Regularization and Model Selection in the Context of Density Estimation

Martin Kreutz[1*] Anja M. Reimetz[2**]
Bernhard Sendhoff[1] Claus Weihs[2]
Werner von Seelen[1]

[1]Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
[2]Fachbereich Statistik, Universität Dortmund, Germany

May 20, 1999

Abstract

We propose a new information theoretically based optimization criterion for the estimation of mixture density models and compare it with other methods based on maximum likelihood and maximum a posterio estimation. For the optimization, we employ an evolutionary algorithm which estimates both structure and parameters of the model. Experimental results show that the chosen approach compares favourably with other methods for estimation problems with few sample data as well as for problems where the underlying density is non-stationary.

1 Introduction

The estimation of the probability density function (pdf) of a data generating process given a finite sample of observations is a very fundamental tool in statistics. The reason to estimate the density is, essentially, twofold: firstly,

---

[*] email: Martin.Kreutz@neuroinformatik.ruhr-uni-bochum.de

[**] email: reimetz@statistik.uni-dortmund.de

the density itself may be the subject of interest, or, secondly it may serve as a base for other statistical tasks including data compression, classification and regression [4, 14, 16]. Over the last decades, a variety of methods have been proposed in the field of statistics and artificial neural networks. However, the task of density estimation still remains to be difficult. This especially holds for the estimation of probability distributions in the case of small sample sizes or non-stationary distributions.

We are concerned with the case in which no prior knowledge about the parametric form of the distribution can be assumed. As a general (semi-)parametric model we consider finite mixtures of normal densities. After a short introduction to mixture models in Sec. 2, we address the problem of model selection, which includes both the estimation of structure and parameters of the model.

If the underlying system is not completely described by the sample data, additional conditions have to be imposed on model selection. A systematical method is the introduction of regularization terms in the optimization criterion. In Sec. 3, we compare three different methods to penalize the log-likelihood function with a regularization term and discuss them in the framework of *penalized maximum likelihood estimation.* A new information theoretically based criterion will be proposed which has the appealing property that no external regularization parameters are required.

The structure estimation problem as well as some of the regularization terms lead to a non-differentiable objective function. Therefore, gradient-based methods are not applicable and direct optimization procedures must be applied. Especially in the context of structure optimization evolutionary algorithms (EAs) have been considered to be a very promising approach to deal with these problems. In Sec. 4, we outline the employed EA. We show that the proposed EA combined with the information criterion outperforms the other presented methods.

2 Gaussian mixture models

Mixtures of densities have been considered as very general and computationally efficient semi-parametric models for density estimation.

They consist of a convex combination of m parametric component densities

$\phi_i(\vec{x}|\theta_i), i = 1, \ldots, m$:

$$\hat{p}(\vec{x}|\vec{\theta}) \;=\; \sum_{i=1}^{m} \alpha_i \, \phi_i(\vec{x}|\theta_i) \;, \quad \vec{x} \in \mathrm{I\!R}^n \;, \quad \vec{\theta} = (\alpha_i, \ldots, \alpha_m, \theta_i, \ldots, \theta_m) \;. \tag{1}$$

$$\sum_{i=1}^{m} \alpha_i = 1 \;, \quad \alpha_i \geq 0 \; \forall \, i = \{1, \ldots, m\} \;. \tag{2}$$

The vector of parameters $\vec{\theta}$ characterizing $\hat{p}$ includes the weighting coefficients α_i and the parameters θ_i of the component densities. In this article, we employ finite mixtures of normal densities:

$$\phi_i(\vec{x}|\theta_i) \;=\; \frac{1}{\sqrt{2\pi}^n \prod_{k=1}^{n} \sigma_{ik}} \exp\left(-\frac{1}{2} \sum_{k=1}^{n} \left(\frac{x_k - \mu_{ik}}{\sigma_{ik}} \right)^2 \right) \tag{3}$$

$$\theta_i \;=\; (\mu_{i1}, \ldots, \mu_{in}, \sigma_{i1}^2, \ldots, \sigma_{in}^2) \in \mathrm{I\!R}^{2n} \;, \tag{4}$$

which have several appealing properties: they are universal in the sense that they can approximate any continuous probability distribution [1] similar to radial basis function networks [18]; they can cope with multi-modal distributions and their complexity can be easily adjusted by the number of components.

3 Regularization

The method of maximum likelihood is widely used in statistical inference. For continuous distributions the likelihood of a sample $X = \{\vec{x}_1, \ldots, \vec{x}_N\}$ is defined by the joint density of X for the chosen probability density model $\hat{p}(\vec{x}|\vec{\theta})$ (characterized by its parameters $\vec{\theta}$). Assuming the $\vec{x}_i$ to be iid. sampled with pdf $p(\vec{x}|\vec{\theta})$ the log-likelihood function reads

$$\ell(\vec{\theta}) \;=\; \log \prod_{k=1}^{N} p(\vec{x}_k|\vec{\theta}) \;=\; \sum_{k=1}^{N} \log \, p(\vec{x}_k|\vec{\theta}). \tag{5}$$

A maximum likelihood estimation corresponds intuitively to the most likely model which would give rise to the data X. A commonly used estimation method is the EM algorithm [8, 19] which is particularly useful for Gaussian mixtures since both steps of the EM can be performed analytically. However, if functions of one or more continuous variables are involved, the principle of maximum likelihood may be inadequate. In these cases the attempt to maximize the likelihood may result in an infinite value of the likelihood and degeneracy of the model. In the context of Gaussian mixtures, Eq. (3), we observe that the likelihood will be infinite, if one of the density functions ϕ_i collapses to a delta function placed on one data point. In structure optimization of Gaussian mixtures, methods solely based on the likelihood, therefore, tend to increase the number of component functions, place their centers on the data points and minimize their widths. A general approach to transform this ill-posed optimization problem into a well-posed problem is the introduction of regularization terms which reflect specific assumptions about the density model. The aim is the simultaneous minimization of bias and variance of the model which is possible in the case of infinite data sets but in practical applications leads to the *bias-variance dilemma* [9]. Essentially, regularization reduces the effective degrees of freedom [17].

In the remainder of this section we review two commonly used regularization methods and propose a new method which is based on the relation between the log-likelihood and the Shannon entropy.

3.1 Roughness penalties

A sensible choice for regularization is to demand a smooth density model. A common choice of a *smoothness* functional $J(\vec{\theta})$ is the integral of the squared second derivative

$$J(\vec{\theta}) \;=\; \int\limits_{-\infty}^{+\infty} p''(x|\vec{\theta})^2 \, dx, \qquad \vec{x} \in \mathbb{R} \;, \tag{6}$$

which has an appealing interpretation as the global measure of curvature of p and can be viewed as a special form of a Tikhonov regularizer [25].

Similar terms have been used in the context of penalized maximum likelihood estimation [11, 23]. The employed terms were derived only for univariate densities. The multivariate case has been treated in the context of artificial neural networks [2, 3]. In this approach, however, the integral in Eq. (6) was approximated by the sum over the training sample leading to a data-dependent smoothness functional. Furthermore, the off-diagonal elements of the Hessian matrix were discarded. In [15] the integral in Eq. (6) has been extended to the multivariate case and solved analytically. Hence, the complete objective function reads

$$F(\vec{\theta}) \;=\; \ell(\vec{\theta}) - \gamma \, J(\vec{\theta}) \ . \tag{7}$$

The *smoothing* parameter γ controls the relative influence of both criteria. However, due to the introduction of $J(\vec{\theta})$ the M-step in the EM algorithm is no longer analytically tractable. Heuristic methods [10, 24] turned out to be numerically instable. Hence, we use a quasi-Newton optimization method in the M-step in order to be still able to apply the EM algorithm.

Another heuristic approach is to use artificial noise during the estimation. In the limit of an infinite sample and under a quadratic approximation this corresponds to a special Tikhonov regularization [20, 5]. In each E-step of the EM the whole sample is corrupted by additive normal distributed noise with variance σ^2 which takes the role of the smoothing parameter γ. In this case the standard EM can be used without any change. However, the choice of the right smoothing parameter still poses a difficult problem.

3.2 Entropy based regularizations

Another sensible measure has been proposed by the authors and is based on the following consideration: Let $p(\vec{x})$ denote the true density and $\hat{p}(\vec{x})$ the estimation of p, respectively. In the case of a continuous distribution with an infinite sample the log-likelihood reads:

$$\ell(\vec{\theta}) \;=\; \int\limits_{-\infty}^{+\infty} \cdots \int\limits_{-\infty}^{+\infty} p(\vec{x}) \, \log \hat{p}(\vec{x}|\vec{\theta}) \, d\vec{x} \ , \qquad \vec{x} \in \mathrm{I\!R}^n \ , \tag{8}$$

If the estimation $\hat{p}$ coincides with p the log-likelihood is just the negative of the Shannon entropy H [7]:

$$\hat{p} \;=\; p \quad \Rightarrow \quad \ell(\vec{\theta}) \;=\; \int\limits_{-\infty}^{+\infty} \cdots \int\limits_{-\infty}^{+\infty} p(\vec{x}) \, \log p(\vec{x}) \, d\vec{x} \;=\; -\, H(p(\vec{x})) \ . \tag{9}$$

The Shannon entropy is a measure of the uncertainty of a random variable and can also be considered as a measure of "disorder" of a probability distribution. Maximization of the entropy represents the minimization of the bias in the choice of the model. In the case of an uniform distribution the entropy reaches its maximum, which corresponds to the fact that instances of an uniformly distributed random variable do not show any order or structure.

On the other hand the entropy attains its minimum for a peaked distribution since such a distribution is highly ordered: all probability mass is concentrated in only one peak This relation between the log-likelihood and the Shannon entropy leads directly to a new optimization criterion which provides a tradeoff between underestimation which corresponds to a low log-likelihood and overestimation which corresponds to a low entropy:

$$Q(\vec{\theta}) \;=\; \frac{1}{N}\sum_{k=1}^{N} \log\, \hat{p}(\vec{x}_k|\vec{\theta}) + H(\hat{p}(\vec{x}|\vec{\theta}))\ . \tag{10}$$

In this sense the criterion $Q(\vec{\theta})$ accounts for what is called the *bias variance dilemma* (which can only be escaped in the limit of an infinite sample) [9, 13].

This criterion includes a very intuitive regularization imposed by the Shannon entropy. Overspecialized solutions would increase the log-likelihood but, on the other hand, would decrease the Shannon entropy. This criterion can be viewed as a penalized maximum likelihood criterion. In the Bayesian sense this corresponds to a maximum a posteriori estimation with an entropic prior [12]. However, in all these approaches some kind of regularization parameters must be set. This is not necessary if we use the new criterion $Q(\vec{\theta})$ in Eq. 10.

Another objective may be the minimization of the absolute value of $Q(\vec{\theta})$. In contrast to the log-likelihood which may reach infinity the criterion $|Q(\vec{\theta})|$ possesses a lower bound. In this paper, we pursue the penalized maximum likelihood approach.

It is not possible to solve the entropy integral analytically for Gaussian mixtures. Therefore, it must be approximated by e.g. Monte Carlo sampling. This fact renders maximization methods like gradient descent or EM algorithms computationally inefficient. On the other hand, since direct optimization methods like EA only rely on the value of the objective function, they provide a well suited framework for optimizing non-differentiable and noisy functions.

4 Evolution of Gaussian mixtures

Two difficulties arise in the context of model estimation: firstly, the optimization of both structure and parameters of a model is carried out in a search space which is generally non-differentiable and multi-modal, and secondly, as shown in Section 3, the introduction of regularizations may lead to complicated and noisy optimization criteria. Evolutionary algorithms have been considered to be a promising method for dealing with this class of optimization problems [26]. However, the representation of solutions and the corresponding evolutionary operators have to be chosen carefully and in general depend on the problem. We will not be able to discuss the specific EA in detail, refer to [15], but point out some conditions, which lead to specific operator settings.

The parameters θ_i of each component are encoded directly in the genome. In the context of recombination the problem of *competing conventions* [21] arises: Due to the invariance to permutations of the components distinct genomes map to functionally equivalent phenotypes (models). This problem is circumvented by using a crossover operator which samples the crossover points in the input space, see Fig. 1 rather than on the position in the genome (a similar operator has been proposed in the context of evolutionary optimization of radial basis function networks [6]).

Several mutation operators for structure modification of the model have been considered. A very straightforward choice are simple insertion and deletion operators. However, the insertion and deletion of components in the model can be very disruptive and violate the principle of a strong causality in EAs [22]. In order to minimize these effects and to better control the impact of mutation, we employ special *merge* and *split* operators which try to increase and decrease, respectively, the number of components while at the same time keeping the effect of mutation as small as possible, see Fig. 2. Details of these operators are discussed in [15].

For the optimization of Gaussian mixtures we employ the operators depicted in Figs. 1 & 2. All operators are followed by a local adaptation of the model parameters which is performed by a single EM iteration. This means, essentially, that undirected structure variations are followed by directed (with respect to the log-likelihood) parameter mutations. In order to be applicable the EM has to act on the unregularized likelihood surface, whereas the EA selects solutions according to the regularized objective function. Since an EM step usually contains a significant component in the direction of an regularized optimum, see Fig. 3, the EA will be able to locate these optima. This separation between the selection and adaptation process makes it possible to combine complicated, non-differentiable and noisy objective functions with fast local estimation methods.

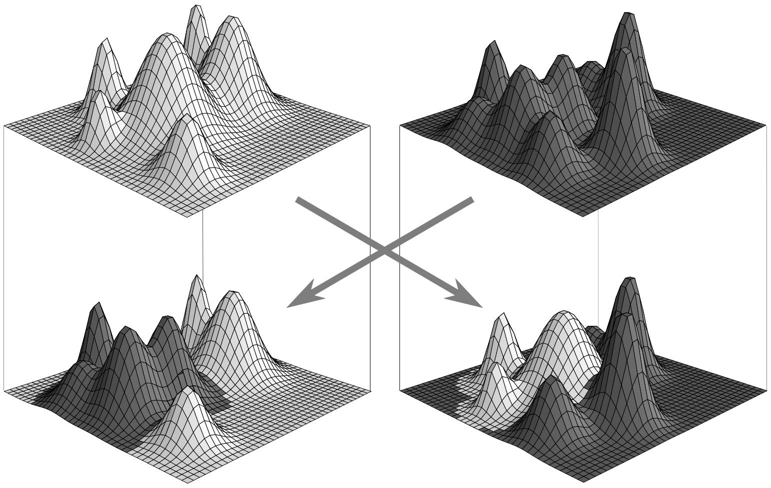

Figure 1: Crossover operator: The portions of mixture components of both parents lying in the crossover section are exchanged.

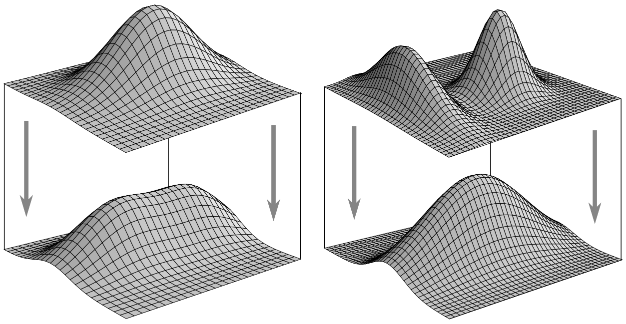

Figure 2: Mutation operator: A randomly selected component is split into two new components (left), two mixture components are merged together to form a new component (right). Both operators try to keep the change in the model and therefore the impact of mutation as small as possible.

5 Experimental results

In order to assess the performance of the proposed method, we employed the EA combined with the information criterion to density estimation problems. The first task was a density estimation problem with few sample data. The sample was generated from a normal distribution with $\sigma = 1/16$ along a ring with radius $r = 1$, see Fig. 4 (left). The sample size for both training and test data set was 500.

In order to give an impression of the effect of roughness penalties, we

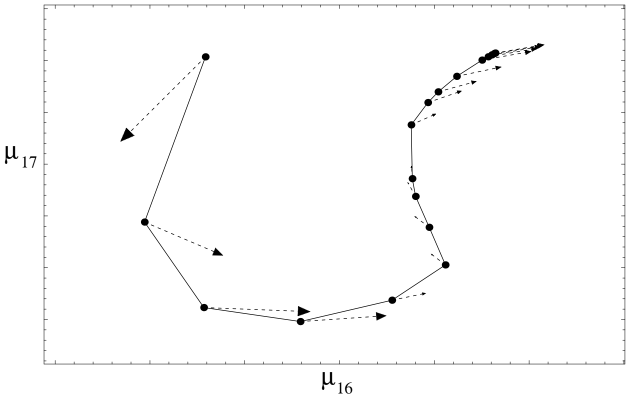

Figure 3: The solid line shows the development of two parameters μ_{16} and μ_{17} of a Gaussian mixture with 20 kernels during a penalized maximum likelihood estimation using the roughness functional $J(\vec{\theta})$. The dashed lines show the direction of the corresponding unregularized EM steps.

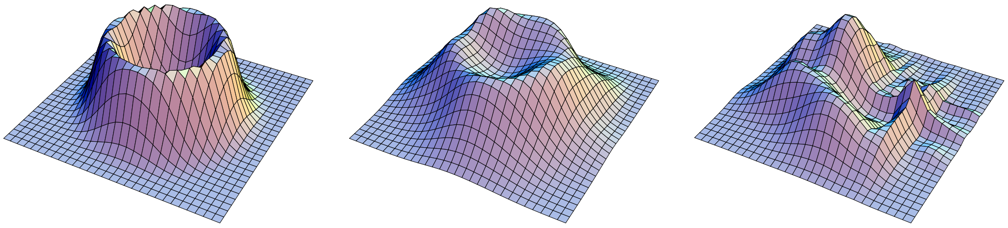

Figure 4: (left) The original density $p(\vec{x})$, (middle) the estimated density using the information criterion Q. (right) The estimated density using $\gamma = 0.001$, which corresponds to a "typical" value from the plateau in Fig. 5(left)

compared both regularization methods described in Sec. 3.1, see Fig. 5. Note, that both methods depend on a regularization parameter.

The information criterion Q does not depend on any regularization parameter. Fig. 6 shows the development of the negative log-likelihood divided by the sample size (training and test) and the entropy during the course of evolution. If no regularization is used the EA produces over-specialized solutions with a poor performance on the test sample, see Fig. 6 (left). With the information criterion the negative log-likelihood on both samples and the entropy remain in the same range. No overfitting can be observed while the performance on the test sample is better than the best performance obtained with an unregularized maximum likelihood estimation. The negative likelihood on the test set (ca. 1260) is higher than for all but the best choices

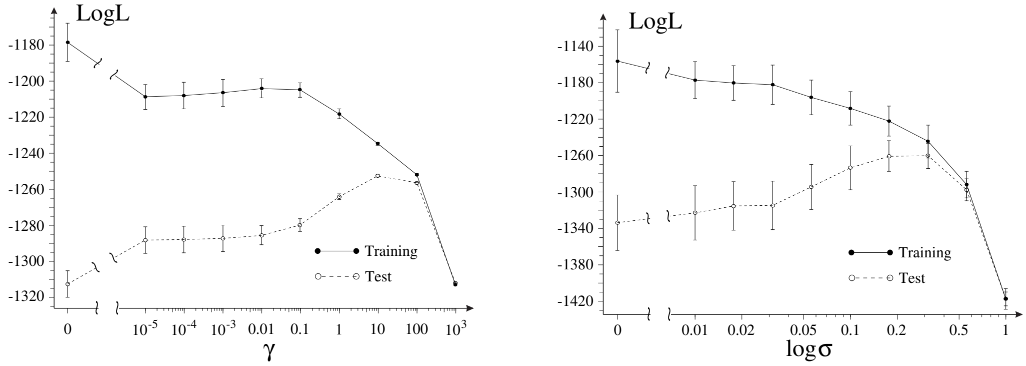

Figure 5: (left) Regularization with a roughness penalty, (right) regularization with noisy training samples. In both cases the log-likelihood of the training sample (solid line) and the test sample (dashed line) is shown for different values of the respective regularization parameter. All results are averaged over 100 runs.

of the regularization parameters. Only the best regularized results, Fig. 5, reach roughly the same value. The final estimate is depicted in Fig. 4 (right).

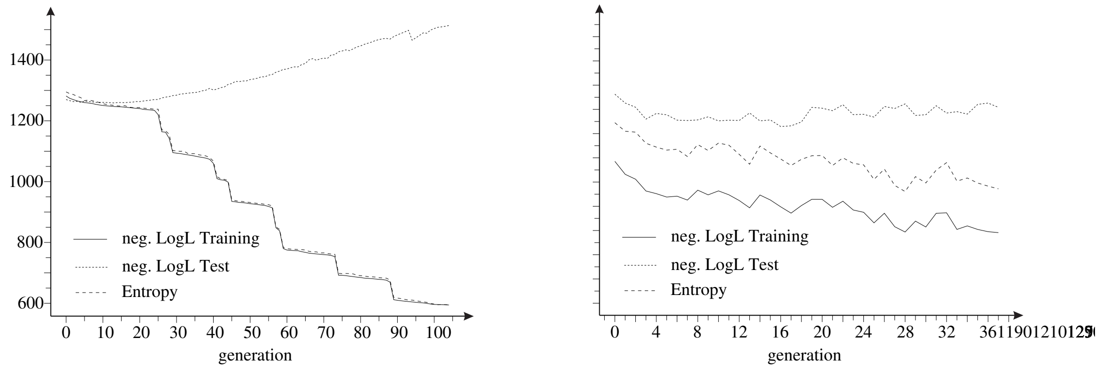

Figure 6: (left) Evolution of a Gaussian mixture without any regularization, (right) using the information criterion. The solid line shows the log-likelihood of the training sample and the dotted line the log-likelihood of the test sample, respectively. For comparison the entropy is shown by the dashed line.

For the second task we applied our method to the estimation of a non-stationary density. Especially for non-stationary densities, it can be difficult to identify one regularization parameter which results in the optimal likelihood on the test set for all possible densities during the estimation. Since the proposed method is independent from any regularization parameters, it is particularly applicable to case where the density is non-stationary. The EA was used to fit the Gaussian mixture model to a normal density with $\sigma = 1/16$ along a ring with radius $r \in [1, 2]$ while the radius is changing over time: The radius r describes one period of a sine wave over the interval of 1000 generations with $r_{min} = 1$ and $r_{max} = 2$, respectively. As depicted in Fig. 7 the EA is able to follow the distribution.

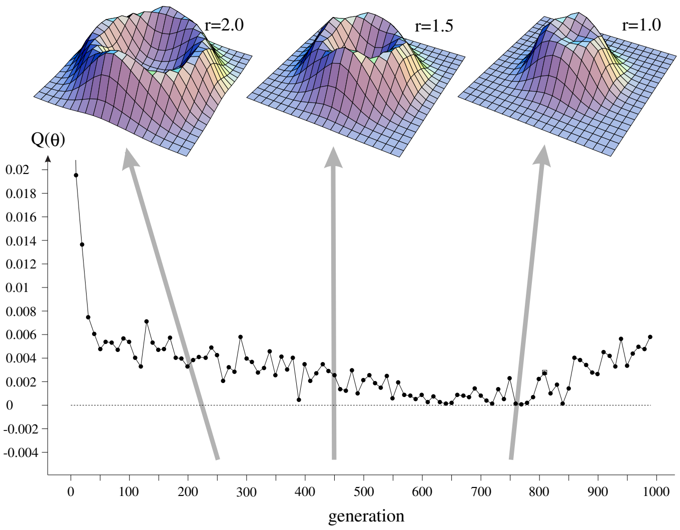

Figure 7: The plot shows the information criterion Q evaluated on the test sample over the course of evolution. For the optimal model Q would be zero, see dotted line. The three small plots above the main plot are snapshots taken from the EA with operator-adaptation at generation 250, 500 and 750, respectively. All values are averaged over 16 runs and 10 generations.

Since Q must be evaluated by stochastic methods like Monte Carlo sampling the curve in Fig. 7 exhibits rather large fluctuations. This can be avoided by increasing the number of Monte Carlo samples.

6 Conclusion

We compared different regularization methods in the context of structure and parameter optimization of density estimation models. Most of them depend on additional parameters which have a strong impact on the performance of the method. We proposed an optimization criterion which was inspired by the relation between the log-likelihood and the Shannon entropy.

The criterion can be interpreted as a maximum a posteriori criterion with an entropic prior. However, it does not depend on any additional parameters. An evolutionary algorithm was applied to the optimization of structure and parameters of Gaussian mixture models in conjunction with the proposed information criterion. Experiments with density estimation problems with small sample sizes as well as with non-stationary distributions show the superiority of the new criterion.

References

[1] A. R. Barron and C. H. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statistics*, 1(3):1347–1369, 1991.

[2] C. M. Bishop. Improving the generalization properties of radial basis function neural networks. *Neural Computation*, 3(4):579–588, 1991.

[3] C. M. Bishop. Curvature-driven smoothing: A learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 4(5):882–884, 1993.

[4] C. M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Dept. of Computer Science and Applied Mathematics, Aston University, Birmingham, U.K., 1994. Available from `www.ncrg.aston.ac.uk`.

[5] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[6] B. Carse and T. C. Fogarty. Tackling the "curse of dimensionality" of radial basis function neural networks using a genetic algorithm. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature IV*, pages 710–719. Springer, 1996.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc., Ser. B*, 39(1):1–38, 1977.

[9] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[10] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Royal Stat. Soc., Ser. B*, 52(3):443–452, 1990.

[11] P. J. Green. Penalized likelihood. *Encyclopaedia of Statistical Sciences, update volume*, 1996.

[12] S. F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 53–71. Kluwer, 1989.

[13] T. Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433, 1998.

[14] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[15] M. Kreutz, A. M. Reimetz, B. Sendhoff, C. Weihs, and W. von Seelen. Optimisation of density estimation models with evolutionary algorithms. In A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature V*, pages 998–1007. Springer, 1998.

[16] M. Kreutz, A. M. Reimetz, B. Sendhoff, C. Weihs, and W. von Seelen. Structure optimization of density estimation models applied to regression problems with dynamic noise. In D. Heckerman and J. Whittaker, editors, *Artificial Intelligence and Statistics 99*, pages 237–242. Morgan Kaufmann, 1999.

[17] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[18] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[19] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[20] R. Reed, R. J. Marks II, and S. Oh. Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter. *IEEE Transactions on Neural Networks*, 6(3):529–538, 1995.

[21] J. D. Schaffer, D. Whitley, and L. J. Eshelman. Combinations of genetic algorithms and neural networks: A survey of the state of the art. In *Proc. Combinations of Genetic Algorithms and Neural Networks (COGANN-92)*, pages 1–37, 1992.

[22] B. Sendhoff, M. Kreutz, and W. von Seelen. A condition for the genotype–phenotype mapping: Causality. In T. Bäck, editor, *Proc. International Conference on Genetic Algorithms*, pages 73–80. Morgan Kaufman, 1997.

[23] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statistics*, 10:795–810, 1982.

[24] B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka. A smoothed em approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Royal Stat. Soc., Ser. B*, 52(2):271–324, 1990.

[25] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill Posed Problems.* Wiley, 1977.

[26] L. D. Whitley. Genetic algorithms and neural networks. In J. Periaux and G. Winter, editors, *Genetic Algorithms in Engineering and Computer Science.* Wiley, 1995.