On Equiprobabilistic Maps and Plausible Density Estimation

Jorge Muruzábal, Susana Vegas-Azcárate
Statistics and Decision Sciences Group
University Rey Juan Carlos
28933 Móstoles, Spain
{jorge.muruzabal, susana.vegas}@urjc.es

Abstract - *Self-organizing maps have long been used for data visualization and clustering. When using the standard SOM training method, it is well-known that an appropriate choice of the final adaptation radius is crucial for obtaining topology-preserving maps. Similar considerations apply to other training schemes. Consequently, various heuristics have been suggested to assist the user in selecting the most appropriate map in each case. With regard to the standard algorithm, equiprobabilistic maps have been argued to provide a more "faithful" fit to the data, although the overfitting problem still remains. Interestingly, certain equiprobabilistic maps also provide generative mixture densities. The purpose of this paper is to explore the potential of the equiprobabilistic map idea as a means to reach plausible predictive densities (capable of avoiding the previous difficulties). A new monitorization idea is introduced, and a few examples are used to illustrate the scope of the approach.*

Key words - Self-organization, map selection, mode estimation, mixture models.

1 Introduction

Self-organizing maps [1] have become standard tools for frequently encountered data-analytic tasks such as visualization, clustering and classification. Unfortunately, however, a complete SOM training methodology is not firmly established yet. For example, when using the standard SOM fitting algorithm, it is well-known that an appropriate choice for the end adaptation radius or final neighborhood width is crucial for obtaining useful results. This parameter controls in effect the degree of smoothness (or the amount of quantization error minimization) in the final map. Several heuristics have been suggested to guide this choice [2, 3, 4, 5]. Still, Lampinen and Kostiainen [4] acknowledge that "the model complexity in SOM is usually chosen ad hoc by the user".

With respect to the standard Kohonen algorithm, *equiprobabilistic* maps have been argued to provide a more faithful representation of the distribution generating the data [6, 7]. Indeed, these maps tend to minimize or eliminate the number of "dead" or "empty" units commonly found in standard fits. The motivating idea is that *all* trained neurons should have the *same* probability of being the best matching unit for a new data vector sampled from the same generating distribution. Here we consider two kernel-based variants of this type of map, namely, the generative topographic mapping (GTM) [8] and the maximum entropy learning

rule (kMER) [6] algorithms. Just like in the case of the standard Kohonen algorithm, relatively organized structures are achieved during the early and intermediate stages of training, whereas more tangled-up structures (typically providing a slight improvement in quantization fit) are obtained when training continues all the way through the latter stages. Thus, an overfitting problem is also latent in equiprobabilistic maps. Van Hulle [7] has recognized the usefulness of certain form of monitorization and early-stopping policy to prevent damage to the organized map once it is achieved.

In this paper we investigate a novel, alternative scheme which is valid for any kind of equiprobabilistic map. Instead of looking directly at measures of self-organization [5], we care only about the extent to which equiprobabilism is achieved. Specifically, the key observation is that we can measure how close to equiprobabilistic a given map is, and we consider to stop training as soon as the map shows the first signs of equiprobabilism. This policy is seen to prevent excessive vector quantization indirectly. Further, the resulting density estimates are also seen to be rather accurate, in the sense that they recover the *modes* of the true generating density in many cases. We show that good results are obtained regardless of the particular training algorithm used.

The organization is as follows. Section 2 revisits the concept of equiprobabilistic map and Section 3 reviews a number of previous ideas on monitorization. The new criterion is presented in Section 4, and the experimental evidence is discussed in Section 5. Finally, Section 6 summarizes some conclusions and points out a few lines for further research.

2 Equiprobabilistic SOMs and density estimation

Inspired by a number of ordered mappings found in certain neural structures, the typical self-organizing map develops, in an unsupervised way, a mapping from a d-dimensional input space $V \subseteq \mathrm{IR}^d$, into an equal or lower-dimensional discrete lattice with regular, fixed topology [1]. For simplicity, in this paper we only consider 2D squared SOMs equipped with the standard topology. Although cluster analysis, data visualization and non-parametric regression are the most common applications, the SOM structure can also be regarded as a tool for generating non-parametric models of the sampling probability density, $p(v)$ say. Models of this sort can be termed explicit or implicit depending on whether a formal density estimate, say $\hat{p}(v)$, is available or not. When such a density is available, standard scoring measures can be readily used to assess its merit. When not, the areal *magnification factor* of a neural map can be used to measure model accuracy, see e.g. [9]. This is the exponent μ verifying $p(w_i) \propto p(v)^\mu$, where $p(w_i)$ is the (asymptotic) neuron weight density in input space. For the standard training algorithm, however, magnification factors are only known in the simplest cases. For example, $\mu = \frac{2}{3}$ in the limit of an infinite density of neurons in a linear map [10]. For higher dimensional maps, the standard algorithm and its variants tend to show the same behaviour, that is, they tend to underestimate higher density regions and therefore fail to provide a *faithful* representation of the sampling distribution [10, 11].

Various strategies have been explored to minimize this problem [9]. In this paper we focus on the idea of *equiprobabilistic* maps. These stick to the information-theoretic optimal $\mu = 1$, so that all neurons have (in the limit) the same probability to be maximally excited. While the issue was initially approached as a limit case of mutual information maximization, a more intuitive approach is perhaps to maximize the entropy of the map's outputs directly. Several

unsupervised learning rules have been developed following this approach [7]. Here we consider two specific algorithms leading to equiprobabilistic maps : GTM [8] and kMER [6]. These kernel-based topographic maps can also be regarded as mixture density models, that is, the target $p(v)$ is approximated via density functions of the form

$$\hat{p}(v) = \sum_{i=1}^{M} q_i K_i(v, w_i, \sigma_i),$$

where M is the total number of neurons in the map, q_i are the mixing parameters, K_i is the kernel function allocated at the neuron weight or centroid w_i, and the radii σ_i determine the spread of the various mixture components. Most kernel-based topographic maps reported in the literature use Gaussian kernels K_i [8, 6, 12, 13]. Both GTM and kMER use *homogeneous* mixtures ($q_i \equiv \frac{1}{M}$). As regards the σ_i, GTM uses a homoscedastic model ($\sigma_i \equiv \sigma$), whereas kMER allows different radii (which are adapted to the local input density together with the weight vectors or kernel centroids w_i). On the other hand, GTM incorporates an optional regularization term λ which can be added to the objective function (a standard likelihood function) to control the topological order in the mapping. This control is accomplished via a joint spherical Gaussian prior (with variance λ^{-1}) on the matrix W defining the non-linear manifold containing the centroids [8].
We use the fitted density estimate $\hat{p}(v)$ as the primary evaluation tool. Specifically, we look at the *modes* of this density as follows. Steepest ascent Hill-Climbing [7] is used over the map lattice, that is, all neighbours are compared and the best is selected (until no further improvement is possible). Since the maps we handle are obtained via the UDL criterion and thus are rather smooth, we can explore the density surface in greater detail by using a star-shaped neighbourhood linking at most 8 neurons to any given unit. More conservative policies can be naturally considered too. Of course, many local maxima may be present, so that the process is restarted from all possible initial neurons. We illustrate below the (gray-valued) DENS-matrix, depicting the $\hat{p}(w_i)$ values at the lattice nodes, and the HC-matrix, highlighting all modes estimated by this procedure.

3 Previous work on map monitorization

The underlying motivations for some kind of monitorization of the training process are as follows. Firstly, the optimal map is that obtained using an "infinitely slow" training process. Although good approximations can be derived for synthetic data with many cycles of training, this long training is not recommended in higher-dimensional real data spaces due to the enormous amount of time consumed by the algorithms and the impossibility of knowing when this "long enough" training has been achieved.
Secondly, for finer density estimation purposes (mode detection) and, more specifically, for its 2D visualization, the output lattice has to be disentangled and well-organized (just like in cluster analysis). Maps not free of topological defects could lead to contiguous high-density regions become split, whereas separated areas could be shown as a single region. In the case of the standard SOM, GTM and other algorithms, it has often been noted that the end adaptation radius or final neighborhood range has a big impact on the final maps obtained. If the final value of this range is too large, neurons will not properly span the input data set, but if it is too small, violations in topographic order will occur. Although it is commonly

suggested that this radius should drop to "small" values by the end of the run, this is not appropriate in many cases [2, 5]. Similar considerations apply to kMER [7, 14].
Hence, most training algorithms need to monitor somehow the degree of topology preservation achieved during training. What varies is the type of heuristic used to carry out this monitorization task. Some heuristics proceed by controlling certain amount of "overlap" among neurons, see e.g. [7]. Others make use of some explicit topology preservation measure [3, 5]. Since the topographic product and other such measures do not involve the training data at all, they can be argued to have some basic limitations for the present task [2, 3]. Thus, Kaski and Lagus [2] propose a goodness of fit measure based on the first and *second* best matching units for a given input vector v ; a similar heuristic is implemented in the well-known SOM_PAK software. Going one step further, Lampinen and Kostiainen [4] propose a *generalization* measure based on the disagreement between the projections of training and *test* data on the trained SOM. We see that these and related ideas differ mainly in the role played by the data. While the measure we introduce below is also based on the training data, it departs substantially from previous approaches. As noted later, however, approaches that consider the test data are also worth-studying in detail.

4 The UDL stopping policy

The number of data projecting onto (won by) each trained unit, which we call the *data load* (DL), is the natural estimate of the probability of activation by new data generated by the sampling distribution, see e.g. [15]. To easily visualize the DL distribution over the map, a gray image is computed, namely, the *DL-matrix*. We consider a criterion based on the uniformity of this DL-matrix, a criterion which we call UDL. In the trained equiprobabilistic map, each neuron covers about the same proportion of the data. Hence, in the truly equiprobabilistic state, the DL vector should be distributed as a Multinomial with equal cell probabilities. In fact, it would follow asymptotically a joint multivariate Gaussian distribution with constant variances and correlations getting weaker with increasing sample size N. For moderate to large N, the equiprobabilistic DL histogram should thus correspond to a common limiting univariate Gaussian. We have verified that this is the case for a 10×10 map and $1,500$ training data. Here the mean is $R = \frac{N}{M} = 15$ and the limit standard deviation is $SD = \sqrt{R(1-\frac{1}{M})}$ or about 3.85. Note that these statistics depend only on lattice size and training sample size, they are otherwise universal over data sets and (equiprobabilistic) training algorithms.
In practice, we have observed that the equiprobabilistic treat tends to obtain rather early in general ; beyond that point on, it is only the quantization error that improves slightly while organization is progressively lost. Hence, the associated UDL stopping policy is viewed as follows : quit as soon as the trained map shows the first signs of having reached the reference Gaussian DL distribution, a moment that we refer to as the "UDL stage". This (fuzzy) stage can be characterized graphically (by a uniform DL matrix) and numerically (by the closeness of the current SD value to the asymptotic level). Graphical diagnostics incorporating lattice information, such as the DL-matrix, may be more suggestive for interactive analysis — for example, we can sometimes tolerate a mild *border effect* in our working maps (see below). However, the DL-matrix should be used in conjunction with the numeric, lattice-independent guidance provided by the SD(t) trajectory. Together with the previous SD function, we also show sometimes the standard mean quantization error (MQE) trajectory.

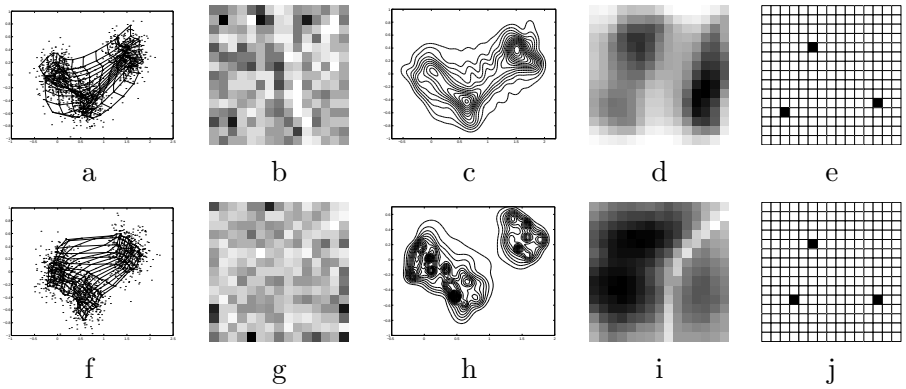

FIG. 1 – Performance by GTM (top) and kMER (bottom) on the trimodal data set : (a,f) trained map with data set highlighted ; (b,g) DL-matrix ; (c,h) 2D isolines generated from the density estimate ; (d,i) DENS-matrix ; (e,j) HC-matrix.

5 Experimental results

We first investigate two artificial data sets of varying nature, then validate the whole approach using a real data set. Our synthetic scenarios concern balanced Gaussian mixtures. We first explore a 2D trimodal data set of size $N = 1,500$. Here we show that the three modes can be recovered using widely different density estimates. We continue with a single 50D Gaussian also of size $N = 1,500$. Here we focus on the training trajectories involved in our UDL criterion. Finally, we examine a highly multimodal problem, the Multiple Features Database or Mfeat (taken from the well-known UCI repository). This data set consists of $d = 649$ features of handwritten numerals $(0, \ldots, 9)$, with 200 patterns per class for a total of $N = 2,000$ patterns. This real data set is useful to verify that the basic procedure still works fine even when the mixture components are far from Gaussian.

In all high-dimensional cases, we show the maps projected via the well-known Sammon's algorithm ; the degree of self-organization in the underlying SOM structure can be assessed informally via this image. `MATLAB`® code for kMER has been developed following the guidelines given in [7]. `MATLAB`® code for GTM is available at `http ://www.ncrg.aston.ac.uk/GTM`. Training parameters are given in the Appendix.

5.1 Three modes in 2D

We first analyze a trimodal 2D data set with two of the modes close enough to illustrate the accuracy of our algorithms, see Figure 1 (here, and in the case of DENS-matrix, darker means larger). The maps discussed here have all been stopped early (so that their DL-matrices exhibit a reasonable amount of uniformity). We first note that good results are obtained using GTM. A precise approximation to the true modes' locations is obtained in the sense that each of the highlighted units in Figure 1e is also the closest to one of the three modes. As regards kMER, we see a mild border effect in Figure 1g, yet mode estimation is very precise again in Figure 1j. Very different densities arise in Figures 1c,h or 1d,i though, kMER appearing better suited to describe the larger gap in these data.

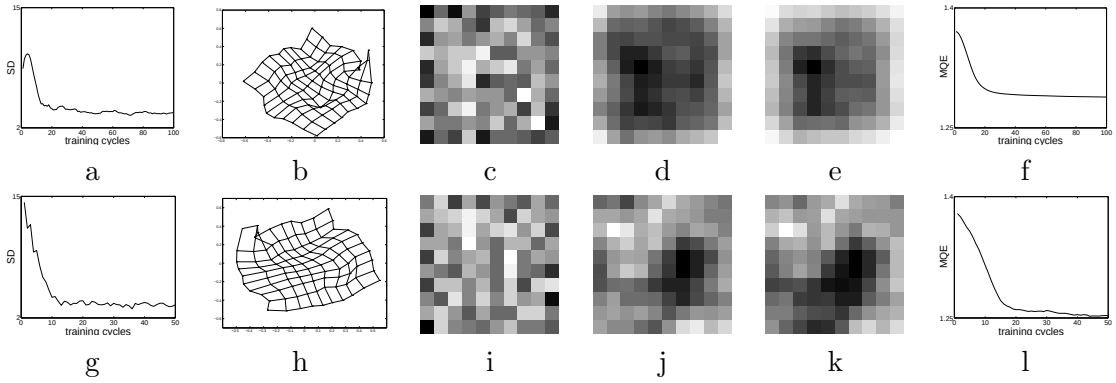

FIG. 2 – Early stopped maps and learning trajectories (top : GTM, bottom : kMER) on single-mode problem. (a,g) SD ; (b,h) Sammon's projected map ; (c,i) DL-matrix ; (d,j) DPO-matrix ; (e,k) DENS-matrix ; (f,l) MQE.

5.2 Single mode in 50D

We consider next a single 50D spherical Gaussian centered at the origin, see Figure 2. Whenever the training data exhibits a single mode at the origin, a gray-scaled matrix based on the $\| w_i \|$ norms is used to assist the assessment of the strategy ; this is termed the Distance from each Pointer to the Origin or DPO-matrix.
In the case of GTM, we see that SD settles around the reference value by $t = 20$ or $t = 22$ training cycles (or around the first tick-mark in Figures 2a,f). The DL-matrix stopped at $t = 22$ shows only a mild border effect in Figure 2c. As a result, the map provides very precise information : the density estimate is unimodal, and the neuron with the highest density is also the closest to the origin — indeed, note the similarity between Figures 2d and 2e. Turning to kMER, we appreciate again a clear UDL-stage at around $t = 15$ (thousand) cycles (or halfway between the first and second tick-marks in Figures 2g,l). The analysis at this UDL-stage leads to a DL-matrix with an appropriate level of uniformity. Therefore, the quality of density estimation is similar and also quite good.

5.3 Real-world example

We finally examine how our criterion does with a highly dimensional real-world example, the Mfeat data set. In this case, $d = 649$, $N = 2,000$ and we look for 10 high-density regions. To aid the assessment, we introduce two additional tools here. The LAB-matrix marks each neuron after the output label that occurs most often within its Voronoi region. To verify the precision of LAB-matrix, CONF-matrix computes the previous label's relative presence in the Voronoi region (so that the higher CONF, the higher the confidence). Only in these two cases, as noted below, darker means smaller.
Figure 3 shows the optimal maps obtained via our UDL criterion. In either case Sammon projections suggest nicely organized maps, and LAB-matrices split the map precisely into ten regions. CONF-matrices suggest that these clusters are well-defined as they highlight the edges of the various regions. Overall, kMER's map seems best, for it leads to the neatest assignment of labels and the highest levels of confidence. kMER's DENS-matrix provides also a fully accurate picture, placing a mode at each of the 10 target regions (GTM misses one).

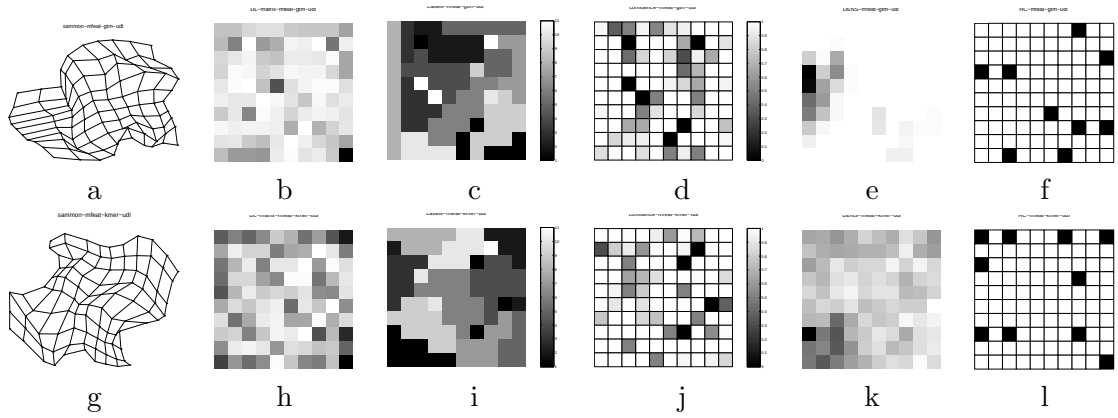

FIG. 3 – Early stopped maps (top : GTM, bottom : kMER) for the Mfeat data. (a,g) Sammon's projected map ; (b,h) DL-matrix ; (c,i) LAB-matrix ; (d,j) CONF-matrix ; (e,k) DENS-matrix ; (f,l) HC-matrix.

6 Summary and conclusions

We have introduced a new monitorization idea for equiprobabilistic maps and we have tested it using two particular learning rules, GTM and kMER. Early stopping seems indeed very appropriate (almost a requirement) if the ultimate goal of the analysis depends on having a faithful approximation to the data-generating distribution. The proposed UDL criterion is easy to apply and has provided sensible answers in our test problems. Hence, the benefits of the new criterion should be more carefully evaluated against other methods. We have briefly discussed here other such stopping criteria based on the resulting density estimate. Further research is needed on this point. Specifically, various ideas related to other uses of data combined with the predictive densities deserve further study. We should also compare our general criterion to kMER's specific criterion [7, 14], as well as to the magnification control schemes developed in [16].

Acknowledgement. The authors are grateful to Drs. M. Van Hulle, T. Gautama and M. Svensén for useful discussions. Our research is funded by Spanish and European agencies. We also appreciate the support by the DMR Foundation's Decision Engineering Lab.

Appendix : Training parameters

Maps have sizes 15×15 in the 3M-2D case and 10×10 otherwise. Initial maps always follow a PCA study in the case of GTM, as suggested by [17], but are randomly created using kMER. GTM involves 100 training cycles for the synthetic data and 2,000 cycles for the real-world data. The size of the basis function grid is set to 6×6 for the 3M-2D data set, 8×8 for 1M-50D and 7×7 for Mfeat. The common width of the basis functions is 1.2, 1.1 and 1 respectively. The regularization term (λ) is set to 2, 100 and 0 (no prior bias at all). kMER involves 500 training cycles for the synthetic data and 2,000 cycles for Mfeat. The initial neighborhood range is set to 10 for 3M-2D and 5 for both 1M-50D and Mfeat. The scale factor and the learning rate are respectively fixed to 1.9, 2.1 and 2 and 0.01, 0.02 and 0.00015.

Références

[1] T. Kohonen, *Self-Organizing Maps*. Berlin : Springer-Verlag, 3rd extended ed., 2001.

[2] S. Kaski and K. Lagus, "Comparing self-organizing maps," in *Proceedings of ICANN'96, International Conference on Artificial Neural Networks, Lecture Notes in Computer Science*, v. S. W. V. J. C. von der Malsburg, C. and B. Sendhoff, Eds., vol. 1112. Springer, Berlin, 1996, pp. 809–814.

[3] T. Villmann, R. Der, M. Herrmann, and T. Martinetz, "Topology preservation in self-organizing feature maps : Exact definition and measurement," *IEEE Trans. Neural Networks*, vol. 8(2), pp. 256–266, 1997.

[4] J. Lampinen and T. Kostiainen, "Overtraining and model selection with the self-organizing map," in *Proc. IJCNN'99, Washington, DC, USA*, 1999.

[5] K. Haese and G. J. Goodhill, "Auto-som : Recursive parameter estimation for guidance of self-organizing feature maps," *Neural Computation*, vol. 13, pp. 595–619, 2001.

[6] M. M. Van Hulle, "Kernel-based equiprobabilistic topographic map formation," *Neural Computation*, vol. 10(7), pp. 1847–1871, 1998.

[7] ——, *Faithful representations and topographic maps : From distortion- to information-based self-organization*. New York : Wiley, 2000.

[8] C. M. Bishop, M. Svensén, and C. K. I. Williams, "Gtm : The generative topographic mapping," *Neural Computation*, vol. 10, pp. 215–235, 1997.

[9] H. U. Bauer, R. Der, and M. Hermann, "Controlling the magnification factor of self-organizing feature maps," in *Neural Computation*, vol. 8, 1996, pp. 757–771.

[10] H. Ritter and K. Schulten, "On the stationary state of kohonen's self-organizing sensory mapping," *Biological Cybernetics*, vol. 54, pp. 99–106, 1986.

[11] J. K. Lin, D. G. Grier, and J. D. Cowan, "Faithful representation of separable distributions," *Neural Computation*, vol. 9, pp. 1305–1320, 1997.

[12] H. Yin and N. Allinson, "Self-organizing mixture networks for probability density estimation," *IEEE Transactions on Neural Networks*, vol. 12(2), pp. 405–411, 2001.

[13] T. Heskes, "Self-organizing maps, vector quantization, and mixture modeling," *IEEE Trans. Neural Networks*, vol. 12(6), pp. 1299–1305, 2001.

[14] M. M. Van Hulle and T. Gautama, "Monitoring the formation of kernel-based topographic maps with application to hierarchical clustering of music signals," *J. VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 32, pp. 119–134, 2002.

[15] E. Bodt, M. Verleysen, and M. Cottrell, "Kohonen maps versus vector quantization for data analysis," in *European Symposium on Artificial Neural Networks, ESANN'97*. Brussels : D-Facto publications, 1997, pp. 211–220.

[16] T. Villmann and J. C. Claussen, "Investigation of magnification control in self-organizing maps and neural gas," in *Proc. WSOM 2003*, Yamakawa, Ed., 2003, pp. 59–64.

[17] K. Kiviluoto and E. Oja, "S-map : A network with a simple self-organization algorithm for generative topographic mappings," in *Advances in Neural Information Processing Systems*, K. M. J. Jordan, M. I. and S. A. Solla, Eds., vol. 10. MIT Press, 1998, pp. 549–555.