On the use of the GTM algorithm for mode detection

Susana Vegas-Azcárate and Jorge Muruzábal
Statistics and Decision Sciences Group
University Rey Juan Carlos, 28936 Móstoles, Spain
s.vegas@escet.urjc.es, j.muruzabal@escet.urjc.es

Abstract. The problem of detecting the modes of the multivariate continuous distribution generating the data is of central interest in various areas of modern statistical analysis. The popular self-organizing map (SOM) structure provides a rough estimate of that underlying density and can therefore be brought to bear with this problem. In this paper we consider the recently proposed, mixture-based generative topographic mapping (GTM) algorithm for SOM training. Our long-term goal is to develop, from a map appropriately trained via GTM, a fast, integrated and reliable strategy involving just a few key statistics. Preliminary simulations with Gaussian data highlight various interesting aspects of our working strategy.

Key words: density estimation, mixture models, latent variables, generative distributions, self-organization.

1 Introduction

Searching the structure of an unknown data-generating distribution is one of the main problems in statistics. Since modes are among the most informative features of the density surface, identifying the number and location of the underlying modal groups is of crucial importance in many areas (such as, e.g., the Bayesian MCMC approach).

A previously proposed idea is to transform mode detection into a mixture problem following either a parametric [9] or a non-parametric [15] approach. Studies based on kernel estimation [7, 16, 18] have also provided useful results. While some methods for testing the veracity of modes in $1D$ [8, 15] and 2D [16] data are available, tools for making exploration possible in higher dimensions are much needed.

Here we inquire about the potential of *self-organizing maps* (SOM) [11] as an approach to multivariate mode detection. The SOM structure derived from the standard fitting algorithm is often found useful for clustering, visualization and other purposes. However, it lacks a statistical model for the data. Recent approaches to SOM training usually incorporate some statistical notions yielding richer models and more principled fitting algorithms.

Specifically, the *generative topographic mapping* (GTM) [1] is a non-linear latent variable model (based on a constrained mixture of Gaussians) in which its parameters can be determined by maximum likelihood via the *expectation-maximization* (EM) algorithm [6]. GTM attempts to combine the topology-preserving trait of SOM structures with a well-defined probabilistic foundation.

The GTM approach provides a number of appealing theoretical properties and is deemed indeed a major candidate to support our mode detection task.

The remainder of the paper is organized as follows. In section 2 we explore basic properties of the GTM model and discuss the reasons for preferring GTM over the original SOM fitting algorithm to deal with multivariate mode detection. Section 3 first presents the basic tools or key statistics extracted from the trained GTM, then sketches our working strategy for mode detection. A basic test of this tentative strategy is provided in section 4, where 2D, 3D and 10D synthetic data following various Gaussian distributions are used to illustrate the prospect of the approach. Finally, section 5 summarizes some conclusions and suggests future research stemming from the work discussed in the paper.

2 The GTM model

The GTM [1] defines a non-linear, parametric mapping $\mathbf{y}(\mathbf{x}, \mathbf{W})$ from an L-dimensional latent space ($\mathbf{x} \in \Re^L$) to a D-dimensional data space ($\mathbf{y} \in \Re^D$) where L < D. The transformation $\mathbf{y}(\mathbf{x}, \mathbf{W})$ maps the latent-variable space into an L-dimensional manifold S embedded within the data space. By suitably constraining the model to a grid in latent space, a posterior distribution over the latent grid is readily obtained for each data point using Bayes' theorem.

As often acknowledged [11], the standard SOM training algorithm suffers from some shortcomings: the absence of a cost function, the lack of a theoretical basis to ensure topographic ordering, the absence of any general proofs of convergence, and the fact that the model does not define a probability density. GTM proceeds by optimizing an objective function via the EM algorithm [6]. Since the cost function is of log-likelihood type, a measure is provided on which a GTM model can be compared to other generative models.

But our main interest in working with GTM algorithm instead of the original SOM fitting algorithm to deal with multivariate mode detection is related to the 'self-organization' and 'smoothness' concepts. While the conditions under which the self-organization of the SOM occurs have not been quantified and empirical confirmation is needed in each case, the neighbourhood-preserving nature of the GTM mapping *is an automatic consequence of the choice of a continuous function* $\mathbf{y}(\mathbf{x}, \mathbf{W})$ [2]. In the same way, the smoothness properties of the original SOM are difficult to control since they are determined indirectly by the neighbourhood function, while basis functions parameters of the GTM algorithm explicitly govern the smoothness of the manifold, see below.

Hence, the GTM algorithm seeks to combine the topology preserving properties of the SOM structure with a well defined probabilistic framework. Moreover, since the evaluation of the Euclidean distances from every data point to every Gaussian centre is the dominant computational cost of GTM and the same calculations must be done for Kohonen's SOM, each iteration of either algorithm takes about the same time.

More specifically, GTM training is based on an optimization procedure aimed at the standard Gaussian mixture log-likelihood [2]

$$\ell(\mathbf{W},\beta)=\sum_{n=1}^{N}\ln\{\frac{1}{K}\sum_{i=1}^{K}(\frac{\beta}{2\pi})^{\frac{D}{2}}\exp\{-\frac{\beta}{2}\parallel\mathbf{W}\phi(\mathbf{x}_i)-\mathbf{t}_n\parallel^2\}\},$$

where a generalized linear regression model is chosen for the embedding map, namely

$$\mathbf{y}(\mathbf{x},\mathbf{W})=\mathbf{W}\phi(\mathbf{x}),\text{with }\phi_m(\mathbf{x})=\exp\{-\frac{\parallel\mathbf{x}-\mu_m\parallel^2}{2\sigma^2}\},\ m=1,...,M.$$

Here $\mathbf{t}_n$ is one of N training points in the D-dimensional data space, $\mathbf{x}_i$ is one of K nodes in the regular grid of a L-dimensional latent space, β is the inverse noise variance of the mixture Gaussian components, $\mathbf{W}$ is the DxM matrix of parameters (weights) that actually govern the mapping and $\phi(\mathbf{x})$ is a set of M fixed (spherical) Gaussian basis functions with common width σ. Some guidelines to deal with GTM training parameters are laid out in [1–3, 17].

3 Basic tools

Once we have trained a GTM model, suitable summaries of its structure will be extracted and analyzed to ascertain the modes' location. In particular, we consider *median interneuron distances*, *dataloads*, *magnification factors* and *Sammon's projections*. The first three quantities are introduced in detail in the next sections.

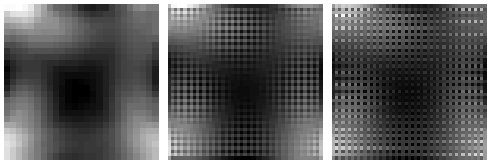
To visualize high-dimensional SOM structures, use of Sammon's projection [14] is customary. Sammon's map provides a useful global image while estimating all pairwise Euclidean distances among SOM pointers and projecting them directly onto 2D space. Thus, since pointer concentrations in data space will tend to be maintained in the projected image, we can proceed to identify high-density regions directly on the projected SOM. Furthermore, by displaying the set of projections together with the connections between immediate neighbours, the degree of self-organization in the underlying SOM structure can be expressed intuitively in terms of the amount of *overcrossing* connections. This aspect of the analysis is rather important as the main problem with the SOM structure, namely *poor organization*, needs to be controlled somehow. As usual, it is crucial to avoid poorly-organized structures (whereby immediate neighbours tend to be relatively distant from each other) but this goal is not so easy when working with high-dimensional data [10, 11]. On this matter, Kiviluoto and Oja [10] suggested the use of PCA-initialization instead of random-initialization to obtain suitably organized GTM structures; they also proposed the *S-MAP* method combining GTM and SOM. In addition, since it is not clear how much organization is possible for a given data set, the amount of connection overcrossing lacks an absolute scale for assessment. On the other hand, if overcrossing in Sammon's projection plot is (closely) null, we can proceed with some confidence.

3.1 Median Interneuron Distances

Since we are interested in regions with higher pointer (or gaussian centre) density, the relative distance from each pointer to its immediate neighbours on the network will provide a useful bit of information. The inspection of pointer interdistances was pioneered by Ultsch [19], who defined the *unified-matrix* (U-matrix) to visualize Euclidean distances between reference vectors in Kohonen's SOM. Emphasis in the U-matrix is on cluster analysis.

Although modes may be associated with clusters [7], a problem exists with high-dimensional data: "while one cluster might be well-represented by a single multivariate Gaussian, another cluster may required dozens of Gaussian components to capture skewness and yet still be unimodal" [16]. So the relationship between the number of modes and the number of mixture components is not straightforward. In particular, when dealing with multivariate data the mixture may have more modes than mixture components [16].

In this paper, we will work with the alternative *median interneuron matrix* (MID-matrix) proposed in [12]. This is a $\sqrt{K}\mathrm{x}\sqrt{K}$ matrix whose (i, j) entry, is the *median* of the Euclidean distances between the gaussian centre and all pointers belonging to a star-shaped, fixed-radius neighbourhood containing typically eight units. To facilitate the visualization of higher pointer concentrations, a linear transformation onto a 256-tone gray scale is standard (here the lower the value, the darker the cell). Figure 1 compares the MID-matrix to two variants of the U-matrix. It appears that the latter images provide a more blurry picture regarding pointer concentration.



a b c

Fig. 1. Data comes from a 2D Gaussian density; (a) MID-matrix showing a higher pointer concentration around the true mode; (b) U-matrix with rectangular topology; (c) U-matrix with hexagonal topology.

3.2 Dataloads

The number of data vectors projecting onto each unit, namely the pointer *dataload* $\widehat{\pi}(i, j)$, is the natural estimate of the weight $\pi(i, j)$ obtained from the true underlying distribution f [4],

$$\pi(i,j) = \int_{V(i,j)} f(\mathbf{x}) d(\mathbf{x}),$$

where $V(i,j)$ collects all input vectors which are closest to unit (i,j). Again, to easily visualize the dataload distribution over the map, a similar gray image is computed, namely, the DL-matrix. Note that, in this case, darker means higher.

It is important to realize that the "density" of pointers in the trained GTM should serve as an estimate of the density underlying the data. In this ideal case, each neuron would cover about the same proportion of data, that is, a uniform DL-matrix should be obtained. Another interesting way to deal with mode detection and density estimation is to obtain uniformly distributed pointers (over the observed range). Now neurons will present markedly different dataloads, higher densities relating intuitively to larger dataloads. Throughout this paper we focus on the first approach, in which mode detection is based on different pointer concentrations. However, the second idea also seems feasible and is under study.

3.3 Magnification Factors

Since the usual Kohonen algorithm tends to underestimate high probability regions and overestimate low probability areas [3], a concept to express the *magnification* between data and pointer density was needed. A basic theorem related to 1D data was presented by Ritter and Schulten [13]. These authors demonstrated that the limiting pointer density is proportional to the data density raised to a *magnification factor* $\kappa = \frac{2}{3}$. This theorem was later complemented by Bodt, Verleysen and Cottrell [4], who recalled a result from *vector quantization* (VQ) theory, but we shall not be concerned with this problem in the present paper.

Thanks to the topology preserving properties of the GTM, nearby points in latent space will map to nearby points in data space. The concept of *magnification factors* came to represent how the topological map was being locally stretched and compressed when embedded in data space [3]. In the context of the original version of the SOM the topological map is represented in terms of a discrete set of reference vectors, so that a non-continuous expression of magnification factors can also be obtained.

We have focused our study on the GTM model, where local magnification factors can be evaluated as continuous functions of the latent space coordinates in terms of the mapping $\mathbf{y}(\mathbf{x}, \mathbf{W})$ [3]. This constitutes one of our main motivations for working with GTM. We obtain the magnification factors as

$$\frac{dV_y}{dV_x} = \sqrt{\det(\mathbf{J}\mathbf{J}^T)} =_{(L=2)}= \sqrt{\| \frac{\partial \mathbf{y}}{\partial x^1} \|^2 \| \frac{\partial \mathbf{y}}{\partial x^2} \|^2 - (\frac{\partial \mathbf{y}}{\partial x^1} \frac{\partial \mathbf{y}}{\partial x^2})^2}$$

where V_x is the volume of an infinitesimal L-dimensional (L=2) hypercuboid in the latent space and V_y is the volume of its image in the data space. Here $\mathbf{J}$ is the Jacobian of the mapping $\mathbf{y}(\mathbf{x}, \mathbf{W})$ and $\frac{\partial \mathbf{y}}{\partial x^l}$ are the partial derivatives of the mapping $\mathbf{y}(\mathbf{x}, \mathbf{W})$ with respect to the latent variable $x^l = (x_1^l, x_2^l)$.

Using techniques of differential geometry it can be shown that the *directions* and *magnitudes* of stretch in latent space are determined by the eigenvectors and eigenvalues of $\mathbf{J}\mathbf{J}^T$ [3]. Only magnitudes of stretch are considered in this paper. Again, a linear transformation on a gray-scale matrix is presented to visualize

the magnitude of that compression, namely the MF-matrix (as in MID-matrix case, darker means lower). Mode detection analysis based on stretch orientations is postponed for future work.

3.4 A strategy for mode detection

The above summaries of the GTM's self-organizing structure constitute the basis of the following scheme for exploring mode estimation.

1. Train a GTM model [17] until the Sammom's projected pointers show a good level of organization and a (nearly) uniform DL-matrix is obtained. Check also the stability of log-likelihood values.
2. Compute the MID and MF matrices. If the existence of more than one mode is suggested, build subsets of data and return to STEP1 to fit individual GTMs at each unimodal subset.
3. Combining MID-matrix's darkest region, pointer concentration on Sammon's projection and stretch location displayed in the MF-matrix, an approximation to the single mode's location can be performed.

This completes our theoretical presentation. We are now ready to examine some empirical evidence.

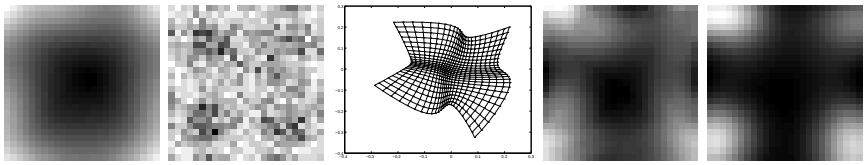


Fig. 2. Data comes from a $N_2(\mathbf{0}, 0.1\mathbf{I}_2)$ density; training parameters are specified in the text; (a) DPO-matrix, showing a map well centered around the mode; (b) DL-matrix, note a tendency towards darker corners; (c) trained map, denser strips are clearly visible; (d) MID-matrix, slightly biased with respect to a); (e) MF-matrix, much coincident with d).

4 Simulation study

As an illustration of the strategy for mode detection presented above, synthetic data samples from several Gaussian distributions with spherical covariance matrices (and often centered at the origin) are considered. Each training set is of size 2500 unless otherwise noted. Since 2-dimensional data allow to examine the GTM directly, a straightforward single Gaussian is tried out first to verify that pointer concentration occurs and can be detected by our summary matrices and

statistics, see Figure 2. For the sake of references, we have also compared, in our 2D cases, the final distribution obtained via the GTM,

$$\frac{1}{K}\sum_{i=1}^{K}(\frac{\beta^*}{2\pi})^{\frac{D}{2}}\exp\{-\frac{\beta^*}{2}\parallel \mathbf{y}^*(\mathbf{x}_i,\mathbf{W})-\mathbf{t}\parallel^2\},$$

where β^* and $\mathbf{y}^*(\mathbf{x}_i,\mathbf{W})$ are the inverse of the variance and the location of the Gaussian centres after training, with the true one, see Figure 3. Whenever synthetic data exhibit a single mode at the origin, a gray-scaled matrix containing the norm of each pointer is also used to test the goodness of our strategy, namely the DPO-matrix, i.e. the Distance from each Pointer to the Origin, see Figure 2.

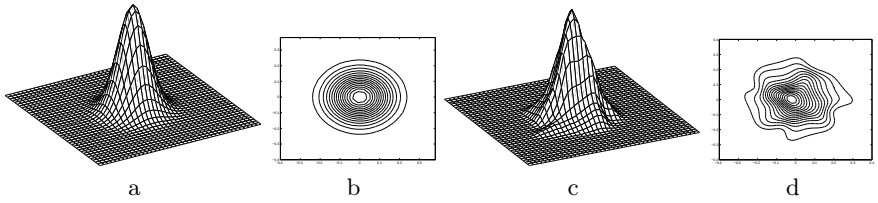


Fig. 3. Data comes from a $N_2(\mathbf{0}, 0.1\mathbf{I}_2)$ density; training parameters are specified in the text; (a) theoretical distribution in 3D; (b) 2D isolines generated from a); (c) mixture distribution obtained after training the GTM; (d) 2D isolines generated from c).

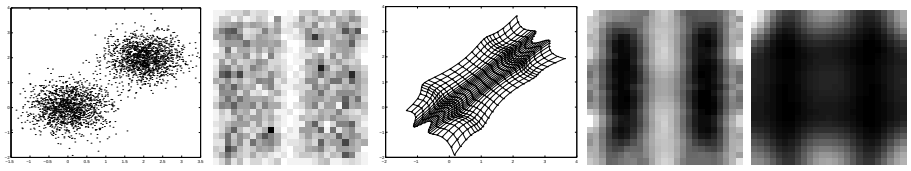


Fig. 4. (a) $N_2(\mathbf{0}, 0.5\mathbf{I}_2) + N_2(\mathbf{2}, 0.5\mathbf{I}_2)$ data set, training parameters: K=625, M=16, $\sigma = 0.4$ and $\tau = 2$; (b) DL-matrix; (c) trained map; (d) MID-matrix; (e) MF-matrix. Note: a) and c) show the data space orientation while b) d) and e) have the latent space one.

To obtain a good level of detail in the study, the number of latent points is fixed to 25x25 unless otherwise noted. Note that having a large number of sample points causes no difficulty beyond increased computational cost [1], so smaller maps, say 15x15 or even 10x10, often provide equally useful information. For these many units, the best values for the number of basis functions M and their

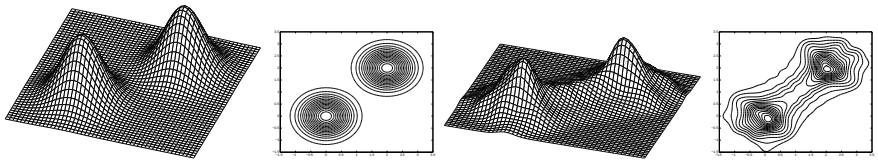


Fig. 5. $N_2(\mathbf{0}, 0.5\mathbf{I}_2)+N_2(\mathbf{2}, 0.5\mathbf{I}_2)$ density, training parameters: K=625, M=16, $\sigma = 0.4$ and $\tau = 2$; (a) theoretical distribution in 3D; (b) 2D isolines generated from a); (c) mixture distribution obtained after training the GTM; (d) 2D isolines generated from c).

common width σ seem to be 16 and 0.5 respectively (they tend to provide a well-organized map and a rather uniform DL-matrix). When selecting the number of training cycles, the evaluation of the log-likelihood can be used to monitor convergence. Hence, we have developed `Matlab` code to train the GTM model until the difference between the log-likelihood values at two consecutive steps is less than a threshold, namely τ. A τ-value smaller than 1.5 is not required for this particular data set, where the training algorithm only needs 20-25 cycles to achieve a good level of convergence.

Special care must be taken with a well-known problem related to the SOM trained structure, namely the *border effect*. By this is meant that units on edges of the network do not stretch out as much as they should [11], which leads to confusing gray-scaled matrices on these map regions, see Figure 2d. Fortunately, these spurious concentrations rarely spread towards the interior of the network, although their traditional presence is somewhat annoying.

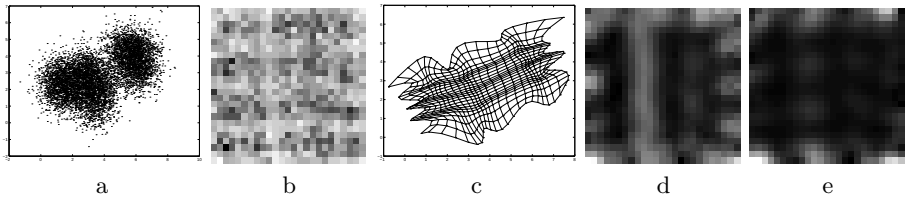


Fig. 6. (a) Carreira data set, N=8000, training parameters: K=625, M=81, $\sigma = 1.0$ and $\tau = 2$; (b) DL-matrix; (c) trained map; (d) MID-matrix; (e) MF-matrix.

To illustrate the strategy in a simple multimodal case, Figure 4 shows an equally weighted mixture of two 2D Gaussians. While all the statistics successfully reveal the existence of two modes, the map does not reflect the support of the true generating distribution. Note also that, somewhat surprisingly, both

horizontal and vertical strips are visible in Figure 4e. As a result note the bulky ridge connecting the two modes in Figure 5c.

We now present a more complex multimodal case, first considered by Carreira [5], involving an equally weighted mixture of eight 2D Gaussians with three modes in total, see Figure 7b. The trained map in Figure 6 shows quite informative lines of concentration and stretch in (d) and (e), revealing plausible horizontal and vertical *split-regions* to separate out the three modes (unfortunately, a strong border effect is also visible in (d)). The fitted density in Figure 7c-d does capture the three modes approximate location (although it tends again to inflate the main gap between modes).

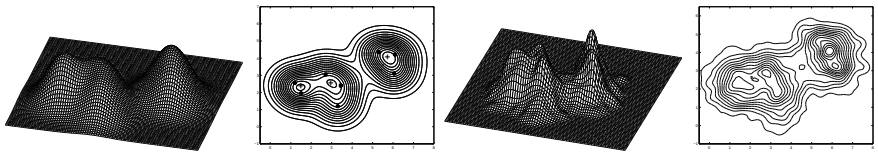


Fig. 7. Carreira data set, training parameters: K=625, M=81, $\sigma = 1.0$ and $\tau = 2$; (a) theoretical distribution in 3D; (b) 2D isolines generated from a), component modes are marked with "•", mixture modes with "+"; (c) mixture distribution obtained after training the GTM; (d) 2D isolines generated from c).

The GTM model in Figure 8 deals with 3D Gaussian data. It reflects how the relationship between adequate organization and uniformly distributed dataloads is already more difficult to obtain. Note that the MID-matrix and MF-matrix show a definite map compression around the center of the map, yet the current pattern is quite different in nature to that found in Figure 2. In these and other higher-dimensional cases, the fitted mixture densities and isolines plots can not be visualized. The main motivation for using SOM structures as an exploratory tool stems from the complexity of the analysis based on a multidimensional fitted density.

When jumping to high dimensional Gaussian data we first note that the probability assigned to the unit radius sphere by the spherical Gaussian distribution goes to zero as the dimension increases. To the extent that lighter concentration should then be expected near the mode, modes in this case may turn out particularly hard to locate [15]. Further, it has been mentioned already that self-organization can be awkward in higher dimensions. Finally, as evidenced in Figure 9, the map may no longer be as centered around the mode as before. Overall, this map is only partially organized, but the patterns in plots (d) and (e) here can be seen as diffuse versions of the inner rings found in the corresponding plots in Figure 8. We stress that the shift in mode location has been consistently observed in many runs with this data set.

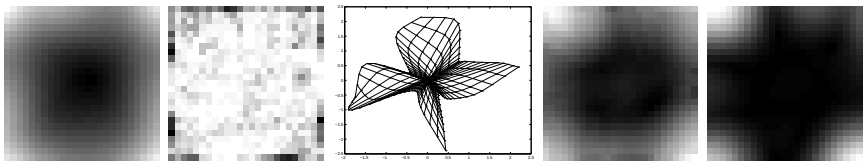


Fig. 8. Data come from a $N_3(\mathbf{0}, \mathbf{I}_3)$ distribution, training parameters: K=625, M=25, $\sigma = 0.5$ and $\tau = 2$; (a) DPO-matrix; (b) DL-matrix; (c) Sammon's projected map; (d) MID-matrix; (e) MF-matrix.

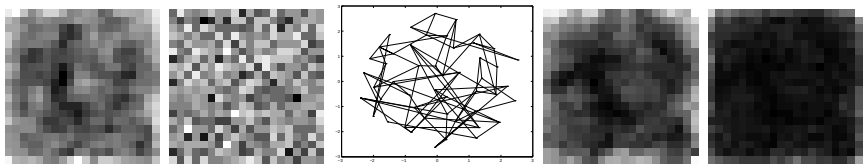


Fig. 9. Data comes from a $N_{10}(\mathbf{0}, \mathbf{I}_{10})$ density, training parameters: K=49, M=144, $\sigma = 1.1$, $\tau = 2$ and 200 Sammon steps; (a) DPO-matrix, showing how the mode is shifted a bit with respect to the 2D Gaussian case; (b) DL-matrix, (c) Sammon's projected map; (d) MID-matrix following the pattern in a); (e) MF-matrix.

5 Summary and discussion

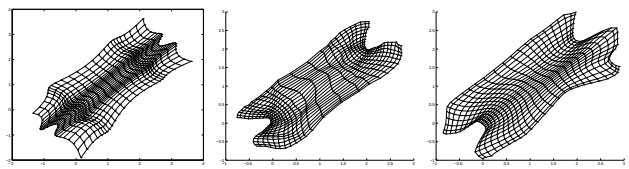
This paper explores the role of the SOM *structure* to deal with the problem of multivariate mode detection. Mode detection differs from cluster detection in that a precise estimation task is faced when tackling unimodal data. The present approach is founded on the tenets that (i) SOM pointers near the modes should lie closer together due to a higher concentration of data in those regions; and (ii), these areas can be highlighted by exploring the structure of pointer inter-distances and other summaries whenever sufficient self-organization is achieved. As the GTM training algorithm explicitly controls the smoothness of the map and also leads to a rich, continuous treatment of local magnification factors, it can be seen to provide a potential advantage over the original SOM fitting algorithm. Hence, we have formulated and investigated a strategy for detecting the modes of a multivariate continuous data-generating distribution based on suitable statistics taken from a trained GTM, namely, median interneuron distances, dataloads, magnification factors together with Sammon's projection.

The strategy works as expected with 2D and 3D Gaussian data. While the maps produced by GTM are quite similar to those found by the standard fitting algorithm in some cases (see Figure 2), they differ markedly in others (Figure 4), which suggests that any detection strategy of the present sort must distinguish the origin of the SOM structure. In a sense, this is the price to be paid for exploiting GTM's richer framework.

The paper also highlights the current risks found when fitting SOMs to high-dimensional data, where larger maps are needed to provide enough detail for the strategy to succeed but turn out to be harder to organize. Some unexpected phenomena have been isolated; for example, in the 10D case modes are typically anchored at noncentral neurons. Additional research is needed to clarify the setting of GTM parameters with an eye put on better self-organization levels.

Future investigations can proceed in various fronts. New diagnostic values can be based on a suitable combination of MID and MF values intended to stress pointer concentration. Directions of stretch deserve also a close look. On the other hand, an additional smoothing construct can be used in GTM, namely, a Gaussian process prior distribution penalizing larger $\mathbf{W}$ entries. Other prior distributions may also be considered, see e.g. [20].

Finally, note that the present strategy for mode detection is based on the assumption that the density of the reference vectors will be similar in some sense to the density of the training data. As mentioned above, an alternative strategy can proceed on the basis of uniformly spaced pointers with widely different dataloads. To this end, some adjustments to the standard algorithm have been proposed to yield a trained map in line with the desired level of pointer concentration [21]. It remains to be seen whether these adjustments have their counterpart in GTM or else can improve detection accuracy. As a first example of the power of the suggestions made in [21], we finally present the alternative maps obtained in a previously discussed example.



a b c

Fig. 10. (a) Map trained via GTM (discussed earlier); (b) the map trained with the convex adjustment reflects more faithfully the sampling density; (c) the map trained with the concave adjustment provides a more uniform distribution of pointers.

Acknowledgement:

We are grateful to Drs. Juan M. Marín and David Ríos for their constructive comments on earlier drafts of this paper. The authors are partially supported by grants from the local CAM Government and the European Council.

References

1. Bishop, C. M., Svensén, M. and Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Computation*, 10: 215-235. 1997a.

2. Bishop, C. M., Svensén, M. and Williams, C. K. I. GTM: A principle alternative to the Self-Organizing Map. *Advances in Neural Information Processing Systems*. vol, 9. MIT Press, 1997.
3. Bishop, C. M., Svensén, M. and Williams, C. K. I. Magnification Factors for the SOM and GTM Algorithms. *Proccedings 1997 Workshop on Self-Organizing Maps*, Helsinki, Finland.
4. Bodt, E., Verleysen, M., and Cottrell, M. Kohonen Maps vs. Vector Quantization for Data Analysis. In *Proceedings of the European Symposium on Artificial Neural Networks*, 221-220. Brusseels: D-Facto Publications. 1997.
5. Carreira-Perpiñán, M. A. Mode-Finding for Mixtures of Gaussians Distributions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(11):1318-1323. 2000.
6. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society, B 39*(1), 1-38. 1977.
7. Good, I. J. and Gaskins, R. A. Density Estimation and Bump-Hunting by Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. *Journal of the American Statistical Association*. 75:42-73, 1980.
8. Hartigan, J. A., and Hartigan, P. M. The DIP test of unimodality. *Ann. Statist. 13*, 70-84. 1985.
9. Izenman, A. J., and Sommer, C. J. Philatelic mixtures and multimodal densities. *J. Am. Statist. Assoc. 83*, 941-953. 1988.
10. Kiviluoto, K. and Oja, E. S-map: A Network with a Simple Self-Organization Algorithm for Generative Topographic Mappings. In M. I. Jordan, M. J. Kearns, and S. A. Solla (Eds.), *Advances in Neural Information Processing Systems*. MIT Press 10:549-555. 1998.
11. Kohonen, T. *Self-Organizing Maps*. Springer-Verlag, Berlin, 2001.
12. Muruzábal, J. and Muñoz, A. On the Visualization of Outliers via Self-Organizing Maps. *Journal of Computational and Graphical Statistics*, 6(4): 355-382. 1997.
13. Ritter, H. and Schulten, K. On the Stationary State of Kohonen's Self-Organizing Sensory Mapping. *Biological Cybernetics*, 54:99-106, 1986.
14. Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, 18(5): 401-409. 1969.
15. Scott, D. W. *Multivariate Density Estimation*. New York: John Wiley and Sons, Inc. 1992.
16. Scott, D. W. and Szewczyk, W. F. The Stochastic Mode Tree and Clustering. To appear in Journal of Computational and Graphical Statistics. 2000
17. Svensén, M. The GTM: Toolbox-User's Guide. 1999. `http://www.ncrg.aston.ac.uk/GTM`.
18. Terrell, G. R. and Scott, D. W. Variable kernel density estimation. *Ann. Statist. 20*, 1236-1265. 1992.
19. Ultsch, A. Self-Organizing Neural Networks for Visualization and Classification. In Opitz, O., Lausen, B., and Klar, R. editors, *Information and Classification*, pages 307-313. Springer-Verlag, Berlin, 1993.
20. Utsugi, A. Bayesian Sampling and Ensemble Learning in Generative Topographic Mapping. *Neural Processing Letters*, 12(3): 277-290. 2000.
21. Zheng, Y. and Greenleaf, J.F. The Effect of Concave and Convex Weight Adjustments on Self-Organizing Maps. *IEEE Transactions on Neural Networks*, 7(1): 87-96, 1996.