Macrostate data clustering

Daniel Korenblum* and David Shalloway†
Biophysics Program, Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853
(Received 10 December 2002; published 15 May 2003)

We develop an effective nonhierarchical data clustering method using an analogy to the dynamic coarse graining of a stochastic system. Analyzing the eigensystem of an interitem transition matrix identifies fuzzy clusters corresponding to the metastable macroscopic states (macrostates) of a diffusive system. A ''minimum uncertainty criterion'' determines the linear transformation from eigenvectors to cluster-defining window functions. Eigenspectrum gap and cluster certainty conditions identify the proper number of clusters. The physically motivated fuzzy representation and associated uncertainty analysis distinguishes macrostate clustering from spectral partitioning methods. Macrostate data clustering solves a variety of test cases that challenge other methods.

DOI: 10.1103/PhysRevE.67.056704 PACS number(s): 02.70.Hm, 02.50.Fz, 89.75.Kd

I. INTRODUCTION

Finding subgroups or *clusters* within large sets of *items* is a problem that occurs in many contexts from astronomy to integrated chip design and pattern recognition (see Refs. [1–4] for reviews). DNA microarray gene expression analysis and bioinformatic sequence comparisons are recent and important applications in molecular biology [3,5].

The clustering problem may be posed in two ways: In the first case (e.g., DNA microarrays), N_M measurements are made on each of the N items. The $N\times N_M$ measurement matrix X is then used in a problem-specific manner to compute a symmetric $N\times N$ *dissimilarity* matrix D. Each off-diagonal element D_{ij} provides an inverse indicator of the correlations between the measurements of items i and j. A straightforward method is to set

$$D_{ij}=\left[\sum_{a,b=1}^{N_M}(X_{ia}-X_{ja})g_{ab}(X_{ib}-X_{jb})\right]^{1/2}, \qquad (1)$$

where g is a problem-specific Euclidean metric tensor. This allows preconditioning of the scales of the different measurements and, by using nondiagonal g, adjustment for measurement correlations (particularly important if some measurements are replicates). Statistical noise and complexity can be reduced by using singular-value decomposition to preidentify principal components of X that span much of the variation in the measurement space. This facilitates identification of clusters ''by eye'' or with various heuristics (e.g., Refs. [6–8]).

In the second case (e.g., pairwise gene sequence comparisons), the primary data are measures of dissimilarities between pairs of items: In this case D is defined, but there is no measurement matrix X and the elements of D may not satisfy the triangle inequality.

Early clustering methods were ''hierarchical,'' generating open binary trees which can (depending on the selected cross-section) dissect the items into any number of clusters between 1 and N. In these methods, the choice of the optimal number of clusters is an independent problem [2,9,10]. ''Agglomerative'' hierarchical methods iteratively join items together to form a decreasing number of larger clusters; ''divisive'' hierarchical methods use successive subdivision to generate an increasing number of smaller clusters. While agglomerative methods can be inexpensive, they usually use only local and not global information, which weakens performance [2]. While divisive methods can use global information, they can have high complexity in N and thus can be too expensive for large problems. A weakness of both types of hierarchical methods is that they cannot repair defects from previous stages of analysis.

Some clustering methods are based on analogies to physical systems, in which macroscopic structure emerges from microscopically defined interactions. A number of them have used analogies to statistical mechanical phase transitions [11–15], while others have used chaotic [16] or quantum mechanical [17] systems as analogs. Most of these have the advantage of being ''fuzzy''—in addition to assigning items to clusters, they provide a continuous measure of the probability or strength of the assignment of each item.

Clustering can also be performed by objective function optimization. If there is an *a priori* model for the structure of the clusters in the measurement space (e.g., as a collection of Gaussians), then a corresponding parametric objective function can be used. Otherwise a nonparametric objective function may be useful. For example, graph theory clustering methods treat the items as nodes of a graph whose interconnecting edges have ''affinities'' or ''weights'' determined from D (See Refs. [18,19] for review). They typically use ''min cut'' or ''normalized cut'' objective functions and search for the (sometimes ''balanced'') graph partitioning that minimizes the (sometimes normalized) sum of the weights of the cut edges. Spectral graph theory [20] methods use the eigenvectors of the affinity matrix (or the closely related generalized Laplacian matrix) to assist the process. Spectral bipartitioning (See Refs. [21] for history and review), which uses one eigenvector, can be applied recursively for hierarchical dissection [22]; and the development of nonhierarchical methods for the concurrent use of multiple eigenvectors is an active topic of research (see Refs. [19,23] for reviews).

*Present address: Gene Network Sciences, Ithaca, NY 14850.
†Electronic address: dis2@cornell.edu

We present here a nonhierarchical, fuzzy clustering method that uses an analogy between data clustering and the coarse graining of a stochastic dynamical system. The items are regarded as microstates that interact via a dynamical *transition matrix* Γ, which is derived from D. Clusters are identified as the slowly relaxing metastable macroscopic states (macrostates) of the system. These are computed by concurrently using multiple eigenvectors of Γ in the same way as the macrostates of a continuous diffusive system are identified from the eigenfunctions of the Smoluchowski operator [24]. The number of clusters is algorithmically determined by the spectral properties and cluster overlap criteria. We demonstrate that the method can solve difficult problems, including ones in which the clusters are irregularly shaped and separated by tortuous boundaries.

II. METHOD

A. Macrostates and stochastic coarse graining—a brief overview

Coarse graining is used in nonequilibrium statistical physics to reduce the dimensionality and complexity of the dynamical description [25]. In the usual situation, the system is initially described microscopically by a fine-grained first-order equation specified over a configuration space of microscopic states (*microstates*). Microscopic degrees of freedom corresponding to very rapid motions are removed by (possibly nonlinear) projection. This generates a coarse-grained master equation with fewer degrees of freedom, which describes the slower dynamics of the system's macrostates. Each macrostate corresponds to a subregion of configuration space that has been projected onto one value of the macroscopic parameters. For example, to describe Brownian motion of pollen in water, the (fast) water molecule degrees of freedom are projected out, leaving only the (slow) coordinate of the pollen to parametrize the macrostates. In this example, the macrostates are continuously parametrized, but they may also be discrete. For example, to describe chemical reactions, each macrostate is a chemical state, a subregion of conformation space which includes all vibrations, translations, and rotations of a specific metastable, covalently bonded arrangement of atoms.

Coarse-graining projections are not arbitrary: the utility of the resultant macroscopic description depends upon the existence of a sufficient disparity between τ^{local}, the time scale of the fast (projected-out) motions (which generate ergodicity within the macrostate), and τ^{global}, the time scale of the remaining slow motions (which are required for ergodicity between macrostates). Appropriate projections can sometimes be chosen heuristically when the disparity between τ^{local} and τ^{global} is large and subjectively obvious. When this is not so, projections into discrete macrostates can be selected by analyzing the eigenspectrum of the microscopic stochastic dynamics. This procedure is described in detail in Refs. [24,26]. We summarize the salient points here.

Consider the example illustrated in Fig. 1(A) of a thermal ensemble of systems having microscopic parameter x and potential energy $V(x)$. The bimodal equilibrium probability density is given by the Gibbs-Boltzmann distribution

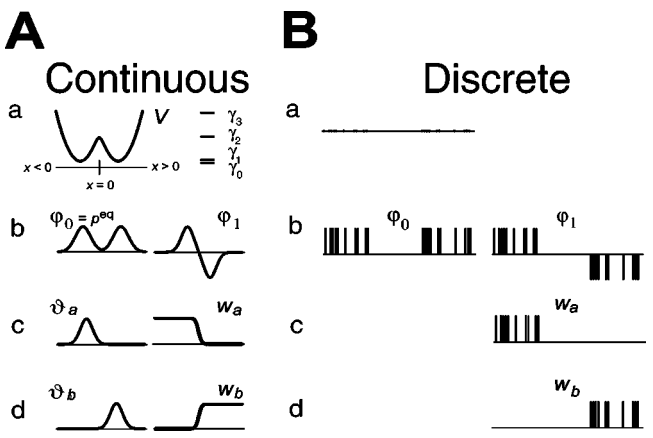

FIG. 1. Heuristic examples. (A) Identifying the macrostates of a continuous stochastic system in one dimension. Panel a: the potential $V(x)$ and eigenvalue spectrum. Panel b: the zeroth and the first excited right eigenfunctions of the corresponding diffusive dynamical (Smoluchowski) equation. Panels c and d: the two macrostate distribution and window functions. (B) Macrostate clustering of items in a one-dimensional space. Panel a: the positions of the items in the univariate measurement space. Panel b: graphical representation of the zeroth and the first eigenvectors of Γ; the height of the bar at the position of item i corresponds to its component within the indicated eigenfunction. Panels c and d: the components of the two window vectors corresponding to the left ($\boldsymbol{w}_a$) and right ($\boldsymbol{w}_b$) clusters.

$p^{\text{eq}}(x) \propto \exp[-\beta V(x)]$, where β is the inverse temperature in inverse energy units. If system dynamics are overdamped (i.e., diffusive), then the nonequilibrium probability distribution $p(x;t)$ evolves in time according to the first-order Smoluchowski equation

$$\frac{\partial p(x;t)}{\partial t} = \int \Gamma(x,x')p(x';t)dx', \tag{2}$$

where Γ is the kernel of an operator determined by V, the temperature, and the diffusion constant. Once the eigenfunctions and eigenvalues of Γ have been determined, the general solution to Eq. (2) can be expanded as

$$p(x;t) = \sum_{n=0}^{\infty} c_n e^{-\gamma_n t}\varphi_n(x), \tag{3}$$

where the eigenvalues and right eigenfunctions of Γ are $-\gamma_n$ and $\varphi_n(x)$, respectively, and the expansion coefficients c_n are determined by the initial conditions $p(x;0)$. (Without loss of generality we normalize φ_0 so that $c_0=1$, and assume that eigenfunctions are ordered according to the magnitudes of their eigenvalues.)

We always have $\gamma_0=0$ and $\varphi_0(x)=p^{\text{eq}}(x)$, corresponding to the stability of the Gibbs-Boltzmann distribution. The other γ_n are non-negative and determine the rates of relaxation towards this equilibrium state. While φ_0 is nonnegative everywhere, the other eigenfunctions take both positive and negative values, and the exponential decays of their amplitudes generate probability fluxes. For illustration, Fig. 1(A), panel b displays (for a selected temperature) φ_0

and the ''slow'' right eigenfunction φ_1. If $c_1>0$, $p(x;0)$ will have a probability excess (relative to p^{eq}) for $x<0$ and a deficiency for $x>0$. These overall deviations from equilibrium will decay away as $\exp(-\gamma_1 t)\to 0$. The ''fast'' eigenfunctions $\varphi_{n>1}$ will have more nodes than φ_1 and their more rapid decays will transport probability over smaller regions.

Sufficiently large potential energy barriers can separate configuration space into localized, dynamically metastable macrostate regions, each having the property that τ^{local}, the time scale for relaxation of $p(x;t)$ towards $p^{\mathrm{eq}}(x)$ within the region, is much less than τ^{global}, the time scale for probability to enter or leave the region. τ^{local} and τ^{global} are determined by the eigenvalues, and a disparity between them will correspond to a gap in the eigenspectrum. If there are m macrostates, a gap will occur between γ_{m-1} and γ_m: there will be m slow modes generating intermacrostate probability equilibration, and the remaining fast modes will generate intramacrostate relaxations.

For example, in Fig. 1(A) the energy barrier centered at $x=0$ separates configuration space into two macrostates a and b (roughly containing the regions $x<0$ or $x>0$, respectively). Correspondingly, there is a spectral gap $\gamma_1 \ll \gamma_2$; so $m=2$. γ_1 is the rate of the slow transfer of probability between a and b,which is generated by the slow decay of the amplitude of φ_1. Thus, $\tau^{\mathrm{global}}\sim \gamma_1^{-1}$. The larger values of the $\gamma_{n>1}$ correspond to the fast decays of the more-rapidly varying $\varphi_{n>1}$, corresponding to intramacrostate probability relaxations. The slowest of these rates, γ_2, determines the time needed for local equilibration. Thus, $\tau^{\mathrm{local}}\sim \gamma_2^{-1}$.

In this simple case, it is tempting to ''crisply'' define the macrostates by inspection as the regions $x>0$ and $x<0$. However, this is inapt for two reasons: (1) A sharp boundary at $x=0$ introduces high-frequency dynamical modes and thus is incompatible with a consistent low-frequency dynamical description; and (2) subjective inspection and barrier identification are not possible in multidimensional problems. Instead, we use this example to show how the correct fuzzy macrostates can be identified (without subjective inspection) by a generalizable algorithm:

The starting point is the recognition of the spectral gap $\gamma_1 \ll \gamma_2$. When $t>\gamma_2^{-1}$, the values of $p(x;t)$ for all $x<0$ or all $x>0$ will be highly correlated, and relative equilibrium within (but not between) these regions will be achieved. Therefore, in this temporal regime, $p(x;t)$ can be well approximated by an expansion within the rank-2 (in general, rank-m) *macrostate subspace* spanned by φ_0 and φ_1, and only the first two terms in the summation in Eq. (3) need to be kept. To obtain a probabilistic description, we replace this truncated eigenfunction expansion by an expansion in the alternative basis provided by the non-negative *macrostate distributions* $\vartheta_a(x)$ and $\vartheta_b(x)$ (to be defined precisely below) shown in Fig. 1(A), panels c and d. ϑ_a (or ϑ_b) is approximately proportional to φ_0 for $x<0$ (or $x>0$) and ≈ 0 for $x>0$ (or $x<0$). Thus, Eq. (3) can be replaced by the *macrostate expansion*

$$p(x;t)\approx \sum_\alpha p_\alpha(t)\,\vartheta_\alpha(x), \tag{4}$$

where greek letters index macrostates and sums over greek letters indicate sums over all macrostates. [We assume the normalization $\int \vartheta_\alpha(x)dx=1$.] Since ϑ_a and ϑ_b have significant support only for $x<0$ and $x>0$, respectively, $p_a(t)$ and $p_b(t)$ specify the time-dependent amounts of probability in these regions. Their dynamics are described by the coarse-grained *macrostate master equation*

$$\frac{dp_\alpha(t)}{dt}=\sum_\beta p_\beta(t)\Gamma_{\beta\alpha}, \tag{5}$$

where $\Gamma_{\beta\alpha}$ is the *macrostate transition matrix*. As $t\to\infty$, Eq. (4) reduces to

$$\lim_{t\to\infty} p(x;t)=\varphi_0=\sum_\alpha p_\alpha^{\mathrm{eq}}\vartheta_\alpha(x). \tag{6}$$

where p_α^{eq} is the total probability contained in the macrostate region α at equilibrium.

The ϑ_α implicitly define the macrostate regions. To make this explicit, we define *macrostate window functions*

$$w_\alpha(x)\equiv \frac{p_\alpha^{\mathrm{eq}}\vartheta_\alpha(x)}{\varphi_0(x)}. \tag{7}$$

Equation (6) and the non-negativity of ϑ_α imply that

$$w_\alpha(x)\geqslant 0 \quad \forall\ \alpha,x \tag{8a}$$

$$\sum_\alpha w_\alpha(x)=1 \quad \forall\ x. \tag{8b}$$

$w_\alpha(x)$ specifies the probability of assignment of microstate x to macrostate α. The window functions corresponding to ϑ_a and ϑ_b are shown in Fig. 1(A). They define a fuzzy dissection of configuration space into the macrostate regions $x<0$ and $x>0$.

Now we can address the precise definition of the ϑ_α themselves. Since they span the macrostate subspace, they must be linear combinations of the slow eigenfunctions:

$$\vartheta_\alpha(x)=\sum_{n=0}^{m-1} M_{\alpha n}\varphi_n(x). \tag{9}$$

Since the φ_n are smooth, the ϑ_α, and hence the w_α, must also be smooth. This induces an unavoidable uncertainty in microstate assignment. For example, in Fig. 1 the assignments are almost certain for $|x|\gg 0$, where $w_\alpha\approx 1$, but are highly uncertain for $x\approx 0$ where $w_\alpha(x)\approx 0.5$. The essential point is to choose M, and hence the ϑ_α and w_α, so as to maximize certainty. We define Υ_α, the *uncertainty* of macrostate α, as the sum of its equilibrium probability-weighted overlaps with the other macrostates, relative to its total probability:[1]

[1]This definition is motivated by an analysis of the experimental macrostate preparation and measurement process [24].

$$\Upsilon_\alpha \equiv \frac{\sum_{\beta\neq\alpha} \int w_\alpha(x) w_\beta(x) p^{\mathrm{eq}}(x) dx}{\int w_\alpha(x) p^{\mathrm{eq}}(x)\, dx}. \tag{10}$$

Using Eqs. (6), (7), and (8b), the *macrostate certainty* $\overline{\Upsilon}_\alpha$ is

$$\overline{\Upsilon}_\alpha \equiv 1-\Upsilon_\alpha = (p_\alpha^{\mathrm{eq}})^{-1} \int w_\alpha^2(x) p^{\mathrm{eq}}(x) dx. \tag{11}$$

We choose M so as to maximize the geometric mean of the $\overline{\Upsilon}_\alpha$ subject to the constraints of Eq. (8). This *minimum uncertainty criterion* minimizes macrostate overlap and, in the example of Fig. 1(A), results in the ϑ_α and w_α shown in panels c and d. The amount of overlap of these optimized macrostate functions depends on the magnitude of the spectral gap.

B. Adapting macrostate coarse graining to data clustering

To adapt the physical coarse graining procedure to data clustering, we make the computational analogy {microstates, macrostates, Γ} $\leftrightarrow$ {items, clusters, $f(D^{-1})$}. In this analogy, the continuous configuration space of microstates x is replaced by a discrete space of items $i: 1 \leqslant i \leqslant N$, and the probability distribution $p(x,t)$ is replaced by $\boldsymbol{p}(t)$, the vector of individual item probabilities $p_i(t)$ [e.g., see the simple example in Fig. 1(B)]. Since $\boldsymbol{p}(t)$ is a probability vector, it must satisfy

$$p_i(t) \geqslant 0 \quad \forall\ i,t, \tag{12a}$$

$$\mathbf{1}\cdot\boldsymbol{p}(t) = 1 \quad \forall\ t, \tag{12b}$$

where

$$\mathbf{1}_i = 1 \quad \forall\ i.$$

By analogy with Eq. (2), we assume that the dynamics in the item space are described by the microscopic master equation

$$\frac{d\boldsymbol{p}(t)}{dt} = \Gamma \boldsymbol{p}(t), \tag{13}$$

where Γ is a first-order transition matrix. Non-negativity of each $p_i(t)$ under time evolution requires that

$$\Gamma_{ij} \geqslant 0, \quad i \neq j \tag{14}$$

and conservation of probability requires that

$$\mathbf{1}\cdot\Gamma = 0. \tag{15}$$

The central assumption of the analogy is to assume that $\Gamma_{ij}(i\neq j)$ depends on D_{ij}, the dissimilarity between items i and j. If D were embedded as a distance matrix in a metric space (e.g., as when it is computed from a measurement matrix X), then Γ could, in principle, be computed by solving a multidimensional diffusion equation in the continuous space. However, this would be extremely expensive. Instead, we model Γ from D using the following heuristic argument: A starting point is to set $\Gamma_{ij} = (D_{ij})^{-2}$ by analogy to the rate of diffusion between two isolated microstates in one dimension. However, this does not account for the interception of probability flux by intervening items. To model interception, we use an exponential cutoff scaled to the mean nearest-neighbor squared distance $\langle d_0^2 \rangle$:

$$\Gamma_{ij} = \frac{e^{-(D_{ij})^2/2\langle d_0^2\rangle}}{(D_{ij})^2}, \quad i \neq j, \tag{16a}$$

$$\langle d_0^2 \rangle = N^{-1} \sum_{i=1}^{N} (D_{i<})^2, \tag{16b}$$

where $D_{i<}$ is the smallest element in the ith row of D. The diagonal elements of Γ are fixed by Eq. (15).

Γ defined by Eq. (16) is symmetric, so its left and right eigenvectors are identical. Therefore, Eq. (15) implies that

$$\Gamma\cdot\mathbf{1} = 0 \tag{17}$$

and the equilibrium probability vector $\boldsymbol{p}^{\mathrm{eq}}$ is

$$\boldsymbol{p}^{\mathrm{eq}} = N^{-1}\mathbf{1}. \tag{18}$$

Equation (14) and the symmetry of Γ imply that all its eigenvectors $\boldsymbol{\varphi}_n$ are orthogonal and that all its eigenvalues $-\gamma_n$ are nonpositive (see Appendix B). It is convenient to use bra-ket notation to indicate the renormalized inner product

$$\langle \boldsymbol{x} | \boldsymbol{y} \rangle \equiv N^{-1} \boldsymbol{x}\cdot\boldsymbol{y}, \tag{19}$$

and to normalize the eigenvectors so that

$$\langle \boldsymbol{\varphi}_n | \boldsymbol{\varphi}_m \rangle = \delta_{nm}. \tag{20}$$

Then,

$$\boldsymbol{\varphi}_0 = \mathbf{1}. \tag{21}$$

Figure 1(B) illustrates $\boldsymbol{\varphi}_0$ and $\boldsymbol{\varphi}_1$ computed in this way for a simple case of $N=20$ items in a one-dimensional measurement space.

Since all the elements of $\boldsymbol{\varphi}_0$ are identical, the vector analog of Eq. (7) is trivial and the macrostate distributions and window functions are directly proportional to each other. Therefore, we simplify by expressing the *m cluster window vectors* directly in terms of the m slow eigenvectors (for now we assume that m has been specified):

$$\boldsymbol{w}_\alpha = \sum_{n=0}^{m-1} M_{\alpha n} \boldsymbol{\varphi}_n. \tag{22}$$

Analogous to Eqs. (8), the $\boldsymbol{w}_\alpha$ satisfy the positivity and summation constraints required for a probabilistic interpretation,

$$(\boldsymbol{w}_\alpha)_i \geqslant 0 \quad \forall\ \alpha,i, \tag{23a}$$

$$\sum_{\alpha} \boldsymbol{w}_{\alpha} = \mathbf{1}. \tag{23b}$$

Equations (21) and (23b) and the orthonormality of the eigenvectors implies the m summation constraints on M

$$\sum_{\alpha} M_{\alpha n} = \delta_{n0}. \tag{24}$$

By analogy to Eq. (11), the certainty of cluster α is

$$\overline{\Upsilon}_{\alpha}(M) = \frac{\langle \boldsymbol{w}_{\alpha} | \boldsymbol{w}_{\alpha} \rangle}{\langle \mathbf{1} | \boldsymbol{w}_{\alpha} \rangle}. \tag{25}$$

As in the continuous case, $0 \leqslant \overline{\Upsilon}_{\alpha} \leqslant 1$. Maximizing the geometric mean of $\overline{\Upsilon}_{\alpha}$ is equivalent to minimizing the objective function

$$\Phi(M) \equiv -\sum_{\alpha} \ln \overline{\Upsilon}_{\alpha}(M). \tag{26}$$

Minimization of Φ, consistent with the linear constraints of Eq. (24) and the linear inequality constraints of Eq. (23a) fixes M, and hence $\boldsymbol{w}_{\alpha}$, for a specified value of m. The solution of this global optimization problem is discussed in Appendix A. There we show that the resultant $\boldsymbol{w}_{\alpha}$ are linearly independent, so they provide a complete basis for the macrostate subspace. Once the $\boldsymbol{w}_{\alpha}$ have been computed, we complete the clustering procedure for m by assigning each item i to the cluster α having the maximal value of $(w_{\alpha})_i$. We say that the assignment is ''strong'' or ''weak'' depending on how close this maximal value, the *item assignment strength*, is to 1. The assignments are extremely strong for the example shown in Fig. 1(B) (re panels c and d) because of the relatively large separation between the two clusters.

In some cases, the procedure may define a cluster with only a single item. In this case τ^{local} is undefined and there is no meaningful dissection of dynamics into slow and fast processes. Therefore, we treat such outliers by a special procedure. When one is identified, it is removed from the dataset and the pruned dataset is reanalyzed. The pruning procedure is repeated if new outliers appear. We designate the final clustering as $\mathcal{C}(m)$.

C. Determining the number of clusters

We use two conditions to determine if $\mathcal{C}(m)$ is an *acceptable clustering*: As motivated above, we examine the eigenspectrum of Γ for spectral gaps, which are defined as

$$\gamma_m / \gamma_{m-1} > \rho_{\gamma}, \tag{27}$$

where ρ_{γ} is the *minimum gap parameter*. However, Eq. (27) alone may accept excessively fuzzy clusters having weak item assignment vectors. To eliminate these, we supplement Eq. (27) with the *minimum macrostate certainty conditions*

$$\overline{\Upsilon}_{\alpha} > \rho_{\Upsilon} \quad \forall \ \alpha. \tag{28}$$

We have found that choosing $\rho_{\gamma} = 3$ and $\rho_{\Upsilon} = 0.68$ (the fraction of the normal distribution contained within one standard deviation of the mean) works well for all the problems that we have tested (see Sec. III).

The complete algorithm is to sequentially compute $\mathcal{C}(m)$ for $m = 2,3, \ldots$ and to test these clusterings for acceptability according to Eq. (28). If multiple clusterings are acceptable, we will usually be most interested in the one of largest m, since it provides the finest resolution. As a practical matter, if $\mathcal{C}(m)$ is not acceptable for three consecutive m's we assume that it will not be acceptable for higher m's and terminate the analysis.

D. Computational implementation

Only two steps in the procedure are potentially expensive: computing the slow eigenvectors and eigenvalues of Γ and the global minimization of $\Phi(M)$. Since we only use a relatively small number (typically $m < 10$) of slow eigenvectors, it suffices to compute only these via the Lanczos method [27] at cost $\sim O(N^2)$. The global minimization of $\Phi(M)$ is a linearly constrained global optimization problem in $m(m-1)$ dimensions. The number of vertices of the feasible region-bounding polytope increases with N, formally as a polynomial dependent on m. However, at least for the problems tested here, a simple minimization algorithm is adequate (see Appendix A).

III. RESULTS

We tested the method on a number of problems that have challenged other clustering methods. Bivariate problems in which the dataset can be graphically displayed in two dimensions were used to enable subjective evaluation of the quality of the results. In addition, to check that performance did not depend on low dimensionality of the data space, we tested problems where the items were embedded in a 20-dimensional space.

A. Bivariate test cases

The algorithm was evaluated on four previously described difficult test cases. In each case, the dataset consisted of $N_M = 2$ measurements on each of N items. These can be represented as N points in a two-dimensional space. For example, the ''crescentic'' clustering problem shown in Fig. 2(a) consists of 52 items, each represented as a point in the two-dimensional measurement space. The two clusters are closely juxtaposed crescents, which makes the problem difficult [28,2]. The D matrix was computed from the coordinates using Eq. (1) with $g_{ab} = \delta_{ab}$, and Γ was computed from D according to Eqs. (16). The slowest eigenvectors, $\boldsymbol{\varphi}_0$, $\boldsymbol{\varphi}_1$, and $\boldsymbol{\varphi}_2$, are graphically displayed in panels b, c, and d, respectively. As per Eq. (18), all components of $\boldsymbol{\varphi}_0$ are identical. It is gratifying to see that $\boldsymbol{\varphi}_1$ clearly reflects the two-cluster structure: the components of all the items in the bottom-right crescent are positive, while the components of all the items in the other crescent are negative. The next eigenvector $\boldsymbol{\varphi}_2$ has three localized regions of same-sign components. Subjectively, it is clear that separating into these

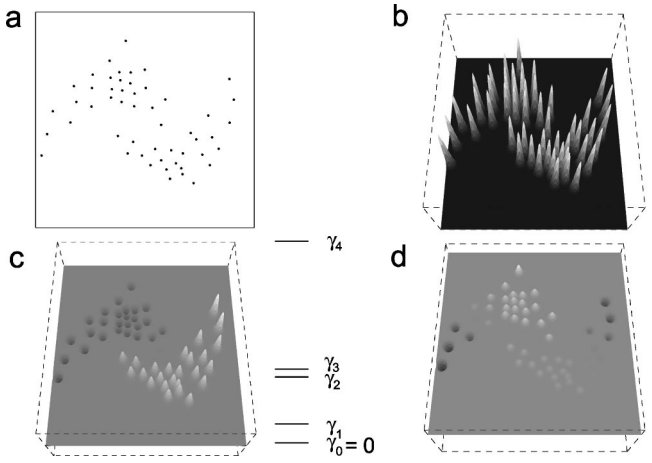

FIG. 2. ''Crescentic'' clustering problem and its slow eigenvectors. (a) The x and y coordinates of each point correspond to two measurement values of the corresponding item. (b)–(d) φ_0, φ_1, and φ_2, respectively. For illustration, the amplitude of the ith component of each φ_n is represented by the height (if positive) or depth (if negative) of a cone centered at position i. The relative magnitudes of the corresponding eigenvalues are indicated.

regions would overdissect the space. As predicted by the discussion above, these eigenvector properties in the spatial domain correspond in the time domain to an eigenspectrum gap between γ_1 and γ_2 (Fig. 2 and Table I). In contrast, there is no gap between γ_2 and γ_3 (Fig. 2). This suggests that the $m=2$ clustering, but not the $m=3$ clustering, will be acceptable.

The task for the algorithm is to recognize that the correct clustering is embedded in the structure of φ_1, and to define the proper clustering. Applying it for $m=2, 3, \ldots$ yields the clusters shown in the top panels of Fig. 3. (For illustration, we display clusterings that do not satisfy the spectral gap condition, even though these would not be computed by an efficient algorithm.) The cluster certainties $\overline{\Upsilon}_\alpha$ are listed in Table I. The $m=2$ clustering satisfies both Eqs. (27) and (28), and all clusterings with $m>2$ fail both criteria. Therefore, the algorithm correctly selects $m=2$ clusters. The individual item assignment strengths for this clustering are displayed in Fig. 4; most are in the range of $0.7-0.9$, indicating that there is significant fuzziness resulting from the close juxtaposition of the clusters. Nonetheless, all the item assignments are made correctly.

TABLE I. Crescentic cluster analysis.

m	$\frac{\gamma_m}{\gamma_{m-1}}$	$\overline{\Upsilon}_\alpha(m)$
2	3.52	0.71
		0.70
3	1.12	0.67
		0.41
		0.53
4	2.73	0.83
		0.81
		0.51
		0.53
5	1.03	0.71
		0.47
		0.55
		0.38
		0.38

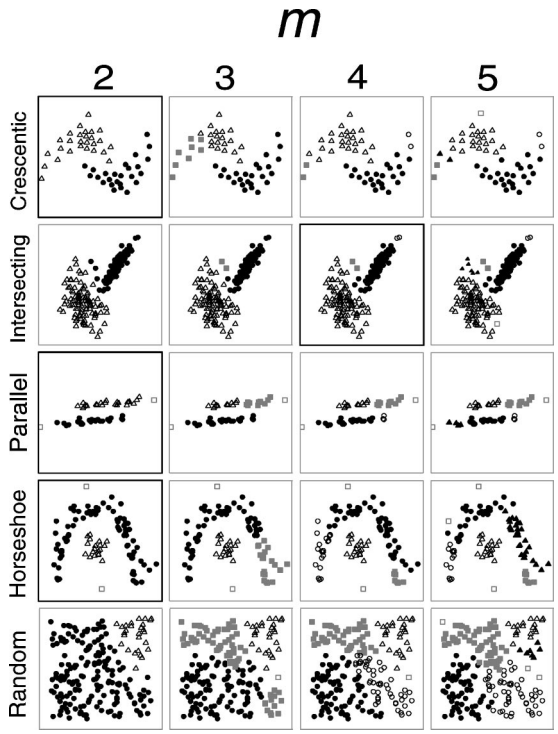

FIG. 3. Bivariate test cases. The algorithmically determined clusterings $\mathcal{C}(m)$ for $2 \leqslant m \leqslant 5$ are displayed for four bivariate examples in which the items are points in a two-dimensional measurement space. Clusters are distinguished by different symbols, except that unfilled squares identify items that were designated as outliers by the algorithm. The acceptable clusterings, which satisfy Eqs. (27) and (28), are outlined by dark boxes.

The following three test problems were analyzed in the same way.

(1) The ''intersecting'' problem consists of two barely contacting sets of items having highly anisotropic Gaussian distributions. It has previously been used to demonstrate the weakness of nonparametric optimization clustering for clusters of greatly different shapes and sizes [2].

(2) The ''parallel'' problem consists of two highly extended, anisotropic sets of items whose separation along the vertical axis is much smaller than their horizontal extent. It has previously been used to demonstrate the failure of agglomerative hierarchical methods [2].

(3) The ''horseshoe'' problem [3] consists of a central cluster of items surrounded by a horseshoe-shaped cluster of items. The center of mass of the outer cluster lies within the inner cluster, increasing difficulty. In addition, a ''random'' test set, in which points were randomly distributed within a square two-dimensional region, was analyzed as a control.

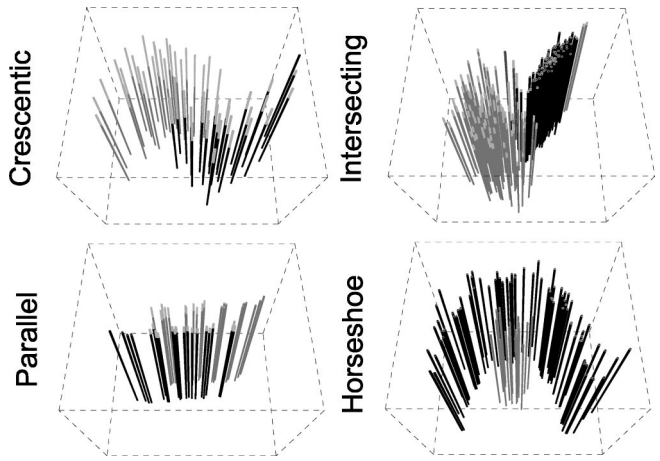

FIG. 4. Item assignment strengths for the acceptable clusterings. The acceptable clusterings for each of the problems in Fig. 3 are shown. The height of the dark section of the bar relative to its total height at the position of an item indicates its assignment strength.

The results obtained for $m=2$, 3, 4, and 5 are illustrated in Fig. 3. The acceptable clusterings that satisfy Eqs. (27) and (28) are outlined by dark boxes. Only a single clustering is acceptable in each case (although this need not be so in general). None of the random control clusterings are acceptable, correctly indicating that it should not be clustered.

As with the crescentic problem, the clustering solution for the ''horseshoe'' test-case (fourth row, Fig. 3) is straightforward, with $m=2$. Cluster certainties (Table II) and item assignment strengths (Fig. 4) are extremely strong (>0.99). The ''parallel'' problem is slightly more challenging, in that two of the items (located at the extreme left and right sides of the item distributions) are identified as outliers. Nonetheless, the algorithm correctly identifies the $m=2$ clustering of the nonoutlying items. As expected, the item assignment strengths are lower for the items in the central overlapping region, and higher for the nonoverlapping items near the left and right edges (Fig. 4).

The solution to the ''intersecting'' problem is more elaborate: While the $m=2$ solution is subjectively acceptable, the

TABLE II. Bivariate test-case analyses.

Problem	m	$\frac{\gamma_m}{\gamma_{m-1}}$	$\overline{\Upsilon}_\alpha(m)$
Crescentic	2	3.52	0.71
			0.70
Intersecting	4	3.82	0.91
			0.95
			0.84
			0.94
Parallel	2	10.68	0.93
			0.93
Horseshoe	2	60.73	0.998
			0.99
Random	1		

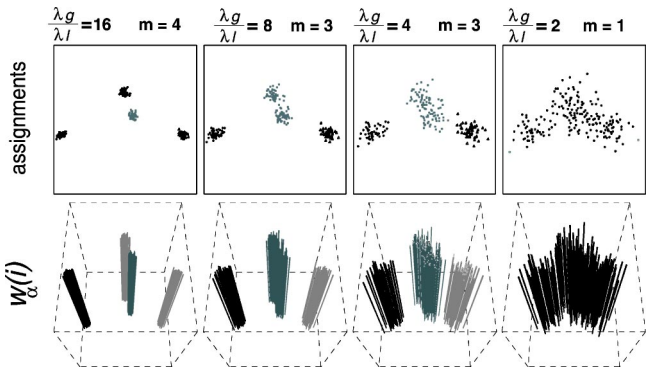

FIG. 5. Clustering of Gaussian-distributed items in two dimensions for various cluster separations. Top: the unique acceptable clustering for each value of λ_g/λ_ℓ is indicated. Bottom: the height of the dark section of the bar at the position of an item indicates its assignment strength. (Most of the strengths are ≈ 1.)

assignment strengths of some of the items in the vicinity of the intersection have weak item assignment strengths. Due to of this, the $m=2$ and $m=3$ clusterings do not satisfy the required assignment certainty condition, Eq. (28), and are rejected by the algorithm. The acceptable $m=4$ clustering resolves this difficulty by segregating these uncertain items into a separate small cluster. It also segregates two outliers (in the top-right corner) while assigning most of the items to two major clusters, as desired. The individual item assignment strengths are strong, except for one item near the intersection of the three clusters (Fig. 4).

None of the $\mathcal{C}(m)$ are acceptable for the ''random'' distribution of items because all of the γ_m/γ_{m-1} were <2.5 for $m>1$. Thus, the algorithm is not misled into spurious clustering.

B. Gaussians with varying overlap in two and 20 dimensions

We systematically tested the performance of the algorithm as a function of the relative distance between clusters. For this purpose, four pseudorandom groups of 50 items were generated with Gaussian kernels having variance λ_ℓ^2. The centers-of-mass of the groups were themselves pseudorandomly selected from a Gaussian kernel having variance λ_g^2 (see Fig. 5). The corresponding ratio of the expected root-mean-square (rms) intercluster item-item separations to the rms intracluster item separations is

$$\sqrt{\frac{\langle(\Delta R)^2\rangle_{\text{inter}}}{\langle(\Delta R)^2\rangle_{\text{intra}}}}=\sqrt{(\lambda_\ell^2+\lambda_g^2)/\lambda_\ell^2}\ . \tag{29}$$

Tests in a bivariate measurement space were conducted for λ_g/λ_ℓ varying from 16 (where the clusters were highly separated) down to 2 (where the clusters were completely overlapping). The algorithm dissects the items into four clusters when $\lambda_g/\lambda_\ell=16$. When $\lambda_g/\lambda_\ell=8$ and $\lambda_g/\lambda_\ell=4$, the top two groups partially merge (see Fig. 5), and the algorithm accordingly reports $m=3$ clusters. The clusters are not subjectively separable for $\lambda_g/\lambda_\ell=2$; correspondingly, the algo-

rithm reports $m=1$ cluster. The assignment strengths for these clusterings are displayed in Fig. 5.

The same test was performed with four groups generated as described above using Gaussian kernels in a 20-dimensional space. The increased dimensionality does not alter Eq. (29). However, the distributions of the intergroup and intragroup squared-distances are narrower: the standard deviations of the intergroup and intragroup $(\Delta R)^2$ normalized by their means are both smaller by factors of $\sqrt{20/2}=\sqrt{10}$. Therefore, for a given value of λ_g/λ_ℓ, clustering is actually easier in higher dimensionality. To compensate and make the 20-dimensional test more challenging, the range of λ_g/λ_ℓ was reduced by a factor of 4 (roughly matching $\sqrt{10}$); i.e., λ_g/λ_ℓ was varied from 4 down to 0.5. The algorithm correctly identifies the four clusters for $\lambda_g/\lambda_\ell=4$ and $\lambda_g/\lambda_\ell=2$. The individual item assignment strengths of these clusterings are displayed in Fig. 6. These are all close to one for $\lambda_g/\lambda_\ell=4$ and $\lambda_g/\lambda_\ell=2$, indicating unambiguous clustering. At smaller values of λ_g/λ_ℓ, the only clustering satisfying both the minimum gap and minimum certainty conditions has one cluster containing all the items. Even so, for $\lambda_g/\lambda_\ell=1$, the (formally unacceptable) $m=3$ clustering correctly reflects some of the group structure (Fig. 6).

IV. DISCUSSION

We have shown that macrostate clustering performs well on a variety of test problems that have challenged other methods. The method only needs a dissimilarity matrix D (not a data matrix X) and has the advantage of being nonhierarchical,[2] which should improve performance, in general. Beyond identifying potential clusterings, it uses internal criteria—the eigenspectrum gaps γ_m/γ_{m-1} and the cluster certainties $\overline{\Upsilon}_\alpha$—to determine the appropriate number of clusters. The corresponding acceptance parameters ρ_γ and ρ_Υ were empirically determined and gave robust performance—a single choice worked well for all the problems tested.

Eigenvectors have previously been used for clustering by many different spectral graph theory (SGT) partitioning methods: SGT bipartitioning methods use the values of $\boldsymbol{\varphi}_1$ to define a one-dimensional ordering of the items which can then be divided by a heuristic. A variety of different approaches have been developed to extend this to multiple eigenvectors and clusters (see Refs. [18,19,21,23] for review). For example, recursive spectral bipartitioning generates a hierarchical binary tree [22]; some methods use k eigenvectors to define 2^k clusters [29]; and many methods project the items into the subspace spanned by k eigenvectors and then use a partitioning heuristic to identify clusters within the subspace (e.g., Refs. [8,19,23,30–32], and references therein).

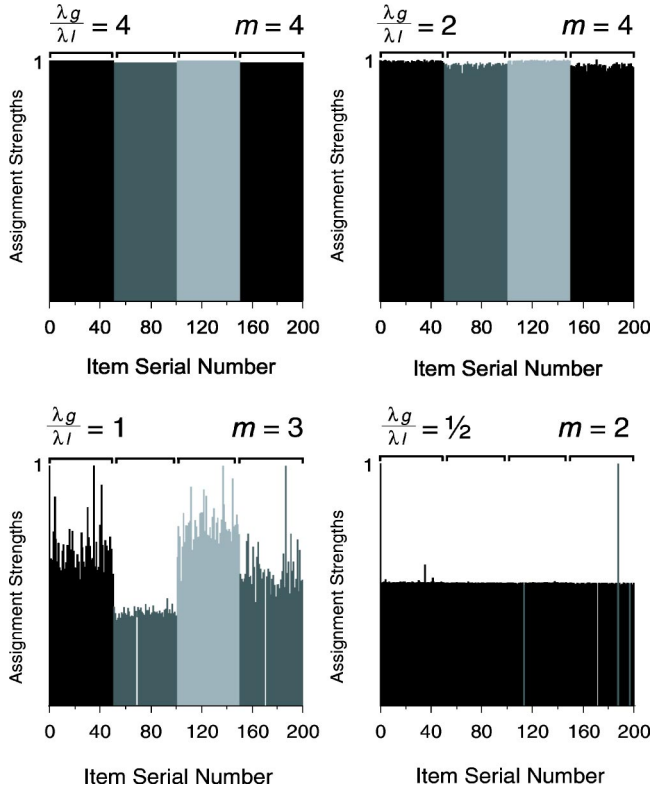

FIG. 6. Item assignment strengths for cluster solutions for various group separations in 20 dimensions. Items were pseudorandomly distributed into four groups in a 20-dimensional measurement space for different values of λ_g/λ_ℓ as described in the text. The items within each group have consecutive serial numbers (i.e., items 1–50 are in the first group, 51–100 are in the second group, etc.). Their assignment strengths for the indicated $\mathcal{C}(m)$ clusterings are displayed in each case. (Item 171 is an outlier for both clusterings shown in the bottom row; hence it is not assigned to any cluster.) However, only the $m=4$ clusterings for $\lambda_g/\lambda_\ell=4$ and $\lambda_g/\lambda_\ell=2$ are acceptable; $\mathcal{C}(3)$ and $\mathcal{C}(2)$ shown in the bottom panels fail the acceptability conditions of Eqs. (27) and (28) because of their low cluster certainties.

Macrostate clustering is different because it computes continuous (fuzzy) assignment window vectors rather than partitionings.[3] This has important ramifications: It permits the window vectors to be expressed as linear combinations of the eigenvectors [see Eq. (22)]; this necessarily results in window function overlap and cluster uncertainty. Combining these concepts with the principle of uncertainty minimization provides a simple prescription for the concurrent use of multiple eigenvectors in clustering. A related difference is that the number of clusters is internally determined in macrostate clustering, while it is usually fixed *a priori* or determined by eigensystem-independent heuristics in SGT methods (e.g.,

[2]For example, the $m=5$ "crescentic" clustering cannot be obtained by subdividing its $m=4$ clustering and the $m=4$ "horseshoe" clustering is not hierarchically related to its $m=3$ clustering. Nevertheless, inherent hierarchical structure can still emerge, and some was evident in all the problems. For example, all the clusterings for $2\leqslant m\leqslant 5$ for the "intersecting" and "parallel" problems are hierarchically related (Fig. 3).

[3]Drineas *et al.* [6] consider real-valued "generalized clusters" within a SGT context, but these are indefinite and do not have a probabilistic interpretation.

Refs. [18,19,23], and references therein). It is perhaps surprising that the spectral gap condition has not been used for this purpose in SGT approaches.[4] This may reflect the fact that it does not work well by itself, and the companion minimum cluster certainty condition is not available when (crisply) partitioning. Macrostate and SGT clustering also differ in the manner in which Γ (or the SGT analog) is computed from the dissimilarity matrix D. SGT methods typically use a weight matrix equivalent to $\Gamma_{ij}=\exp(-D_{ij}/\sigma)$, $i \neq j$, where σ is an empirically-determined scale constant. In contrast, motivated by the analogy to a diffusive system, we used Eqs. (16). While this difference is not of fundamental significance, the relationship between Γ and D can affect performance. Thus, it may be helpful to test the use of Eqs. (16) in SGT methods or the SGT relationship in macrostate clustering.

The use of a linear transformation from indefinite, orthogonal eigenvectors to semidefinite, nonorthogonal window vectors is fundamental, but some freedom remains in the choice of the objective function used to determine the optimal transformation and in the conditions used to determine acceptable clusterings. An uncertainty minimization criterion is a natural choice, since it is (in an information-theoretic sense) the entropic counterpart to the (implicit) ''energy'' minimization criterion that focuses attention on the slow eigenvectors (see Sec. II C of Ref. [26]). On the other hand, the definition of uncertainty could be modified and tested for improved performance. Similarly, while we believe that it is advantageous to combine energetic (spectral gap) and entropic (cluster certainty) conditions in determining the number of clusters, it may be possible to improve upon the specific criteria used here.

Other improvements and extensions requiring attention are the following.

(1) While we accepted or rejected each clustering *in toto*, it may be useful in some cases to examine incomplete clusterings in which only some of the clusters satisfy the cluster certainty condition. This modification would enable the algorithm to resolve all four clusters for the case of $\lambda_g/\lambda_\ell=8$ in Fig. 5.[5]

(2) The individual item assignment strengths $(\boldsymbol{w}_\alpha)_i$ measure the certainty of each assignment, but their precise statistical significance is not known. It would be helpful to have a model for assessing this.

(3) The *cluster transition matrix* $\gamma_{\beta\alpha}=\langle \boldsymbol{w}_\beta|\Gamma|\boldsymbol{w}_\alpha\rangle$ can be used to assess the strength of the relationship between the clusters and may be useful in setting the cluster acceptance criteria.

(4) We have defined Γ as a symmetric matrix, which implies that $\boldsymbol{p}^{\rm eq}\propto\mathbf{1}$. However, this restriction is not required: The generalization to asymmetric Γ is straightforward [24] and it could be used to incorporate additional experimental information. For example, if item i is known *a priori* to be partially redundant with other items (e.g., when analyzing expression levels of members of gene families), it may be given reduced weight in the analysis by setting $\boldsymbol{p}_i^{\rm eq}<1$.

Our main goal has been a proof-of-principle demonstration of the high quality of the clusterings provided by the dynamical macrostate approach. The current implementation is sufficiently efficient for problems where $N\sim O(10^2)$, but we have not examined performance for very large problems. The continuous formulation replaces the nonpolynomial-hard (NP-hard) combinatoric SGT partitioning problem with a global minimization problem having polynomial complexity in N. However, the order of the polynomial can be very large for large m (Appendix A) so, formally, this is not much of an improvement. Nonetheless, as discussed in Appendix A, since the objective function is smooth and the constraints are highly degenerate, a simple minimization algorithm has worked well and we believe that it will be possible to obtain adequate approximate solutions efficiently, even for very large problems. This remains to be examined.

ACKNOWLEDGMENTS

We thank Bruce Church, Jason Gans, Ron Elber, Jon Kleinberg, and Golan Yona for helpful conversations and the NSF (Grant No. CCR9988519) and the NIH (training Grant No. T32GM08267) for support.

APPENDIX A: MINIMIZING $\Phi(M)$

$\Phi(M)$ is to be minimized as a function of the m^2 elements of $M_{\alpha n}$ within the feasible region specified by the $m \times N$ linear inequality constraints of Eq. (23a). The rows of M can be regarded as the coordinates of m particles in the m-dimensional space of the slow eigenvectors. Designating the coordinate row vector of particle α as $\vec{M}_\alpha=(M_{\alpha 0},M_{\alpha 1},\ldots M_{\alpha(m-1)})$, M is the outer product of the $\vec{M}_\alpha$'s:

$$M=\bigotimes_\alpha \vec{M}_\alpha . \tag{A1}$$

The equality constraints of Eq. (24) imply that the center of mass of the m particles is at position

$$\frac{1}{m}\sum_\alpha \vec{M}_\alpha=\frac{\hat{\varepsilon}_0}{m}, \tag{A2}$$

where $\hat{\varepsilon}_0$ is the unit vector in the zeroth direction:

$$\hat{\varepsilon}_0=(1,0,\ldots,0). \tag{A3}$$

[Equation (A3) must be modified when there is more than one stationary eigenvector; see Appendix B.] The feasible region is a polytope in the $m(m-1)$-dimensional subspace where Eq. (A2) is satisfied.

[4]However, spectral gaps have been used heuristically to determine the appropriate dimensionality of singular subspaces in data mining [33].

[5]The $m=5$ solution identifies the four major clusters with strong certainty, but also groups three items (located near the boundary between the two top clusters) into a fifth cluster which has $\overline{\Upsilon}_\alpha<\rho_\Upsilon$. In an incomplete clustering, all but these three items would be unambiguously assigned.

The minimum of $\Phi(M)$ must lie at a vertex of this polytope.

Proof. The gradient of Φ with respect to $\vec{M}_\alpha$ is

$$\vec{\nabla}_\alpha \Phi \equiv \frac{\delta\Phi}{\delta\vec{M}_\alpha} = -2\frac{\vec{M}_\alpha}{|\vec{M}_\alpha|^2} + \frac{\hat{\varepsilon}_0}{\vec{M}_\alpha \circ \hat{\varepsilon}_0} \tag{A4}$$

and the Hessian is

$$\vec{\nabla}_\alpha \otimes \vec{\nabla}_\beta \Phi \equiv \frac{\delta^2\Phi}{\delta\vec{M}_\alpha \delta\vec{M}_\beta} = -\delta_{\alpha\beta}\left[\frac{2I}{|\vec{M}_\alpha|^2} - 4\frac{\vec{M}_\alpha \otimes \vec{M}_\alpha}{|\vec{M}_\alpha|^4} + \frac{\hat{\varepsilon}_0 \otimes \hat{\varepsilon}_0}{(\vec{M}_\alpha \circ \hat{\varepsilon}_0)^2}\right], \tag{A5}$$

where I is the $m \times m$ identity matrix and $\circ$ denotes the inner product over the eigenvector indices,

$$\vec{x} \circ \vec{y} \equiv \sum_{n=0}^{m-1} x_n y_n .$$

The gradient does not vanish anywhere, so Φ has no minimum in the absence of constraints.

In fact, a minimum will occur only when *all* m^2 degrees of freedom are constrained by the m equality constraints and $m(m-1)$ inequality constraints. To see this, consider the situation without the equality constraints, but with some number $c \leqslant m(m-1)$ of active inequality constraints. Each active inequality constraint acts (identified by item index i) on a single $\boldsymbol{w}_\alpha$, so by Eq. (22) it acts on a single $\vec{M}_\alpha$ to enforce

$$\vec{M}_\alpha \circ \vec{\boldsymbol{\varphi}}_i = 0, \tag{A6}$$

where $\vec{\boldsymbol{\varphi}}$ is the supervector having components $(\boldsymbol{\varphi}_0, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{m-1})$. Therefore, the inequality constraints are separable and, similar to Eq. (A1), the space of inequality constrained M's can be expressed as the outer product of the subspaces of inequality constrained $\vec{M}_\alpha$'s. Thus, if $M^c = \otimes_\alpha \vec{M}^c_\alpha$ is an inequality-constrained minimizer of Φ, it must be stable with respect to independent variations of each of the inequality constrained $\vec{M}^c_\alpha$. However, this is not possible: For any such variation $\vec{M}^c_\alpha \to \vec{M}^c_\alpha + \vec{\delta}_\alpha$, the existence of a minimum would require that

$$\vec{\delta}_\alpha \circ \vec{\nabla}_\alpha \Phi = 0 \tag{A7}$$

and

$$\vec{\delta}_\alpha \circ (\vec{\nabla}_\alpha \otimes \vec{\nabla}_\alpha)\Phi \circ \vec{\delta}_\alpha > 0. \tag{A8}$$

However, Eqs. (A4) and (A7) imply that

$$\frac{\vec{M}_\alpha \circ \vec{\delta}_\alpha}{|\vec{M}_\alpha|^2} = \frac{\vec{\delta}_\alpha \circ \hat{\varepsilon}_0}{2\vec{M}_\alpha \circ \hat{\varepsilon}_0},$$

and combining this with Eq. (A5) implies that

$$\vec{\delta}_\alpha \circ (\vec{\nabla}_\alpha \otimes \vec{\nabla}_\alpha)\Phi \circ \vec{\delta}_\alpha = -\frac{2|\vec{\delta}_\alpha|^2}{|\vec{M}_\alpha|^2} < 0.$$

Thus, Eqs. (A7) and (A8) cannot both be true. Therefore, a minimum can occur only if *all* variations of the $\vec{M}_\alpha$ are prevented by a combination of inequality and equality constraints. Since there are only m equality constraints, we must have $c = m(m-1)$ active inequality constraints. This corresponds to a vertex of the feasible region.

Note also that the minimizing $\{\vec{M}^c_\alpha\}$ must be linearly independent within the m-dimensional slow eigenvector space. This implies that the associated $\{\boldsymbol{w}_\alpha\}$ must span the macrostate subspace.

Proof. If the $\{\vec{M}^c_\alpha\}$ are not independent, there would exist a linear combination of vectors such that

$$\sum_\alpha \xi_\alpha \vec{M}^c_\alpha = 0.$$

Then, the combined variation

$$\vec{M}^c_\alpha \to \vec{M}^c_\alpha + \delta\xi_\alpha \vec{M}^c_\alpha, \quad \forall\,\alpha,$$

where δ is a small number, will not affect the equality constraint, Eq. (A2). As proven above, all the components of the constrained minimum must be fixed by constraints, so this variation must be excluded by an inequality constraint. However, this variation only rescales each $\vec{M}^c_\alpha$ and hence each $\boldsymbol{w}_\alpha$. Therefore, it also will not affect the inequality constraints and is permitted, contrary to assumption. *Reductio ad absurdum.*

To find the vertex with the lowest value of Φ, we used a simple minimizer that operates in the $m(m-1)$-dimensional subspace that remains after one of the M_α has been explicitly eliminated using Eq. (A2). The minimizer starts from $\vec{M}_\alpha = \hat{\varepsilon}_0/m \quad \forall\alpha$ chooses a random direction in the $m(m-1)$-dimensional space, proceeds to the nearest inequality constraint, and then proceeds along faces of the feasible region (of decreasing dimensionality) until a vertex is reached. This process was repeated until the same extremal minima was found three times or for a minimum of 500 000 trials, whichever was greater.

Accounting for the separability of the inequality constraints and assuming no degeneracies between the values of the $\vec{\varphi}_n$ (the usual case), the number of vertices of the constraining polytope might grow as rapidly as $O(N^m)$. However, we expect that most of the inequality constraints of Eq. (23a) will be almost degenerate because of the relatively small differences between the components of the eigenvectors at different items within a cluster. Moreover, the objective function Φ is smooth, so we expect that the variation of Φ over nearby vertices will be small. Therefore, it will not affect $\boldsymbol{w}_\alpha$ much if a neighbor, rather than the global minimizer itself, is found. Thus, we anticipate that the practical growth in computational cost with N will be much less than the worst-case bound. These considerations also suggest that it will always be advantageous to use solvers that move

through the $[m(m-1)-1]$-dimensional space of search-space directions rather than between vertices of the constraining polytope.

APPENDIX B: DEGENERATE "ZERO" EIGENVALUES

Since Γ is a symmetric matrix that satisfies Eqs. (15) and (17),

$$-\boldsymbol{x}\cdot\Gamma\cdot\boldsymbol{x}=\sum_{\substack{j>i\\ i}}\Gamma_{ij}(x_i-x_j)^2 \tag{B1}$$

for any vector $\boldsymbol{x}$. The right-hand side (rhs) can be viewed as the potential energy of N particles having pairwise quadratic interactions in one dimension. Since all the off-diagonal elements of Γ are positive, the rhs must be non-negative. The implied nonpositivity of $\boldsymbol{x}\cdot\Gamma\cdot\boldsymbol{x}$ for all $\boldsymbol{x}$ implies that all the eigenvalues of Γ must be nonpositive. Furthermore, the isomorphism makes it evident that $\boldsymbol{x}=\mathbf{1}$ is the only stationary eigenvector (up to a multiplicative constant) unless the dataset contains an *isolated subset* $\mathcal{S}$, which has $\Gamma_{ij}=0$ if $i \in \mathcal{S}$ and $j \notin \mathcal{S}$. In this case, Γ will have multiple zero eigenvalues, and there will be one stationary eigenvector corresponding to each isolated subset. This degeneracy can be removed by analyzing each isolated subset independently.

It is more common to encounter approximate isolation in which none of the Γ_{ij} is exactly zero but in which there are multiple small eigenvalues that are 0 on the scale of numerical accuracy. (This occurs in the Gaussian clustering problem shown in Fig. 5 when λ_g/λ_ℓ is large.) This can cause numerical problems: $\vec{\varphi}_0$ returned by a numerical eigensystem solver will not necessarily satisfy Eq. (21), but instead will be a linear combination of the approximately degenerate eigenvectors. Due to this, Eq. (21), and hence Eq. (24), may not be true.

The simplest resolution of this numerical problem is to replace Eq. (24) with Eq. (A2) and to replace Eq. (A3) with

$$\hat{\varepsilon}_0=\langle\mathbf{1}|\vec{\varphi}\rangle. \tag{B2}$$

This does not require the numerical validity of Eq. (21).

[1] B. Mirkin, *Mathematical Classification and Clustering* (Kluwer Academic, Boston, 1996).

[2] B. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, 4th ed. (Arnold, London, 2001), pp. 1–10.

[3] Z. Szallasi, in *Proceeding of the Second International Conference on Systems Biology* (California Institute of Technology, Pasadena, CA, 2001), URL http://www.icsb2001.org/SzallasiTutorial.pdf

[4] A. Jain and R. Dubes, *Algorithms for Clustering Data* (Prentice-Hall, Englewood Cliffs, NJ, 1981).

[5] R.B. Altman and S. Raychaudhuri, Curr. Opin. Struct. Biol. **11**, 340 (2001).

[6] P. Drineas, R. Kannan, A. Frieze, S. Vempala, and V. Vinay, in *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms* (ACM Press, New York, 1999), URL http://doi.acm.org/10.1145/314500.314576

[7] O. Alter, P.O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. U.S.A. **97**, 10101 (2000).

[8] R. Kannan, S. Vempala, and A. Vetta, in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, 2001), pp. 367–377, URL http://citeseer.nj.nec.com/495691.html

[9] G. Milligan and M. Cooper, Psychometrika **50**, 159 (1985).

[10] A.D. Gordon, in *Data Science, Classification and Related Methods*, edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Springer-Verlag, Tokyo, 1998), pp. 22–39.

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

[12] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).

[13] S. Wiseman, M. Blatt, and E. Domany, Phys. Rev. E **57**, 3767 (1998).

[14] L. Kullmann, J. Kertesz, and R.N. Mantegna, Physica A **287**, 412 (2000).

[15] L. Giada and M. Marsili, Phys. Rev. E **63**, 061101 (2001).

[16] L. Angelini, F. DeCarlo, C. Marangi, M. Pellicoro, and S. Stramaglia, Phys. Rev. Lett. **85**, 554 (2000).

[17] D. Horn and A. Gottlieb, Phys. Rev. Lett. **88**, 18702 (2002).

[18] A.J. Seary and W.D. Richards, in *Proceedings of the International Conference on Social Networks*, edited by M.G. Everett and K. Rennel (European Network Conference, London, 1996), Vol. 1, pp. 47–58, URL http://www.sfu.ca/richards/Pdf-ZipFiles/london98.pdf

[19] Y. Weiss, in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, edited by B. Werner (IEEE Computer Society, Los Alamitos, CA, 1999), Vol. II, pp. 975–982, URL http://citeseer.nj.nec.com/weiss99segmentation.html

[20] F.R.K. Chung, *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics Vol. 92 (American Mathematical Society, Providence, RI, 1997).

[21] D.A. Spielman and S.-H. Teng, in *Proceedings of the 37th Annual Symposium on Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, CA, 1996), pp. 96–105, URL http://citeseer.nj.nec. com/spielman96spectral. html

[22] S.T. Barnard and H.D. Simon, Concurrency: pract. ex. **6**, 101 (1994).

[23] C.J. Alpert, A.B. Kahng, and S.-Z. Yao, Discrete Appl. Math. **90**, 3 (1999).

[24] D. Shalloway, J. Chem. Phys. **105**, 9986 (1996).

[25] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II :Nonequilibrium Statistical Mechanics* (Springer-Verlag, New York, 1985).

[26] A. Ulitsky and D. Shalloway, J. Chem. Phys. **109**, 1670 (1998).

[27] G.H. Golub and C.F. VanLoan, *Matrix Computations*, 2nd ed. (Hopkins University Press, Baltimore, MD, 1989).

[28] M. Wong and T. Lane, J. R. Stat. Soc. Ser. B. Methodol. **45**, 362 (1983).

[29] B. Hendrickson and R. Leland, in *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing* (SIAM, Philadelphia, PA, 1993), pp. 953–961, URL http://citeseer.nj.nec.com/hendrickson93multilevel.html

[30] A. Pothen, H.D. Simon, and K.P. Liou, SIAM J. Matrix Anal. Appl. **11**, 430 (1990).

[31] M. Meila and J. Shi, in *Advances in Neural Information Processing Systems 13*, edited by T. Leen, T.G. Dietterich, and V. Tresp (MIT Press, Cambridge, MA, 2001), pp. 873–879, URL http://citeseer.nj.nec.com/meil01learning.html

[32] A. Ng, M. Jordan, and Y. Weiss, in *Advances in Neural Information Processing Systems 14*, edited by T.G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, MA, 2001), URL http://citeseer.nj.nec.com/ng01spectral.html

[33] Y. Azar, A. Fiat, A.R. Karlin, F. McSherry, and J. Saia, in *Proceedings of the ACM Symposium on Theory of Computing* (ACM Press, New York, 2001), pp. 619–626, URL http://citeseer.nj.nec.com/azar00spectral.html