Pairwise Data Clustering by Deterministic Annealing

Thomas Hofmann, *Student Member, IEEE,* and Joachim M. Buhmann, *Member, IEEE*

Abstract—Partitioning a data set and extracting hidden structure from the data arises in different application areas of pattern recognition, speech and image processing. *Pairwise data clustering* is a combinatorial optimization method for data grouping which extracts hidden structure from proximity data. We describe a *deterministic annealing* approach to pairwise clustering which shares the robustness properties of maximum entropy inference. The resulting Gibbs probability distributions are estimated by mean–field approximation. A new structure-preserving algorithm to cluster dissimilarity data and to simultaneously embed these data in a Euclidian vector space is discussed which can be used for dimensionality reduction and data visualization. The suggested embedding algorithm which outperforms conventional approaches has been implemented to analyze dissimilarity data from protein analysis and from linguistics. The algorithm for pairwise data clustering is used to segment textured images.

Index Terms—pairwise data clustering, maximum entropy method, multidimensional scaling, exploratory data analysis

1 Introduction

Modern information and communication technology confronts us with massive amounts of data. The primary goal of pattern recognition is to extract hidden structure from data in order to generate a compact data representation and to enable symbolic data processing concepts. One of the basic problems in pattern recognition is concerned with the detection of clusters in data sets. The potential applications of clustering algorithms cover a wide range from data compression of video and audio signals to structure detection and automatic inference engines in machine learning and artificial intelligence. We will describe a stochastic optimization approach to data clustering which relies on the well-known robustness of maximum entropy inference [1], [2], [3][1].

The problem of optimally partitioning a data set arises in two different forms dependent on the data representation as vectorial or proximity data. A partitioning approach known as *central clustering* derives a set of reference or prototype vectors which quantize a set of vectorial data with minimal quantization error [6], [7]. Data compression is achieved by transmission and storage of the indices of reference vectors rather than the original data vectors. The second approach to data clustering, referred to as *pairwise data clustering* [8], partitions a set of data into clusters in which the data are indirectly characterized by pairwise comparisons instead of explicit coordinates. The characteristics of the data set are hidden in these pairwise relations or proximity values which frequently violate the requirements of a distance measure, i.e., the triangular inequality does not necessarily hold, the self-dissimilarity may not vanish and the proximity values might be negative. The grouping of proximity data is mathematically formulated as a combinatorial optimization problem which we solve with a minimization heuristic called *deterministic annealing*. Sets of relational data are abundant in many applications, e.g., in molecular biology, psychology, linguistics, economics and image processing.

Data clustering as a problem in pattern recognition and statistics belongs to the class of unsupervised learning problems. There is a large body of literature available on this topic and the reader is referred to the text books of Duda and Hart [8] and of Jain and Dubes [4] for an overview. The method of deterministic annealing is described in various papers mostly in the literature on neural networks [9], [10], [11], [12] and on computer vision [13], [14], [15]. Deterministic annealing applied to central clustering has been discussed by Rose et al. in a series of papers [16], [17], [18], [19]. A solution of the deterministic annealing procedure for vector quantization with different rate constraints was suggested in [20], [21]. This work, as well as Chou et al. [22], emphasized the design question of how large the code book should be chosen dependent on prespecified costs per code vector.

The remainder of this paper is structured in the following way: We discuss the advantage of a maximum entropy based search heuristic in Section 2. A discussion of cost functions for central and pairwise data clustering is presented in Section 3. Approximation techniques to calculate expectation values for the data assignments are discussed in Section 4. The widely used estimation technique called mean–field approximation is derived by variational techniques and, alternatively, by an expansion for small fluctuations. An extension of pairwise data clustering to data

• *The authors are with Rheinische Friedrich-Wilhelms-Universität, Institut für Informatik III, Römerstraße 164, D-53117 Bonn, Germany.*
E-mail: {th, jb}@informatik.uni-bonn.de.

Manuscript received June 3, 1995; revised Sept. 26, 1996. Recommended for acceptance by D.M. Titterington.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96100.

1. Data clustering is viewed as a partitioning problem throughout this paper and not as a density estimation problem of a mixture model [4], [5] in the sense of parametric statistics.

visualization is described in Section 5. The results of central clustering (Section 3) are employed to simultaneously group and embed proximity data in a low dimensional Euclidian space. Simulation results of clustering problems in molecular biology and linguistics, a performance comparison between deterministic annealing and a conventional, gradient descend technique for clustering as well as an application of pairwise data clustering to image segmentation are summarized in Section 6.

2 Stochastic Optimization by Maximum Entropy Inference

2.1 Simulated Annealing

In seminal papers Kirkpatrick et al. [23] and, independently, Černy [24] have proposed the stochastic optimization strategy *Simulated Annealing*. By analogy to an experimental annealing procedure where the stability of metal or glass is improved by heating and cooling, solutions for an optimization problem are heated and cooled in simulations to find one with very low costs. The search for good solutions is implemented by a Markov process which stochastically samples the solution space Ω of an optimization problem. The optimization problem is characterized by a cost function $\mathcal{H} : \Omega \mapsto \mathbb{R}, \omega \in \Omega$ denoting an admissible solution of the optimization problem. A new solution is accepted or rejected according to the Metropolis algorithm, i.e., new solutions with decreased costs are always accepted and solutions with increased costs are accepted with an exponentially weighted probability, i.e.,

$$\mathbf{P}\left(\omega^{\text{old}} \rightarrow \omega^{\text{new}}\right) = \begin{cases} 1 & \text{if } \Delta\mathcal{H} \leq 0, \\ \exp\left(-\Delta\mathcal{H} \,/\, T\right) & \text{else.} \end{cases} \tag{1}$$

with $\Delta\mathcal{H} \equiv \mathcal{H}\left(\omega^{\text{new}}\right) - \mathcal{H}\left(\omega^{\text{old}}\right)$. The parameter T is called the computational temperature. The philosophy of simulated annealing is to gradually reduce the temperature during the search process, thereby forcing the system into solutions with low costs. Mathematically, the stochastic search process for the optimal solution is a random walk in the solution space. Cost differences between neighboring states act as a force field. The effect of the temperature can be interpreted as a random force with an amplitude proportional to T. Valleys and peaks with a cost difference less than T are smeared out and vanish in the stochastic search. A Markov process with a transition matrix (1) converges to an equilibrium probability distribution [25]

$$\begin{aligned} \mathbf{P}^{\text{Gb}}(\omega) &= \frac{\exp\left(-\mathcal{H}(\omega) \,/\, T\right)}{\sum_{\omega' \in \Omega} \exp\left(-\mathcal{H}(\omega') \,/\, T\right)} \\ &\equiv \exp\left(-\left(\mathcal{H}(\omega) - \mathcal{F}\left(\mathcal{H}\right)\right) \,/\, T\right) \end{aligned} \tag{2}$$

which is known as the Gibbs distribution. The quantity $\mathcal{F}\left(\mathcal{H}\right) \equiv -T \log \sum_{\omega' \in \Omega} \exp\left(-\mathcal{H}(\omega') \,/\, T\right)$ denotes the Gibbs free energy. The temperature T formally plays the role of a Lagrange parameter to enforce a constraint on the expected costs

$$\left\langle \mathcal{H} \right\rangle \equiv \sum_{\omega \in \Omega} \mathbf{P}^{\text{Gb}}(\omega)\, \mathcal{H}(\omega). \tag{3}$$

The Gibbs free energy $\mathcal{F}(\mathcal{H})$ is related to the expected costs $\langle\mathcal{H}\rangle$ and to the entropy S by

$$S\left(\mathbf{P}^{\text{Gb}}\right) = -\sum_{\omega \in \Omega} \mathbf{P}^{\text{Gb}}(\omega) \log \mathbf{P}^{\text{Gb}}(\omega) = \frac{1}{T}\left\langle \mathcal{H} \right\rangle - \frac{1}{T}\mathcal{F}\left(\mathcal{H}\right). \tag{4}$$

2.2 Deterministic Annealing

A stochastic search according to a Markov process with a transition matrix (1) allows us to estimate expectation values of system parameters by computing time averages in a Monte Carlo simulation, e.g., the variables of the optimization problem are drawn according to $\mathbf{P}^{\text{Gb}}(\omega)$. This random, sequential sampling of the solution space, however, is slow compared to deterministic optimization techniques due to the diffusive nature of the search process. A deterministic variant of simulated annealing, "*deterministic annealing*," analytically estimates relevant expectation values of system parameters, e.g., the variables of the optimization problem. We introduce the generalized free energy

$$\begin{aligned} \tilde{\mathcal{F}}(\mathbf{P}) \equiv \left\langle \mathcal{H} \right\rangle_{\mathbf{P}} - TS(\mathbf{P}) = \\ \sum_{\omega \in \Omega} \mathbf{P}(\omega)\mathcal{H}(\omega) + T\sum_{\omega \in \Omega} \mathbf{P}(\omega)\log\mathbf{P}(\omega) \end{aligned} \tag{5}$$

which has to be minimized over a (tractable) subspace of probability distributions. The inequality $\tilde{\mathcal{F}}(\mathbf{P}) \geq \tilde{\mathcal{F}}\left(\mathbf{P}^{\text{Gb}}\right) \equiv \mathcal{F}\left(\mathcal{H}\right)$ holds since the Gibbs distribution maximizes entropy [1], [2] for $\langle\mathcal{H}\rangle_{\mathbf{P}}$ kept fixed. The search space of probability densities is defined in order to analytically approximate expectation values of the optimization parameters. The temperature parametrizes a family of generalized free energies with increasing complexity for $T \rightarrow 0$, i.e., high temperature smoothes the cost function and low temperature reveals the full complexity of the original optimization problem, i.e., recovering it for $T \rightarrow 0$. Deterministic annealing algorithms track good solutions from high to low temperature in a similar way to cooling in simulated annealing. We will discuss this technique in detail in Section 4.

Why should we consider a stochastic or deterministic search strategy based on principles from statistical physics? The fundamental relationship between statistical physics and robust statistics has been established by Jaynes [1], [2], [3] who postulated the principle of maximum entropy inference. Maximizing the entropy yields the least biased inference method being *maximally noncommittal with respect to missing data*. In the context of data clustering the missing information are the assignments of data to clusters. Another important argument in favor of the maximum entropy method stresses the robustness of this inference technique. Tikochinsky et al. [26] have proven that the maximum entropy probability distribution is maximally stable in terms of the L_2 norm if the expected cost $\langle\mathcal{H}\rangle$ is lowered or raised by changes of the temperature. The family of Gibbs distributions for a given cost function possesses the optimality property to induce the least variations if $\langle\mathcal{H}\rangle$ is reduced. Using concepts from differential geometry, the family of Gibbs distributions parameterized by the temperature

forms a trajectory in the space of probability distributions which has minimal length [27]. We conclude from these facts that a stochastic search heuristic which starts with a large noise level and which gradually reduces stochasticity to zero should be based on the family of Gibbs distributions with decreasing temperature. This strategy guarantees maximal robustness with respect to noise.

3 Cost Functions for Data Clustering

3.1 Central Clustering

The most widely used nonparametric technique for finding data prototypes is central clustering or vector quantization. Given a set of d-dimensional data vectors $\Xi = \{x_i \in \mathbb{R}^d : i \in \{1, \ldots, N\}\}$, central clustering poses the problem of determining an optimal set of d-dimensional reference vectors or prototypes $\Upsilon = \{\mathbf{y}_\nu \in \mathbb{R}^d : \nu \in \{1, \ldots, K\}\}$. To specify a data partition we introduce Boolean assignment variables $\mathbf{M}$ and a configuration space $\mathcal{M}$,

$$\mathbf{M} = \left(M_{i\nu}\right)\Big|_{\substack{i=1,\ldots,N \\ \nu=1,\ldots,K}} \in \{0,1\}^{N\times K}, \tag{6}$$

$$\mathcal{M} = \left\{\mathbf{M} \in \{0,1\}^{N\times K} : \sum_{\nu=1}^{K} M_{i\nu} = 1,\ \forall i\right\}. \tag{7}$$

$M_{i\nu} \equiv 1$ if the data point $\mathbf{x}_i$ is assigned to reference vector $\mathbf{y}_\nu$, while $M_{i\nu} \equiv 0$ otherwise. The solution space is defined as the set of admissible configurations in (7) with the constraints $\sum_{\nu=1}^{K} M_{i\nu} = 1,\ \forall i$ assuring that each data point is represented by a unique reference vector $\mathbf{y}_\alpha \in \Upsilon$. The quality of a set of reference vectors is assessed by the objective function for central clustering which sums up the average distortion error $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)$ between a data vector $\mathbf{x}_i$ and the corresponding reference vector $\mathbf{y}_\nu$, i.e.,

$$\mathcal{H}^{\mathrm{cc}}(\mathbf{M}) = \sum_{i=1}^{N}\sum_{\nu=1}^{K} M_{i\nu}\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu). \tag{8}$$

An appropriate distortion measure $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)$ depends on the application domain, with the most common choice being the squared Euclidian distance $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) \equiv \|\mathbf{x}_i - \mathbf{y}_\nu\|^2$ between the data vector and its reference vector. Applications with a topological ordering of the reference vectors as for source-channel coding demand a distortion measure which considers the topological organization of the reference vectors, e.g., $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha) \equiv \sum_{\nu=1}^{K} T_{\alpha\nu}\|\mathbf{x}_i - \mathbf{y}_\nu\|^2$, where $T_{\alpha\nu}$ specifies the probability that index α is confused with index ν due to transmission noise. Distortions with a low-dimensional, topological arrangement defining a chain or a two-dimensional grid are very popular such as self-organizing topological maps in the area of neural computing [28], [29]. The number of clusters can be limited by additional rate distortion constraints, e.g., Shannon entropy or penalties for small clusters, rather than postulating a fixed number K [21].

Stochastic optimization of the cost function (8) requires us to determine the probability distribution of assignments $\mathbf{M}$. The maximum entropy principle, originally suggested by Rose et al. [16] for central clustering, states that the assignments are distributed according to the Gibbs distribution

$$\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\right) = \exp\left(-\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M}) - \mathcal{F}\left(\mathcal{H}^{\mathrm{cc}}\right)\right) / T\right), \tag{9}$$

$$\mathcal{F}\left(\mathcal{H}^{\mathrm{cc}}\right) = -T\log\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{cc}}(\mathbf{M}) / T\right). \tag{10}$$

As pointed out in Section 2, the free energy $\mathcal{F}(\mathcal{H}^{\mathrm{cc}})$ in (10) can be interpreted as a smoothed version of the original cost function $\mathcal{H}^{\mathrm{cc}}$. The factor $\exp(\mathcal{F}(\mathcal{H}^{\mathrm{cc}})/T)$ normalizes the exponential weights $\exp(-\mathcal{H}^{\mathrm{cc}}(\mathbf{M})/T)$. Its inverse can be rewritten as

$$\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\sum_{i=1}^{N}\sum_{\nu=1}^{K} M_{i\nu}\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) / T\right) = \prod_{i=1}^{N}\sum_{\nu=1}^{K}\exp\left(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) / T\right), \tag{11}$$

since the sum over assignments is constrained by $\sum_{\nu=1}^{K} M_{i\nu} = 1$. The cost function $\mathcal{H}^{\mathrm{cc}}$ which is linear in $M_{i\nu}$ yields a factorized Gibbs distribution

$$\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\right) = \prod_{i=1}^{N}\frac{\exp\left(-\sum_{\nu=1}^{K} M_{i\nu}\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) / T\right)}{\sum_{\mu=1}^{K}\exp\left(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\mu) / T\right)} \tag{12}$$

for predefined reference vectors $\Upsilon = \{\mathbf{y}_\nu\}$. This Gibbs distribution can also be interpreted as the complete data likelihood for mixture models with parameters Υ.

Following Rose et al. [16], the optimal reference vectors $\{\mathbf{y}_\nu^*\}$ are derived by maximizing the entropy of the Gibbs distribution, keeping the average costs $\langle\mathcal{H}^{\mathrm{cc}}\rangle$ fixed, i.e.,

$$\begin{aligned}\Upsilon^* &= \arg\max_{\Upsilon}\left(-\sum_{\mathbf{M}\in\mathcal{M}}\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\right)\log\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\right)\right) \\ &= \arg\max_{\Upsilon}\left(\sum_{\mathbf{M}\in\mathcal{M}}\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}(\mathbf{M})\right) / T\right. \\ &\quad \left. + \sum_{i=1}^{N}\log\sum_{\mu=1}^{K}\exp\left(-\mathcal{D}\left(\mathbf{x}_i, \mathbf{y}_\mu\right) / T\right)\right).\end{aligned} \tag{13}$$

$\mathbf{P}^{\mathrm{Gb}}(\mathbf{M})$ is the Gibbs distribution of the assignments for a set Υ of fixed reference vectors. To determine closed equations for the optimal reference vectors $\mathbf{y}_\nu^*$ we differentiate the argument in (13) with the expected costs kept constant. The resulting equation

$$0 = \sum_{i=1}^{N}\left\langle\mathbf{M}_{i\nu}\right\rangle\frac{\partial}{\partial\mathbf{y}_\nu}\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) \tag{14}$$

$$\left\langle M_{i\nu}\right\rangle \equiv \frac{\exp\left(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu) / T\right)}{\sum_{\mu}\exp\left(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\mu) / T\right)} \quad \forall\nu \in \{1, \ldots, K\} \tag{15}$$

is known as the centroid equation in signal processing. The

angular brackets denote Gibbs expectation values, i.e., $\langle f(\mathbf{M})\rangle \equiv \sum_{\mathbf{M}\in\mathcal{M}} f(\mathbf{M})\mathbf{P}^{\mathrm{Gb}}(\mathbf{M})$. The reader should realize that entropy maximization implies $\mathbf{y}_\nu$ to be a centroid which is also optimal in the sense of rate distortion theory [30].

Equations (14) and (15) are efficiently solved in an iterative fashion using the expectation maximization (EM) algorithm [30]. The EM algorithm alternates an expectation step to determine the expected assignments $\langle M_{i\nu}\rangle$ with a maximization step to estimate maximum likelihood values for the cluster centers $\mathbf{y}_\nu$. Dempster et al. [31] have proven that the likelihood increases monotonically under this alternation scheme which demonstrates convergence of the algorithm toward a local maximum of the likelihood function. The log-likelihood is up to a factor $(-T)$ equivalent to the free energy for central clustering. In Section 5 we will use the solutions of central clustering with squared Euclidian distances to simultaneously group a data set and embed it in the Euclidian space $\mathbb{R}^d$ by preserving the cluster structure.

3.2 Pairwise Clustering

Central clustering requires that the data can be characterized by feature values $\mathbf{x}_i \in \mathbb{R}^d$ in a d-dimensional Euclidian space. Frequently in empirical sciences, however, the only available information source about a data set are comparisons between data pairs; these dissimilarity values are denoted in the following by $\mathbf{D} = \left(\mathcal{D}_{ik}\right)\Big|_{\substack{i=1,\ldots,N\\ k=1,\ldots,N}} \in \mathbb{R}^{N\times N}$. Clustering on the basis of this data description is achievable by grouping the data to clusters such that the sum of dissimilarities between data of the same cluster is minimized [4], [8]. This criterion favors compact and coherent groups over heterogeneous data collections. We again use the set of assignment variables $\mathbf{M}$ as defined in (6) and denote by $M_{i\nu}$ the indicator function of an assignment of datum i to cluster ν. To compensate for different numbers of data per cluster the costs of a particular cluster ν are normalized by the percentage $p_\nu = \sum_{i=1}^{N} M_{i\nu} \,/\, N$ of data in that cluster. Without this normalization, the undesirable and often detrimental tendency can be observed that clusters with few data grow at the expense of equally coherent clusters with many data. The cost function for pairwise clustering with K clusters

$$\mathcal{H}^{\mathrm{pc}}(\mathbf{M}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{\mathcal{D}_{ik}}{N}\left(\sum_{\nu=1}^{K}\frac{M_{i\nu}M_{k\nu}}{p_\nu} - 1\right) \tag{16}$$

stresses cluster coherency. Alternative clustering costs have been proposed [8], but have not found wide spread acceptance in pattern recognition applications. The constant term $\sum_{i=1}^{N}\sum_{k=1}^{N}\mathcal{D}_{ik} \,/\, (2N)$ has been subtracted in (16) to emphasize the independence of the clustering cost function on the absolute dissimilarity scale, i.e.,

$$\mathcal{H}^{\mathrm{pc}}\left(\mathbf{M}\middle|\left(\mathcal{D}_{ik} - \mathcal{D}_0\right)\right) = \mathcal{H}^{\mathrm{pc}}\left(\mathbf{M}\middle|\left(\mathcal{D}_{ik}\right)\right), \tag{17}$$

since $\mathcal{D}_0\sum_{i=1}^{N}\sum_{k=1}^{N}\left(\sum_{\nu=1}^{K} M_{i\nu}M_{k\nu} \,/\, p_\nu - 1\right) / N = 0$. A uniform shift of the dissimilarity values by an offset $\mathcal{D}_0$ does not change the clustering costs and, consequently, has no influence on the statistics of the assignments. Another cost function with this offset invariance is

$$\mathcal{H}'(\mathbf{M}) = \tfrac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathcal{D}_{ik}\left(\sum_{\nu=1}^{K} M_{i\nu}M_{k\nu} \,/\, p_\nu^2 - K\right) / N$$

which, however, displays a tendency for extremely heterogeneous partitionings with very large and very small clusters [32]. A second important property of the proposed cost function concerns non–symmetric dissimilarities, $\mathcal{D}_{ik} \neq \mathcal{D}_{ki}$. $\mathcal{H}^{\mathrm{pc}}$ is not changed if all dissimilarities are replaced by the arithmetic mean, $\mathcal{D}_{ik} \leftarrow (\mathcal{D}_{ik} + \mathcal{D}_{ki})/2$. For reasons of simplicity, we henceforth assume symmetric $\mathcal{D}_{ik}$. Furthermore, $\mathcal{H}^{\mathrm{pc}}$ is also invariant under an arbitrary permutation of the cluster indices $\nu \to \pi(\nu)$[2].

An important, although often ignored consideration for stochastic optimization problems is the scaling of the $\mathcal{D}_{ik}$ values with the number N of data. The correct scaling should yield constant costs per data point to achieve independence of annealing schedules and stochastic search heuristics from the instance size N. In the case of completely consistent dissimilarities, i.e., data i, k in different clusters have large $\mathcal{D}_{ik}$ and data i, k in the same clusters have small $\mathcal{D}_{ik}$, constant costs per datum require a scaling $\mathcal{D}_{ik} \sim \mathcal{O}(1)$. In the opposite case of random dissimilarity values averaging effects necessitate a scaling $\mathcal{D}_{ik} \sim \mathcal{O}\left(\sqrt{N}\right)$. A thorough discussion of this point can be found in the statistical physics literature of optimization problems [33].

4 Mean-Field Approximation of Pairwise Clustering

Following the strategy of stochastic optimization as discussed in Section 3.1 for central clustering, we estimate the expectation values for the assignment of data to clusters at a specified uncertainty level parametrized by the computational temperature T. Assignments $\mathbf{M}$ of data to clusters are randomly drawn from the set of admissible configurations (7) according to the Gibbs distribution

$$\begin{aligned}\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\right) &= \frac{\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M}) \,/\, T\right)}{\sum_{\mathbf{M}'\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M}') \,/\, T\right)}\\ &\equiv \exp\left(-\left(\mathcal{H}^{\mathrm{pc}}(\mathbf{M}) - \mathcal{F}\left(\mathcal{H}^{\mathrm{pc}}\right)\right) / T\right),\end{aligned} \tag{18}$$

where $\mathcal{H}^{\mathrm{pc}}$ are the costs for a pairwise clustering solution (16). Contrary to the Gibbs distribution for central clustering, the data assignments $\mathbf{M}$ in pairwise clustering are statistically dependent and the Gibbs distribution (18) cannot be exactly rewritten in factorized form. Each assignment variable $M_{i\nu}$ interacts with all other assignment variables. These cost contributions, however, converge to averages in the limit of large data sets and reduce the influence of correlations on individual data assignments. We, therefore,

2. The permutation symmetry can be removed by adding a small perturbation $\delta\mathcal{H} := \sum_{\nu=1}^{K} R_\nu(N)\sum_{i=1}^{N} M_{i\nu}$ to the cost function (16) with $0 < R_1(N) < \ldots < R_K(N)$, $\lim_{N\to\infty} R_1(N) = 0$, $\lim_{N\to\infty} N\,R_1(N) = \infty$. The perturbations $R_\nu(N)$ favor an indexing of the clusters according to their size $(p_1 > p_2 > \ldots > p_K)$.

approximate the average interaction of $M_{i\nu}$ with other assignment variables by a *mean–field* $\mathcal{E}_{i\nu}$. The following two sections present a variational technique and a perturbation expansion to derive the mean–field approximation and corrections to the assignment correlations. The method is, however, not restricted to clustering problems and can be applied to many combinatorial optimization problems.

4.1 Mean–Field Approximation as Minimization of KL-Divergence

A mean–field approximation of the Gibbs distribution $\mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^{\mathrm{pc}})$ neglects the correlations between the stochastic variables in the pairwise clustering cost function $\mathcal{H}^{\mathrm{pc}}$ and determines the "most similar" factorized distribution within an $\mathcal{E}$–parametrized family of distributions $\mathbf{P}^0(\boldsymbol{\mathcal{E}})$. The distribution $\mathbf{P}^0(\mathcal{E}^*)$ which represents most accurately the statistics of the original problem is specified by the minimum of the Kullback–Leibler divergence to the original Gibbs distribution, i.e.,

$$\mathcal{E}^* = \arg\min_{\mathcal{E}} \mathcal{I}\left(\mathbf{P}^0(\mathcal{E}) \middle\| \mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{pc}}\right)\right). \tag{19}$$

In the pairwise clustering case we define an approximating family of distributions introducing potentials $\mathcal{E} = \left(\mathcal{E}_{i\nu}\right)\Big|_{\substack{i=1,\ldots,N\\ \nu=1,\ldots,K}} \in \mathbb{R}^{\mathrm{N}\times\mathrm{K}}$ for the effective interactions, where $\mathcal{E}_{i\nu}$ represents the partial costs for assigning datum i to cluster ν. Summing up the partial costs we arrive at a family of cost functions without correlations between the assignments, i.e.,

$$\mathcal{H}^0(\mathbf{M}, \mathcal{E}) = \sum_{i=1}^{N}\sum_{\nu=1}^{K} M_{i\nu}\mathcal{E}_{i\nu}. \tag{20}$$

The linearity in the assignments of (20) reflects the fact that we assume statistical independence between the assignments, i.e., they are distributed according to a factorized Gibbs distribution $\mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^0) \equiv \mathbf{P}^0(\boldsymbol{\mathcal{E}})$.

An equivalent minimization condition for the free energy can be derived from (19) by the following algebraic transformations

$$\begin{aligned}\mathcal{I}\left(\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^0\right)\middle\|\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{pc}}\right)\right) &= \sum_{\mathbf{M}\in\mathcal{M}} \mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^0(\mathbf{M})\right) \\ &\quad \log\frac{\exp\left[-\mathcal{H}^0(\mathbf{M}) / T\right]\sum_{\mathbf{M}'\in\mathcal{M}}\exp\left[-\mathcal{H}^{\mathrm{pc}}(\mathbf{M}') / T\right]}{\exp\left[-\mathcal{H}^{\mathrm{pc}}(\mathbf{M}) / T\right]\sum_{\mathbf{M}'\in\mathcal{M}}\exp\left[-\mathcal{H}^0(\mathbf{M}') / T\right]} \\ &= \frac{1}{T}\left[\mathcal{F}\left(\mathcal{H}^0\right) - \mathcal{F}\left(\mathcal{H}^{pc}\right) + \left\langle\mathcal{H}^{pc} - \mathcal{H}^0\right\rangle\right].\end{aligned} \tag{21}$$

The averaging brackets $\langle\cdot\rangle$ denote the average with respect to $\mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^0)$. Since the KL–divergence is always positive and vanishes only for $\mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^0) \equiv \mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^{\mathrm{pc}})$ we obtain the well-known upper bound first derived by Peierls [34]:

$$\mathcal{F}(\mathcal{H}^{\mathrm{pc}}) \le \mathcal{F}(\mathcal{H}^0) + \langle\mathcal{H}^{\mathrm{pc}} - \mathcal{H}^0\rangle.$$

In summary, the optimal mean–fields $\boldsymbol{\mathcal{E}}^*$ result from a variational approach to minimizing the upper bound (22) on the free energy and thus to minimizing the KL-divergence (19). The upper bound can be interpreted as the generalized free energy (5) which is defined in the restricted space of factorized probability distributions for $\mathbf{M}$.

The minimization of the upper bound on the free energy yields the "optimal" potentials $\mathcal{E}^*_{i\nu}$ for assigning datum i to cluster ν:

$$\begin{aligned}&\frac{\partial}{\partial\mathcal{E}_{i\nu}}\left(\mathcal{F}\left(\mathcal{H}^0\right) + \left\langle\mathcal{H}^{\mathrm{pc}} - \mathcal{H}^0\right\rangle\right)\Big|_{\mathcal{E}_{i\nu}=\mathcal{E}^*_{i\nu}} = 0 \\ &\qquad \Rightarrow \mathcal{E}^*_{i\nu} \equiv \left\langle\tilde{\mathcal{E}}_{i\nu}\right\rangle \qquad \forall\nu\in\{1,\ldots,K\},\end{aligned} \tag{23}$$

with

$$\begin{aligned}\tilde{\mathcal{E}}_{i\nu} &\equiv \frac{1}{\sum_{j\neq i}^{N} M_{j\nu} + 1} \\ &\left[\frac{1}{2}\mathcal{D}_{ii} + \sum_{k\neq i}^{N} M_{k\nu}\left(\mathcal{D}_{ik} - \frac{1}{2}\sum_{j\neq i}^{N}\frac{M_{j\nu}}{\sum_{l\neq i}^{N} M_{l\nu}}\mathcal{D}_{jk}\right)\right].\end{aligned} \tag{24}$$

The resulting optimal (with respect to (21)) assignments are given by

$$\left\langle M_{i\alpha}\right\rangle = \frac{\exp\left(-\mathcal{E}^*_{i\alpha} / T\right)}{\sum_{\nu=1}^{K}\exp\left(-\mathcal{E}^*_{i\nu} / T\right)}. \tag{25}$$

The technical details can be found in Appendix A. The reader should note that the potentials $\mathcal{E}^*_{i\nu}$ do not depend on the variables $(\langle M_{i1}\rangle, \ldots, \langle M_{iK}\rangle)$.

We introduce an approximation which neglects terms of order $\mathcal{O}(1/N)$ to simplify the potentials $\mathcal{E}^*_{i\nu}$. The approximations $\left\langle 1 / \left(\sum_{j\neq i}^{N} M_{j\nu} + 1\right)\right\rangle \approx 1 / \left(\sum_{j\neq i}^{N}\left\langle M_{j\nu}\right\rangle + 1\right)$ and $\left\langle 1 / \left(\sum_{j\neq i}^{N} M_{j\nu}\right)\right\rangle \approx 1 / \left(\sum_{j\neq i}^{N}\left\langle M_{j\nu}\right\rangle\right)$ are correct in the limit of large N. To simplify the presentation further, we assume zero self–dissimilarities, $\mathcal{D}_{ii} = 0$, $\forall i$. The simplified "optimal" potentials

$$\mathcal{E}^*_{i\nu} = \frac{1}{\sum_{j\neq i}^{N}\left\langle M_{j\nu}\right\rangle + 1}\sum_{k=1}^{N}\left\langle M_{k\nu}\right\rangle\left(\mathcal{D}_{ik} - \frac{1}{2\sum_{j\neq i}^{N}\left\langle M_{j\nu}\right\rangle}\sum_{j=1}^{N}\left\langle M_{j\nu}\right\rangle\mathcal{D}_{jk}\right) \tag{26}$$

depend on the given distance matrix, the averaged assignment variables and the cluster probabilities. The following algorithm estimates the assignment probabilities $\langle M_{i\nu}\rangle$ and the optimal potentials $\mathcal{E}^*_{i\nu}$ (defined in (26)) iteratively.

Algorithm I

```
INITIALIZE E*_iν^(0) and ⟨M_iν⟩^(0) randomly;
   temperature T ← T_0;
WHILE T > T_FINAL
 t ← 0;
 REPEAT
   E-like step: estimate ⟨M_iν⟩^(t+1)
     as a function of E*_iν^(t);
   M-like step: calculate E*_iν^(t+1)
     for given ⟨M_iν⟩^(t+1);
   t ← t + 1;
 UNTIL all {⟨M_iν⟩^(t), E*_iν^(t)} satisfy (26);
```

$$T \leftarrow \eta T;\ \langle M_{i\nu}\rangle^{(0)} \leftarrow \langle M_{i\nu}\rangle^{(t)};\ \mathcal{E}_{i\nu}^{*(0)} \leftarrow \mathcal{E}_{i\nu}^{*(t)};$$

The algorithm decreases the temperature exponentially ($0 < \eta < 1$) and alternates the estimation of data assignments for given potentials (E-like step) with an estimate of potentials for given assignments. This estimation procedure can be carried out sequentially or in parallel.[3] For the assignments in the E-like step and the potentials in the M-like step. A sequential version where the E-like step and the M-like step are performed for a randomly selected datum i converges to a local minimum (with respect to $(\mathcal{E}_{i1}, \ldots \mathcal{E}_{iK})$) of the upper bound of the free energy, since $\mathcal{E}_{i\nu}^{*(t+1)}$ is uniquely determined by $\left\{\left\langle M_{k\nu}\right\rangle^{(t+1)}\right\}\Big|_{\substack{k=1\\k\neq i}}^{N}$, which have no explicit dependency on $\mathcal{E}_{i\nu}^{*(t+1)}$. The upper bound in (22) plays the role of a Lyapunov function for the update dynamics of the potentials $\mathcal{E}_{i\nu}^{*(t+1)}$ [32]. The sequential update scheme has been implemented in the clustering experiments (see Section 6). The outer loop of the algorithm reduces the temperature in an exponential fashion, i.e., we choose $T \leftarrow \eta T$. Other choices such as linear annealing schedules to lower the temperature could be used as well and they might yield superior optimization results since the search process is extended by a slower temperature reduction.

4.2 Equations for Expected Data Assignments

The variational approach with a family of factorized distributions implicitly assumes that correlations between assignments can be neglected. A direct estimate of the average assignments allows us to check how well this assumption holds and what estimation errors are introduced by the underlying independence hypothesis. The detailed derivations of the equations (27), (28), and (29) are summarized in Appendix B. The true expected assignments are given by

$$\left\langle M_{i\alpha}\right\rangle = \left\langle \frac{\exp\left(-\tilde{\mathcal{E}}_{i\alpha} / T\right)}{\sum_{\nu}\exp\left(-\tilde{\mathcal{E}}_{i\nu} / T\right)}\right\rangle, \tag{27}$$

$\tilde{\mathcal{E}}_{i\alpha}$ being defined in (24). The fraction $\exp\left(-\tilde{\mathcal{E}}_{i\alpha} / T\right) / \sum_{\nu}\exp\left(-\tilde{\mathcal{E}}_{i\nu} / T\right)$ in (27) implements a partition of unity. The system of the $N \times K$ equations (27) is computationally intractable since we have to carry out the averaging of the partition of unity over an exponential number of assignment configurations. The smoothness of a transition from one cell of the partition to a neighboring cell is controlled by the inverse temperature $1/T$.

Naively interchanging the averaging brackets with the nonlinear function in (27) yields the equations (25), (26); a refined mean–field approach which is known as the TAP approach [35] models the feedback effects in strongly disordered clustering instances more faithfully than the naive approach. The refined expected assignments are

$$\left\langle M_{i\alpha}\right\rangle = \frac{\exp\left(-\left(\left\langle\tilde{\mathcal{E}}_{i\alpha}\right\rangle - \tilde{h}_{i\alpha}\right) / T\right)}{\sum_{\nu}\exp\left(-\left(\left\langle\tilde{\mathcal{E}}_{i\nu}\right\rangle - \tilde{h}_{i\nu}\right) / T\right)} \tag{28}$$

$$\tilde{h}_{i\alpha} = \frac{1}{2T}\sum_{k=1}^{N}\mathcal{D}_{ik}^{2}\sum_{\mu=1}^{K}\frac{\left\langle M_{k\alpha}\right\rangle\left(\delta_{\alpha\mu} - \left\langle M_{k\mu}\right\rangle\right)}{p_{\alpha}p_{\mu}N^{2}}\left(\delta_{\alpha\mu} - 2\left\langle M_{i\mu}\right\rangle\right) \tag{29}$$

where $\delta_{\alpha\mu}$ denotes the Kronecker delta. The corrections $\tilde{h}_{i\alpha}$ are also called cavity fields.

The question about the range of validity of (28) is subtle, and it has been studied extensively in the statistical physics of disordered systems [33]. Empirically, we can measure average values of the assignments by Monte Carlo simulation. These estimates are inserted into (28) or into (25), which yields residual errors, i.e., the difference between the right and the left side of both equations. The residual errors determine the quality of the TAP approximation compared to the naive mean field approximation. According to our Monte Carlo experiments with matrices of Gaussian distributed random dissimilarity values (N = 1,200), the TAP equations (28) estimate the average assignments $\langle M_{i\nu}\rangle$ with a reduced residual error of up to 50% less compared with the naive mean field approximation. The difference reaches a maximum for temperatures near the phase transition point, i.e., when degenerate clusters split into separate clusters. The naive mean–field equation is superior in the low temperature range. Furthermore, we observed that the improvements achieved by the TAP equations can be neglected for small problems ($N < 100$) because of the $N \to \infty$ asymptotics.

5 Pairwise Clustering and Embedding

Grouping data into clusters is an important concept in discovering structure. Apart from partitioning data into classes, the data analyst often relies on visual inspection of data to recognize correlations and deviations from randomness. The task of embedding given dissimilarity data **D** in a d-dimensional Euclidian space, a prerequisite for visual inspection, is known as multidimensional scaling [8], [36]. Usually, multidimensional scaling is formulated as an optimization problem for the coordinates $\{\mathbf{x}_i\}$ with costs

$$\mathcal{H}^{\mathrm{mds}}\left(\{\mathbf{x}_i\}\right) = \frac{1}{2N}\sum_{i,k=1}^{N}\left(\frac{\left\|\mathbf{x}_i - \mathbf{x}_k\right\|^{2} - \mathcal{D}_{ik}}{\hat{\mathcal{D}}_{ik}}\right)^{2}. \tag{30}$$

The so-called stress function $\mathcal{H}^{\mathrm{mds}}$ was introduced by Kruskal in [37]. $\mathcal{H}^{\mathrm{mds}}$ with a constant normalization $\hat{\mathcal{D}}_{ik} = 1$ measures the absolute stress and $\hat{\mathcal{D}}_{ik} = \mathcal{D}_{ik}$ penalizes relative stress.

5.1 Mean Field Approximation of Pairwise Clustering by Central Clustering

In this section, we establish a connection between the clustering and the multidimensional scaling problem. The strategy of combining data clustering and data embedding in a Euclidian space is based on a variational approach to

3. Experimentally, we observed oscillation for the parallel $\mathcal{E}_{i\nu}^{*(t)}$ update as it is known from parallel update of neural networks.

maximum entropy estimation as discussed in Section 4.1. The coordinates of data points in the embedding space are estimated in such a way that the statistics of the resulting cluster structure matches the statistics of the original pairwise clustering solution. The relation of this new principle for structure preserving data embedding to standard multidimensional scaling is summarized in the following diagram:

$$\begin{array}{ll} \{\mathcal{D}_{ik}\} & \rightarrow \mathcal{H}^{\mathrm{pc}}\left(\mathbf{M}\middle|\{\mathcal{D}_{ik}\}\right) \rightarrow \mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{pc}}\left(\mathbf{M}\middle|\{\mathcal{D}_{ik}\}\right)\right) \\ \downarrow \mathcal{H}^{\mathrm{mds}} & \downarrow I\left(\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}\right)\middle\|\mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{pc}}\right)\right) \\ \left\{\|\mathbf{x}_i-\mathbf{x}_k\|^2\right\} \rightarrow \mathcal{H}^{\mathrm{cc}}\left(\mathbf{M}\middle|\{\mathbf{x}_i\}\right) & \rightarrow \mathbf{P}^{\mathrm{Gb}}\left(\mathcal{H}^{\mathrm{cc}}\left(\mathbf{M}\middle|\{\mathbf{x}_i\}\right)\right). \end{array}$$

Multidimensional scaling offers the left path from dissimilarities to coordinates whereas we advocate the right path. The variational approach to mean–field approximation involved in the right path requires us to specify a parametrized family of factorized Gibbs distributions. We choose the factorized Gibbs distributions (12) based on the cost function for central clustering $\mathcal{H}^{\mathrm{cc}}(\mathbf{M}\,|\,\{\mathbf{x}_i\})$ and use the embedding coordinates $\{\mathbf{x}_i\}$ as the variational parameters. This approach is motivated by the identity

$$\sum_{i=1}^{N} M_{i\nu}\|\mathbf{x}_i-\mathbf{y}_\nu\|^2 = \frac{1}{2Np_\nu}\sum_{i=1}^{N}\sum_{k=1}^{N} M_{i\nu}M_{k\nu}\|\mathbf{x}_i-\mathbf{x}_k\|^2 \qquad (31)$$

with

$$\mathbf{y}_\nu = \frac{\sum_{i=1}^{N} M_{i\nu}\mathbf{x}_i}{\sum_{i=1}^{N} M_{i\nu}} \qquad (32)$$

which yields a correct approximation for pairwise clustering instances with $\mathcal{D}_{ik}=\|\mathbf{x}_i-\mathbf{x}_k\|^2$.

Suppose we have found a stationary solution of the mean–field equations (25), (26). For the clustering problem it suffices to consider the mean assignments $\langle M_{i\nu}\rangle$ with the parameters $\mathcal{E}^*_{i\nu}$ being auxiliary variables. The identity (31) allows us to interpret these variables as the squared distance to the cluster centroid under the assumption of Euclidian data. In the multidimensional scaling problem the coordinates $\mathbf{x}_i$ are the unknown quantities. If we restrict the potentials $\mathcal{E}_{i\upsilon}$ to be of the form $\|\mathbf{x}_i-\mathbf{y}_\nu\|^2$ with the centroid definition (32) we have specified a new family of approximating distributions defined in (12) with parameters $\{\mathbf{x}_i\in\mathbb{R}^d : 1\le i\le N\}$. The effective dimensionality of the parameter space is $\min\{d,(K-1)\}\times N$ instead of $(K-1)\times N$, which is a significant reduction, especially in the case of low–dimensional embeddings ($d \ll K$). The criterion for determining the embedding coordinates is

$$\frac{\partial}{\partial\mathbf{x}_i}\left[\mathcal{F}\left(\mathcal{H}^{\mathrm{cc}}\right)+\left\langle\mathcal{H}^{\mathrm{pc}}-\mathcal{H}^{\mathrm{cc}}\right\rangle\right]=0, \qquad (33)$$

which approximately yields the coordinates

$$\mathbf{K}_i\mathbf{x}_i \approx \frac{1}{2}\sum_{\nu=1}^{K}\left\langle M_{i\upsilon}\right\rangle\left(\|\mathbf{y}_\nu\|^2-\mathcal{E}^*_{i\nu}\right)\left(\mathbf{y}_\nu-\sum_{\mu=1}^{K}\left\langle M_{i\mu}\right\rangle\mathbf{y}_\mu\right), \qquad (34)$$

$$\mathbf{K}_i=\left(\left\langle\mathbf{y}\mathbf{y}^T\right\rangle_i-\langle\mathbf{y}\rangle_i\langle\mathbf{y}\rangle_i^T\right), \qquad \langle\mathbf{y}\rangle_i=\sum_{\nu=1}^{K}\left\langle M_{i\nu}\right\rangle\mathbf{y}_\nu. \qquad (35)$$

Details of the derivation are summarized in Appendix C. The coordinates $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_\nu\}$ are determined by iteratively solving (34) according to the following algorithm:

Algorithm II: Structure Preserving MDS

```
INITIALIZE x_i^(0), y_ν^(0) and ⟨M_iν⟩^(0) ∈ (0,1)
  randomly;
  temperature T ← T_0;
WHILE T > T_FINAL
    t = 0;
  REPEAT
    E-like step: estimate ⟨M_iν⟩^(t+1) as a
      function of {x_i, y_ν};
    M-like step:
      REPEAT
        calculate x_i^(t+1) given ⟨M_iν⟩^(t+1) y_ν^(t+1).
        update y_ν^(t+1) to fulfill the
          centroid condition;
      UNTIL convergence;
    t ← t + 1;
  UNTIL convergence;
  T ← ηT; ⟨M_iν⟩^(0) ← ⟨M_iν⟩^(t); x_i^(0) ← x_i^(t); y_ν^(0) ← y_ν^(t);
```

To understand the properties of the algorithm we have to recollect the key idea for deriving the mean–field approximation. The statistics of the approximating system with the cost function $\mathcal{H}^{\mathrm{cc}}$ has to be optimally adjusted to the statistics of the original system. This fact implies that we are not able to determine the variational parameters in the limit of fixed statistics, e.g., in the limit of zero temperature. As can be easily seen, equations (34) is singular for $T = 0$ and asymptotic results require us to apply l'Hospital's Rule.

The derived system of transcendental equations given by (15) with quadratic distortions, by (34) and by the centroid condition explicitly reflects the dependencies between the clustering procedure and the Euclidian representation. Simultaneous solution of these equations leads to an efficient algorithm which interleaves the multidimensional scaling process and the clustering process, and which avoids an artificial separation into two uncorrelated data processing steps.

6 Results

We demonstrate the properties of the proposed clustering Algorithms I and II by three classes of experiments:

i) benchmark optimization experiments compare deterministic annealing with a greedy gradient descent method and a linkage algorithm for pairwise clustering in Section 6.1;
ii) simultaneous pairwise clustering and embedding is performed on artificial and real-world data in Section 6.2;
iii) pairwise clustering as a segmentation technique for textured images is discussed in Section 6.3.

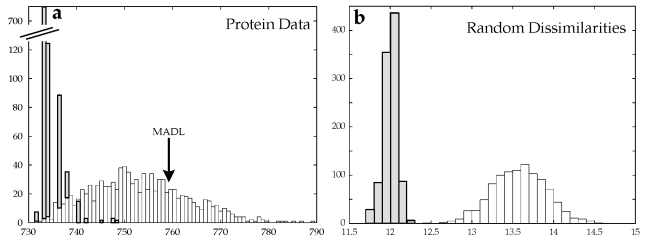


Fig. 1. Histograms of clustering costs for a protein data set (a) and a random clustering instance (b). The gray and white bins denote the results of optimization with deterministic annealing and gradient descent, respectively. The **M**ean **A**verage **D**issimilarity **L**inkage solution has costs of $\mathcal{H}^{pc} = 759.1$.

6.1 Benchmark Experiments for Deterministic Annealing

The theoretical derivations of the deterministic annealing Algorithms I and II are motivated by the known robustness properties of maximum entropy inference. To test this claim, a large number of randomly initialized clustering experiments has been performed on (i) dissimilarities taken from protein sequences and on (ii) dissimilarities which were randomly drawn from a uniform distribution on [0, 1.0]. The dissimilarity values between pairs of protein sequences are determined by a sequence alignment program which takes biochemical and structural information into account. In essence, the alignment program measures the number of amino acids which have to be exchanged to transform the first sequence into the second. The sequences belong to different protein families like hemoglobin, myoglobin and other globins. The protein dissimilarities, sorted according to a clustering solution with $N = 226$, $K = 9$ clusters, are displayed in Fig. 3. The two cases, dissimilarities from protein sequence comparisons and random dissimilarities, span the spectrum between ordered and random clustering instances. The benchmark clustering experiments are designed to validate the claim that superior clustering results are achieved by deterministic annealing compared to standard clustering techniques based on gradient descent. The histograms of 1,000 clustering runs with different initializations are summarized in Fig. 1 for (a) the protein dissimilarity data ($K = 9$) and for (b) the random data ($N = 100$, $K = 10$). Deterministic annealing clearly outperformed the conventional gradient descent method in the random case with even the worst deterministic annealing solution being better than the best gradient descent solution. In the case of the protein data the average costs of a deterministic annealing solution is in the best one percent of the gradient descent solutions, e.g., an average deterministic annealing solution is better than the best out of 100 gradient descent solutions. The standard **M**ean **A**verage **D**issimilarity **L**inkage algorithms (MADL), also known as Ward's method (see [4], Sec. 3.2.7), yields a clustering result with costs $\mathcal{H}^{pc} = 759.1$ compared to the best (experimentally achieved) result with $\mathcal{H}^{pc} = 730.9$. All experiments support our claim that deterministic annealing yields substantially better solutions for comparable computing time.

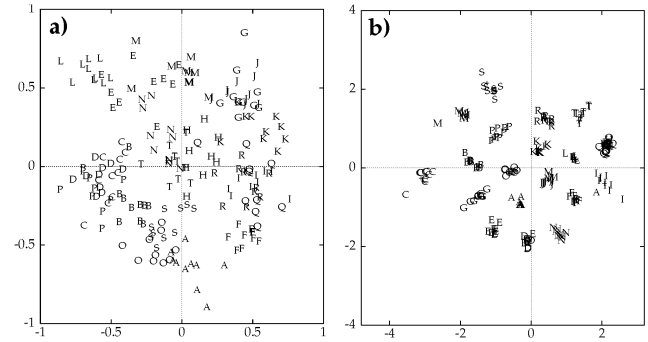


Fig. 2. Embedding of 20-dimensional data into two dimension: (a) projection of the data onto the first two principle components; (b) cluster preserving embedding with Algorithm II. Only 10% of the data are shown.

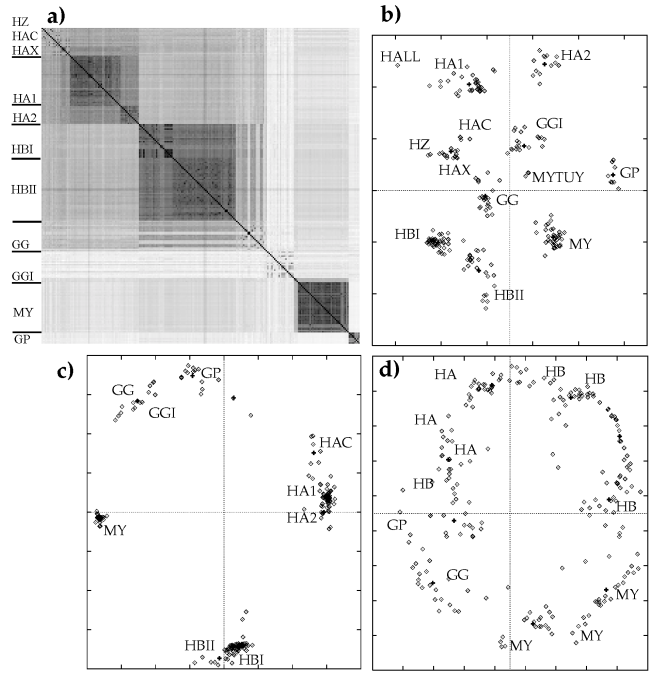


Fig. 3. Similarity matrix of 226 protein sequences of the globin family (a): dark gray levels correspond to high similarity values. Clustering with embedding in two dimensions (b); clustering of MDS embeddings found by global (c) or local (d) stress minimization.

6.2 Clustering and Embedding Results

The properties of the described algorithm for simultaneous Euclidian embedding and data clustering are illustrated by two different experiments:

1) Clustering and dimension reduction of inhomogeneously distributed data.
2) Clustering of real-world proximity data from protein sequences.

The capacity for finding low dimensional representations for high dimensional data is demonstrated with a data set drawn from a mixture of 20 Gaussians in 20 dimensions. The centers of the Gaussians are randomly distributed on the unit sphere. The covariance matrices are diagonal with

values being randomly drawn from the set {0.1, 0.2, 0.4}. The best linear projection according to principal component analysis is shown in Fig. 2a. The positions of the data points are denoted by the letters which name the respective mixture component. Other linear projection methods like projection pursuit [38] yield comparable results since no direction is distinguished in the data generation procedure. Simultaneous clustering and embedding by Algorithm II distributes the data in two dimension with approximately the same group structure as in the high dimensional space. All but cluster (A) are preserved and well separated. A few data points are assigned to the wrong cluster. The algorithm selects a representation in a completely unsupervised fashion and preserves the "essential" structure present in the grouping formation and in the topology of the data set. This procedure for dimension reduction is weakly related to the idea of principal curves [39] or principal surfaces.

Fig. 3 summarizes the clustering result (K = 9) for a real–world data set of 226 protein sequences. Families of protein sequences are abbreviated by capital letters. The gray level visualization of the dissimilarity matrix with dark values for similar protein sequences shows the formation of distinct "squares" along the main diagonal. These squares correspond to the discovered partition after clustering, the resulting clustering costs being $\mathcal{H}^{pc} = 735.2$. The embedding in two dimensions (Fig. 3b) shows intercluster distances which are in good agreement with the similarity values of the data. The best experimentally determined solution ($\mathcal{H}^{pc} = 730.9$) without the embedding constraint exceeded the quality of the solution in Fig. 3b only by 0.83 percent. The results are consistent with the biological classification. The labels HALL and MYTUY in Fig. 3b are individual globin sequences which are known to play an "outlier role" in the globin family. The corrections by the cavity fields (29) are in the range of 10 to 20 percent of the assignment costs $\mathcal{E}_{i\nu}$ (18.0% for a Monte Carlo simulation at $1/T = 2.5$). We have compared the clustering solutions of Algorithm II with results of a two step procedure, i.e., first to embed the data using Kruskal's multidimensional scaling criterion (30) and then to cluster the embedded data by the EM procedure of Algorithm I. Depending on the embedding criterion as absolute (Fig. 3c) or relative (Fig. 3d) stress, the visualizations of the protein dissimilarities reveal little to almost no cluster structure. This fact is reflected in high clustering costs $\mathcal{H}^{pc} = 782.7$, (833.7) for the embedding guided by absolute and relative stress, respectively. It is obvious from Figs. 3b-3d that simultaneous clustering and embedding by the structure preserving MDS algorithm preserves the characteristics of the original cluster structure much better than the classical MDS techniques with subsequent central clustering.

An application of pairwise clustering to a linguistic data set is shown in Fig. 4, in which 825 word fragments have been compared by a dynamic programming algorithm [40]. The dissimilarity matrix is visualized on the left side. Dark gray values denote high similarity values. The matrix is ordered according to the determined clustering solution with eleven clusters (K = 11). Word fragments with similar beginning or ending have a high likelihood to be grouped together, as can be seen from the labels in Fig. 4b. The corrections by the cavity fields are again in the ten percent range (7.7% for $1/T = 3.0$).

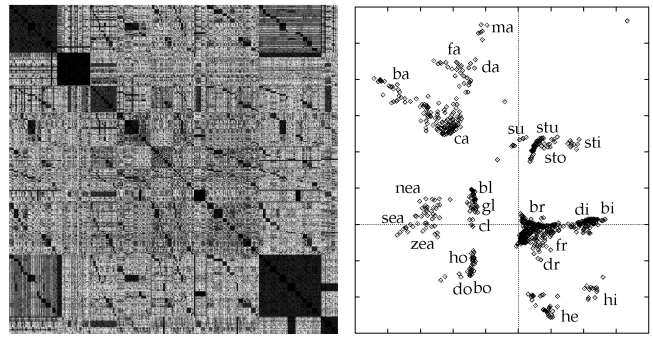


Fig. 4. Similarity matrix for a data set with 825 word fragments (a). The calculated clustering solution with embedding in two dimensions (b). The labels denote the groups by common word beginnings.

6.3 Unsupervised Texture Segmentation by Pairwise Clustering

Segmenting a digital image into homogenous regions, e.g. regions of constant or slowly varying intensity, constant color or uniform texture, arises as a fundamental problem in image processing. Following Geman et al. [41] we formulate texture segmentation as a grouping problem with constraints about valid region shapes. The grouping problem is based on pairwise dissimilarities between texture patches which correspond to pixel blocks of the image. Three major modifications compared to [41] have been introduced:

1) Dissimilarities are calculated based on a Gabor wavelet scale–space representation.
2) The normalized pairwise clustering cost function (16) is used as an objective function for image segmentation.
3) The presented deterministic annealing algorithm replaces the Monte Carlo method proposed in [41].

The calculation of dissimilarity matrices from textured images can be separated into three stages. In the first stage, the image I is transformed in a Gabor wavelet representation. The Gabor transformation possesses a bandpass characteristic and is known to display good texture discrimination properties [42], [43]. We have used four orientations at three different scales, separated by a full octave, resulting in L = 12 feature Images $I^{(l)}$, $1 \le l \le L$ and the raw gray scale image $I^{(0)} = I$. In a second step, the empirical feature distribution function $F_i^{(l)}$ is calculated separately for every feature image $I^{(l)}$ and every image block B_i, $1 \le i \le N$. The blocks B_i are centered on a regular grid and can overlap with each other. In the third stage, pairs of empirical distribution functions belonging to the same feature image are compared using the Kolmogorov–Smirnov distance. For a pair of blocks (B_i, B_j) the latter is defined by

$$D_{ik}^{(l)} \equiv D^{(l)}\left(F_i^{(l)}, F_k^{(l)}\right) := \max_x \left| F_i^{(l)}(x) - F_k^{(l)}(x) \right| \in [0; 1]. \qquad (36)$$

Following the three stage procedure, a set of $L + 1$ independently calculated dissimilarity matrices has been generated, which are combined with a simple maximum rule

$D_{ik} = \max_{0 \le l \le L} D_{ik}^{(l)}$. This is reminiscent of Julesz' theory of texture perception [44], conjecturing that a dissimilarity in a single feature channel is sufficient to discriminate textures. The procedure for generating dissimilarity data from images is schematically summarized in Fig. 5.

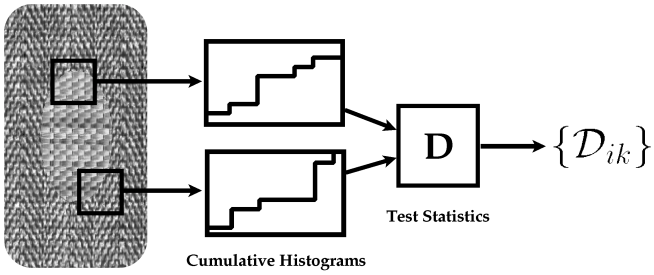


Fig. 5. Texture segmentation by pairwise clustering: local properties of image patches, e.g., intensity differences and local frequencies, are extracted. The respective empirical feature distribution functions are compared with the Kolmogorov-Smirnov statistics to yield dissimilarity values between image blocks.

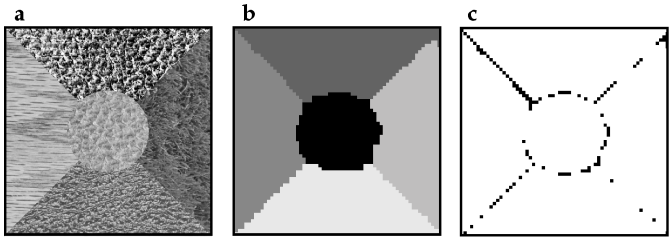


Fig. 6. An image of size 512×512 and with five different textures (a) is segmented by pairwise data clustering. The segmentation result by deterministic pairwise clustering (25) is shown in (b). Segmentation errors displayed by black pixels in (c) are located at segment boundaries.

We have applied the algorithm to over a hundred randomly composed texture images, one being depicted in Fig. 6a. The resulting segmentation based on a deterministic annealing algorithm for pairwise clustering (Fig. 6b) shows, that the five different textures are well discriminated. The difference image (Fig. 6c) between the resulting segmentation and the ground truth demonstrates that incorrect assignments are only observed in the border regions, where statistics belonging to different textures are mixed together. Corrections by the cavity field terms range around six to eight percent changes in the assignments. The segmentation was postprocessed with additional penalties for thin regions as suggested in [41] to enforce local texture consistency and to prevent a too large fragmentation of the texture regions. Moreover only a small fraction (<1%) of dissimilarities calculated from 64×64 blocks was actually processed, including all pairs of adjoined blocks and a small random neighborhood. More details on the neighborhood selection, the adapted mean-field approach to sparse clustering and performance statistics for a large number of textured images can be found in [45] and [32].

7 Discussion

The problem of grouping data can be regarded as one of the initial, although fundamental, steps of information processing and data analysis. Concepts in artificial intelligence as well as in pattern recognition and signal processing are dependent on robust and reliable data clustering principles, with robustness being mandatory with respect to unobservable, as well as to noisy events. In this paper, we have developed a maximum entropy framework for pairwise data clustering. A well-known approximation scheme from statistical physics—the mean-field approximation—has been derived in two different ways:

1) a variational method minimizes the Kullback–Leibler divergence between the original Gibbs distribution for data assignments and a parametrized family of factorized distributions;
2) the expectation values of the data assignments are calculated in a direct fashion.

This technique allows us to correct the influence of small fluctuations in data assignments. The variational approximation of pairwise clustering with central clustering yields a structure preserving, multidimensional scaling algorithm which simultaneously clusters data and embeds them in a Euclidian space. This algorithm can be used for non-linear dimension reduction and for visualization purposes. Results of the pairwise data clustering algorithms in analyzing protein and linguistic data and in segmenting textured images have been reported. Benchmark clustering experiments support our claim that deterministic annealing yields substantially better results than conventional clustering concepts based on gradient descent minimization. The outlined strategy for analyzing stochastic algorithms for pairwise clustering should be considered as a general program for deriving robust optimization algorithms which are based on the maximum entropy principle; analogous results for the metric multidimensional scaling problem and for the hierarchical data clustering problem [46] will be reported elsewhere.

Acknowledgments

It is a pleasure to thank M. Vingron and D. Bavelier for providing the protein data and the linguistic data, respectively. We thank J. Puzicha for the segmentation experiments and H.-J. Klock for the MDS experiments. This work was supported by the Federal Ministry of Education and Science BMBF.

Appendix A

In this appendix, we derive the mean-field equations for the pairwise data clustering problem by minimizing the upper bound on the free energy given in (22) with respect to the variational parameters $\mathcal{E}_{i\alpha}$. Taking derivatives of the upper bound on the free energy yields

$$\frac{\partial}{\partial \mathcal{E}_{i\alpha}} \left(\mathcal{F}\left(\mathcal{H}^0\right) + \left\langle \mathcal{H}^{\mathrm{pc}} - \mathcal{H}^0 \right\rangle \right) = \frac{\partial \left\langle \mathcal{H}^{\mathrm{pc}} \right\rangle}{\partial \mathcal{E}_{i\alpha}} - \sum_{\nu=1}^{K} \frac{\partial \left\langle M_{i\nu} \right\rangle}{\partial \mathcal{E}_{i\alpha}} \mathcal{E}_{i\nu} \quad (37)$$

with

$$\frac{\partial\left\langle \mathcal{H}^{\mathrm{pc}}\right\rangle}{\partial\mathcal{E}_{i\alpha}} = \frac{1}{2}\sum_{v=1}^{K}\sum_{j=1}^{N}\sum_{k=1}^{N}\frac{\partial}{\partial\mathcal{E}_{i\alpha}}\left\langle\frac{M_{jv}M_{kv}}{Np_v}\right\rangle\mathcal{D}_{jk}$$

$$= \sum_{v=1}^{K}\frac{\partial\left\langle M_{iv}\right\rangle}{\partial\mathcal{E}_{i\alpha}}\left[\sum_{\substack{k=1\\k\neq i}}^{N}\left\langle\frac{M_{kv}}{\sum_{j\neq i}^{N}M_{jv}+1}\right\rangle\mathcal{D}_{ik} + \left\langle\frac{1}{\sum_{j\neq i}^{N}M_{jv}+1}\right\rangle\frac{\mathcal{D}_{ii}}{2}\right]$$

$$-\frac{1}{2}\sum_{v=1}^{K}\frac{\partial\left\langle M_{iv}\right\rangle}{\partial\mathcal{E}_{i\alpha}}\sum_{j\neq i}^{N}\sum_{k\neq i}^{N}\left\langle\frac{M_{jv}M_{kv}}{\left(\sum_{\substack{l=1\\l\neq i}}^{N}M_{lv}\right)\left(\sum_{\substack{l=1\\l\neq i}}^{N}M_{lv}+1\right)}\right\rangle\mathcal{D}_{jk}, \tag{38}$$

where we have used the identities

$$\frac{M_{iv}}{Np_v} = \frac{M_{iv}}{\sum_{j\neq i}^{N}M_{jv}+1}, \tag{39}$$

$$\frac{1}{Np_v} = \frac{1}{\sum_{j\neq i}^{N}M_{jv}+M_{iv}} = \frac{1}{\sum_{j\neq i}^{N}M_{jv}} - \frac{M_{iv}}{\left(\sum_{j\neq i}^{N}M_{jv}\right)\left(\sum_{j\neq i}^{N}M_{jv}+1\right)}. \tag{40}$$

Inserting the derivatives gives a necessary condition for a minimum of the upper bound on the free energy,

$$\sum_{v=1}^{K}\frac{\partial\left\langle M_{iv}\right\rangle}{\partial\mathcal{E}_{i\alpha}}\left[\mathcal{E}_{iv} - \left\langle\tilde{\mathcal{E}}_{iv}\right\rangle\right] = -\frac{1}{T}\left\langle M_{i\alpha}\right\rangle\left[\left(\mathcal{E}_{i\alpha} - \left\langle\tilde{\mathcal{E}}_{i\alpha}\right\rangle\right) - \sum_{v=1}^{K}\left\langle M_{iv}\right\rangle\left(\mathcal{E}_{iv} - \left\langle\tilde{\mathcal{E}}_{iv}\right\rangle\right)\right] = 0, \tag{41}$$

with

$$\tilde{\mathcal{E}}_{iv} = \frac{1}{\sum_{j\neq i}^{N}M_{jv}+1}\left[\frac{1}{2}\mathcal{D}_{ii} + \sum_{k\neq i}^{N}M_{kv}\left(\mathcal{D}_{ik} - \frac{1}{2}\sum_{j\neq i}^{N}\frac{M_{jv}}{\sum_{l\neq i}^{N}M_{lv}}\mathcal{D}_{jk}\right)\right]. \tag{42}$$

The K equations (41) are only fulfilled for all values $\alpha = 1, \ldots, K$ simultaneously if

$$\mathcal{E}_{iv} \equiv \left\langle\tilde{\mathcal{E}}_{iv}\right\rangle + c_i, \qquad \forall v = 1, \ldots, K. \tag{43}$$

c_i being N arbitrary constants.

APPENDIX B

In this appendix, we derive the mean field equations of data assignments and fluctuation corrections in the case of strongly disordered clustering problems. The dissimilarities scale as $D_{ik} \sim \mathcal{O}(\sqrt{N})$. This alternative derivation is necessary since the variational approach of Section 4.1 does not capture these fluctuations adequately, as is known from statistical physics [33]. In the following, data assignments are considered to be randomly drawn from the set of admissible configurations $\mathcal{M}$ according to the Gibbs distribution $\mathbf{P}^{\mathrm{Gb}}(\mathcal{H}^{\mathrm{pc}}(\mathbf{M}))$ see (18). Therefore, the expected assignment of datum i to cluster α is

$$\left\langle M_{i\alpha}\right\rangle = \frac{\sum_{\mathbf{M}\in\mathcal{M}}M_{i\alpha}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right)}{\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right)} \tag{44}$$

$$= \sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right)\frac{\sum_{\{\vec{M}_i\}}M_{i\alpha}\exp\left(-\mathcal{H}^{\mathrm{pc}}\left(\vec{M}_i,\hat{\mathbf{M}}\right)\,/\,T\right)}{\sum_{\{\vec{M}_i\}}\exp\left(-\mathcal{H}^{\mathrm{pc}}\left(\vec{M}_i,\hat{\mathbf{M}}\right)\,/\,T\right)}\Bigg/\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right) \tag{45}$$

where $\hat{\mathbf{M}}$ denotes the set of assignments without $\vec{M}_i$. The partial summation over the admissible states $\left\{\vec{M}_i\right\} = \left\{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, \ldots, 0, 1)\right\}$ can be carried out analytically. The first step is the separation of the clustering costs $\mathcal{H}^{\mathrm{pc}}$ in a term $\mathcal{H}_{\dot{+}}$ without any contribution from $\vec{M}_i$ and costs which are related to $\vec{M}_{i\cdot}$. $\mathcal{H}_{\dot{+}}$ is given by

$$\mathcal{H}_{\dot{+}} \equiv \frac{1}{2}\sum_{\substack{j=1\\j\neq i}}^{N}\sum_{\substack{k=1\\k\neq i}}^{N}\mathcal{D}_{jk}\sum_{v=1}^{K}\frac{M_{jv}M_{kv}}{\sum_{\substack{l=1\\l\neq i}}^{N}M_{lv}}. \tag{46}$$

The summation over the admissible states $\left\{\vec{M}_i\right\}$ yields

$$\frac{\sum_{\{\vec{M}_i\}}M_{i\alpha}\exp\left(-\mathcal{H}^{\mathrm{pc}}\left(\vec{M}_i,\hat{\mathbf{M}}\right)\,/\,T\right)}{\sum_{\{\vec{M}_i\}}\exp\left(-\mathcal{H}^{\mathrm{pc}}\left(\vec{M}_i,\hat{\mathbf{M}}\right)\,/\,T\right)} = \frac{\sum_{\{\vec{M}_i\}}M_{i\alpha}\exp\left(-\left(\mathcal{H}_{\dot{-}i} + \sum_{v=1}^{K}M_{iv}\tilde{\mathcal{E}}_{iv}\right)\,/\,T\right)}{\sum_{\{\vec{M}_i\}}\exp\left(-\left(\mathcal{H}_{\dot{-}i} + \sum_{v=1}^{K}M_{iv}\tilde{\mathcal{E}}_{iv}\right)\,/\,T\right)} = \frac{\exp\left(-\tilde{\mathcal{E}}_{i\alpha}\,/\,T\right)}{\sum_{v=1}^{K}\exp\left(-\tilde{\mathcal{E}}_{iv}\,/\,T\right)}; \tag{47}$$

$\tilde{\mathcal{E}}_{i\alpha}$ has been defined in (42). In summary, the expected assignments are

$$\left\langle M_{i\alpha}\right\rangle = \frac{\sum_{\mathbf{M}\in\mathcal{M}}M_{i\alpha}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right)}{\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\mathrm{pc}}(\mathbf{M})\,/\,T\right)} = \left\langle\frac{\exp\left(-\tilde{\mathcal{E}}_{i\alpha}\,/\,T\right)}{\sum_{v}\exp\left(-\tilde{\mathcal{E}}_{iv}\,/\,T\right)}\right\rangle. \tag{48}$$

Equation (48), known as the Markov blanket identity, is analogous to the Callen equation for Ising spins in the theory of magnetic systems (see [47], Section 3.2).

A Taylor expansion of (48) in small fluctuations $\Delta\tilde{\mathcal{E}}_{iv} = \tilde{\mathcal{E}}_{iv} - \left\langle\tilde{\mathcal{E}}_{iv}\right\rangle$ renders a closed system of equations exclusively depending on the averaged assignments $\langle M_{iv}\rangle$. The expected assignments are

$$\langle M_{i\alpha}\rangle = \left\langle f_\alpha\left(\left\{\tilde{\mathcal{E}}_{iv}\right\}\right)\right\rangle = \left\langle f_\alpha\left(\left\{\langle\tilde{\mathcal{E}}_{iv}\rangle + \Delta\tilde{\mathcal{E}}_{iv}\right\}\right)\right\rangle$$

$$= f_\alpha\left(\left\{\langle\tilde{\mathcal{E}}_{iv}\rangle\right\}\right) + \frac{1}{2}\sum_{v=1}^{K}\sum_{\mu=1}^{K}\frac{\partial^2 f_\alpha}{\partial\tilde{\mathcal{E}}_{iv}\partial\tilde{\mathcal{E}}_{i\mu}}\langle\Delta\tilde{\mathcal{E}}_{iv}\Delta\tilde{\mathcal{E}}_{i\mu}\rangle + o\left(\langle\Delta\tilde{\mathcal{E}}_{iv}^3\rangle\right), \quad (49)$$

with $f_\alpha\left(\left\{\tilde{\mathcal{E}}_{iv}\right\}\right) = \exp\left(-\tilde{\mathcal{E}}_{i\alpha} / T\right) / \sum_v \exp\left(-\tilde{\mathcal{E}}_{iv} / T\right)$. Neglecting the second order terms of the expansion we receive a closed system of N × K transcendental equations for the expected assignments

$$\langle M_{i\alpha}\rangle = \frac{\exp\left(-\langle\tilde{\mathcal{E}}_{i\alpha}\rangle / T\right)}{\sum_v \exp\left(-\langle\tilde{\mathcal{E}}_{iv}\rangle / T\right)}. \quad (50)$$

The derivation of (50) tacitly assumes that the assignment correlation function scales as

$$\langle M_{kv}M_{l\mu}\rangle - \langle M_{kv}\rangle\langle M_{l\mu}\rangle = \begin{cases}\langle M_{kv}\rangle\left(\delta_{v\mu} - \langle M_{k\mu}\rangle\right) & \text{for } k = l\\ o\left(1/\sqrt{N}\right) & \text{for } k \neq l\end{cases} \quad (51)$$

Fluctuations of the data assignments for consistent dissimilarities are averaged in the limit $N \to \infty$, in view of the central limit theorem. In the case of random dissimilarities these fluctuations do not vanish for large N and they are captured by the quadratic terms in the Taylor expansion. We introduce an effective internal field $\tilde{h}_{iv}$ which simulates the indirect influence of the disorder on the data assignments (see Thouless, Anderson and Palmer [35]). Without loss of generality the dissimilarity values are assumed to have vanishing expected values, i.e., $\langle\mathcal{D}_{ik}\rangle \equiv 0$. The scaling of the dissimilarities is assumed to be $\mathcal{D}_{ik} \sim o\left(\sqrt{N}\right)$. This shift of the dissimilarity values and their random nature allows us to neglect the second term in (26) since

$$\sum_{i=1}^{N}\sum_{k=1}^{N} M_{iv}M_{kv}\mathcal{D}_{ik} / \left(2p_v^2N^2\right) \sim N\sqrt{N} / N^2 = 1/\sqrt{N}.$$

The Ansatz for the expected assignments with effective internal field is

$$\langle M_{i\alpha}\rangle = f_\alpha\left(\left\{\langle\tilde{\mathcal{E}}_{iv}\rangle - \tilde{h}_{iv}\right\}\right)$$
$$= f_\alpha\left(\left\{\langle\tilde{\mathcal{E}}_{iv}\rangle\right\}\right) - \sum_{v=1}^{K}\frac{\partial f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{iv}\rangle}\tilde{h}_{iv} + o\left(\max_{i,v}\tilde{h}_{iv}^2\right). \quad (52)$$

The term linear in $\tilde{h}_{iv}$ (52) has to capture all fluctuation contributions of (49). A comparison of the coefficients yields

$$-\sum_{v=1}^{K}\frac{\partial f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{iv}\rangle}\tilde{h}_{iv} = \frac{1}{2}\sum_{v=1}^{K}\sum_{\mu=1}^{K}\frac{\partial^2 f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{iv}\rangle\partial\langle\tilde{\mathcal{E}}_{i\mu}\rangle}\langle\Delta\tilde{\mathcal{E}}_{iv}\Delta\tilde{\mathcal{E}}_{i\mu}\rangle. \quad (53)$$

Inserting the partial derivatives and dividing by $\langle M_{i\alpha}\rangle / T$ yields

$$\sum_{v=1}^{K}\left(\delta_{\alpha v} - \langle M_{iv}\rangle\right)\tilde{h}_{iv} =$$
$$\frac{1}{2T}\sum_{v=1}^{K}\sum_{\mu=1}^{K}\Gamma_{iv\mu}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\mathcal{D}_{ik}\mathcal{D}_{il}}{p_v p_\mu N^2}\left(\langle M_{kv}M_{l\mu}\rangle - \langle M_{kv}\rangle\langle M_{k\mu}\rangle\right)$$
$$= \frac{1}{2T}\sum_{v=1}^{K}\left(\delta_{\alpha v} - \langle M_{iv}\rangle\right)$$
$$\sum_{k=1}^{N}\sum_{\mu=1}^{K}\mathcal{D}_{ik}^2\frac{\langle M_{kv}\rangle\left(\delta_{v\mu} - \langle M_{k\mu}\rangle\right)}{p_v p_\mu N^2}\left(\delta_{v\mu} - 2\langle M_{i\mu}\rangle\right)$$
$$+ \text{ terms with } k \neq l \quad (54)$$

$$\Gamma_{iv\mu} = \left(\delta_{\alpha v} - \langle M_{iv}\rangle\right)\left(\delta_{\alpha\mu} - \langle M_{i\mu}\rangle\right) - \langle M_{i\mu}\rangle\left(\delta_{v\mu} - \langle M_{iv}\rangle\right). \quad (55)$$

If assumption (51) holds, the terms $k \neq l$ in (54) vanish as $\sqrt{N^2} / \left(\sqrt{N^3}\right)$. An in depth discussion when the assumption (51) is valid can be found in [33]. Assuming the validity of (51), the refined mean–field equations are

$$\langle M_{i\alpha}\rangle = \frac{\exp\left(-\left(\langle\tilde{\mathcal{E}}_{i\alpha}\rangle - \tilde{h}_{i\alpha}\right) / T\right)}{\sum_v \exp\left(-\left(\langle\tilde{\mathcal{E}}_{iv}\rangle - \tilde{h}_{iv}\right) / T\right)}, \quad (56)$$

$$\tilde{h}_{i\alpha} = \frac{1}{2T}\sum_{k=1}^{N}\mathcal{D}_{ik}^2\sum_{\mu=1}^{K}\frac{\langle M_{k\alpha}\rangle\left(\delta_{\alpha\mu} - \langle M_{k\mu}\rangle\right)}{p_\alpha p_\mu N^2}\left(\delta_{\alpha\mu} - 2\langle M_{i\mu}\rangle\right). \quad (57)$$

APPENDIX C

The chain rule yields the derivatives of the upper bound (22) with respect to the variational parameters $\mathbf{x}_i$:

$$\frac{\partial}{\partial\mathbf{x}_i}\left[\mathcal{F}\left(\mathcal{H}^{cc}\right) + \left\langle\mathcal{H}^{pc} - \mathcal{H}^{cc}\right\rangle\right] =$$
$$\sum_{k=1}^{N}\sum_{\mu=1}^{K}\frac{\partial}{\partial\mathcal{E}_{k\mu}}\left[\mathcal{F}\left(\mathcal{H}^{cc}\right) + \left\langle\mathcal{H}^{pc} - \mathcal{H}^{cc}\right\rangle\right]\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i}$$
$$= -\frac{1}{T}\sum_{k=1}^{N}\sum_{\mu=1}^{K}\sum_{v=1}^{K}\langle M_{kv}\rangle\left(\delta_{v\mu} - \langle M_{k\mu}\rangle\right)\Delta\mathcal{E}_{kv}\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i}, \quad (58)$$

where $\Delta\mathcal{E}_{k\alpha} = \mathcal{E}_{k\alpha} - \mathcal{E}^*_{k\alpha}$ and $\mathcal{E}_{k\alpha} = \|\mathbf{x}_k - \mathbf{y}_\alpha\|^2$. The derivatives (58) are given by

$$\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i} =$$
$$2\left(\mathbf{x}_k - \mathbf{y}_\mu\right)^T\left[\delta_{ik} - \frac{\langle M_{i\mu}\rangle}{Np_\mu} - \frac{1}{Np_\mu}\sum_{l=1}^{N}\left(\mathbf{x}_l - \mathbf{y}_\mu\right)\frac{\partial\langle M_{l\mu}\rangle}{\partial\mathbf{x}_i}\right]. \quad (59)$$

Setting (58) equal to zero, results in the exact stationary conditions

$$\sum_{v,\mu=1}^{K}\langle M_{iv}\rangle\langle M_{i\mu}\rangle\left(\Delta\mathcal{E}_{i\mu} - \Delta\mathcal{E}_{iv}\right)\mathbf{y}_\mu =$$
$$\sum_{k=1}^{N}\sum_{v,\mu=1}^{K}\frac{\langle M_{iv}\rangle\langle M_{k\mu}\rangle}{Np_\mu}\left(\Delta\mathcal{E}_{i\mu} - \Delta\mathcal{E}_{iv}\right)\left[\langle M_{i\mu}\rangle + \sum_{l=1}^{N}\left(\mathbf{x}_l - \mathbf{y}_\mu\right)\frac{\partial\langle M_{l\mu}\rangle}{\partial\mathbf{x}_i}\right]^T\left(\mathbf{x}_k - \mathbf{y}_\mu\right). \quad (60)$$

The left-hand side can be further reduced to an expression explicit in $\mathbf{x}_i$:

$$\sum_{\nu,\mu=1}^{K} \langle M_{i\nu}\rangle\langle M_{i\mu}\rangle\left(\Delta\mathcal{E}_{i\mu} - \Delta\mathcal{E}_{i\nu}\right)\mathbf{y}_\mu =$$

$$\sum_{\nu=1}^{K} \langle M_{i\nu}\rangle\Delta\mathcal{E}_{i\nu}\left(\mathbf{y}_\nu - \sum_{\mu=1}^{K}\langle M_{i\mu}\rangle\mathbf{y}_\mu\right)$$

$$= -2\mathbf{K}_i\mathbf{x}_i + \sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\left(\|\mathbf{y}_\nu\|^2 - \mathcal{E}^*_{i\nu}\right)\left(\mathbf{y}_\nu - \sum_{\alpha=1}^{K}\langle M_{i\alpha}\rangle\mathbf{y}_\alpha\right), \quad (61)$$

where $\mathbf{K}_i = \left(\left\langle \mathbf{y}\mathbf{y}^T\right\rangle_i - \langle\mathbf{y}\rangle_i\langle\mathbf{y}\rangle_i^T\right)$ is a d × d covariance matrix, $\langle\mathbf{y}\rangle_i = \sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\mathbf{y}_\nu$. Note that there still exist implicit dependencies, since $\mathbf{y}_\nu$ depends on $\mathbf{x}_i$.

The derivatives $\partial\langle M_{k\alpha}\rangle/\partial\mathbf{x}_i$ on the right hand side of (60) can be exactly calculated, since they are given as the solutions of a linear equation system with $N \times K$ unknowns for every $\mathbf{x}_i$. However, to reduce the computational complexity we perform an approximation under the assumption of $\partial\mathbf{y}_\mu/\partial\mathbf{x}_i \approx 0$, treating $\mathbf{y}_\mu$ as an independent variable. Equation (61) simplifies to a vector equation for every $\mathbf{x}_i$:

$$\mathbf{K}_i\mathbf{x}_i \approx \frac{1}{2}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\left(\|\mathbf{y}_\nu\|^2 - \mathcal{E}^*_{i\nu}\right)\left(\mathbf{y}_\nu - \sum_{\mu=1}^{K}\langle M_{i\mu}\rangle\mathbf{y}_\mu\right). \quad (62)$$

REFERENCES

[1] E.T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review*, vol. 106, pp. 620-630, 1957.

[2] E.T. Jaynes, "Information Theory and Statistical Mechanics II," *Physical Review*, vol. 108, pp. 171-190, 1957.

[3] E.T. Jaynes, "On the Rationale of Maximum-Entropy Methods," *Proc IEEE*, vol. 70, pp. 939-952, 1982.

[4] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[5] G.J. McLachlan and K.E. Basford, *Mixture Models*. New York, Basel: Marcel Dekker, Inc., 1988.

[6] R.M. Gray, "Vector Quantization," *IEEE Acoustics, Speech and Signal Processing*, pp. 4-29, Apr. 1984.

[7] A. Gersho and R.M. Gray, *Vector Quantization and Signal Processing*. Boston: Kluwer Academic Publisher, 1992.

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

[9] P. Simic, "Statistical Mechanics as the Underlying Theory of "Elastic" and "Neural" Optimizations," *Network*, vol. 1, pp. 89-103, 1990.

[10] P. Simic, "Constrained Nets for Graph Matching and Other Quadratic Assignment Problems," *Neural Computation*, vol. 3, pp. 268-281, 1991.

[11] A. Yuille, P. Stolorz, and J. Utans, "Statistical Physics, Mixtures of Distributions and the EM Algorithm," *Neural Computation*, vol. 6, pp. 334-340, 1994.

[12] S. Gold and A. Rangarajan, "A Graduated Assignment Algorithm for Graph Matching," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 377-388, 1996.

[13] D. Geiger and F. Girosi, "Parallel and Deterministic Algorithms from MRFs: Surface Reconstruction," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 13, pp. 401-412, May 1991.

[14] A.L. Yuille, "Generalized Deformable Models, Statistical Physics and Matching Problems," *Neural Computation*, vol. 2, no. 1, pp. 1-24, 1990.

[15] C. Bregler and S. Omohundro, "Surface Learning with Applications to Lipreading," *Advances in Neural Information Processing Systems* vol. 6, J. Cowan, G. Tesauro, and J. Alspector, eds., 1994.

[16] K. Rose, E. Gurewitz, and G. Fox, "Statistical Mechanics and Phase Transitions in Clustering," *Physical Review Letters*, vol. 65, no. 8, pp. 945-948, 1990.

[17] K. Rose, E. Gurewitz, and G. Fox, "A Deterministic Annealing Approach to Clustering," *Pattern Recognition Letters*, vol. 11, no. 11, pp. 589-594, 1990.

[18] K. Rose, E. Gurewitz, and G. Fox, "Vector Quantization by Deterministic Annealing," *IEEE Trans Information Theory*, vol. 38, no. 4, pp. 1249-1257, 1992.

[19] K. Rose, E. Gurewitz, and G. Fox, "Constrained Clustering as an Optimization Method," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 15, no. 8, pp. 785-794, 1993.

[20] J.M. Buhmann and H. Kühnel, "Complexity Optimized Data Clustering by Competitive Neural Networks," *Neural Computation*, vol. 5, pp. 75-88, 1993.

[21] J.M. Buhmann and H. Kühnel, "Vector Quantization with Complexity Costs," *IEEE Trans Information Theory*, vol. 39, pp. 1,133-1,145, July 1993.

[22] P.A. Chou, T. Lookabaugh, and R.M. Gray, "Entropy-Constrained Vector Quantization," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 37, pp. 31-42, 1989.

[23] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, pp. 671-680, 1983.

[24] V. Černy, "Thermodynamical Approach to the Traveling Salesman Problem: an Efficient Simulation Algorithm," *J. Optimization Theory and Applications*, vol. 45, pp. 41-51, 1985.

[25] C.W. Gardiner, *Handbook of Stochastic Methods*. Berlin: Springer, 1983.

[26] Y. Tikochinsky, N. Tishby, and R.D. Levine, "Alternative Approach to Maximum-Entropy Inference," *Physical Review A*, vol. 30, pp. 2638-2644, 1984.

[27] I. Csiszár, "I-Divergence, Geometry of Probability Distributions and Minimization Problems," Annals Of Probability, vol. 3, pp. 146-158, 1975.

[28] T. Kohonen, *Self-Organization and Associative Memory*. Berlin: Springer, 1984.

[29] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps*. New York: Addison Wesley, 1992.

[30] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[31] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Society Ser. B* (methodological), vol. 39, pp. 1-38, 1977.

[32] T. Hofmann, J. Puzicha, and J. M. Buhmann, "Unsupervised Segmentation of Textured Images by Pairwise Data Clustering," Technical Report IAI-TR-96-2, Rheinische Friedrich-Wilhelms-Universität Bonn, Institut für Informatik III, Feb. 1996.

[33] M. Mezard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond*. Singapore: World Scientific, 1987.

[34] R.E. Peierls, "On a Minimum Property of the Free Energy," *Physical Review*, vol. 54, p. 918, 1938.

[35] D.J. Thouless, P.W. Anderson, and R.G. Palmer, "A Solution to a "Solvable" Model of a Spin Glass," *Philosophical Magazine*, vol. 35, p. 593, 1977.

[36] J.W. Sammon Jr, "A Non-Linear Mapping for Data Structure Analysis," *IEEE Trans. Computers*, vol. 18, pp. 401-409, 1969.

[37] J.B. Kruskal, "Nonmetric Multidimensional Scaling: a Numerical Method," *Psychometrika*, vol. 29, pp. 115-129, 1964.

[38] P. Huber, "Projection Pursuit," *Annals of Statistics*, vol. 13, pp. 435-475, 1985.

[39] T. Hastie and W. Stuetzle, "Principal Curves," *J. American Statistical Ass'n*, vol. 84, pp. 502-516, 1989.

[40] D. Bavelier and M. Jordan, "Representing Words in Connectionist Models." *Abstract 34th Ann. Meeting Psychonomics Society*, Washington, DC., 1993.

[41] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary Detection by Constrained Optimization," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 12, pp. 609-628, July 1990.

[42] I. Fogel and D. Sagi, "Gabor Filters as Texture Discriminators," *Biological Cybernetics*, vol. 61, pp. 103-113, 1989.

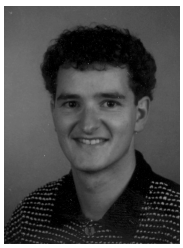
[43] J. Daugman, "Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters," *J. Optical Society of America A*, vol. 2, no. 7, pp. 1,160-1,169, 1985.

[44] B. Julesz, "Visual Pattern Discrimination," *IRE Transactions on Information Theory*, pp. 84-92, Feb. 1961.

[45] T. Hofmann, J. Puzicha, and J.M. Buhmann, "Unsupervised Segmentation of Textured Images by Pairwise Data Clustering," *Proc. Int'l Conf. Image Processing* Lausanne, 1996.

[46] T. Hofmann and J.M. Buhmann, "Infering Hierarchical Clustering Structures by Deterministic Annealing," *Proc. Knowledge Discovery and Data Mining Conf.*, Portland, 1996.
[47] G. Parisi, *Statistical Field Theory*. Redwood City, Calif.: Addison Wesley, 1988.

Joachim M. Buhmann received a PhD degree in theoretical physics from the Technical University of Munich in 1988. He held postdoctoral positions at the University of Southern California and at the Lawrence Livermore National Laboratory. Currently, he works as an associate professor of computer science at the University of Bonn, Germany, where he heads the research group on computer vision and pattern recognition. His current research interests cover the theory of neural networks and their applications to image understanding and signal processing. Special research topics include data clustering and data visualization, active data selection, stochastic optimization techniques, video compression and sensor fusion for autonomous robots.



Thomas Hofmann received the Diplom degree in computer science from the University of Bonn, Germany in 1993. In October 1993, he joined the Computer Vision and Pattern Recognition Group at the University of Bonn, where he is currently completing his PhD thesis on deterministic annealing algorithms for exploratory data analysis. His research interests include computer vision, neural networks, graphical models, machine learning, and autonomous robots.