First-order phase transitions in clustering

J. Schneider
Lehrstuhl für Experimentelle Physik V, Universität Würzburg, 97074 Würzburg, Germany
(Received 16 June 1997)

A characterization of phase transition in a hierarchical clustering process is achieved by first minimizing the free-energy function and then calculating the splitting direction of the cluster vectors. Starting from this derivation, this paper gives physical and mathematical criteria for the characterization of the phase splitting of one into two distinguishable cluster centers. It is particularly shown that, in general, a first-order phase transition must be expected. If the investigated data set fulfills a certain symmetry, a splitting of one into more than two distinguishable cluster centers will occur, which is also in all cases a first-order phase transition. [S1063-651X(98)08402-5]

PACS number(s): 05.70.Fh, 02.50.−r, 89.70.+c

The investigation of data sets without *a priori* knowledge of their distribution is an important task of data analysis. Clustering methods are major tools here (see, e.g., [1,2]). This paper refers to a hierarchical clustering algorithm that, without making any further assumptions, has been derived from the maximum entropy principle [3,4]. Clusters of data points are represented by S parameter vectors $\mathbf{y}_a$. It can be shown [5] that the same results may be obtained from the maximum-likelihood principle and the corresponding expectation-maximization (EM) algorithm [5,6], if a mixture of a Gaussian distribution with fixed an equal variances in each iteration step is assumed. Therefore, the annealing parameter β in [3] corresponds to the inverse of the variance used in the EM algorithm.

For theoretical formulation of clustering a set of p data points $\{\mathbf{x}^\nu\}_{\nu=1}^p$ and S clusters C_a has been considered. In order to compare the thermodynamical quantities such as free energy and entropy, the number of clusters S was kept fixed. Describing the state of the system by a set of probability distribution P_a^ν for associating data points $\mathbf{x}^\nu$ with clusters C_a (specified by the parameter vector $\mathbf{y}_a$), we were interested in the most probable set of cluster vectors. Introducing a squared distance cost $E_a^\nu=|\mathbf{x}^\nu-\mathbf{y}_a|^2$ for assigning data point $\mathbf{x}^\nu$ to the cluster C_a and applying the maximum-entropy principle, the association probability is given then by

$$P_a^\nu=\frac{e^{-\beta|\mathbf{x}^\nu-\mathbf{y}_a|^2}}{\sum_{a=1}^{S} e^{-\beta|\mathbf{x}^\nu-\mathbf{y}_a|^2}}=\frac{1}{Z^\nu}\,e^{-\beta|\mathbf{x}^\nu-\mathbf{y}_a|^2}, \tag{1}$$

with the partition function Z^ν [3]. The corresponding free energy can be derived as

$$F=-\frac{1}{\beta}\ln Z=-\frac{1}{\beta}\sum_{\nu=1}^{p}\ln\left[\sum_{a=1}^{S} e^{-\beta|\mathbf{x}^\nu-\mathbf{y}_a|^2}\right], \tag{2}$$

with the total partition function $Z=\Pi_{\nu=1}^p Z^\nu$. Considering that the most probable cluster vector set is the one that minimizes the free-energy function (2), one obtains a fixed-point iteration for the cluster centers

$$\mathbf{y}_a=\frac{\sum_{\nu=1}^{p}\mathbf{x}^\nu P_a^\nu}{\sum_{\nu=1}^{p} P_a^\nu}\quad\forall a=1,\ldots,S, \tag{3}$$

with P_a^ν from Eq. (1). Equation (3) was solved numerically in a deterministic annealing process for a data set similar to the one of [3], using $p=1000$ data points distributed on four Gaussian clouds in a two-dimensional Euclidean space ($N=2$). Starting at small-β values ($\beta\simeq 0$), β was increased stepwise by 0.001. At each iteration step, the number of distinguishable cluster vectors S_C and their degeneration were calculated. The results differ from those of [3] in two essential points. First, after the ''true'' number of centers had been reached, no ''cluster explosion'' was found. (The term ''cluster explosion'' is used in [3] for splitting of n into a number greater than $n+1$ distinguishable cluster vectors.) We observed a successive splitting of the cluster centers like $1\rightarrow 2\rightarrow 3\rightarrow 4\rightarrow 5\cdots$ different cluster vectors. This may be explained by the smaller stepsize in the annealing process. (Rose, Gurewitz, and Fox increased β exponentially by 10% at each iteration step [4].) The second and more interesting result is the splitting of one cluster vector in the center of mass of the data set (c.m. solution) at small β into two different clusters.

We have shown analytically that this transition can generally be expected to be a first-order phase transition. Without loss of generality, the following assumptions are made: A data set with p data points in an N-dimensional Euclidian space is considered, where the center of mass is taken to be the origin. The eigenvectors of the correlation matrix $\mathcal{C}$ with elements $\mathcal{C}_{ij}=(1/p)\Sigma_{\nu=1}^p x_i^\nu x_j^\nu$ are chosen as the orthonormal basis and the eigenvalues are arranged as follows: $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_N$. The number S of cluster vectors is fixed ($S>1$). Investigating the stability of the c.m. solution, the free-energy function (2) can be expanded as a function of the $\mathbf{y}_a$ in an $(N\times S)$-dimensional Taylor series about the origin. The second term (Hessian matrix) of the expansion reads

$$\mathcal{H}=\frac{2p}{S}\left[\mathbb{1}-2\beta\left(\mathcal{A}-\frac{1}{S}\mathcal{B}\right)\otimes\mathcal{C}\right], \tag{4}$$

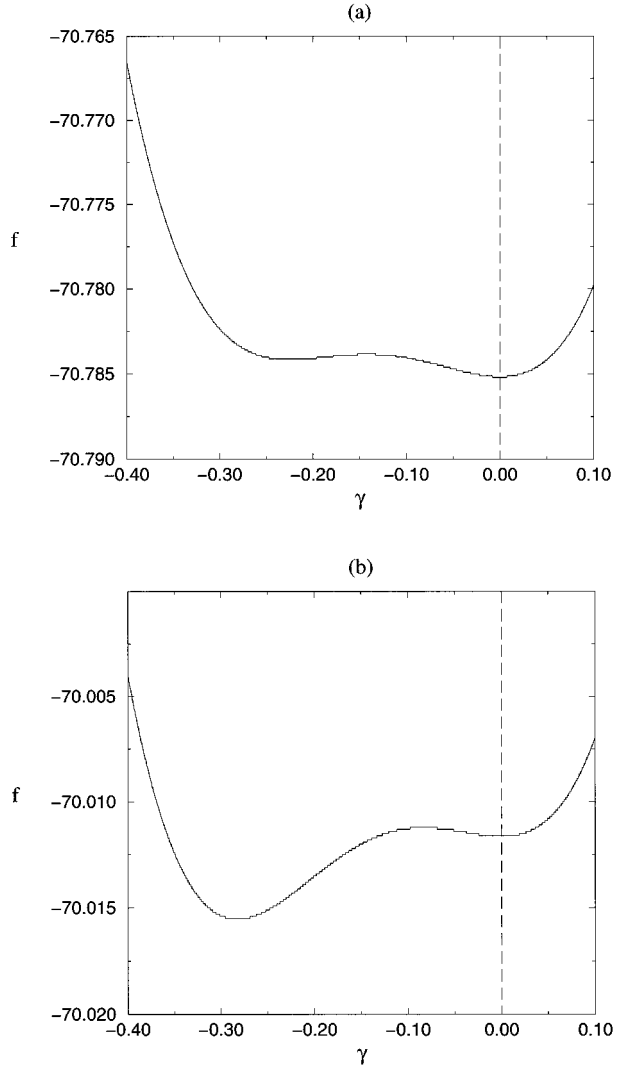

FIG. 1. Free energy per data point ($f=F/p$) projected on the splitting direction as a function of the distance to the center of mass for two different β values. (a) The c.m. solution ($\gamma=0$) gives the global minimum ($\beta<\beta^{\mathrm{split}}$) and (b) for $\beta\simeq\beta^{\mathrm{split}}$ the c.m. solution gives only a local minimum.

with

$$\mathcal{A}_{ab}=\delta_{ab}, \quad \mathcal{B}_{ab}\equiv 1 \quad \forall a,b\in 1,\ldots,S. \tag{5}$$

An upper bound β^{sp} up to which the Hessian matrix is positive definite and the c.m. solution remains a minimum can be calculated if the eigenvector $\mathbf{Y}^*=(\mathbf{y}_1^*,\ldots,\mathbf{y}_S^*)^T$ of the Hessian matrix satisfies the conditions $(\mathcal{A}\otimes\mathcal{C})\cdot\mathbf{Y}^*=\lambda_1\mathbf{Y}^*$ and $(\mathcal{B}\otimes\mathcal{C})\cdot\mathbf{Y}^*=0$. Thus β^{sp} is given by

$$\beta^{\mathrm{sp}}=\frac{1}{2\lambda_{\max}}=\frac{1}{2\lambda_1}, \tag{6}$$

which is identical to the β_C in [3]. As the second term of the series expansion must vanish at $\beta\equiv\beta^{\mathrm{split}}$, the cluster vectors fulfill the relations

$$y_a^i=\delta_{i1}y_a^1, \quad \sum_{a=1}^{S} y_a^1=0. \tag{7}$$

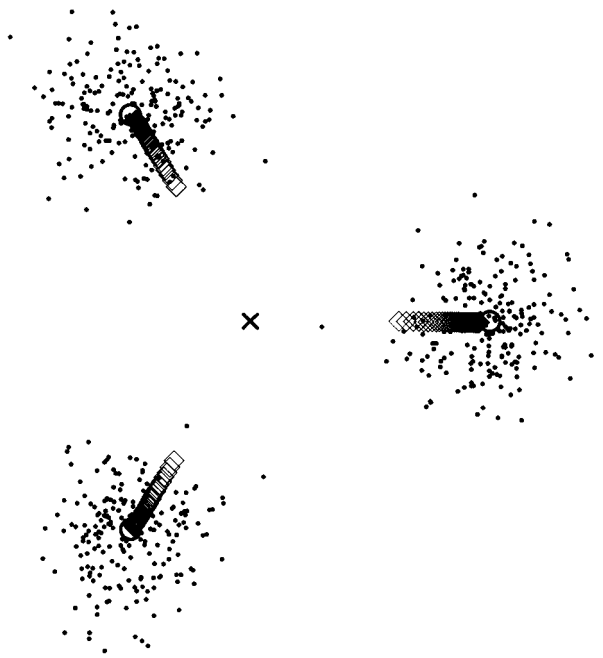

FIG. 2. Evolution of the cluster vectors with increasing β (600 data points). The circles indicate the center of the data clouds. The c.m. solution directly splits from one into three distinguishable cluster vectors.

Furthermore, the direction of the cluster splitting can be calculated by maximizing the curvature of the free-energy function (2). Following this, the split cluster vectors have to satisfy the condition

$$\frac{y_1}{y_2}=\frac{1}{S-1}. \tag{8}$$

With Eq. (7) this results in an asymmetry in the degeneration of these vectors within the eigenvector $\mathbf{Y}^*$. Calculating now the third term of the series expansion, which can be written with Eq. (7) as

$$\sum_{a,b,c}\sum_{i,j,k}\left.\frac{\partial^3 F}{\partial y_a^i\partial y_b^j\partial y_c^k}\right|_{\mathbf{Y}=\mathbf{0}} y_a^i y_b^j y_c^k = -\frac{8\beta^2}{S}\left(\sum_{\nu}(x_1^{\nu})^3\right)\sum_a (y_a^1)^3, \tag{9}$$

it is obvious that in the case of an asymmetric distribution of the data points and/or finite data sets as well as an asymmetric splitting of the cluster vectors (e.g., $S>2$), the terms on the right-hand side of Eq. (9) do not vanish. Thus the free-energy function has a saddle point at $\beta=\beta^{\mathrm{split}}$, which automatically leads to a first-order phase transition in cluster splitting. Moreover, with Eqs. (7) and (8) a projection of the free-energy function on the splitting direction can be calculated by parametrizing the cluster vectors as

$$y_1=-\gamma, \quad y_2=(S-1)\gamma; \tag{10}$$

The results are illustrated in Figs. 1(a) and 1(b) where the parametrized free energy is plotted as a function of γ for two different values of β (the number of assumed clusters S was

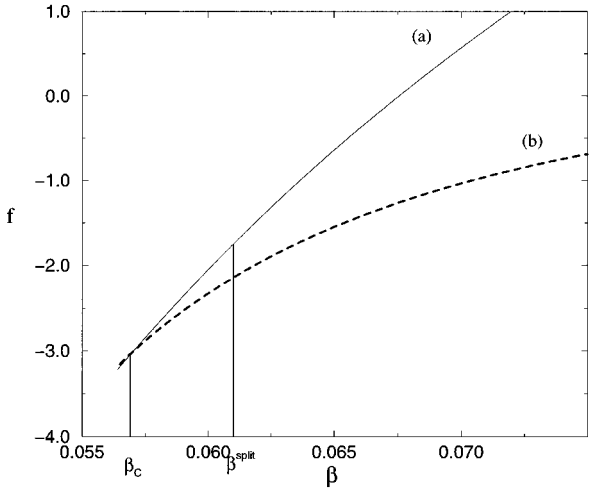

FIG. 3. Free energy per data point ($f=F/p$) as a function of the annealing parameter β for the (a) c.m. solution and (b) splitting in the direction of the center of the data clouds. β_C indicates the critical value at which the splitting should occur; at β^{split} the algorithm performs the splitting.

10). In Fig. 1(a) the global minimum is represented by the c.m. solution ($\gamma=0$), but there also exists a local minimum for $\gamma<0$. A new global minimum can be found at $\beta\simeq\beta^{\text{split}}$ [Eq. (6)], whereas the c.m. solution remains only a local minimum [Fig. 1(b)]. If $S=2$, only a symmetrical splitting is possible. As the third [Eq. (9)] but not the fourth term of the Taylor expansion vanishes, the splitting of the cluster vectors is continuous in this case, in analogy to the spontaneous magnetization of a ferromagnet [7]. These analytical results show complete agreement with our numerical simulations where the clustering algorithm has been applied on a data set, as mentioned above [8].

Finally, another interesting question was if and under what circumstances a splitting of one cluster center into more than two distinguishable cluster centers would occur. The answer to this question can be given in the following way. Consider a data set with a symmetrical arrangement of n ($n>2$) clouds, where the centers of the clouds form a regular polygon (see Fig. 2 for $n=3$). If the variances of the clouds are small enough so that the symmetry of the arrangement is clearly recognizable, one would expect a splitting of one cluster vector into n cluster vectors. This expectation proves to be true, as it is shown in Fig. 2. In this case, due to the discontinuity of the transition, no critical value for β can be calculated analytically. The algorithm persists in a local minimum (c.m. solution) until the barrier between the local and global minimum can be crossed (see Fig. 3). The ''true'' β_C can be numerically calculated by reversing the annealing process (starting at a β value, where n different cluster vectors exist) and comparing the free-energy functions (Fig. 3). As shown above, it is obvious that this splitting is also a first-order phase transition. Furthermore, these results have been proved to hold for $n>3$ qualitatively.

Concluding the results, it is now possible to give clear criteria to determine whether or not a splitting of the c.m. solution is a continuous phase transition. It is also evident from Eq. (9) that, generally, a first-order transition can be expected. Therefore, the minimization of the free-energy function becomes more difficult and the probability of achieving only local minima rises. Nevertheless, in the case of a ''supervised'' clustering, i.e., using a small amount of *a priori* knowledge, the ''true parameters'' can be estimated in a sufficient way.

Special thanks go to G. Reents for his assistance and helpful discussions and to M. Biehl and W. Kinzel for a critical reading of the manuscript.

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
[2] Barkei and H. Sompolinsky, Phys. Rev. E **50**, 1766 (1994).
[3] K. Rose, E. Gurewitz, and G. C. Fox, Phys. Rev. Lett. **65**, 495 (1990).
[4] K. Rose, E. Gurewitz, and G. C. Fox, Pattern Recogn. Lett., **11**, 589 (1990).
[5] C. M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon, Oxford, 1995).
[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc. B **39**, 1 (1977).
[7] L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics, Vol. 5 Statistical Physics* (Pergamon, London, 1959).
[8] J. Schneider, Diplomathesis, Universität Würzburg, 1995 (unpublished).