

# Overtraining and model selection with the self-organizing map

Jouko Lampinen and Timo Kostiainen  
Laboratory of Computational Engineering  
Helsinki University of Technology  
P.O. Box 9400, FIN-02015 Espoo, FINLAND  
{Jouko.Lampinen,Timo.Kostiainen}@hut.fi

## Abstract

*We discuss the importance of finding the correct model complexity, or regularization level, in the self-organizing map (SOM) algorithm. The complexity of the SOM is determined mainly by the width of the final neighborhood, which is usually chosen ad hoc or set to zero for optimal quantization error. However, if the SOM is used for visualizing the joint probability distribution of the data, then care must be taken not to overfit the model to the data sample, similarly as with any statistical model. We propose a heuristic criterion for model selection in SOM, and demonstrate by simulations that the criterion can be used for selecting the neighborhood that suppresses overfitting.*

## 1 Introduction

A key issue in statistical model fitting is to select the correct model complexity. In the self-organizing map (SOM [4]) the model complexity, that is, the effective number of parameters, is determined by the number of nodes in the map and the regularization effect caused by the final neighborhood, which roughly corresponds to a smoothing prior on the positions of the units on the map. The model complexity in SOM is usually chosen *ad hoc* by the user. Often training is run down to zero neighborhood, to obtain maximal fit to the data. Most of the procedures proposed for model selection in SOM aim to select the neighborhood, or lattice dimension, so that the resulting map optimally preserves topological relations in the training data, see e.g. [1], [8] and [3]. This approach of searching for the maximum fit to the training data is reasonable in data-analysis and visualization when we are analyzing the population instead of a random sample of it, so that any phenomena in the data are relevant.

However, if the SOM is used as a representation of the joint probability density of the data, then care must be taken not to overfit the model to the data sample, similarly as with any statistical model. Fig. 1 shows an

example of overfitting in the SOM. The map was trained on real-world data consisting of eleven variables, four of which are shown in the figure.

The nearest neighbor type density estimate in the SOM renders the method difficult to analyze. Replacing the Voronoi regions with regular kernels facilitates more advanced methods, shortly reviewed below. In this contribution, we propose an error criterion for the basic SOM algorithm that can be used to select the correct map complexity using cross-validation or any other similar procedure, and demonstrate the performance of the method.

Algorithms similar to the SOM have been presented in which the nearest neighbor rule has been replaced with soft probabilistic assignments to units. The units are then components in a somehow modified Gaussian mixture density model. The parameters of such a model can be optimized using the framework of probability theory.

Utsugi [6, 7] has presented a density model whose mixture components constitute an elastic net. A smoothing prior regularizes the ordering of the components. This approach enables the use of efficient estimation methods: Utsugi derives a method, similar to Bayesian evidence framework, for obtaining maximum a posteriori estimates for the parameters that control the topological constraint and noise variance. Also, a Bayesian approach to the comparison of different topologies is presented.

In the generative topographic mapping [2], the Gaussian mixture model is constrained by a nonlinear mapping from a regularly organized distribution in a latent space to the component centroids in data space. Hyperparameters of the model, which control noise variance, stiffness of the nonlinear mapping, and the prior distribution of weights, can be optimized based on log-likelihood or Bayesian evidence [5].

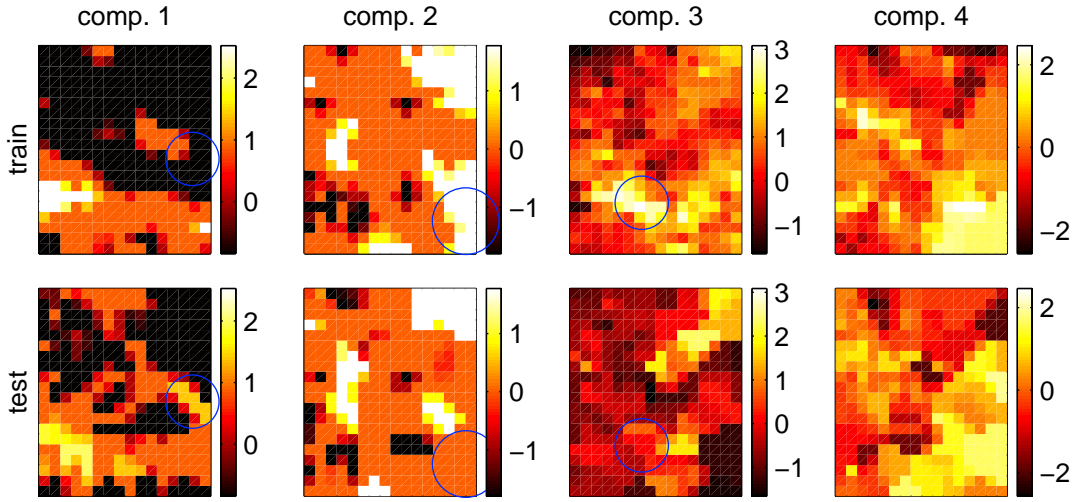


Figure 1: Example of overfitting in the SOM. The top row shows means of training data in the Voronoi cells around the reference vectors, and the bottom row shows the respective means for a separate test data. The circles point two areas where overfitting of a variable has occurred: the units on the map have large values, which are lacking from the test data. Thus, the conclusion that the area represents a cluster with a large value for the variable would be quite erroneous.

## 2 Measure for generalization in the SOM

In this section, we propose a heuristic measure for comparing the generalization capabilities of different SOM models. Note that the quantization error for validation data is an inappropriate measure, because it ignores map units that are far away from the validation data, i.e., the overfitted units that easily cause false interpretation on the map. The same applies also to the error function that is stochastically minimized by the SOM training

$$E = \sum_k \sum_i h_{ck} \|\mathbf{x}^i - \mathbf{m}^k\|^2, \quad (1)$$

where  $c$  denotes the best matching unit for input  $\mathbf{x}^i$ ,  $\mathbf{m}^k$  is the reference vector of unit  $k$  and  $h_{ck}$  is the value of the neighborhood function between units  $c$  and  $k$ .

The error measure should have a term that measures how accurately the data distribution is modeled by the map, as well as a term that penalizes units for drifting away from the data.

Let the  $d$ -dimensional data samples  $\mathcal{X}$  be divided into a training set  $\mathcal{X}_T$  and a validation set  $\mathcal{X}_V$ . The SOM is trained by the data  $\mathcal{X}_T$ . The proposed measure for the

generalization error is

$$\epsilon = \frac{1}{d} \frac{1}{M} \sum_{j=1}^d \sum_{i=1}^M \left\{ \left[ E(\mathbf{x}_T^j | \mathbf{x}_T \in V_i) - E(\mathbf{x}_V^j | \mathbf{x}_V \in V_i) \right]^2 + \frac{1}{2} \left[ Var(\mathbf{x}_T^j | \mathbf{x}_T \in V_i) + Var(\mathbf{x}_V^j | \mathbf{x}_V \in V_i) \right] \right\}, \quad (2)$$

where  $V_i$  denotes the receptive region or Voronoi cell associated with map unit  $i$ . The first term in the formula is the squared distance between the means of training data and validation data within each Voronoi cell. It measures overfitting as systematic deviation in the projections of training and validation data onto the map. The second term is the average variance of training and validation data within the cells. For cells with no hits we interpolate the means and variances from the nearest topological neighbors.

The proposed error can be used for selecting the best generalizing map from several candidates during the search for the optimal hyperparameters. The neighborhood function can vary from one map unit to another, corresponding to input dependent smoothing, or each variable can have its own neighborhood width, corresponding to determination of relevance of the variables. However, adjusting a large number of hyperparameters so as to minimize the validation error may easily cause the map to be overfitted to the validation data. Also,

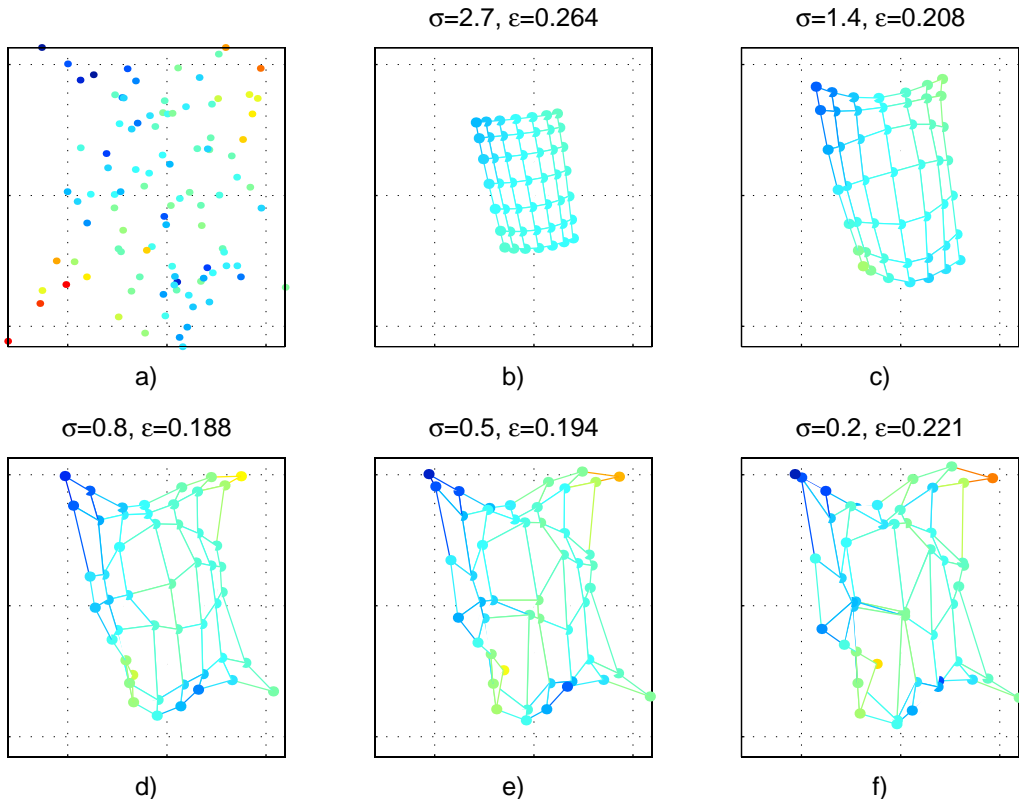


Figure 2: Generalization error of SOM at different phases of training. Value of third data component is shown on color scale running from blue to green to red. *a)* Both training and validation data points. *b) – f)* The maps.  $\sigma$  is the width parameter of Gaussian neighborhood function and  $\epsilon$  is the value of the generalization error.

the error measure provides no direct method for estimating the hyperparameters, contrary to the Bayesian evidence framework in [5, 6], so the optimization of a large number of parameters may be computationally expensive.

In practice we reduce the neighborhood width monotonically and choose the width that gives the minimum error averaged over all variables. Another, more flexible procedure is to record the optimal width for each variable separately during the decreasing of the neighborhood, and retrain the map with the individually selected neighborhood widths.

### 3 Experiments

Fig. 2 illustrates the variation of the generalization error at different phases of training in an artificial problem. The data was generated from a saddle-like surface ( $z = xy$  around the origin). As the neighborhood gets smaller the stiffness of the map decreases and the folds in the map become smaller. Below the neighborhood width

$\sigma = 0.8$  (graph *d*) the error of validation data starts to grow, indicating that the folds follow random variations in the training data.

Fig. 3 shows an example of feature selection based on various error measures of the SOM. The proposed measure is compared to the quantization error and the measure for topology preservation presented in [3]. The artificial data consisted of eight features, one of which was purely random while the others shared certain functional dependencies, as shown below:

$$\begin{aligned}
 x_1 &\sim \text{Uniform}(-1, 1) \\
 x_2 &\sim \text{Uniform}(-1, 1) \\
 x_3 &= x_1 x_2 \\
 x_4 &= \sin[2(x_2 + x_3)] \\
 x_5 &= \exp(x_1) - x_2 \\
 x_6 &\sim \text{Uniform}(-1, 1) \\
 x_7 &= x_6^2 \\
 x_8 &\sim \text{Uniform}(-1, 1).
 \end{aligned}$$

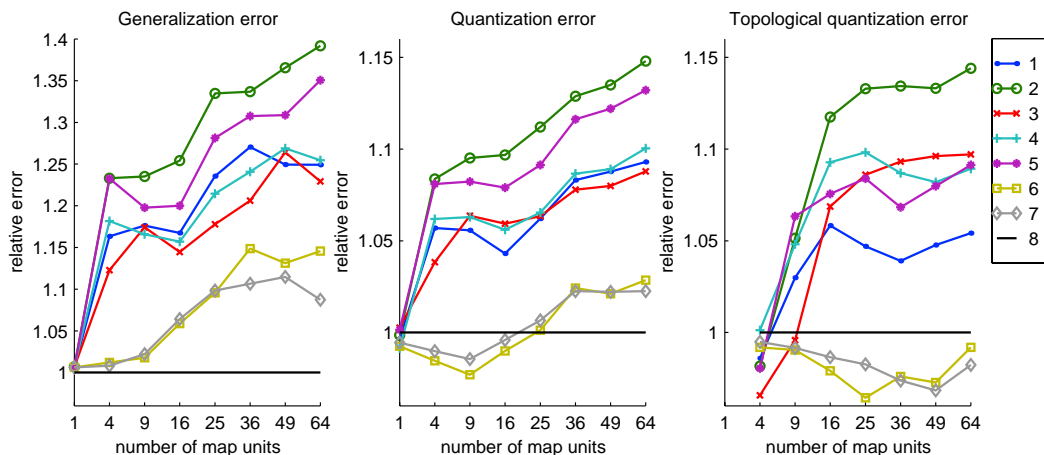


Figure 3: Discrimination of random feature. Values of different SOM error measures for data sets with one variable excluded in turn (see text). Left: the proposed generalization error. Middle: quantization error. Right: a measure of topology preservation proposed in [3]. The values are normalized so that the data set with the irrelevant feature excluded (curve 8) gives unity error.

All the features were scaled to have equal variance. Different size two-dimensional SOMs were trained with the intent to distinguish the random feature  $x_8$  from the others. Each feature in turn was left out from training, with the hypothesis that the map that did not contain the irrelevant variable would have smaller error than the others. The proposed generalization error appears more reliable than the other error measures in picking out the random feature. Note that features  $x_6$  and  $x_7$  were only related to each other while the other dependencies were more complicated.

Fig. 4 shows an example from a real data analysis task where the overfitting in SOM was a crucial problem. One of the goals in the project was to analyze the effect of different factors to the subjective perception of indoor air [9]. In the project we are using SOM as one tool (together with factor analysis, Bayesian generalized linear models and non-linear regression by MLP) for clustering the variables related to physical and chemical measurements, psychosocial working environment and occupants' medical records. In the bottom row of Fig. 4 separate neighborhood widths for each variable were estimated as explained in section 2. Since the optimal neighborhood was very large for some of the features the map can follow the other features rather accurately, as can be seen in the bottom map where the first three components resemble the test data closely.

## 4 Conclusion

We have proposed a heuristic measure for the generalization error in SOM. The measure can be used for

choosing the map that gives the most reliable view on the data, suppressing overfitted units that represent only random variations in the training data. In addition, by estimating different neighborhood sizes for each variable the SOM can organize according to variables that have statistical dependencies and smooth out the model with large neighborhoods for the variables that are random or cannot be modeled due to an insufficient training data set.

However, experiments have suggested that the width of the neighborhood is often an insufficient regularization method to prevent overfitting; the map may start to overfit as soon as the neighborhood is reduced to any practical level, indicating that some other forms of regularization may also be needed.

Another problem that should be further considered is the border effect that causes maps with large neighborhoods to shrink, as in Fig. 2. This causes large quantization errors for stiff maps, leading to unnecessarily small neighborhoods even when the underlying distribution is smooth.

## References

- [1] Bauer, H.-U., Pawelzik, K. & Geisel, T. (1992). A topographic product for the optimization of self-organizing feature maps. In J. E. Moody, S. J. Hanson & R. P. Lippmann, eds., *Advances in Neural Information Processing Systems 4*, pp. 1141–1147. Morgan Kaufmann, San Mateo, CA.
- [2] Bishop, C., Svensén, M. & Williams, C. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1).
- [3] Kaski, S. & Lagus, K. (1996). Comparing self-organizing

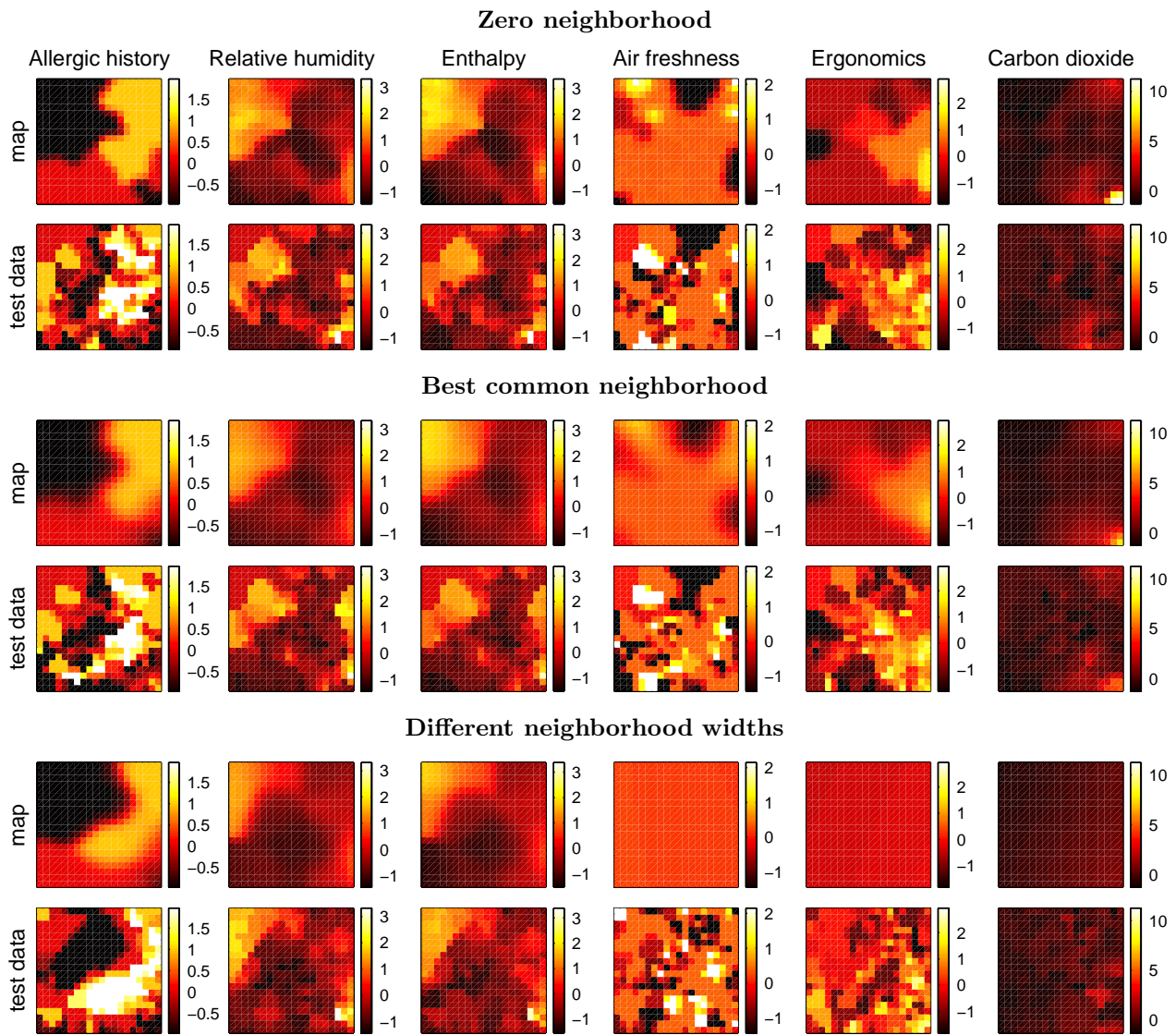


Figure 4: SOM model selection in real data analysis task. The top of the figure shows the final map with zero neighborhood and test data mapped onto it. The middle rows shows the map with the minimum total error (neighborhood width  $\sigma = 2$ ). The bottom map is the result when different neighborhood widths were selected. Here the components *Air freshness*, *Ergonomics* and *Carbon dioxide* have very large neighborhoods, indicating that the data do not cluster well with respect to those variables, and the map is organized according to the other components, which are now fitted comparatively well.

- maps. In *Proc. ICANN'96, International Conference on Artificial Neural Networks*, vol. 1112, pp. 809–814. Springer, Berlin.
- [4] Kohonen, T. (1997). *Self-Organizing Maps*, vol. 30 of *Lecture Notes in Information Sciences*. Springer, 2nd edn.
- [5] Svensén, M. (1998). *GTM: The Generative Topographic Mapping*. Ph.D. thesis, Aston University, Birmingham, UK.
- [6] Utsugi, A. (1996). Topology selection for self-organizing maps. *Network: Computation in Neural Systems*, 7:pp. 727–740.
- [7] Utsugi, A. (1997). Hyperparameter selection for self-organizing maps. *Neural Computation*, 9(3):pp. 623–635.
- [8] Villmann, T., Der, R., Herrmann, M. & Martinez, T. M. (1997). Topology preservation in self-organizing feature maps: exact definition and measurement. *IEEE Transactions on Neural Networks*, 8(2):pp. 256–66.
- [9] Welling, I., Kähkönen, E., Lahtinen, M., Valkonen, J., Lampinen, J., Varsta, M. & Kostianen, T. (1999). Real time indoor air monitoring system and analysing method. *American Journal of Industrial Medicine*. To appear.