of γ, is plotted versus γ. (c) The size of the two output clusters versus γ.

tion. Scale-free cost functions depend on the exponent γ, while scale-dependent ones depend on the scale-parameter a. Varying a, i.e., changing the resolution at which the data-set is processed, may give rise to transitions between different partitions; in the scale-free case, the clustering output is fairly stable, with respect to γ, in a wide range.

Further work will be devoted to test these new cost functions on other real applications and to study related issues, such as the introduction of an *adaptive* relation between distorsion and similarity, i.e., the function $s = F(d)$ might be depending on the properties of the data-set in a neighbourhood of the pair of points under consideration. It will be also important to de-

velop cluster-validity criterions to provide a means to choose an optimal q value in situations where the number of classes is ambiguous.

References

[1] B.D. Rypley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, 1996.
[2] R.O. Duda, P.E. Hart, Pattern Recognition and Scene Analysis, Wiley, New York, 1973.
[3] A. Dekel, M.J. West, Astrophys. J. 228 (1985) 411.
[4] Y. Linde, A. Buzo, R.M. Gray, IEEE Trans. Commun. 28 (1980) 84.
[5] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Proc. Natl. Acad. Sci. USA 96 (1999) 6745.
[6] L. Kullmann, J. Kertesz, R.N. Mantegna, Physica A 287 (2000) 412.
[7] M. Blatt, S. Wiseman, E. Domany, Phys. Rev. Lett. 76 (1996) 3251.
[8] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, S. Stramaglia, Phys. Rev. Lett. 85 (2000) 554.
[9] C.M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, Oxford, 1995.
[10] A. Utsugi, Network 7 (1996) 727.
[11] L. Giada, M. Marsili, cond-mat/0101237.
[12] K. Rose, E. Gurewitz, G. Fox, Phys. Rev. Lett. 65 (1990) 945.
[13] T. Hofman, J.M. Buhmann, IEEE Trans. PAMI 19 (1997) 1.
[14] S.P. Luttrel, Neural Comput. 6 (1994) 767.
[15] C.M. Bishop, M. Svensen, C.K.I. Williams, Neural Comput. 10 (1997) 215.
[16] T.Graepel, Statistical physics of clustering algorithms, Diplomarbeit, Technique Universität, FB Physik, Institut für Theoretishe Physik, Berlin, April 1998.
[17] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Science 220 (1983) 671.
[18] See, e.g., A.L. Yuille, J.J. Kosowsky, Neural Comput. 6 (1994) 341, and references therein.
[19] G. Parisi, Statistical Field Theory, Addison-Wesley, 1988.
[20] A.P. Dempster, N.M. Laird, D.B. Rubin, J. R. Statist. Soc. 39 (1977) 1.
[21] In the text we use an operational definition of *ground state* as the configuration of minimum energy over a number (10–50) of simulated annealing runs. The true ground state might be found only by an unpractical exhaustive search.
[22] E. Anderson, Bull. Am. Iris Soc. 59 (1935) 2.