Pairwise Data Clustering by Deterministic Annealing

Thomas Hofmann & Joachim M. Buhmann

Rheinische Friedrich–Wilhelms–Universität
Institut für Informatik III, Römerstraße 164
D-53117 Bonn, Germany

email: {th,jb}@informatik.uni-bonn.de

March 14, 1996

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 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 analyse dissimilarity data from protein analysis and from linguistics. The algorithm for pairwise data clustering is used to segment textured images.

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].

The term *data clustering* gives rise to different interpretations and expectations depending on the goal in mind [4, 5]. A definition of data clustering which is motivated by statistics is the inference of a probability distribution for a set of data vectors. Parameter estimation in classical statistics and model comparison in Bayesian statistics provide an answer to such data clustering questions. The most frequently used approach are mixture density models, e.g., Gaussian mixture models with the EM algorithm to infer mixture parameters which locally maximize the likelihood function [6]. Alternatively to probability density estimation, data clustering can be viewed as a data partitioning problem. Given a data set we might ask the question how the data can be split into different partitions dependent on a quality criterion. In this paper we adopt the second view.

The problem of optimally partitioning a data set arises in two different forms dependent on the data representation. Let us first consider a set of data vectors $\Xi = \left\{ \mathbf{x}_i \in \mathbb{R}^d : i \in \{1, \ldots, N\} \right\}$. A partitioning approach known as *central clustering* derives a set of reference or prototype vectors $\Upsilon = \left\{ \mathbf{y}_\nu \in \mathbb{R}^d : \nu \in \{1, \ldots, K\} \right\}$ which represent the data set efficiently. This approach is also discussed as vector quantization [7] in the data compression literature, Υ being called a codebook. The set of reference vectors defines a Voronoi tesselation of the data space with each data vector being assigned to its nearest reference vector. Data compression is achieved by transmission and storage of the indices of reference vectors $\{\mathbf{y}_\alpha\}$ rather than the original data vectors $\{\mathbf{x}_i\}$. The distortion which is generated by the mapping of $\mathbf{x}_i \rightarrow \mathbf{y}_\alpha$ has to be minimized by a proper choice of the reference vector set for a given size of the codebook.

The second approach to data clustering is concerned with data sets which are indirectly characterized by pairwise comparisons instead of explicit coordinates. In many applications, e.g., in molecular biology, psychology, linguistics, economics and image processing, a set of N data is represented by pairwise dissimilarity values $\mathbf{D} = (\mathcal{D}_{ik})\Big|_{\substack{i=1,\ldots,N \\ k=1,\ldots,N}} \in \mathbb{R}^{N \times N}$.

The characteristics of the data set are hidden in the numbers $\mathcal{D}_{ik}$ which indicate proximity or dissimilarity of items. These proximity values frequently violate the triangular inequality, their self-dissimilarity does not necessarily vanish and they might be negative. The grouping of proximity data is mathematically formulated as a combinatorial optimization problem which we address with a minimization heuristic called *deterministic annealing*.

Data clustering as a problem in pattern recognition and statistics falls into the class of unsupervised learning problems. There is a large body of literature available on this topic and the reader is refered to the textbooks of Duda & Hart and of Jain & Dubes [8, 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 data clustering has been discussed by Rose et al. in a series of papers [16, 17, 18, 19]. A general solution of the deterministic annealing procedure for vector quantization was suggested in [20, 21]. This work as well as Chou et al. [22] emphasized the complexity issue in the codebook design.

The remainder of the paper is structured in the following way: We discuss the advantage of a maximum entropy based search heuristic in Sect. 2. A discussion of cost functions for central and pairwise data clustering is presented in Sect. 3. Approximation techniques to calculate expectation values for the data assignments are discussed in Sect. 4. The widely used estimation technique called mean–field approximation is derived by variational techniques and, alternatively, by an expansion for small fluctuations. The first derivation is necessary to adapt parameterized models to the data clustering problem, whereas the second derivation yields a systematic way to compute fluctuation corrections. An extension of pairwise data clustering to data visualization is described in Sect. 5. The results of central clustering (Sect. 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 anealing and a conventional, gradient descend technique for clustering as well as an application of pairwise data clustering to image segmentation are summarized in Sect. 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*. In 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.,

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

The parameter T is called computational temperature. The philosophy of simulated annealing is to gradually reduce the temperature during the search process which forces 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]

$$\mathbf{P}^{\text{Gibbs}}(\omega) = \frac{\exp(-\mathcal{H}(\omega)/T)}{\sum_{\bar{\omega}\in\Omega}\exp(-\mathcal{H}(\bar{\omega})/T)} \equiv \exp(-(\mathcal{H}(\omega) - \mathcal{F}(\mathcal{H}))/T) \tag{2}$$

which is known as the Gibbs distribution. The quantity $\mathcal{F}(\mathcal{H}) \equiv -T\log\sum_{\Omega}\exp(-\mathcal{H}(\omega)/T)$ denotes the Gibbs free energy. The reader should realize that the definition of a cost function $\mathcal{H}(\omega)$ implies the Gibbs probability distribution if state transitions take place according to the Metropolis algorithm. The inverse temperature $\beta \equiv 1/T$ formally plays the role of a Lagrange parameter to enforce a constraint on the expected costs

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

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

$$\mathcal{S}(\mathbf{P}^{\text{Gibbs}}) \;=\; -\sum_{\omega\in\Omega}\mathbf{P}^{\text{Gibbs}}(\omega)\log\mathbf{P}^{\text{Gibbs}}(\omega) \;=\; \frac{1}{T}\sum_{\omega\in\Omega}\mathbf{P}^{\text{Gibbs}}(\omega)\mathcal{H}(\omega) - \frac{1}{T}\mathcal{F}(\mathcal{H}). \tag{4}$$

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 Monte Carlo simulation, e.g., the variables of the optimization problem are drawn according to $\mathbf{P}^{\text{Gibbs}}(\omega)$. This sequential sampling of the solution space, however, is slow compared to deterministic optimization techniques. 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

$$\tilde{\mathcal{F}}(\mathbf{P}) \;\equiv\; \langle\mathcal{H}\rangle_{\mathbf{P}} - T\mathcal{S}(\mathbf{P}) \;=\; \sum_{\omega\in\Omega}\mathbf{P}(\omega)\mathcal{H}(\omega) - T\sum_{\omega\in\Omega}\mathbf{P}(\omega)\log\mathbf{P}(\omega) \tag{5}$$

which has to be minimized over a (tractable) subspace of probability distributions. The inequality $\tilde{\mathcal{F}}(\mathbf{P}) \geq \mathcal{F}(\mathbf{P}^{\text{Gibbs}})$ holds since the Gibbs distribution maximizes entropy [1, 2] for $\langle\mathcal{H}\rangle_{\mathbf{P}}$ being 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 $T \to 0$, i.e., high temperature smoothes the cost function and low temperature reveals the full complexity of the original optimization problem. The original optimization problem is recovered for $T \to 0$. Deterministic annealing algorithms track good solutions from high to low temperature similar to cooling in simulated annealing. We will discuss this technique in detail in Sect. 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 $\langle \mathcal{H} \rangle$ (Eq. 3) 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 w.r.t. noise.

3 Cost Functions for Data Clustering

3.1 Central Clustering

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

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

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

$M_{i\nu} \equiv 1$ denotes that the data point $\mathbf{x}_i$ is assigned to reference vector $\mathbf{y}_\nu$, while $M_{i\nu} \equiv 0$ denotes that $\mathbf{x}_i$ is not assigned to $\mathbf{y}_\nu$. The solution space is given as the set of admissible configurations (7). The constraints $\sum_{\nu=1}^{K} M_{i\nu} = 1, \ \forall i$ assure that each data point is represented by a unique reference vector. The quality of the set of reference vectors is assessed by a clustering cost function. The objective function for central clustering sums up the average distortion error $\mathcal{D}(\mathbf{x}_k, \mathbf{y}_\nu)$ between a data vector $\mathbf{x}_i$ and the corresponding reference vector $\mathbf{y}_\nu$, i.e.,

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

An appropriate distortion measure $\mathcal{D}(\mathbf{x}_k, \mathbf{y}_\nu)$ depends on the application domain. The most common choice is the squared Euclidian distance $\mathcal{D}(\mathbf{x}_k, \mathbf{y}_\nu) \equiv \|\mathbf{x}_k - \mathbf{y}_\nu\|^2$ between the data vector and its reference vector. Applications with a topological ordering of the reference vectors

as for source-channel coding favor a distortion measure which considers the topological organization of the reference vectors, e.g., $\mathcal{D}(\mathbf{x}_k,\mathbf{y}_\alpha) \equiv \sum_{\nu=1}^{K} T_{\alpha\nu}\|\mathbf{x}_k-\mathbf{y}_\nu\|^2$, 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 as selforganizing topological maps in the area of Neural Computing [28, 29]. The number of clusters can be limited by additional complexity costs rather than postulating a fixed number K. Well-known cases are constant clustering costs or complexity costs proportional to the Shannon entropy of the cluster set. A detailed discussion of complexity issues in vector quantization can be found in [21].

Stochastic optimization of the cost function (8) requires to determine the probability distribution of assignments $\mathbf{M}$. The maximum entropy principle states that the assignments are distributed according to the Gibbs distribution

$$\begin{aligned} \mathbf{P}(\mathbf{M}) &= \exp\left(-(\mathcal{H}^{\text{cc}}(\mathbf{M})-\mathcal{F}(\mathcal{H}^{\text{cc}}))/T\right), &(9)\\ \mathcal{F}(\mathcal{H}^{\text{cc}}) &= -T\ln\sum_{\mathbf{M}\in\mathcal{M}}\exp\left(-\mathcal{H}^{\text{cc}}(\mathbf{M})/T\right). &(10)\end{aligned}$$

As pointed out in Sect. 2, the free energy $\mathcal{F}(\mathcal{H}^{\text{cc}})$ in Eq. (10) can be interpreted as a smoothed version of the original cost function (8). The factor $\exp(\mathcal{F}(\mathcal{H}^{\text{cc}})/T)$ normalizes the exponential weights $\exp\left(-\mathcal{H}^{\text{cc}}(\mathbf{M})/T\right)$. Its inverse can be rewritten as

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

since the sum over assignments is constrained by $\sum_{\nu=1}^{K} M_{i\nu}=1$. The cost function (8) which is linear in $M_{i\nu}$ yields a factorized Gibbs distribution

$$\mathbf{P}(\mathbf{M}) = \prod_{k=1}^{N}\frac{\exp(-\sum_{\nu=1}^{K}M_{k\nu}\mathcal{D}(\mathbf{x}_k,\mathbf{y}_\nu)/T)}{\sum_{\mu=1}^{K}\exp(-\mathcal{D}(\mathbf{x}_k,\mathbf{y}_\mu)/T)}. \qquad (12)$$

The distribution (12) can be interpreted as the complete data likelihood for mixture models with parameters $\{\mathbf{y}_\nu\}$. This correspondence only holds for cost functions which are linear in the dynamic variables.

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

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

To determine closed equations for the reference vectors $\mathbf{y}_\nu$ we differentiate the argument in Eq. (13) with the expected costs being kept fixed. The resulting equation

$$0 = \sum_{i=1}^{N}\langle M_{i\nu}\rangle\frac{\partial}{\partial\mathbf{y}_\nu}\mathcal{D}(\mathbf{x}_i,\mathbf{y}_\nu) \quad \text{with} \qquad (14)$$

$$\langle M_{i\nu} \rangle \;\equiv\; \frac{\exp(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)/T)}{\sum_\mu \exp(-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\mu)/T)} \tag{15}$$

is known as the centroid equation in signal processing. The angular brackets denote expectation values, i.e., $\langle f(\mathbf{M}) \rangle \equiv \sum_{\mathcal{M}} f(\mathbf{M}) \mathbf{P}(\mathbf{M})$. The reader should realize that entropy maximization implies $\mathbf{y}_\alpha$ to be a centroid. This choice is also optimal in the sense of rate distortion theory [30].

The equations (14,15) are efficiently solved in an iterative fashion using the expectation maximization (EM) algorithm [6].

Algorithm I

```
INITIALIZE y_ν^(0) and ⟨M_iν⟩^(0) ∈ (0,1) randomly;
      temperature T ← T_0;
WHILE T > T_FINAL
   t ← 0;
   REPEAT
      E-step:  estimate ⟨M_iν⟩^(t+1) as a function of y_ν^(t);
      M-step:  calculate y_ν^(t+1) for given ⟨M_iν⟩^(t+1);
      t ← t + 1;
   UNTIL all {⟨M_iν⟩^(t), y_ν^(t)} satisfy Eqs.(14,15)
   T ← T/2;  ⟨M_iν⟩^(0) ← ⟨M_iν⟩^(t);  y_ν^(0) ← y_ν^(t);
```

The EM algorithm alternates an estimation 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. [6] 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. The outer loop of the algorithm reduces the temperature in an exponential fashion, i.e., we halve the temperature T. Other choices, e.g., 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. In Sect. 5 we will use the solutions of central clustering with squared Euclidian distances to simultaneously find a grouping of the data and an embedding in the Euclidian space $\mathbb{R}^d$.

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. 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 [8, 4]. 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) to denote the 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

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

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

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

since $\mathcal{D}_0 \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\sum_{\nu=1}^{K} M_{k\nu}M_{l\nu}/p_\nu - 1\right)/N = 0$. A uniform shift $\mathcal{D}_0$ of the dissimilarity values does not change the clustering costs and, consequently, has no influence on the statistics of the assignments. Another important property of the proposed cost function concerns non–symmetric dissimilarities, $\mathcal{D}_{ik} \neq \mathcal{D}_{ki}$. $\mathcal{H}^{\text{pc}}$ 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 assume symmetric $\mathcal{D}_{ik}$. Furthermore, $\mathcal{H}^{\text{pc}}$ is also invariant under an arbitrary permutation of the cluster indices $\nu \to \pi(\nu)$[1].

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., large $\mathcal{D}_{ik}$ for data i, k in different clusters and small $\mathcal{D}_{ik}$ for data i, k in the same clusters, constant costs per data point require a scaling $\mathcal{D}_{ik} \sim \mathcal{O}(1)$. In the opposite case of random dissimilarity values averaging effects necessitate a scaling $\mathcal{D}_{ik} \sim \mathcal{O}(\sqrt{N})$. A thorough discussion of this point can be found in the statistical physics literature of optimization problems [31].

4 Mean-Field Approximation to Pairwise Clustering

Following the strategy of stochastic optimization as discussed in Sect. 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

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

randomly drawn from the set of admissible configurations (7) according to the Gibbs distribution

$$\mathbf{P}(\mathbf{M}) \;=\; \frac{\exp\left(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T\right)}{\sum_{\mathcal{M}} \exp\left(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T\right)} \;\equiv\; \exp(-(\mathcal{H}^{\text{pc}}(\mathbf{M}) - \mathcal{F}(\mathcal{H}^{\text{pc}}))/T), \tag{18}$$

$\mathcal{H}^{\text{pc}}$ being the costs for a pairwise clustering solution (see Eq. 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 factorial form. Each assignment variable $M_{i\nu}$ interacts with all other assignment variables. These cost contributions, however, average in the limit of large data sets and reduce the influence of correlations on individual data assignments. Therefore, we 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, however, is 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 a Gibbs distribution neglects the correlations between the stochastic variables in the pairwise clustering cost function $\mathcal{H}^{\text{pc}}$. The Gibbs distribution $\mathbf{P}(\mathcal{H}^{\text{pc}})$ corresponding to the clustering cost function with interactions is approximated within an $\mathcal{E}$–parametrized family of distributions $\mathbf{P}^0(\mathcal{E})$. The distribution $\mathbf{P}^0(\mathcal{E}^*)$ which represents most accurately the statistics of the original problem is determined by the minimum of the Kullback–Leibler divergence to the original Gibbs distribution, i.e.,

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

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

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

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

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

$$\begin{aligned} \mathcal{I}\left(\mathbf{P}(\mathcal{H}^0) \middle\| \mathbf{P}(\mathcal{H}^{\text{pc}})\right) \;&=\; \sum_{\mathbf{M}\in\mathcal{M}} \mathbf{P}(\mathbf{M}; \mathcal{H}^0) \log \frac{\exp\left[-\mathcal{H}^0(\mathbf{M})/T\right] \sum_{\bar{\mathbf{M}}\in\mathcal{M}} \exp\left[-\mathcal{H}^{\text{pc}}(\bar{\mathbf{M}})/T\right]}{\exp\left[-\mathcal{H}^{\text{pc}}(\mathbf{M})/T\right] \sum_{\bar{\mathbf{M}}\in\mathcal{M}} \exp\left[-\mathcal{H}^0(\bar{\mathbf{M}})/T\right]} \\ &=\; \frac{1}{T}\left[\mathcal{F}(\mathcal{H}^0) - \mathcal{F}(\mathcal{H}^{\text{pc}}) + \langle \mathcal{H}^{\text{pc}} - \mathcal{H}^0 \rangle\right]. \end{aligned} \tag{21}$$

The averaging brackets $\langle . \rangle$ denote the average with respect to $\mathbf{P}(\mathcal{H}^0)$. Since the KL–divergence is always positive and vanishes only for $\mathbf{P}(\mathcal{H}^0) \equiv \mathbf{P}(\mathcal{H}^{\rm pc})$ we receive the well-known upper bound first derived by Peierls [32]

$$\mathcal{F}(\mathcal{H}^{\rm pc}) \leq \mathcal{F}(\mathcal{H}^0) + \langle \mathcal{H}^{\rm pc} - \mathcal{H}^0 \rangle. \tag{22}$$

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

$$\frac{\partial}{\partial \mathcal{E}_{i\nu}} \Big(\mathcal{F}(\mathcal{H}^0) + \langle \mathcal{H}^{\rm pc} - \mathcal{H}^0 \rangle \Big)\Big|_{\mathcal{E}_{i\nu}=\mathcal{E}^*_{i\nu}} = 0 \quad \Longrightarrow \quad \mathcal{E}^*_{i\nu} \equiv \langle \tilde{\mathcal{E}}_{i\nu} \rangle \quad \forall \nu \in \{1,\ldots,K\}, \tag{23}$$

$$\text{with} \qquad \tilde{\mathcal{E}}_{i\nu} \equiv \frac{1}{\sum_{\substack{j=1 \\ j\neq i}}^{N} M_{j\nu} + 1} \left[\frac{1}{2}\mathcal{D}_{ii} + \sum_{\substack{k=1 \\ k\neq i}}^{N} M_{k\nu} \left(\mathcal{D}_{ik} - \frac{1}{2} \sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{M_{j\nu}}{\sum_{\substack{j=1 \\ j\neq i}}^{N} M_{j\nu}} \mathcal{D}_{jk} \right) \right]. \tag{24}$$

The resulting optimal assignments are given by

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

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 $\langle \vec{M}_i \rangle$.

We introduce an approximation which neglects terms of the order $\mathcal{O}(1/N)$ to simplify the potentials $\mathcal{E}^*_{i\nu}$. The approximations $\sum_{\substack{j=1 \\ j\neq i}}^{N} M_{j\nu} + 1 \approx \sum_{\substack{j=1 \\ j\neq i}}^{N} M_{j\nu} \approx p_\nu N$ and $\langle 1/p_\nu \rangle \approx 1/\langle p_\nu \rangle$ 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

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

depend on the given distance matrix, the averaged assignment variables and the cluster probabilities. Contrary to Eq. (23), a weak ($\mathcal{O}(1/N)$) dependence of $\mathcal{E}^*_{i\nu}$ on the variables $\langle \vec{M}_i \rangle$ has been introduced by these approximations. The following algorithm estimates the assignments and the optimal potentials $\mathcal{E}^*_{i\nu}$ (defined in Eq. (26)) and assignments $\langle M_{i\nu} \rangle$ iteratively.

Algorithm II

```
INITIALIZE ℰ*_iν^(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 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 Eq.(26)
    T ← T/2;  ⟨M_iν⟩^(0) ← ⟨M_iν⟩^(t);  ℰ*_iν^(0) ← ℰ*_iν^(t);
```

The algorithm decreases the temperature exponentially 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 [2] 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 carried out for a randomly selected datum i converges to a local minimum of the upper bound of the free energy since $\mathcal{E}_{i\nu}^{*\,(t+1)}$ is uniquely determined by the $\langle M_{k\nu}\rangle^{(t+1)}, k \neq i$ which have no implicit dependency on $\mathcal{E}_{i\nu}^{*\,(t+1)}$. The upper bound in Eq. (22) plays the role of a Lyapunov function for the update dynamics of the potentials $\mathcal{E}_{i\nu}^{*\,(t+1)}$. The sequential update scheme has been implemented in the clustering experiments (see Sect. 6).

4.2 Equations for Expected Data Assignments

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

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

$\tilde{\mathcal{E}}_{i\alpha}$ being defined in Eq. (24). The fraction $\exp(-\tilde{\mathcal{E}}_{i\alpha}/T)/\sum_\nu \exp(-\tilde{\mathcal{E}}_{i\nu}/T)$ in Eq. (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 $1/T$.

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

Naively interchanging the averaging brackets with the nonlinear function in Eq. (27) yields Eqs. (25,26); a refined mean–field approach which is known as the TAP approach [33] models the feedback effects in strongly disordered clustering instances more faithfully than the naive approach. The refined expected assignments are

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

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

$\delta_{\alpha\mu}$ denotes the Kronecker symbol, i.e., $\delta_{\alpha\mu} = 1$ if $\alpha = \mu$ and $\delta_{\alpha\mu} = 0$ otherwise. The corrections $\tilde{h}_{i\alpha}$ are also called cavity fields.

The question about the range of validity of Eq. (28) is subtle and it has been studied extensively in statistical physics of disordered systems [31]. Empirically, we can measure average values of the assignments by Monte Carlo simulation. These estimates are inserted into Eq. (28) or into Eq. (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 in comparison to the naive mean–field approximation. According to our Monte Carlo experiments with matrices of Gaussian distributed random dissimilarity values ($N = 1200$), the TAP equations (28) estimate the average assignments $\langle M_{i\nu}\rangle$ with a reduced residual error of up to 50% compared to 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 by the TAP equations can be neglected for small problems ($N < 100$) due to the $N \to \infty$ asymptotics.

5 Pairwise Clustering and Embedding

Grouping data to clusters is an important concept to discover 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 $\mathbf{D}$ in a d-dimensional Euclidian space, a prerequisite for visual inspection, is known as multidimensional scaling [8, 34]. Usually, multidimensional scaling is formulated as an optimization problem for the coordinates $\{\mathbf{x}_i\}$ with costs

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

The stress $\mathcal{H}^{\text{MDS}}$ was introduced by Kruskal in [35]. $\mathcal{H}^{\text{MDS}}$ 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 to combine data clustering and data embedding in a Euclidian space is based on a variational approach to maximum entropy estimation as discussed in subsection 4.1. The coordinates of data points in the embedding space are estimated such that the statistics of the resulting cluster structure matches the statistics of the original pairwise clustering solution. The variational approach to mean–field approximation requires to specify a parametrized family of factorial Gibbs distributions. We choose the factorized Gibbs distributions (12) based on the cost function for central clustering $\mathcal{H}^{\text{cc}}$ and use the embedding coordinates $\{\mathbf{x}_i\}$ as the variational parameters. This approach is motivated by the identity

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

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 $\langle M_{k\nu}\rangle$. The parameters $\mathcal{E}^*_{k\nu}$ are auxiliary variables. The identity (31) demonstrates the meaning of 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 $\mathcal{E}_{i\nu}$ to be of the form $\|\mathbf{x}_i - \mathbf{y}_\nu\|^2$ with the centroid definition $\mathbf{y}_\nu = \sum_{k=1}^{N} M_{k\nu}\mathbf{x}_k / \sum_{k=1}^{N} M_{k\nu}$, 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 to determine the embedding coordinates is

$$\frac{\partial}{\partial \mathbf{x}_i}\left[\mathcal{F}(\mathcal{H}^{\text{cc}}) + \langle \mathcal{H}^{\text{pc}} - \mathcal{H}^{\text{cc}}\rangle\right] = 0 \tag{32}$$

which approximately yields the coordinates

$$\mathbf{K}_i\mathbf{x}_i \approx \frac{1}{2}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\left(\|\mathbf{y}_\nu\|^2 - \mathcal{E}^*_{i\nu}\right)\Big(\mathbf{y}_\nu - \sum_{\mu=1}^{K}\langle M_{i\mu}\rangle\mathbf{y}_\mu\Big), \tag{33}$$

$$\mathbf{K}_i = \left(\langle \mathbf{y}\mathbf{y}^T\rangle_i - \langle\mathbf{y}\rangle_i\langle\mathbf{y}\rangle_i^T\right) \tag{34}$$

Details of the derivation are summarized in Appendix C. The coordinates $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_\nu\}$ are determined by solving Eqs. (33) in an iterative fashion. The following algorithm summarizes the important steps:

Algorithm III

```
INITIALIZE x̂_i^(0) and ⟨M_iν⟩^(0) ∈ (0,1) randomly ; t = 0
      temperature T ← T_0;
WHILE T > T_FINAL
   REPEAT
      E-like step:  estimate ⟨M_iν⟩^(t+1) as a function of {x̂_i, ŷ_ν}
      M-like step:
         REPEAT
            calculate x̂_i^(t+1) given ⟨M_iν⟩^(t+1) and ŷ_ν^(t+1)
            update ŷ_ν^(t+1) to fulfill the centroid condition
         UNTIL convergence
      t ← t + 1
   UNTIL convergence
   T ← T/2; ⟨M_iν⟩^(0) ← ⟨M_iν⟩^(t); x̂^(0) ← x̂^(t); ŷ_ν^(0) ← ŷ_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 $\mathcal{H}^{\text{cc}}$ 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, the equations (33) are singular for $T = 0$ and asymptotic results require to apply l'Hospital's rule.

The derived system of transcendental equations given by (25), (33) and the centroid condition explicitly reflects the dependencies between the clustering procedure and the Euclidian representation. Solving these equations simultaneously 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, II and III 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 Sect. 6.1; (ii) simultaneous pairwise clustering and embedding is performed on artificial and real-world data in Sect. 6.2; (iii) pairwise clustering as a segmentation technique for textured images is discussed in Sect 6.3.

6.1 Benchmark Experiments for Deterministic Annealing

The theoretical derivations of the deterministic annealing algorithms I, II and III are motivated by the known robustness results of maximum entropy inference. To test this claim, a large

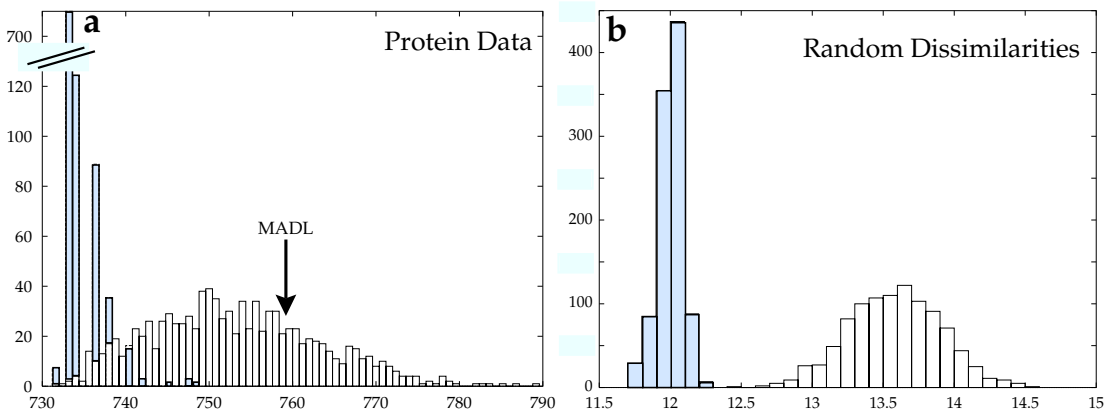

Figure 1: Histograms of clustering costs for a protein data set (**a**) and a random clustering instance (**b**). The gray and white bins denote the results of optimization with deterministic annealing and gradient descent, respectively. The **M**ean **A**verage **D**issimilarity **L**inkage solution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

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 cases. 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 1000 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. Even the worst deterministic annealing solution was better than the best gradient descent solution. In the clustering instance for protein dissimilarities 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 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 $\mathcal{H}^{\text{pc}} = 759.1$ compared to the best (experimentally achieved) result with $\mathcal{H}^{\text{pc}} = 730.9$. All experiments support our claim that deterministic annealing yields substantially better solutions for comparable computing time.

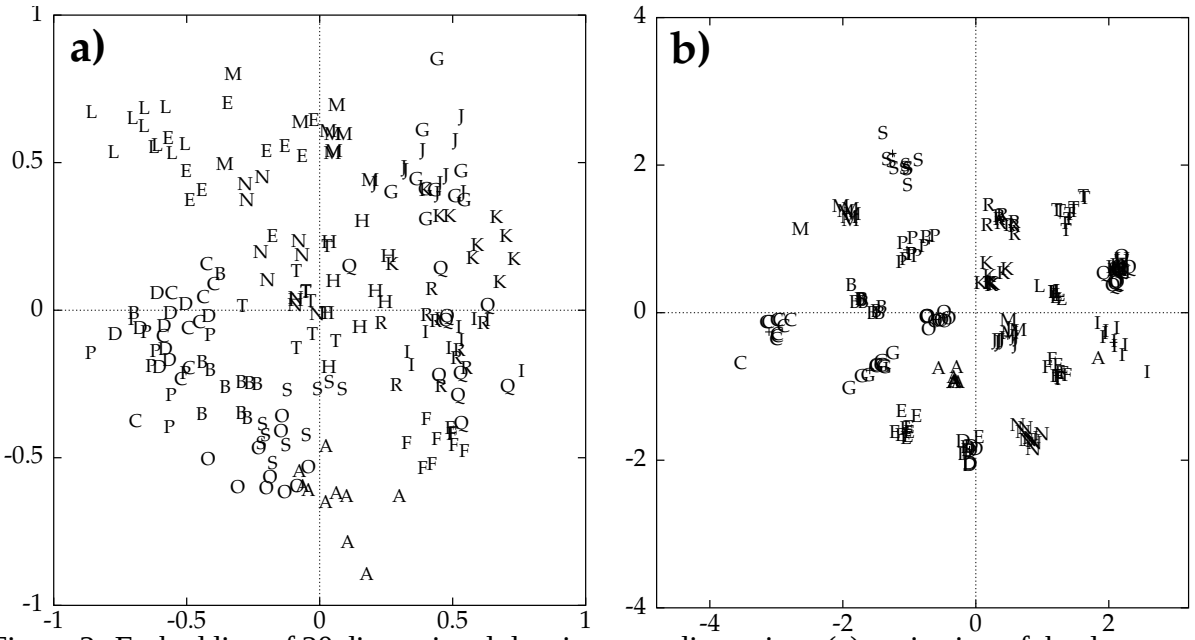

Figure 2: Embedding of 20 dimensional data into two dimension: (a) projection of the data onto the first two principle components; (b) cluster preserving embedding with algorithm III. Only 10% of the data are shown.

6.2 Clustering and Embedding Results

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

1. Clustering and dimension reduction of inhomogeneously distributed data.

2. Clustering of real-world proximity data from protein sequences.

The capability of 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 [36] yield comparable results since no direction is distinguished in the data generation procedure. Simultaneous clustering and embedding by Algorithm III 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 similar to the idea of principal curves [37] or principal surfaces.

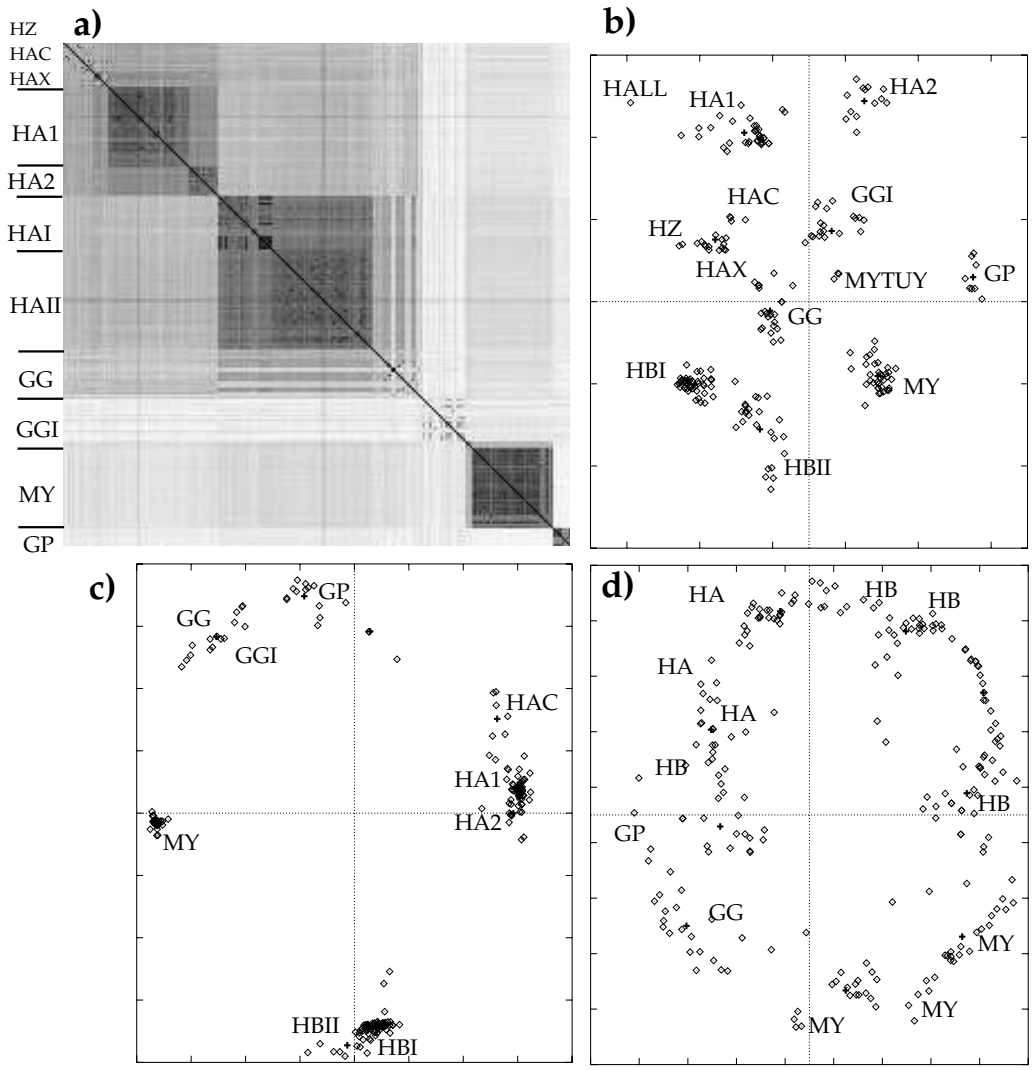

Figure 3: Similarity matrix of 226 protein sequences of the globin family (a): dark gray levels correspond to high similarity values. Clustering with embedding in two dimensions (b); clustering of MDS embeddings found by global (c) or local (d) stress minimization.

Figure 3 summarizes the clustering result ($K = 9$) for a real–world data set of 226 protein sequences. The protein sequences are abbreviated with the displayed 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 $\mathcal{H}^{\text{pc}} = 735.2$. The embedding in two dimensions (Fig. 3b) shows inter–cluster distances which are in good agreement with the similarity values of the data. The best experimentally determined solution ($\mathcal{H}^{\text{pc}} = 730.9$) 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 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%

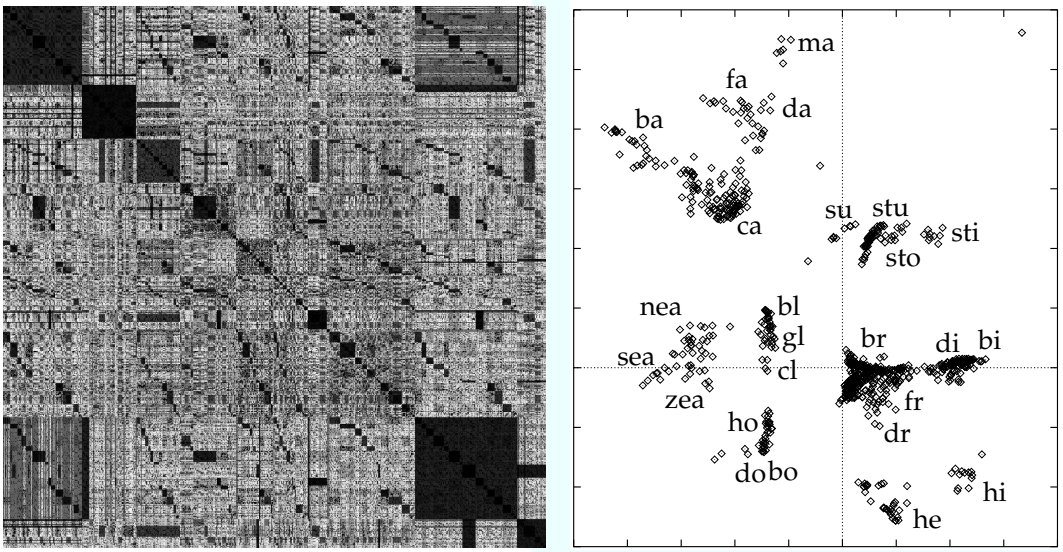

Figure 4: Similarity matrix for a data set with 825 word fragments (a). The calculated clustering solution with embedding in two dimensions (b). The labels denote the groups by common word beginnings.

for a Monte Carlo simulation at $1/T = 2.5$). We have compared the clustering solutions of algorithm III with results of a two step procedure, i.e., first to embed the data with 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 $\mathcal{H}^{\text{pc}} = 782.7\,(833.7)$ for the embedding guided by absolute (relative) stress, respectively. It is obvious from Fig. 3b-d that simultaneous clustering and embedding 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. 825 word fragments have been compared by a dynamic programming algorithm. 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 $1/T = 3.0$).

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 [38] we formulate texture segmentation as a grouping problem with constraints about valid region shapes. The grouping problem is based on *pairwise dissimilarities* between texture

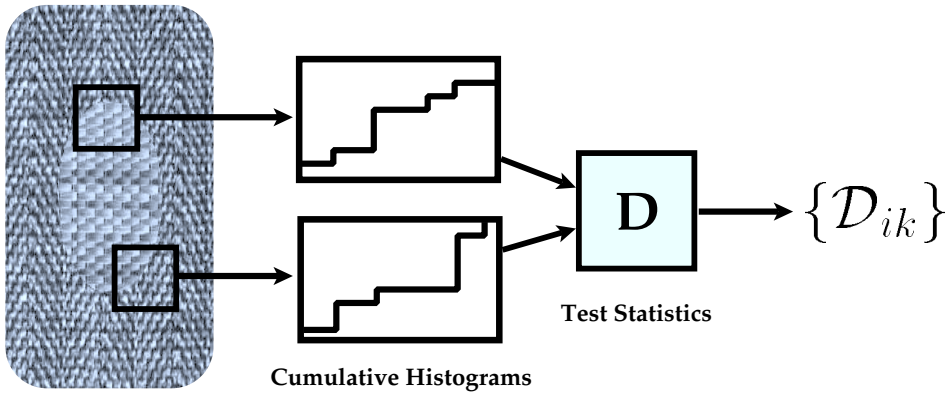

Figure 5: Texture segmentation by pairwise clustering: local properties of image patches, e.g., intensity differences and local frequencies, are extracted. The respective cumulative histograms are compared with the Kolmogorov-Smirnov statistics which yields dissimilarity values to group image regions.

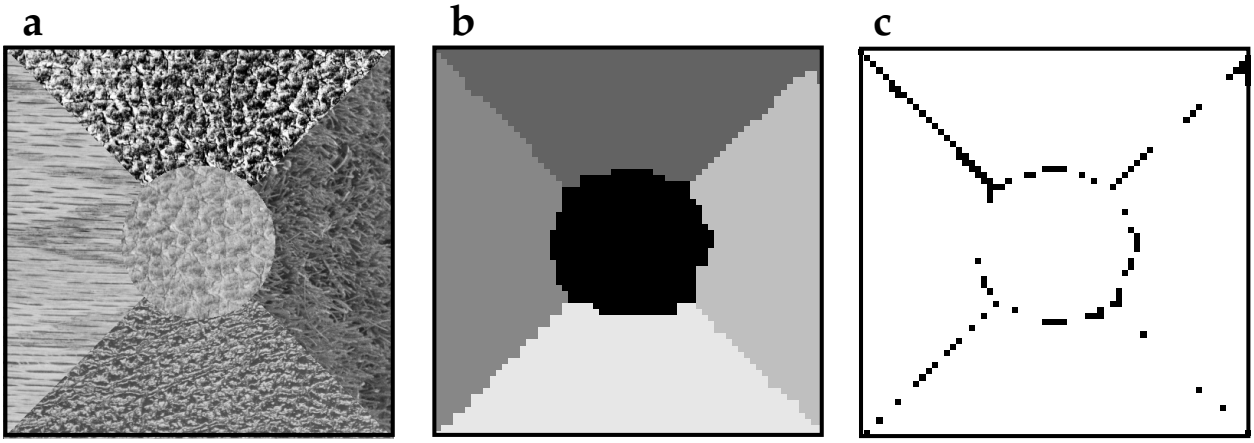

Figure 6: An image of size 512×512 and with five different textures (a) is segmented by pairwise data clustering. The segmentation result by deterministic pairwise clustering (25) is shown in (b). Segmentation errors displayed by black pixels in (c) are located at segment boundaries.

patches which correspond to pixel blocks of the image. Three major modifications compared to [38] 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 [38].

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 characterics and is known to display good texture disrimination properties [39, 40]. We have used four orientations at three different scales, separated by a full octave, resulting in $L = 12$ feature images $I^{(l)}$, $1 \leq l \leq L$ and the raw gray scale image $I^{(0)} = I$. In a second step, the empirical distribution function $F_i^{(l)}$ (cumulative histogram) is calculated separately for every feature image $I^{(l)}$ and every image block B_i, $1 \leq i \leq N$. 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

$$D_{ij}^{(l)} := D^{(l)}(F_i^{(l)}, F_j^{(l)}) := \max_x |F_i^{(l)}(x) - F_j^{(l)}(x)| \in [0; 1]. \tag{35}$$

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 $D_{ij} = \max_{0 \leq l \leq L} D_{ij}^{(l)}$. This is reminiscent of Julesz' theory of texture perception [41], conjecturing that a dissimilarity in a single feature channel is sufficient to discriminate textures. The procedure to generate dissimilarity data from images is schematically summarized in Fig. 5.

We have applied the algorithm to the image shown 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 found 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 [38] 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 and a performance statistics for a large number of textured images can be found in [42].

7 Discussion

The problem of grouping data can be regarded as one of the initial, although fundamental step 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. Robustness with respect to unobservable, as well as to noise events is mandatory for grouping algorithms. In this paper, we have developed a maximum entropy framework for central and pairwise data clustering. A well-known approximation scheme from statistical physics — the mean-field approximation — has been derived in two different ways: (i) a variational method minimizes the Kullback–Leibler divergence between the original Gibbs distribution for data assignments and a parametrized family of factorial distributions; (ii) the expectation values of the data assignments are calculated in a direct fashion. This technique allows us to correct the

influence of small fluctations in data assignments. The variational approximation of pairwise clustering with central clustering yields an algorithm which simultaneously clusters data and embeds them in a Euclidian space. This algorithms can be used for non-linear dimension reduction and for visualization purposes. Results of the pairwise data clustering algorithms to analyse protein and linguistic data and to segment 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 to analyse 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.

Acknowledgement: 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 meanfield equations for the pairwise data clustering problem by minimizing the upper bound on the free energy given in Eq. (22) w.r.t. the variational parameters $\mathcal{E}_{i\alpha}$. Taking derivatives of the upper bound on the free energy yields

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

with

$$\begin{aligned}
\frac{\partial \langle \mathcal{H}^{\text{pc}} \rangle}{\partial \mathcal{E}_{i\alpha}} &= \frac{1}{2} \sum_{\nu=1}^{K} \sum_{j=1}^{N} \sum_{k=1}^{N} \frac{\partial}{\partial \mathcal{E}_{i\alpha}} \left\langle \frac{M_{j\nu} M_{k\nu}}{N p_\nu} \right\rangle \mathcal{D}_{jk} \\
&= \sum_{\nu=1}^{K} \frac{\partial \langle M_{i\nu} \rangle}{\partial \mathcal{E}_{i\alpha}} \left[\sum_{\substack{k=1 \\ k \neq i}}^{N} \left\langle \frac{M_{k\nu}}{\sum_{\substack{j=1 \\ j \neq i}}^{K} M_{j\nu} + 1} \right\rangle \mathcal{D}_{ik} + \left\langle \frac{1}{\sum_{\substack{j=1 \\ j \neq i}}^{K} M_{j\nu} + 1} \right\rangle \frac{\mathcal{D}_{ii}}{2} \right] \\
&\quad - \frac{1}{2} \sum_{\nu=1}^{K} \frac{\partial \langle M_{i\nu} \rangle}{\partial \mathcal{E}_{i\alpha}} \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\substack{k=1 \\ k \neq i}}^{N} \left\langle \frac{M_{j\nu} M_{k\nu}}{\left(\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} \right) \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} + 1 \right)} \right\rangle \mathcal{D}_{jk}, \qquad (36)
\end{aligned}$$

where we have used the identities

$$\frac{M_{i\nu}}{N p_\nu} \;=\; \frac{M_{i\nu}}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} + 1}, \qquad (37)$$

$$\frac{1}{N p_\nu} \;=\; \frac{1}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} + M_{i\nu}} = \frac{1}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu}} - \frac{M_{i\nu}}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} (\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} + 1)}. \qquad (38)$$

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

$$\sum_{\nu=1}^{K} \frac{\partial \langle M_{i\nu} \rangle}{\partial \mathcal{E}_{i\alpha}} \left[\mathcal{E}_{i\nu} - \langle \tilde{\mathcal{E}}_{i\nu} \rangle \right] = -\frac{1}{T} \langle M_{i\alpha} \rangle \left[\left(\mathcal{E}_{i\alpha} - \langle \tilde{\mathcal{E}}_{i\alpha} \rangle \right) - \sum_{\nu=1}^{K} \langle M_{i\nu} \rangle \left(\mathcal{E}_{i\nu} - \langle \tilde{\mathcal{E}}_{i\nu} \rangle \right) \right] = 0, \qquad (39)$$

$$\text{with} \qquad \tilde{\mathcal{E}}_{i\nu} = \frac{1}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu} + 1} \left[\frac{1}{2} \mathcal{D}_{ii} + \sum_{\substack{k=1 \\ k \neq i}}^{N} M_{k\nu} \left(\mathcal{D}_{ik} - \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{M_{j\nu}}{\sum_{\substack{j=1 \\ j \neq i}}^{N} M_{j\nu}} \mathcal{D}_{jk} \right) \right]. \qquad (40)$$

The K equations (39) are only fulfilled for all values $\alpha = 1, \ldots, K$ simultaneously if

$$\mathcal{E}_{i\nu} \;\equiv\; \langle \tilde{\mathcal{E}}_{i\nu} \rangle + c_i, \qquad \forall \nu = 1, \ldots, K. \qquad (41)$$

Appendix B

In this appendix, we derive the meanfield equations of data assignments and fluctuation corrections in the case of strongly disordered clustering problems. The dissimilarities scale as

$\mathcal{D}_{ik} \sim \mathcal{O}(\sqrt{N})$. This alternative derivation is necessary since the variational approach of Sect. [4.1] does not capture these fluctuations adequately, as is known from statistical physics [31]. In the following, data assignments are considered to be randomly drawn from the set of admissible configurations $\mathcal{M}$ according to the Gibbs distribution

$$\mathbf{P}(\mathbf{M}) \;=\; \frac{\exp\left(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T\right)}{\sum_{\bar{\mathbf{M}}\in\mathcal{M}} \exp\left(-\mathcal{H}^{\text{pc}}(\bar{\mathbf{M}})/T\right)}. \tag{42}$$

Therefore, the expected assignment of datum i to cluster α is

$$\langle M_{i\alpha}\rangle \;=\; \frac{\sum_{\mathcal{M}} M_{i\alpha}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)}{\sum_{\mathcal{M}}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)} \tag{43}$$

$$= \frac{\displaystyle\sum_{\mathcal{M}}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)\;\frac{\displaystyle\sum_{\{\vec{M}_i\}} M_{i\alpha}\exp(-\mathcal{H}^{\text{pc}}(\vec{M}_i,\hat{\mathbf{M}})/T)}{\displaystyle\sum_{\{\vec{M}_i\}}\exp(-\mathcal{H}^{\text{pc}}(\vec{M}_i,\hat{\mathbf{M}})/T)}}{\displaystyle\sum_{\mathcal{M}}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)} \tag{44}$$

where $\hat{\mathbf{M}}$ denotes the set of assignments without $\vec{M}_i$. The partial summation over the admissible states $\{\vec{M}_i\} = \{(1,0,\ldots,0),(0,1,\ldots,0),\ldots(0,\ldots,0,1)\}$ can be carried out analytically. The first step is the separation of the clustering costs $\mathcal{H}^{\text{pc}}$ in a term $\mathcal{H}_{i\not}$without any contribution from $\vec{M}_i$ and costs which are related to $\vec{M}_i$. $\mathcal{H}_{i\not}$is given by

$$\mathcal{H}_{i\not} \equiv \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N}\mathcal{D}_{jk}\sum_{\nu=1}^{K}\frac{M_{j\nu}M_{k\nu}}{\sum_{\substack{l=1\\ l\neq i}}^{N} M_{l\nu}}. \tag{45}$$

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

$$\frac{\displaystyle\sum_{\{\vec{M}_i\}} M_{i\alpha}\exp(-\mathcal{H}^{\text{pc}}(\vec{M}_i,\hat{\mathbf{M}})/T)}{\displaystyle\sum_{\{\vec{M}_i\}}\exp(-\mathcal{H}^{\text{pc}}(\vec{M}_i,\hat{\mathbf{M}})/T)} \;=\; \frac{\displaystyle\sum_{\{\vec{M}_i\}} M_{i\alpha}\exp\left(-\Big(\mathcal{H}_{i\not}+\sum_{\nu=1}^{K} M_{i\nu}\tilde{\mathcal{E}}_{i\nu}\Big)/T\right)}{\displaystyle\sum_{\{\vec{M}_i\}}\exp\left(-\Big(\mathcal{H}_{i\not}+\sum_{\nu=1}^{K} M_{i\nu}\tilde{\mathcal{E}}_{i\nu}\Big)/T\right)}$$

$$= \frac{\exp\left(-\tilde{\mathcal{E}}_{i\alpha}/T\right)}{\displaystyle\sum_{\nu=1}^{K}\exp\left(-\tilde{\mathcal{E}}_{i\nu}/T\right)} \tag{46}$$

$\tilde{\mathcal{E}}_{i\alpha}$ has been defined in Eq. (40). In summary, the expected assignments are

$$\langle M_{i\alpha}\rangle = \frac{\sum_{\mathcal{M}} M_{i\alpha}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)}{\sum_{\mathcal{M}}\exp(-\mathcal{H}^{\text{pc}}(\mathbf{M})/T)} \;=\; \left\langle \frac{\exp(-\tilde{\mathcal{E}}_{i\alpha}/T)}{\sum_{\nu}\exp(-\tilde{\mathcal{E}}_{i\nu}/T)}\right\rangle. \tag{47}$$

Equation (47) is analogous to the Callen equation for Ising spins in the theory of magnetic systems (see [43], Sect 3.2).

A Taylor expansion of Eq. (47) in small fluctuations $\Delta\tilde{\mathcal{E}}_{i\nu} = \tilde{\mathcal{E}}_{i\nu} - \langle\tilde{\mathcal{E}}_{i\nu}\rangle$ renders a closed system of equations exclusively depending on the averaged assignments $\langle M_{i\nu}\rangle$. The expected assignments are

$$\begin{aligned}\langle M_{i\alpha}\rangle &= \Big\langle f_\alpha(\{\tilde{\mathcal{E}}_{i\nu}\})\Big\rangle = \Big\langle f_\alpha(\{\langle\tilde{\mathcal{E}}_{i\nu}\rangle + \Delta\tilde{\mathcal{E}}_{i\nu}\})\Big\rangle \\ &= f_\alpha(\{\langle\tilde{\mathcal{E}}_{i\nu}\rangle\}) + \frac{1}{2}\sum_{\nu=1}^{K}\sum_{\mu=1}^{K}\frac{\partial^2 f_\alpha}{\partial\tilde{\mathcal{E}}_{i\nu}\partial\tilde{\mathcal{E}}_{i\mu}}\langle\Delta\tilde{\mathcal{E}}_{i\nu}\Delta\tilde{\mathcal{E}}_{i\mu}\rangle + \mathcal{O}(\langle\Delta\tilde{\mathcal{E}}^3_{i\nu}\rangle), \end{aligned} \tag{48}$$

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

$$\langle M_{i\alpha}\rangle = \frac{\exp(-\langle\tilde{\mathcal{E}}_{i\alpha}\rangle/T)}{\sum_\nu \exp(-\langle\tilde{\mathcal{E}}_{i\nu}\rangle/T)}. \tag{49}$$

The derivation of Eq. (49) tacitly assumes that the assignment correlation function scales as

$$\langle M_{k\nu}M_{l\mu}\rangle - \langle M_{k\nu}\rangle\langle M_{l\mu}\rangle = \begin{cases}\langle M_{k\nu}\rangle(\delta_{\nu\mu} - \langle M_{k\mu}\rangle) & \text{for } k = l \\ \mathcal{O}(1/\sqrt{N}) & \text{for } k \neq l\end{cases}. \tag{50}$$

Fluctuations of the data assignments for consistent dissimilarities are averaged in the limit $N \to \infty$ due to 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 $\tilde{h}_{i\nu}$ which simulates the indirect influence of the disorder on the data assignments (see Thouless, Anderson and Palmer [33]). Without loss of generality the dissimilarity values are assumed to have a vanishing expectation value, i.e., $\langle\mathcal{D}_{ik}\rangle \equiv 0$. The scaling of the dissimilarities is assumed to be $\mathcal{D}_{ik} \sim \mathcal{O}(\sqrt{N})$. This shift of the dissimilarity values and their random nature allows us to neglect the second term in Eq. (26) since $\sum_{k=1}^{N}\sum_{l=1}^{N} M_{k\nu}M_{l\nu}\mathcal{D}_{kl}/(2p_\nu^2 N^2) \sim N\sqrt{N}/N^2 = 1/\sqrt{N}$.

The Ansatz for the expected assignments with effective internal field is

$$\begin{aligned}\langle M_{i\alpha}\rangle &= f_\alpha(\{\langle\tilde{\mathcal{E}}_{i\nu}\rangle - \tilde{h}_{i\nu}\}) \\ &= f_\alpha(\{\langle\tilde{\mathcal{E}}_{i\nu}\rangle\}) - \sum_{\nu=1}^{K}\frac{\partial f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{i\nu}\rangle}\tilde{h}_{i\nu} + \mathcal{O}(\max_{i\nu}\tilde{h}^2_{i\nu}) \end{aligned} \tag{51}$$

The term linear in $\tilde{h}_{i\nu}$ (Eq. 51) has to capture all fluctuation contributions of Eq. (48). A comparison of the coefficients yields

$$-\sum_{\nu=1}^{K}\frac{\partial f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{i\nu}\rangle}\tilde{h}_{i\nu} = \frac{1}{2}\sum_{\nu=1}^{K}\sum_{\mu=1}^{K}\frac{\partial^2 f_\alpha}{\partial\langle\tilde{\mathcal{E}}_{i\nu}\rangle\partial\langle\tilde{\mathcal{E}}_{i\mu}\rangle}\langle\Delta\tilde{\mathcal{E}}_{i\nu}\Delta\tilde{\mathcal{E}}_{i\mu}\rangle \tag{52}$$

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

$$\sum_{\nu=1}^{k}(\delta_{\alpha\nu}-\langle M_{i\nu}\rangle)\,\tilde{h}_{i\nu} \;=\; \frac{1}{2T}\sum_{\nu=1}^{K}\sum_{\mu=1}^{K}\Gamma_{i\nu\mu}\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\mathcal{D}_{ik}\mathcal{D}_{il}}{p_\nu p_\mu N^2}\Big(\langle M_{k\nu}M_{l\mu}\rangle-\langle M_{k\nu}\rangle\langle M_{l\mu}\rangle\Big)$$

$$=\frac{1}{2T}\sum_{\nu=1}^{K}(\delta_{\alpha\nu}-\langle M_{i\nu}\rangle)\sum_{k=1}^{N}\sum_{\mu=1}^{K}\mathcal{D}_{ik}^2\frac{\langle M_{k\nu}\rangle(\delta_{\nu\mu}-\langle M_{k\mu}\rangle)}{p_\nu p_\mu N^2}(\delta_{\nu\mu}-2\langle M_{i\mu}\rangle) \tag{53}$$

$$+\text{ terms with } k\neq l$$

$$\Gamma_{i\nu\mu}=(\delta_{\alpha\nu}-\langle M_{i\nu}\rangle)(\delta_{\alpha\mu}-\langle M_{i\mu}\rangle)-\langle M_{i\mu}\rangle(\delta_{\nu\mu}-\langle M_{i\nu}\rangle) \tag{54}$$

If assumption (50) holds, the terms $k\neq l$ in Eq. (53) vanish as $\sqrt{N^2}/(\sqrt{N}^3)$. An in depth discussion when the assumption (50) is valid can be found in [31]. Assuming the validity of Eq. (50) the refined mean–field equations are

$$\langle M_{i\alpha}\rangle \;=\; \frac{\exp\Big(-(\langle\tilde{\mathcal{E}}_{i\alpha}\rangle-\tilde{h}_{i\alpha})/T\Big)}{\sum_\nu\exp\Big(-(\langle\tilde{\mathcal{E}}_{i\nu}\rangle-\tilde{h}_{i\nu})/T\Big)}, \tag{55}$$

$$\tilde{h}_{i\alpha} \;=\; \frac{1}{2T}\sum_{k=1}^{N}\mathcal{D}_{ik}^2\sum_{\mu=1}^{K}\frac{\langle M_{k\alpha}\rangle(\delta_{\alpha\mu}-\langle M_{k\mu}\rangle)}{p_\alpha p_\mu N^2}\,(\delta_{\alpha\mu}-2\langle M_{i\mu}\rangle)\,. \tag{56}$$

Appendix C

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

$$\frac{\partial}{\partial\mathbf{x}_i}\left[\mathcal{F}(\mathcal{H}^{\text{cc}})+\langle\mathcal{H}^{\text{pc}}-\mathcal{H}^{\text{cc}}\rangle\right] \;=\; \sum_{k=1}^{N}\sum_{\mu=1}^{K}\frac{\partial}{\partial\mathcal{E}_{k\mu}}\left[\mathcal{F}(\mathcal{H}^{\text{cc}})+\langle\mathcal{H}^{\text{pc}}-\mathcal{H}^{\text{cc}}\rangle\right]\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i}$$

$$=\; -\frac{1}{T}\sum_{k=1}^{N}\sum_{\mu=1}^{K}\sum_{\nu=1}^{K}\langle M_{k\nu}\rangle\,(\delta_{\nu\mu}-\langle M_{k\mu}\rangle)\,\Delta\mathcal{E}_{k\nu}\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i}, \tag{57}$$

where $\Delta\mathcal{E}_{k\alpha}=\mathcal{E}_{k\alpha}-\mathcal{E}^*_{k\alpha}$ and $\mathcal{E}_{k\mu}=\|\mathbf{x}_k-\mathbf{y}_\mu\|^2$. The derivatives (57) are given by

$$\frac{\partial\mathcal{E}_{k\mu}}{\partial\mathbf{x}_i}=2\,(\mathbf{x}_k-\mathbf{y}_\mu)^T\left[\delta_{ik}-\frac{\langle M_{i\mu}\rangle}{Np_\mu}-\frac{1}{Np_\mu}\sum_{l=1}^{N}(\mathbf{x}_l-\mathbf{y}_\mu)\,\frac{\partial\langle M_{l\mu}\rangle}{\partial\mathbf{x}_i}\right]. \tag{58}$$

Setting Eq. (57) equal to zero results in the exact stationary conditions

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

$$\sum_{k=1}^{N}\sum_{\nu,\mu=1}^{K}\frac{\langle M_{k\nu}\rangle\langle M_{k\mu}\rangle}{Np_\mu}\,(\Delta\mathcal{E}_{i\mu}-\Delta\mathcal{E}_{i\nu})\left[\langle M_{i\mu}\rangle+\sum_{l=1}^{N}\left((\mathbf{x}_l-\mathbf{y}_\mu)\frac{\partial\langle M_{l\mu}\rangle}{\partial\mathbf{x}_i}\right)^T\right](\mathbf{x}_k-\mathbf{y}_\mu)\,. \tag{59}$$

The lefthand side can be further reduced to an expression explicit in $\mathbf{x}_i$

$$\sum_{\nu,\mu=1}^{K} \langle M_{i\nu}\rangle\langle M_{i\mu}\rangle \left(\Delta\mathcal{E}_{i\mu} - \Delta\mathcal{E}_{i\nu}\right) \mathbf{y}_\mu \;=\; \sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\Delta\mathcal{E}_{i\nu}\left(\mathbf{y}_\nu - \sum_{\mu=1}^{K}\langle M_{i\mu}\rangle\mathbf{y}_\mu\right)$$
$$= -2\mathbf{K}_i\mathbf{x}_i + \sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\left(\|\mathbf{y}_\nu\|^2 - \mathcal{E}^*_{i\nu}\right)\left(\mathbf{y}_\nu - \sum_{\alpha=1}^{K}\langle M_{i\alpha}\rangle\mathbf{y}_\alpha\right), \tag{60}$$

where $\mathbf{K}_i = \left(\langle \mathbf{y}\mathbf{y}^T\rangle_i - \langle\mathbf{y}\rangle_i\langle\mathbf{y}\rangle_i^T\right)$ is a $d\times d$ 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 Eq. (59) can be exactly calculated, since they are given as the solutions of a linear equation system with $N\times K$ unknowns for every $\mathbf{x}_i$. But 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 (60) simplifies to a vector equation for every $\mathbf{x}_i$,

$$\mathbf{K}_i\mathbf{x}_i \approx \frac{1}{2}\sum_{\nu=1}^{K}\langle M_{i\nu}\rangle\left(\|\mathbf{y}_\nu\|^2 - \mathcal{E}^*_{i\nu}\right)\left(\mathbf{y}_\nu - \sum_{\mu=1}^{K}\langle M_{i\mu}\rangle\mathbf{y}_\mu\right). \tag{61}$$

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," *Proceedings of the IEEE*, vol. 70, pp. 939–952, 1982.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ 07632: Prentice Hall, 1988.

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

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society Ser. B (methodological)*, vol. 39, pp. 1–38, 1977.

[7] R. M. Gray, "Vector quantization," *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4–29, April 1984.

[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 Transactions on Pattern Analysis and Machine Intelligence*, 1996. in press.

[13] G. D. and G. F., "Parallel and deterministic algorithms from MRF's: surface reconstruction," *IEEE Transactions on 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," in *Advances in Neural Information Processing Systems 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 Transactions on 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 Transactions on 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 Transactions on Information Theory*, vol. 39, pp. 1133–1145, July 1993.

[22] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Transactions on 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 travelling salesman problem: an efficient simulation algorithm," *Journal of 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 & Sons, 1991.

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

[32] R. E. Peierls, "On a minimum property of the free energy," *Physical Review*, vol. 54, p. 918, 1938.

[33] 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.

[34] J. W. Sammon Jr, "A non-linear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, pp. 401–409, 1969.

[35] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, pp. 115–129, 1964.

[36] P. Huber, "Projection pursuit," *Annals of Statistics*, vol. 13, pp. 435–475, 1985.

[37] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, pp. 502–516, 1989.

[38] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 609–628, July 1990.

[39] I. Fogel and D. Sagi, "Gabor filters as texture discriminators," *Biological Cybernetics*, vol. 61, pp. 103–113, 1989.

[40] J. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *Journal of the Optical Society Am. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[41] B. Julesz, "Visual pattern discrimination," *IRE Transactions on Information Theory*, pp. 84–92, February 1961.

[42] T. Hofmann, J. Puzicha, and J. M. Buhmann, "Unsupervised segmentation of textured images by pairwise data clustering," Tech. Rep. IAI-TR-96-2, Rheinische Friedrich–Wilhelms–Universität Bonn, Institut für Informatik III, February 1996.

[43] G. Parisi, *Statistical Field Theory*. Redwood City, CA: Addison Wesley, 1988.