EM algorithms for self-organizing maps

Tom Heskes, Jan-Joost Spanjers, and Wim Wiegerinck
RWCP Theoretical Foundation SNN University of Nijmegen
Geert Grooteplein 21, 6252 EZ, Nijmegen, The Netherlands
{tom,janj,wimw}@mbfys.kun.nl

Abstract

Self-organizing maps are popular algorithms for unsupervised learning and data visualization. Exploiting the link between vector quantization and mixture modeling, we derive EM algorithms for self-organizing maps with and without missing values. We compare self-organizing maps with the elastic-net approach and explain why the former is better suited for the visualization of high-dimensional data. Several extensions and improvements are discussed.

1 Introduction

Self-organizing maps are popular tools for clustering and visualization of high-dimensional data [8, 13]. To derive an error function for the self-organizing map, we will follow the vector quantization interpretation given in, among others, [9].

A self-organizing map consists of a set of nodes r with corresponding weight vectors $\mathbf{w}_r$. The quantization error of the node with weight $\mathbf{w}_r$ given a particular input $\mathbf{x}^\mu$ reads

$$D(\mathbf{x}^\mu, \mathbf{w}_r) = \frac{1}{2}||\mathbf{x}^\mu - \mathbf{w}_r||^2 .$$

Given a set of inputs $\mathcal{X}$ and weights $\mathcal{W}$, let p_r^μ denote the probability that input $\mathbf{x}^\mu$ is assigned to the node with weight $\mathbf{w}_r$. It is constrained by $\sum_r p_r^\mu = 1$ and $p_r^\mu \geq 0$. Even if we assign input μ to node r, there is a confusion probability h_{rs} that input μ is instead quantized by the weight vector $\mathbf{w}_s$ corresponding to node s. h_{rs} corresponds to the lateral-interaction strength and defines the underlying manifold: usually it is a decreasing function of the distance between nodes r and s on a two-dimensional grid. Given the data $\mathcal{X}$, the goal is now to find the probability assignments $\mathcal{P}$ and weights $\mathcal{W}$ minimizing the error

$$F_{\text{quantization}}(\mathcal{P}, \mathcal{W}) = \sum_\mu \sum_r p_r^\mu \sum_s h_{rs} D(\mathbf{x}^\mu, \mathbf{w}_s) .$$

An annealed variant of the self-organizing map is obtained if we add an entropy term of the form

$$F_{\text{entropy}}(\mathcal{P}) = \sum_\mu \sum_r p_r^\mu \log \left[\frac{p_r^\mu}{q_r} \right] ,$$

where q_r can be interpreted as prior probability assignments. The usual choice is $q_r = 1/K$ with K the number of nodes, but for later purposes we will consider here the general situation. This entropy term favors probability assigments that are similar to q_r, i.e., maximize the entropy for homogeneous q_r. Annealed vector quantization has been introduced in [12] and applied to self-organizing maps in e.g. [4].

The final "free energy functional" now follows from a weighted combination of the quantization and entropy term:

$$F(\mathcal{P}, \mathcal{W}) = \beta F_{\text{quantization}}(\mathcal{P}, \mathcal{W}) + F_{\text{entropy}}(\mathcal{P}) . \tag{1}$$

β plays the role of an inverse temperature: the larger β, the smaller the influence of the entropy term. Formulation of the optimization criterion in terms of this free energy functional will be very convenient in the derivation of EM algorithms later on.

The dependency on the assignments $\mathcal{P}$ can be removed by computing the optimal assignments $\mathcal{P}(\mathcal{W})$ given a particular set of weights $\mathcal{W}$:

$$p_r^\mu(\mathcal{W}) = \frac{q_r \exp\left[-\beta \sum_t h_{rt} D(\mathbf{x}^\mu, \mathbf{w}_t)\right]}{\sum_s q_s \exp\left[-\beta \sum_t h_{st} D(\mathbf{x}^\mu, \mathbf{w}_t)\right]} \,. \tag{2}$$

With $D(\mathbf{x}, \mathbf{w})$ continuous in $\mathbf{w}$, these assignments are unique. Substitution into (1) then yields

$$E(\mathcal{W}) \equiv \min_{\mathcal{P}} F(\mathcal{P}, \mathcal{W}) = -\sum_\mu \log \sum_r q_r \exp\left[-\beta \sum_t h_{rt} D(\mathbf{x}^\mu, \mathbf{w}_t)\right] \,. \tag{3}$$

This error function (with $q_r = 1/K$) corresponds to an annealed version of a closely related variant of Kohonen's original self-organizing map algorithm [7]. The (small) differences are discussed in [6, 4]. It can be shown (e.g. following a proof in [10]) that any (locally) optimal solution of $F(\mathcal{W}, \mathcal{P})$ corresponds to a (locally) optimal solution of $E(\mathcal{W})$ and vice versa. In other words, we can exchange the two optimization criteria, as we will do throughout the paper.

In the following, we will sometimes compare with the elastic-net approach [2, 14]. Topology is introduced by adding a penalty term to an (annealed) vector quantization error, i.e., the goal is to minimize an error function of the form

$$E_{\text{elastic}}(\mathcal{W}) = -\sum_\mu \log \sum_r q_r \exp\left[-\beta D(\mathbf{x}^\mu, \mathbf{w}_r)\right] + \sum_{r,s} h_{rs} \|\mathbf{w}_r - \mathbf{w}_s\|^2 \,. \tag{4}$$

Also here the standard choice is $q_r = 1/K$.

2 EM algorithm without missing values

The free energy functional (1) allows for an extremely straightforward derivation of an EM algorithm. Both the expectation and the maximization step can be seen as minimizing this same functional [10].

The *expectation step* in the full EM algorithm follows by minimizing $F(\mathcal{P}, \mathcal{W})$ with respect to the assignments $\mathcal{P}$, given the current set of parameters $\mathcal{W}$. We immediately obtain (2).

The *maximization step* in the full EM algorithm follows by minimizing $F(\mathcal{P}, \mathcal{W})$ with respect to the parameters $\mathcal{W}$, given the current set of assignments $\mathcal{P}$. For sum-squared $D(\mathbf{x}, \mathbf{w})$, we easily find

$$\mathbf{w}_s(\mathcal{P}) = \frac{\sum_\mu \sum_r p_r^\mu h_{rs} \mathbf{x}^\mu}{\sum_\mu \sum_r p_r^\mu h_{rs}} \,. \tag{5}$$

This EM algorithm (in the limit $\beta \to \infty$) is referred to as the batch-map algorithm in [1, 8]. Compared with the elastic-net approach, the EM batch-map algorithm is surprisingly efficient: incorporation of tolopology only requires extra summations over nodes, limited to the width of the lateral interaction h_{rs}. On the other hand, the additive penalty term in the elastic-net approach [see (4)] makes that the M-step amounts to solving a set of K linear equations [14].

3 A mixture-modeling interpretation

There is a close link between vector quantization and mixture modeling. Saying that a particular $\mathbf{w}_r$ is a good quantizer for a pattern $\mathbf{x}^\mu$ because of a low quantization error $D(\mathbf{x}^\mu, \mathbf{w}_r)$ is similar to stating that some probability $G(\mathbf{x}^\mu|\mathbf{w}_r)$ of finding $\mathbf{x}^\mu$ given $\mathbf{w}_r$ is quite high. The obvious choice for this probability in the case of a sum-squared error $D(\mathbf{x}, \mathbf{w})$ is a Gaussian.

Let us first consider the case of no lateral interaction, i.e., $h_{rs} = \delta_{rs}$. As a mixture model, we take

$$P(\mathbf{x}|\mathcal{W}) = \sum_r q_r G(\mathbf{x}|\mathbf{w}_r) \text{ with } G(x|w) = \sqrt{\frac{\beta}{2\pi}} \mathrm{e}^{-\beta(x-w)^2/2} \text{ and } G(\mathbf{x}|\mathbf{w}) = \prod_\alpha G(x_\alpha|w_\alpha) \,.$$

Through simple substitution it is easy to show that, with this particular choice of mixture model and up to irrelevant constants, we have

$$L(\mathcal{W}) \equiv \sum_{\mu} \log P(\mathbf{x}^{\mu}|\mathcal{W}) = -E(\mathcal{W}) \,, \tag{6}$$

i.e., the optimization criterion for annealed vector quantization corresponds to a maximum likelihood procedure for a mixture of Gaussians.

With lateral interaction, the link is not so obvious. To simplify the term in the exponent in (3), we need the "bias-variance decomposition"

$$\sum_{s} h_{rs} D(x, w_s) = D(x, \tilde{w}_r) + \sum_{s} h_{rs} D(\tilde{w}_r, w_s) \text{ with } \tilde{w}_r = \sum_{s} h_{rs} w_s \,. \tag{7}$$

The essence here is that the average error on the righthand side can decomposed into an error of an average weight $\tilde{w}_r$ and a variance term *independent* of the input x. The variance $V_r(\mathcal{W}) = \sum_s h_{rs} D(\tilde{\mathbf{w}}_r, \mathbf{w}_s)$ measures to what extent the weights vary around node r.

The decomposition is used to proof that self-organizing maps can be interpreted as mixture models with an added regularization term. If we take the mixture model

$$P(\mathbf{x}|\mathcal{W}) = \sum_{r} \tilde{q}_r(\mathcal{W}) G(\mathbf{x}|\tilde{\mathbf{w}}_r) \text{ with } \tilde{q}_r(\mathcal{W}) \equiv \frac{q_r \mathrm{e}^{-\beta V_r(\mathcal{W})}}{\sum_s q_s \mathrm{e}^{-\beta V_s(\mathcal{W})}} \,, \tag{8}$$

and compute the loglikelihood, we obtain, after some rewriting, the self-organizing map error (3), except for a term independent of the patterns $\mathcal{X}$. That is, neglecting irrelevant constants independent of $\mathcal{W}$, we have

$$E(\mathcal{W}) = -L(\mathcal{W}) + E_{\text{regularization}}(\mathcal{W}) \,, \tag{9}$$

with $L(\mathcal{W})$ the loglikelihood as in (6) and the regularization term

$$E_{\text{regularization}}(\mathcal{W}) \equiv -\sum_{\mu} \log \sum_{r} q_r \mathrm{e}^{-\beta V_r(\mathcal{W})} \,. \tag{10}$$

4 Self-organizing maps and elastic nets

The correspondence between the self-organizing map error (9) and the elastic-net error (4) is striking. The important difference between the two is that fixed $q_r = 1/K$ in (4), the standard choice in the elastic-net approach, corresponds to fixed marginals $P(r|\mathcal{W})$. This yields a tendency to make all nodes equally important. In the self-organizing map approach, with $q_r = 1/K$ in (3), one can still have nodes with a low marginal $P(r|\mathcal{W})$, namely those with a high local variance $V_r(\mathcal{W})$. These variances are similar to what in the literature on self-organizing maps is called the "U-matrix" [13]. The U-matrix is often visualized as a surface on the two-dimensional topology of the self-organizing map and indicates clusters, with different clusters separated by barriers. In mixture-modeling terms these barriers correspond to nodes with a low marginal $P(r|\mathcal{W})$, which can focus on interpolating between different clusters. An elastic-net algorithm does not have this flexibility and therefore seems less suited for the visualization of high-dimensional data.

The regularization term (10) aims at low variances, that is, small differences between weight vectors of neighboring nodes. This is the term that explains the self-organizing property of self-organizing maps: it implements the tendency for neighboring nodes to represent similar input patterns. Note that the regularization term scales with the number of patterns $N = \sum_{\mu}$. It can therefore not be truly interpreted as resulting from a kind of Bayesian prior, since such a term would become less and less important with growing N.

5 EM algorithm with missing values

The basic idea in EM algorithms for mixture models is to extend the distribution $P(\mathbf{x}|\mathcal{W})$ to a joint distribution $P(\mathbf{x}, r|\mathcal{W})$, where the states of the nodes r are considered hidden. The extra set of parameters

$\mathcal{P}$ is introduced to represent the probabilities of these states. Another important application of EM is to learning with truly missing (input) values. The combination of both missing inputs and mixture models is pursued in [3]. The probabilistic interpretation of self-organizing maps derived in this paper allows for a similar combination. We consider the standard situation $q_r = 1/K$.

We assume that for each pattern μ some inputs are known, indicated by the lower index k, and some may be missing, indicated by m. We should in fact write k^μ and m^μ, but for the sake of clarity we will leave it at k and m. From the definition of the error $E(\mathcal{W})$ in vector-quantization terms, as in (3), it is not so obvious how to incorporate missing values. The link (9), where the vector-quantization error is decomposed in a loglikelihood term $-L(\mathcal{W})$ and a regularization term $E_{\text{regularization}}(\mathcal{W})$, provides a solution. The regularization term is independent of the data $\mathcal{X}$ and thus unaffected by the presence of missing values. The loglikelihood term, on the other hand, can only look at the known components and thus becomes

$$L(\mathcal{W}) = \sum_\mu \log P(\mathbf{x}_k^\mu|\mathcal{W}) \text{ with } P(\mathbf{x}_k|\mathcal{W}) = \int dx_m P(\mathbf{x}|\mathcal{W}) \,,$$

and $P(\mathbf{x}|\mathcal{W})$ from (8). Following [10] and similar to the above link between the error (3) and the free energy (1), the free-energy functional corresponding to the error $E(\mathcal{W}) = -L(\mathcal{W}) + E_{\text{regularization}}(\mathcal{W})$ can be written

$$F(\mathcal{P},\mathcal{W}) = -\sum_r \int dx_m p_r^\mu(\mathbf{x}_m) \log P(\mathbf{x}_k^\mu, \mathbf{x}_m|\tilde{\mathbf{w}}_r) + \sum_r \int dx_m p_r^\mu(\mathbf{x}_m) \log p_r^\mu(\mathbf{x}_m) + E_{\text{regularization}}(\mathcal{W}) \,, \tag{11}$$

with for each μ a joint distribution $p_r^\mu(\mathbf{x}_m)$ over both the state of the nodes and the missing inputs.

The E-step follows by minimizing free energy (11) with respect to these distributions for given $\mathcal{W}$. We state the result:

$$p_r^\mu(\mathbf{x}_m|\mathcal{W}) = p_r^\mu(\mathcal{W}) G(\mathbf{x}_m|\tilde{\mathbf{w}}_{rm}) \,, \tag{12}$$

where we have defined

$$p_r^\mu(\mathcal{W}) = \frac{q_r \mathrm{e}^{-\beta \sum_t h_{rt} D(\hat{\mathbf{x}}_r^\mu, \mathbf{w}_t)}}{\sum_s q_s \mathrm{e}^{-\beta \sum_t h_{st} D(\hat{\mathbf{x}}_s^\mu, \mathbf{w}_t)}} \text{ with } \hat{x}_{r\alpha}^\mu = \begin{cases} x_\alpha^\mu \,, & \text{if } \alpha \text{ known for } \mu \,; \\ \tilde{w}_{r\alpha} \,, & \text{if } \alpha \text{ missing in } \mu \,. \end{cases} \tag{13}$$

In other words, the E-step in the case of missing values yields (12) with $G(\mathbf{x}_m|\tilde{\mathbf{w}}_{rm})$ a Gaussian probability distribution over the missing inputs given the current average weight $\tilde{\mathbf{w}}_{rm}$, and $p_r^\mu(\mathcal{W})$ equivalent to (2) for the case of no missing values with missing $\mathbf{x}_m^\mu$ replaced by $\tilde{\mathbf{w}}_r$.

The M-step is based on the minimization of the free energy (11) with respect to the parameters $\mathcal{W}$ for fixed $\mathcal{P}$. After straightforward manipulations we obtain

$$\mathbf{w}_s(\mathcal{P}) = \frac{\sum_\mu \sum_r p_r^\mu h_{rs} \hat{\mathbf{x}}_r^\mu}{\sum_\mu \sum_r p_r^\mu h_{rs}} \,, \tag{14}$$

with $\hat{x}_{r\alpha}^\mu$ as in (13). The M-step with missing values is equivalent to the M-step (5) without missing values, using the same substitution as in the E-step for the missing $\mathbf{x}_m^\mu$. Really, this is what one could have expected from the start, except that it is important to realize that the parameter used for filling in the unknown input x_α is the average $\tilde{w}_{r\alpha}$, and not the original $w_{r\alpha}$.

6 Discussion

The batch-map algorithm corresponding to Kohonen's original learning (see e.g. [8]) differs from the EM algorithm discussed here in two aspects: it corresponds to the limit $\beta \to \infty$ and has no neighbor averaging in the E-step. The limit $\beta \to \infty$ is just a special case of the analysis in this paper and contains no further pecularities except that some of the proofs may be technically more involved because of discontinuities. The simpler E-step can be interpreted as an approximation to the one derived in (2) which does involve neighbor averaging. The simpler E-step is faster, but the connection with a global error function like (3) is lost. This

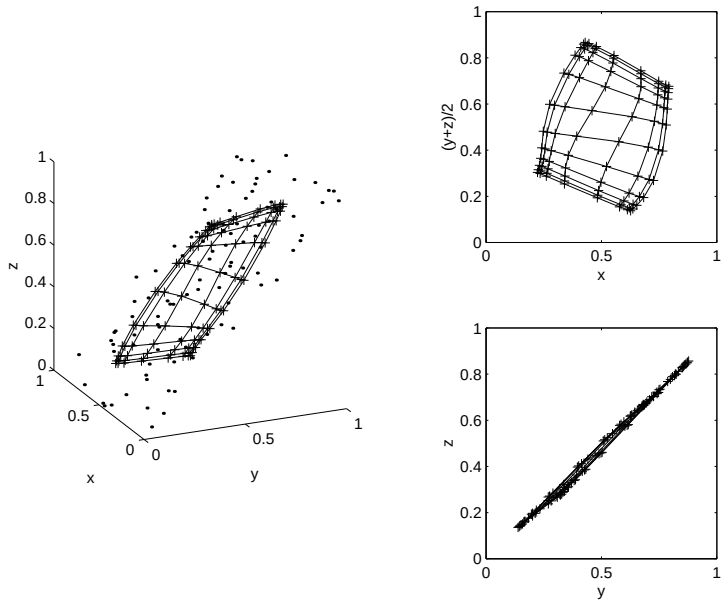

Figure 1: Self-organizing map learned on data with missing values. Training set consists of 500 points in 3 dimensions, all close to the plane $y = z$, but with 50% of all values missing (not shown). The self-organizing map has 8×12 nodes with lateral interactions $h_{rs} \propto \exp[-d_{rs}/2\sigma^2]$ where d_{rs} refers to the node distance on a two-dimensional grid. Parameters $\sigma = 0.5$ and $\beta = 100$. The map manages to unfold and, as can also be seen from the projections, finally represents the data quite well.

makes it difficult to check and proof convergence. Maps obtained through application of both winner mechanisms are roughly the same (for example, tested on WEBSOM, Prof. Kohonen, private communication). The constraint $\sum_s h_{rs} = 1$ facilitates a probabilistic interpretation of the lateral interactions, but has further no consequences.

The analysis presented in this paper focussed on sum-squared error $D(x, w)$ and corresponding Gaussian probability $G(x|w)$. It can be easily extended to quantization errors that can be derived from probability distributions in the exponential family. For distributions of the form

$$G(x|w) = \exp\left[c(w)T(x) + d(w) + S(x)\right] \ ,$$

the quantization error is the deviance

$$D(x, w) = -\log G(x|w) + \log G(x|x) = [c(x) - c(w)]T(x) + d(x) - d(w) \ .$$

$c(w)$ is called the canonical link, $T(x)$ the sufficient statistic. Examples include the Gamma distribution, multinomial, and Poisson. The bias-variance decomposition (7), and thus the correspondence of self-organizing maps to regularized mixture modeling, still holds if we define the average weight by averaging the canonical links (see e.g. [5]):

$$c(\tilde{w}_r) = \sum_s h_{rs} c(w_s) \ .$$

Furthermore, if the sufficient statistic is linear, i.e., $T(x) = x$ as for most common distributions, we still have the simple form (5) for the M-step. The EM algorithm for missing values stays the same for continuous distributions where the annealing parameter β can be interpreted as a dispersion parameter and for all other distributions if $\beta = 1$. There hardly seems to be a reason to restrict self-organizing maps to sum-squared errors. Depending on the format of the data and the underlying assumptions, more appropriate quantization errors can be chosen, perhaps even different ones for different dimensions.

The EM algorithms presented here are the standard versions. There are many different ways to speed them up. Especially attractive and relatively simple is the "accelerated" version. The idea is to take the

new weight vectors w_r^{new} "beyond" the optimal $w_r(\mathcal{P})$ given in (5) and (14):

$$w_r^{\text{new}} = \eta w_r(\mathcal{P}) + (1-\eta) w_r^{\text{old}} ,$$

with $1 \leq \eta < 2$. The same can be done for the probabilities $\mathcal{P}$ in the E-step. Here we might take, for all μ,

$$\log p_r^{\text{new}} \propto \eta \log p_r(\mathcal{W}) + (1-\eta) \log p_r^{\text{old}} ,$$

where the proportionality constant follows from the normalization $\sum_r p_r^{\text{new}} = 1$. This logarithmic averaging seems to work a little better than simple linear averaging and explicitely constrains the probabilities to positive numbers. By applying the same reasoning as in [11], it can be shown that accelerated EM is locally contractive (converges to a local minimum if starting sufficiently close to this minimum) for $\eta < 2$. In practice, $\eta \approx 1.3$ seems to work fine and speeds up the convergence of the EM algorithm considerably (roughly a factor 2).

References

[1] Y. Cheng. Convergence and ordering of Kohonen's batch map. *Neural Computation*, 9:1667–1676, 1997.

[2] R. Durbin and Willshaw D. An analogue approach to the traveling salesman problem using an elastic net method. *Nature*, 326:689–691, 1987.

[3] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an EM approach. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 120–127, San Mateo, 1994. Morgan Kaufmann.

[4] T. Graepel, M. Burger, and K. Obermayer. Self-organizing maps: generalizations and new optimization techniques. *Neurocomputing*, 21:173–190, 1998.

[5] J. Hansen and T. Heskes. General bias/variance decomposition with target independent variance of error functions derived from the exponential family of distributions. *Submitted*, 1999.

[6] T. Heskes and B. Kappen. Error potentials for self-organization. In *International Conference on Neural Networks, San Francisco*, volume 3, pages 1219–1223, New York, 1993. IEEE.

[7] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[8] T. Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.

[9] S. Luttrell. Self-organisation: A derivation from first principles of a class of learning algorithms. In *International Joint Conference on Neural Networks*, volume 2, pages 495–498. IEEE Computer Society Press, 1989.

[10] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, Dordrecht, 1998.

[11] B. Peters and H. Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal of Applied Mathematics*, 35:362–378, 1987.

[12] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics of phase transitions in clustering. *Physical Review Letters*, 65:945–948, 1990.

[13] A. Ultsch. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 33–45, Amsterdam, 1999. Elsevier.

[14] A. Yuille, P. Stolorz, and J. Utans. Statistical physics, mixtures of distributions, and the EM algorithm. *Neural Computation*, 6:334–340, 1994.