Deterministic annealing for density estimation by multivariate normal mixtures

Martin Kloppenburg and Paul Tavan
Institut für Medizinische Optik, Ludwig-Maximilians-Universität München, Theresienstraße 37, D-80333 München, Germany
(Received 27 September 1996)

An approach to maximum-likelihood density estimation by mixtures of multivariate normal distributions for large high-dimensional data sets is presented. Conventionally that problem is tackled by notoriously unstable expectation-maximization (EM) algorithms. We remove these instabilities by the introduction of soft constraints, enabling deterministic annealing. Our developments are motivated by the proof that algorithmically stable fuzzy clustering methods that are derived from statistical physics analogs are special cases of EM procedures. [S1063-651X(97)50803-8]

PACS number(s): 02.50.−r, 07.05.Mh, 02.60.Pn, 89.70.+c

I. INTRODUCTION

The identification of principal structures and features within large sets of high-dimensional data and the generation of simplified models for such data distributions are important tasks, which arise in many technical fields including pattern recognition [1] or the study of complex physical systems, e.g. protein dynamics [2]. However, the design of mathematical methods, which are capable of solving these tasks, remains a problem with many pitfalls despite the considerable attention which it has received over the last decades. Various approaches have been proposed [1], among which density estimation is a most important one, as it rests on fundamental statistical notions. That method aims at representing the data by a model probability density and requires adjustment of a parameter set; usually mixtures of multivariate normal distributions are chosen as model densities. A second important approach is clustering; here, data points are grouped into clusters or classes, which are represented by a prototypical member, e.g., by their centroid. More recently, artificial neural networks have also been shown to qualify as possible tools, since they can perform tasks like clustering or the extraction of principal components [3].

Usually the task of finding dimension-reduced descriptions of high-dimensional data sets is formulated in terms of optimization of a cost function, and iterative gradient-based algorithms are applied. For density estimation, the likelihood to draw the given data sample from the model density is such a function [1,4]. In clustering, the error associated with the representation of the data points by their corresponding prototypes is a suitable choice [1,5]; also learning rules of neural networks frequently have been obtained from related cost functions [6]. Note, however, that there are some biologically inspired neural learning algorithms, which succeed to determine useful descriptions despite the fact that cost functions or other formal quality criteria are lacking [7,8].

All quoted cost functions have a common drawback: generally they are not convex and the search for their global optimum is highly difficult; therefore, iterative gradient-based algorithms get easily caught in local extrema. In order to avoid this difficulty, annealing strategies derived from statistical mechanics concepts have been successfully applied [9,10]. Here, the cost function is conceived of as the energy of an analog physical system, and the optimization process is mapped to the ongoing relaxation of a canonical ensemble towards thermal equilibrium at decreasing temperatures. As a result, at finite temperatures, the original corrugated cost function is effectively replaced by a new and smoother one representing an analog free energy. Both detailed simulated annealing procedures, which safely but slowly converge [11,12], and mean-field type deterministic annealing methods [10,13], which are faster but may lead to suboptimal solutions [14,15], have been used.

Taking the socalled ''elastic net'' algorithm as a prototypical example [16], Yuille, Stolorz, and Utans have recently elaborated the close connection between a special case of density estimation and the deterministic annealing approach to clustering [17]. Extending that type of reasoning, in this paper we will derive a class of algorithms which solve the general problem of density estimation by multivariate normal mixtures and avoid the hitherto inevitable difficulties of singular solutions [1]. That progress will be achieved by identifying and combining the respective advantages of two mutually interconnected approaches. To present our arguments, we will first state the task to be solved from the classical statistical point of view and then sketch its relations to clustering procedures derived from statistical mechanics analogs. This discussion serves to motivate our algorithmic procedures, which will be presented and subsequently illustrated using a simple example.

II. DENSITY ESTIMATION FOR NORMAL MIXTURES

Consider a set of D-dimensional data $\mathcal{X}=\{\boldsymbol{x}_n|n=1,\ldots,N\}\subset\mathcal{R}^D$, which is to be represented by a model density

$$p(\boldsymbol{x}|\Theta)=\sum_{r=1}^{K} P_r p(\boldsymbol{x}|r,\theta_r), \tag{1}$$

composed of K multivariate normal distributions

$$p(\boldsymbol{x}|r,\theta_r)=\frac{\exp[-(\boldsymbol{x}-\boldsymbol{y}_r)^t\ \Sigma_r^{-1}(\boldsymbol{x}-\boldsymbol{y}_r)/2]}{(2\pi)^{d/2}(\det\Sigma_r)^{1/2}}. \tag{2}$$

The set of adjustable parameters Θ includes the statistical weights P_r, the means $\boldsymbol{y}_r$, and the covariance matrices Σ_r of the normal distributions r.

The logarithm of the *likelihood* $P(\mathcal{X}|\Theta)$ that the sample $\mathcal{X}$ is drawn from this density is

$$l(\mathcal{X}|\Theta)=\sum_{n=1}^{N}\ln p(\boldsymbol{x}_n|\Theta). \tag{3}$$

In the classical statistics approach one tries to obtain suitable parameters Θ by maximizing the log-likelihood $l(\mathcal{X}|\Theta)$. Taking derivatives of Eq. (3) one finds a set of necessary conditions for the optimal parameters [1]. For a most simple specification of these conditions we introduce the Bayesian conditional probability

$$P(r|\boldsymbol{x}_n,\Theta)=\frac{P_r p(\boldsymbol{x}_n|r,\theta_r)}{p(\boldsymbol{x}_n|\Theta)}, \tag{4}$$

that the data point $\boldsymbol{x}_n$ is generated by the normal distribution r, and the global expectation value $\langle\ldots\rangle$ of the probability

$$\langle P(r|\mathbf{x},\Theta)\rangle=\frac{1}{N}\sum_{n=1}^{N}P(r|\boldsymbol{x}_n,\Theta), \tag{5}$$

that a data point is due to r. Then one can define local expectation values for classes r by

$$\langle f(\boldsymbol{x})\rangle_{r,\Theta}=\langle P(r|\boldsymbol{x},\Theta)f(\boldsymbol{x})\rangle/\langle P(r|\boldsymbol{x},\Theta)\rangle, \tag{6}$$

and the stationarity conditions for the parameters read

$$P_r=\langle P(r|\mathbf{x},\Theta)\rangle, \tag{7}$$

$$\boldsymbol{y}_r=\langle\boldsymbol{x}\rangle_{r,\Theta}, \tag{8}$$

$$\Sigma_r=C_{r,\Theta}, \tag{9}$$

where $C_{r,\Theta}\equiv\langle(\boldsymbol{x}-\boldsymbol{y}_r)(\boldsymbol{x}-\boldsymbol{y}_r)^t\rangle_{r,\Theta}$ are the r-local covariance matrices.

In order to determine a set of parameters Θ satisfying these conditions one can apply the so-called expectation-maximization(EM) algorithm [4]. Starting with some initial estimates of the parameters, one first calculates the conditional probabilities (4) and subsequently uses Eqs. (5)–(9) for an iterative update of the estimates until self-consistency is reached. Generally one finds, that the results strongly depend on the choice of the initial estimates, represent suboptimal solutions, and are frequently even singular, i.e., for some of the normal distributions the ranges $\det\Sigma_r$ or the statistical weights P_r become very small, whereas for others they become very large. These findings signify the corrugated structure of the log-likelihood functional within parameter space and testify, that a naive application of the EM algorithm is inadequate.

Unfortunately, the statistical approach sketched above does not provide any clues as to how one can systematically avoid convergence towards suboptimal solutions, e.g., by imposing suitable constraints on the variation of the parameters. In contrast, such clues naturally show up if one considers the seemingly unrelated clustering problem from the point of view of an appropriate physical analog, which allows analysis in terms of statistical mechanics concepts [10].

III. STATISTICAL MECHANICS AND CLUSTERING

In clustering a given data sample $\mathcal{X}$ is to be represented by a *codebook* $(\mathcal{Y},\mathcal{V})$, such that a suitable error functional $U(\mathcal{Y},\mathcal{V})$ becomes minimal. Here, the codebook consists of a set of K prototypes, $\mathcal{Y}=\{\boldsymbol{y}_r\}\subset\mathcal{R}^D$, and of a set of $N\times K$ binary variables, $\mathcal{V}=\{v_{nr}\in\{0,1\}\}$, which associate each data point $\boldsymbol{x}_n$ to exactly one codebook vector $\boldsymbol{y}_r$. Chosing as an error measure the squared distance within data space, the total error for the representation of the sample $\mathcal{X}$ by the codebook $(\mathcal{Y},\mathcal{V})$ is

$$U(\mathcal{Y},\mathcal{V})=\sum_{n=1}^{N}\sum_{r=1}^{K}v_{nr}(\boldsymbol{x}_n-\boldsymbol{y}_r)^2. \tag{10}$$

At given $\mathcal{V}$ the optimal $\boldsymbol{y}_r$ are the centroids $\boldsymbol{y}_r=\Sigma_n v_{nr}\boldsymbol{x}_n/\Sigma_{n'}v_{n'r}$. However, concerning the choice of $\mathcal{V}$ the clustering problem stated above is a hard, so-called NP-complete optimization problem [18]. To tackle that problem one may interpret $U(\mathcal{Y},\mathcal{V})$ as the *energy* of an analog physical system with dynamical variables $\boldsymbol{y}_r$ and v_{nr}; considering a *canonical ensemble* of such systems with microstates $(\mathcal{Y},\mathcal{V})$, i.e., maximizing the entropy under the constraint of a given average energy U one obtains the partition function

$$Z=\int\exp[-\beta\widetilde{F}(\mathcal{Y})]\mathrm{d}^{DK}y, \tag{11}$$

where

$$\widetilde{F}(\mathcal{Y})=-\frac{1}{\beta}\sum_{n=1}^{N}\ln\left(\sum_{r=1}^{K}\exp[-\beta(\boldsymbol{x}_n-\boldsymbol{y}_r)^2]\right). \tag{12}$$

Since the partition function (11) is not easily calculated, one applies the mean-field approximation, within which the integral is replaced by the maximum of its integrand assuming that the latter is strongly peaked. The corresponding minimum of $\widetilde{F}(\mathcal{Y})$, which is the mean-field free energy, is determined by solving the saddle point equations for the ensemble expectation values $P(r|\boldsymbol{x}_n,\beta)\equiv\overline{v_{nr}}$ and $\bar{\boldsymbol{y}}_r$ of the dynamical variables. The saddle point equations are

$$\bar{\boldsymbol{y}}_r=\frac{\sum_{n=1}^{N}P(r|\boldsymbol{x}_n,\beta)\boldsymbol{x}_n}{\sum_{n=1}^{N}P(r|\boldsymbol{x}_n,\beta)} \tag{13}$$

and

$$P(r|\boldsymbol{x}_n,\beta)=\frac{\exp[-\beta(\boldsymbol{x}_n-\bar{\boldsymbol{y}}_r)^2]}{\sum_{r'=1}^{K}\exp[-\beta(\boldsymbol{x}_n-\bar{\boldsymbol{y}}_r)^2]}. \tag{14}$$

These equations are intimately related to some of the expressions presented earlier for the maximum likelihood esti-

mation by normal mixtures. For instance, the stationarity conditions (8) for the $\boldsymbol{y}_r$ exactly reduce to the mean-field centroid conditions (13) in the special case of *univariate* normal distributions with identical statistical weights and variances, i.e., if $P_r = 1/K$ and $\Sigma_r^{-1} = 2\beta\mathbb{1}$ [cf. Eqs. (1), (2), (6)]. Similarly the expressions (4) for the Bayesian conditional probabilities $P(r|\boldsymbol{x}_n,\Theta)$ reduce to the mean-field equations (14) for the ensemble expectation values $P(r|\boldsymbol{x}_n,\beta)$ of the association variables v_{nr}. Furthermore, concerning the dependence on the $\boldsymbol{y}_r$ the log-likelihood (3) is equivalent to the free energy (12) and, with respect to these parameters, maximizing the log-likelihood amounts to minimizing the free energy. Thus free energy clustering is a special case of maximum likelihood density estimation.

Note however, that the classical statistics approach provides an optimality criterion for the variance σ^2 of the univariate normal distributions; in that case conditions (9) reduce to $\sigma^2 = (1/DK)\Sigma_r\langle(\mathbf{x}-\boldsymbol{y}_r)^2\rangle_{r,\Theta}$. No such criterion is obtained in free energy clustering. Here, the variance $\sigma^2 = 1/2\beta$ is a global parameter measuring the temperature of the physical analog system.

Now it might seem, that the restriction of the model density to a mixture of univariate normal distributions with identical weigths and variances, as well as the absence of a prescription for an optimal choice of σ^2, represent distinct disadvantages of the free energy approach. But (i) the reduction of the parameter set to the $\boldsymbol{y}_r$ and (ii) the use of σ^2 as a fixed steering parameter actually generate its main advantage, i.e., a stable algorithmic scheme: Constraining the parameters Σ_r and P_r to predefined values excludes nasty singular solutions. Furthermore, the interpretation of $2\sigma^2$ as a temperature leads to a *annealing scheme* for the optimization procedure upon which the emerging solutions become *independent* of the initial conditions. The properties of this annealing process, which involves a hierarchically ordered series of data representations at increasing resolutions, have been analyzed in detail by various authors [10,15]. Although the solutions obtained at small or vanishing σ do not necessarily represent the global optimum of $\widetilde{F}(\mathcal{Y})$ or $U(\mathcal{Y},\mathcal{V})$, respectively, they usually are quite satisfactory (for a modified and more safely converging algorithm see Ref. [15]).

The above analysis of the sources of algorithmic stability in maximum likelihood density estimation by univariate normal mixtures, i.e., in free energy clustering, has inspired us to develop related algorithmic procedures also for the more general multivariate case. Here, the rigid constraints on the parameters Σ_r and P_r will be replaced by soft ones, such that the possibility to define an annealing procedure is preserved.

IV. ANNEALING SCHEMES FOR MULTIVARIATE GAUSSIAN MIXTURES

In order to impose suitable soft constraints on the covariance matrices Σ_r we represent them in terms of their eigenvectors $\boldsymbol{w}_{ir}$ and eigenvalues σ_{ir}^2. Then the inverse matrices Σ_r^{-1} can be expressed in terms of the orthogonal diagonalizing transformations $W_r = (\boldsymbol{w}_{1r}, \ldots, \boldsymbol{w}_{dr})$ and of the diagonal matrices $\hat{\Sigma}_r$ of eigenvalues as

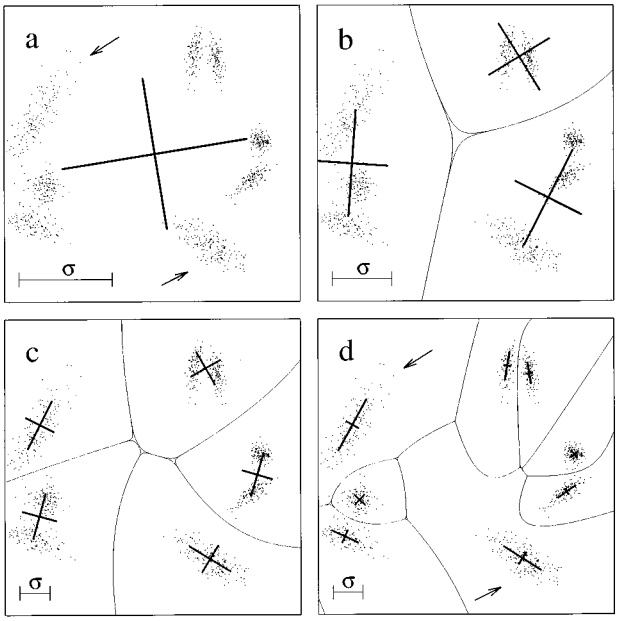

FIG. 1. Annealing of a normal mixture with $K = 10$ bivariate components for a two-dimensional data set; 1000 data points are drawn (black pixels); the crosses measure the ranges $2\sigma_{ri}$ of the component densities in directions $\mathbf{w}_{ri}$; thin lines indicate fuzzy boundaries, at which the association probabilities (4) to the corresponding mixture components have the value 1/2; (a) initial state at annealing parameters $(\sigma,\mu,\lambda) = (1.1\sigma_0, 0.5, 1.0)$; (b) $(0.65\sigma_0, 0.5, 1.0)$; (c) $(0.32\sigma_0, 0.5, 1.0)$; (d) final state $(0.32\sigma_0, 0.0, 0.0)$.

$$\Sigma_r^{-1} = W_r \hat{\Sigma}_r^{-1} W_r^t . \tag{15}$$

Using this representation of Σ_r^{-1} for maximization of the log-likelihood (3) and extending that cost function by conditions, which guarantee the normalization of the eigenvectors $\mathbf{w}_{ir}$, the stationarity conditions (9) separate into two sets of equations. According to the first set

$$\sigma_{ir}^2 = \boldsymbol{w}_{ir}^t C_{r,\Theta} \boldsymbol{w}_{ir} \tag{16}$$

the σ_{ir}^2 should be the r-local variances in the directions $\boldsymbol{w}_{ir}$, whereas according to the second set

$$\sigma_{ir}^2 \boldsymbol{w}_{ir} = C_{r,\Theta} \boldsymbol{w}_{ir} , \tag{17}$$

the $\boldsymbol{w}_{ir}$ should be the eigenvectors of the r-local covariance matrices $C_{r,\Theta}$.

Now the stationarity conditions (16) for the eigenvalues σ_{ir}^2 enable to add the desired constraints. A possible choice is

$$\sigma_{ir}^2 = \boldsymbol{w}_{ir}^t C_{r,\Theta} \boldsymbol{w}_{ir} + \mu(\sigma^2 - \sigma_{ir}^2)/\langle P(r|\boldsymbol{x},\Theta)\rangle . \tag{18}$$

Here, the σ_{ir}^2 are coupled to an annealing parameter σ^2 and μ determines the rigidity of coupling. Note, that the constraints can be derived by adding the regularization term $V(\hat{\Sigma}_r;\sigma) = -(\mu/2)\Sigma_{r,i}(\ln\sigma_{ir}^2 + \sigma^2/\sigma_{ir}^2)$, which has a quadratic maximum at $\sigma_{ir} = \sigma$, to the log-likelihood (3). Thus, in the strong coupling limit $(\mu \to \infty)$ free energy clustering is recovered.

Applying the EM algorithm we use Eqs. (17) and (18) for an iterative parameter update and assure the required orthogonality of the vectors $\boldsymbol{w}_{ir}$ hierarchically by Schmidt's method. Vector $\boldsymbol{w}_{1r}$ then converges to the eigenvector of $C_{r,\Theta}$ with the largest eigenvalue, $\boldsymbol{w}_{2r}$ to one with the second largest, etc. (for the mathematics of that type of diagonalization of covariance matrices see, e.g., Ref. [19]).

In contrast to the case of the Σ_r, introduction of soft constraints for the weights P_r is trivial. One may simply replace the EM equations (7) by

$$P_r = \langle P(r|\mathbf{x},\Theta)\rangle + \lambda(1/K - P_r) \tag{19}$$

in order to keep the weights of the local distributions approximately balanced at $1/K$. Like in free energy clustering that balance ensures the stability of the algorithm. Note, that the constraint derives from adding the log-likelihood $\Sigma_r(1/K)\ln P_r$, that the P_r are uniformly distributed, weighted by λ to the original log-likelihood $l(\mathcal{X}|\Theta)$.

The annealing schedule, which accompanies the EM parameter update according to Eqs. (8), (17), (18), and (19) in our algorithm, involves a reduction of the parameter σ from large to small values and a subsequent or concomitant lifting of constraints by decreasing μ and λ to zero. The progress of optimization is monitored by the value of $l(\mathcal{X}|\Theta,\sigma,\mu,\lambda)$. A sequential, stochastic version of the algorithm, in which data points are presented one by one for parameter optimization, has also been implemented and the following simple example has actually been computed using that version.

V. EXAMPLE

Figures 1 illustrate the annealing process and the corresponding dimension-reduced descriptions for a simple two-dimensional data set $\mathcal{X}$ composed of $N=500\,000$ data points with a maximal variance σ_0. That data set is distributed according to an *a priori* mixture density composed of eight bivariate Gaussians; the weights of two of the data clusters [marked by arrows in Fig. 1(a)] have been chosen 1.5 times larger than those of the other clusters. For our model density (1) we have chosen $K=10$ components. The annealing is initialized at a high temperature ($\sigma=1.1\sigma_0$); as shown in Fig. 1(a) all components are degenerate at the center of the data distribution and the description represents that of a global principal component analysis [1]. Lowering the temperature leads to a splitting into three [Fig. 1(b)] and, subsequently, five components [Fig. 1(c)], which are still degenerate; that process uncovers the hierarchical distance relations among the clusters of the data set. Finally, for Fig. 1(d) the constraints on the variances and weights have been removed at constant temperature; as a result, the substructures within the three small data clusters become resolved by a lifting of the corresponding degeneracies whereas the coherence of the extended clusters (marked by arrows) is retained by preservation of (twofold) degeneracy. The resulting model density represents the optimal solution with eight effective components of correct covariances and weights. Note, that the thin lines in the figures illustrate the respective fuzzy partitions $P(r|\boldsymbol{x},\Theta)$ of the data set.

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).

[2] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).

[3] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, New York, 1991).

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Statist. Soc. Ser. B **39**, 1 (1977).

[5] Y. Linde, A. Buzo, and R. M. Gray, IEEE Trans. Commun. Technol. **28,** 84 (1980).

[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature (London) **323**, 533 (1986).

[7] C. v. d. Malsburg and D. J. Willshaw, Proc. Natl. Acad. Sci. USA **74**, 5176 (1977).

[8] T. Kohonen, Biol. Cybern. **43**, 59 (1982).

[9] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognitive Sci. **9**, 147 (1985).

[10] K. Rose, E. Gurewitz, and G. Fox, Phys. Rev. Lett. **65**, 945 (1990).

[11] S. Kirkpatrick, C. Gelatt, and M. Vecchi, Science **220**, 671 (1983).

[12] S. Geman and D. Geman, IEEE Trans. Pattern. Anal. Mach. Intell. **6**, 721 (1984).

[13] C. Peterson and J. Anderson, Complex Syst. **1**, 995 (1987).

[14] R. Durbin, R. Szeliski, and A. Yuille, Neural Comput. **1**, 348 (1989).

[15] D. R. Dersch and P. Tavan, in *Proceedings of the IEEE International Conference on Neural Networks ICNN'94* (IEEE, Piscataway, 1994), pp. 698–703.

[16] R. Durbin and D. Willshaw, Nature **326**, 689 (1987).

[17] A. L. Yuille, P. Stolorz, and J. Utans, Neural Comput. **6**, 334 (1994).

[18] M. R. Garey, D. S. Johnson, and H. S. Witsenhausen, IEEE Trans. Inf. Theory **28**, 255 (1982).

[19] J. Rubner and P. Tavan, Europhys. Lett. **10**, 693 (1989).