Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Evolution of adaptive nonlinear models

Martin Kreutz, Anja Busse, Bernhard Sendhoff

2000

Preprint:

This is an accepted article published in Seventh International Conference on Neural Information Processing – Proceedings. The final authenticated version is available online at: https://doi.org/[DOI not available]

Evolution of adaptive nonlinear models*

Martin Kreutz†, Anja M. Busse‡, Bernhard Sendhoff§,

† kreutz@zn-gmbh.com, ZN GmbH, Bochum, Germany
‡ FB Statistik, Universität Dortmund, Germany &
Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
§ Future Technology Research (FTR), HONDA R&D EUROPE GmbH, Germany

Abstract

The evolution of adaptive models as well as the interaction between learning and evolution has elicited much interest in recent time. In this paper we focus on evolutionary structure optimization of nonlinear statistical models and the combination of structure optimization and learning in these models. For the experimental evidence the optimized models are applied to the task of time series prediction where the data stems from a real world system consisting of physiological data of apnea patients.

1 Introduction

The structure of an adaptive model has a direct impact on training time, convergence, generalisation ability, robustness, etc. of the model. Therefore, the choice of at least a near optimal structure is one of the key issues in model building. The optimization of the structure of a model, however, represents a difficult problem since the search space is in general non-differentiable and multimodal. In recent years evolutionary algorithms turned out to be successful for the task of structure optimization [21, 25]. Yet, the right choice of the encoding of the structure in combination with the evolutionary operators acting upon it still constitutes a difficult task. In this paper we consider the optimization of a particular type of statistical models, called mixture of densities, which is introduced in Sec. 2. The reasons for the choice of a density estimation model are twofolded. First, the density of a data distribution (if it exists) is the most basic probabilistic description of the data generating process. All relevant values, e.g. conditional moments, can be derived. Second, a probabilistic approach allows for the systematic inclusion of all additional information that can be modelled probabilistically like assumptions of special distributions of the noise, missing data, etc. In Sec. 3 we propose an evolutionary algorithm that is employed for the structure optimization as well as for the optimization of the parameters of the model. We introduce the respective operators and discuss the role of learning in this approach. Experimental evidence for the strength of the proposed approach is given in Sec. 4 by means of a real world example. The paper ends with a conclusion in Sec. 5.

* Partially supported by the BMBF under Grant. No. 01IB802C4 and 01IB802A9 (LEONET).

2 Mixture of Densities

In order to estimate the unknown density of a data distribution we employ mixtures of densities which have been considered as very general and computationally efficient semi-parametric models. They consist of a convex combination of m parametric component densities $\phi_i(\boldsymbol{x}|\theta_i), i = 1, \ldots, m$:

$$\begin{aligned} \hat{p}(\boldsymbol{x}|\boldsymbol{\theta}) &= \sum_{i=1}^{m} \alpha_i \, \phi_i(\boldsymbol{x}|\theta_i), \\ \boldsymbol{x} \in \mathbb{R}^n \quad &, \quad \boldsymbol{\theta} = (\alpha_i, \ldots, \alpha_m, \theta_i, \ldots, \theta_m) \ , \\ \sum_{i=1}^{m} \alpha_i = 1 \ , \quad &\alpha_i \geq 0 \ \forall \ i = \{1, \ldots, m\} \ . \end{aligned}$$

The vector of parameters $\boldsymbol{\theta}$ characterizing $\hat{p}$ includes the weighting coefficients α_i and the parameters θ_i of the component densities which are in this article chosen as normal densities:

$$\begin{aligned} \phi_i(\boldsymbol{x}|\theta_i) &= \frac{1}{\sqrt{2\pi}^n \prod_{k=1}^{n} \sigma_{ik}} \exp\left(-\frac{1}{2} \sum_{k=1}^{n} \left(\frac{x_k - \mu_{ik}}{\sigma_{ik}} \right)^2 \right) \\ \theta_i &= (\mu_{i1}, \ldots, \mu_{in}, \sigma_{i1}^2, \ldots, \sigma_{in}^2) \in \mathbb{R}^{2n} \ , \end{aligned}$$

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 [16]; they can cope with multi-modal distributions and their complexity can be easily adjusted by the number of components. In the remainder these models are denoted as normal mixtures.

2.1 Regularization

In order to estimate the unknown parameters of the model based on observed data the method of maximum likelihood is widely used in statistical inference. For continuous distributions the likelihood of a sample $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is defined by the

joint density of X for the chosen probability density model $\hat{p}(\boldsymbol{x}|\boldsymbol{\theta})$ (characterized by its parameters $\boldsymbol{\theta}$). Assuming the $\boldsymbol{x}_i$ to be iid. sampled with pdf $p(\boldsymbol{x}|\boldsymbol{\theta})$ the log-likelihood function reads

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

A maximum likelihood estimation corresponds intuitively to the most likely model which would give rise to the data X. With regard to normal mixtures the EM algorithm [6, 17] is commonly used since in this case both steps of the EM can be performed analytically. However, in structure optimization of normal mixtures the maximization problem using the likelihood criterion is ill-posed, since the likelihood would increase to infinity, if there would be no upper bound for the number of components and no lower bound for the variances of the normals [14]. A general approach to transform this ill-posed optimization problem into a well-posed problem is the introduction of regularization terms that reflect specific assumptions about the density model [13].

2.2 Roughness penalties

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

$$J(\boldsymbol{\theta}) \;=\; \int_{-\infty}^{+\infty} p''(x|\boldsymbol{\theta})^2 \, dx, \qquad \boldsymbol{x} \in \mathbb{R} \;, \quad (2)$$

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 [24].

Similar terms have been used in the context of penalized maximum likelihood estimation [8, 22]. The employed terms were derived only for univariate densities. The multivariate case has been treated in the context of artificial neural networks [2]. In this approach, however, the integral in Eq. (2) 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 [12] the integral in Eq. (2) has been extended to the multivariate case and solved analytically. Thus, the complete objective function reads

$$F_{smooth}(\boldsymbol{\theta}) \;=\; \ell(\boldsymbol{\theta}) - \gamma\, J(\boldsymbol{\theta}) \;. \quad (3)$$

The *smoothing* parameter γ controls the relative influence of both criteria. However, due to the introduction of $J(\boldsymbol{\theta})$ the M-step in the EM algorithm is no longer analytically tractable. Hence, we use a quasi-Newton optimization method in the M-step in order to be still able to apply the EM algorithm (Refer to [13] for further discussion).

2.3 Entropy based regularizations

Another sensible measure has been proposed by the authors [13] and is based on the following consideration: Let $p(\boldsymbol{x})$ denote the true density and $\hat{p}(\boldsymbol{x})$ the estimation of p, respectively. In the case of a continuous distribution with an infinite sample the log-likelihood reads:

$$\ell(\boldsymbol{\theta}) \;=\; \int_{-\infty}^{+\infty} p(\boldsymbol{x})\, \log \hat{p}(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} \;, \qquad \boldsymbol{x} \in \mathbb{R} \;,$$

If the estimation $\hat{p}$ coincides with p the log-likelihood is just the negative of the Shannon entropy H [5]:

$$\hat{p} = p \;\;\Rightarrow\;\; \ell(\boldsymbol{\theta}) = \int_{-\infty}^{+\infty} p(\boldsymbol{x})\, \log p(\boldsymbol{x})\, d\boldsymbol{x} = -H(X)) \,.$$

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:

$$F_{entr}(\boldsymbol{\theta}) \;=\; \frac{1}{N}\ell(\boldsymbol{\theta}) + H(X)) \;. \quad (4)$$

In this sense the criterion $F_{entr}(\boldsymbol{\theta})$ accounts for what is called the *bias variance dilemma* (which can only be escaped in the limit of an infinite sample) [7, 10].

This criterion includes a very intuitive regularization imposed by the Shannon entropy. Over-specialized 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 [9]. However, in these approaches some kind of regularization parameters must be set. This is not necessary if we use the new criterion $F_{entr}(\boldsymbol{\theta})$ in Eq. 4.

3 Evolution of mixture models

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 Sec. 2.1, the introduction of regularizations may lead to complicated and noisy optimization criteria. Evolutionary algorithms (EAs) have been considered to be a promising method for dealing with this class of optimization problems [25]. However, the representation of solutions and the corresponding evolutionary operators have to be chosen carefully and in general depend on the problem. In the following we motivate a direct representation of mixture models and outline the basic concepts of the used operators.

3.1 Representation

Due to the local behavior of each component of the mixture the output of the model given a certain input only depends on few components. The modification of a single component, therefore, is very likely to cause only a small change in the model. Hence, a direct encoding of all parameters θ_i of each component in a linear chromosome of variable length will lead to a low *epistasis*. Let m denote the number of components a chromosome is constituted by

$$\mathcal{C} \;=\; (\alpha_1, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2, \ldots, \alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m^2) \;\; .$$

3.2 Recombination

In the context of recombination the problem of *competing conventions* [15, 19] arises: Due to the invariance to permutations of the components distinct genomes map to functionally equivalent phenotypes (models). Besides the redundancy in the encoding introduced by the $m!$ equivalent permutations for each model (with m components) the recombination of two models that differ in their permutation is very unlikely to lead to meaningful results. Even the recombination of two equivalent models with different permutation in general lead to functionally different models. The permutation problem can be circumvented by using a crossover operator that samples the crossover points in the input space 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 [4]). Two points in the input space are randomly sampled and all components that have a center $\boldsymbol{\mu}_i$ lying inside the hypervolume spanned by the two points are exchanged, see Fig. 1 for illustration.

An alternative approach is based on *gene pooling* [3] which also recombines normal mixtures and RBFNs irrespective to the permutation of its components.

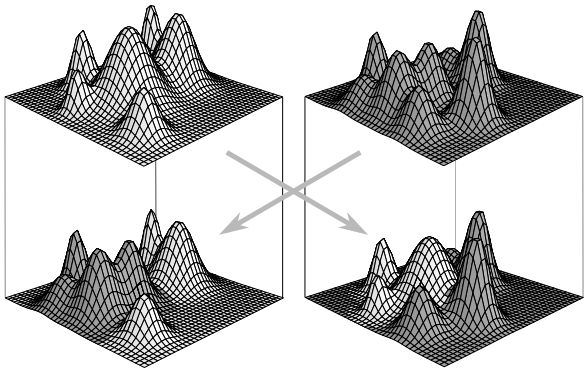

Figure 1: Crossover operator: The portions of mixture components of both parents lying in the crossover section are exchanged.

3.3 Mutation

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 [20]. 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 the number of components while at the same time keeping the effect of mutation as small as possible. The split operator replaces a component by two new components that are the best approximation for the old one under the constraint that their centers $\boldsymbol{\mu}_i$, $\boldsymbol{\mu}_j$ are significantly separated. The merge operator finds the best approximation of two neighboring components by a new component and replaces the old ones. Details of these operators are discussed in [12].

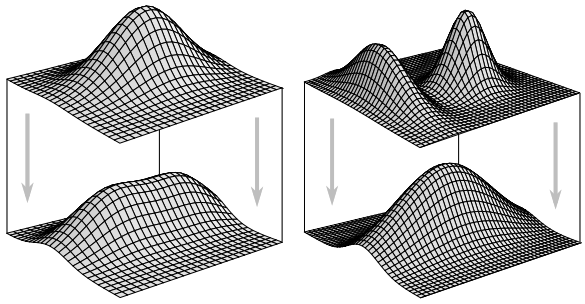

Figure 2: Mutation operator: A randomly selected component is split into two new components (left), two mixture components are merged together to form a new component (right). Both operators try to keep the change in the model and therefore the impact of mutation as small as possible.

3.4 Interaction with Learning

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.

The M-step of the EM is not analytically tractable for the proposed objective functions $F_{smooth}(\boldsymbol{\theta})$ and $F_{entr}(\boldsymbol{\theta})$. In order to avoid slow iterative methods 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, an evolutionary progress towards the correct optimum will be possible.

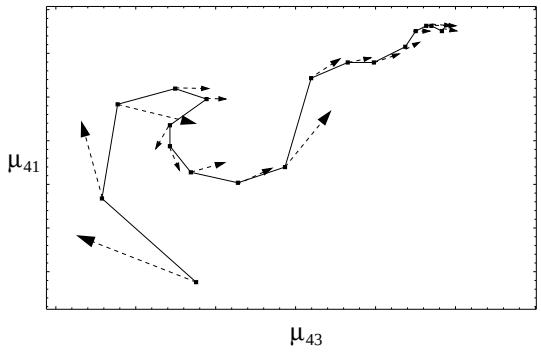

Figure 3: The solid lines show the trajectory of parameter change in the mixture model during maximization of the regularized log-likelihood by the EM algorithm. The dashed lines show the direction of the corresponding parameter changes with respect to the unregularized log-likelihood. For visualization a projection to two parameters is shown.

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.

3.5 Operator adaptation

The evolutionary optimization of normal mixtures using the operators outlined in the previous sections has been successfully applied to both density estimation [12] and nonlinear regression [14]. In these experiments the evolutionary setting (internal parameters of the operators and rates of their application) was held fixed. Yet, it has been long recognized that the choice of the evolutionary operators and rates at which they are applied in an EA have a significant impact on the performance of the EA. The optimal rate of application of each operator, however, depends not only on the given problem but also on the current state of the search [23]. This means that in order to maintain good or near optimal rates, these rates have to be tracked by some kind of adaptation mechanism during search. An intuitive approach is to keep track of the relative performance of the different operators by comparing their "success" in producing good offspring. The rates of application are then adjusted according to these performance values (refer to [11] for further details).

For the proposed EA, in addition to the mutation operators introduced in Sec. 3.3, different merge and split strategies, combinations of merge and split etc. are used. Initially, all operators are applied at the same rate which is periodically updated according to the estimated success of the respective operator. As a selection scheme the (μ, λ)-selection from evolution strategies is borrowed.

4 Experimental Results

The prediction of physiological data or specific patterns in physiological data constitute difficult problems even if some features are missing due to unrecorded information or the measurements are affected by a substantial amount of noise. We used the *data set B* from the Santa Fe time series prediction competition [18] for our experiments. The data was recorded from a patient in a sleep laboratory and consists of measurements of the heart rate, the chest volume (respiration force) and the blood oxygen concentration (measured by ear oximetry), which were taken every 0.5 sec. The data set contains 34.000 measurements of each signal in total. The patient shows periods of sleep apnea, during which respiration does not occur. This interval is followed by a period of several very small movements of the chest and finally by roughly four deep breath, see Fig. 4. This pattern of *no breathing – heavy breathing* repeats itself periodically, which is associated with a periodic change of the blood oxygen concentration (see [14, 18] for a detailed discussion)[1]. The physiological time series has several properties that make it difficult to forecast, especially in the case of models which can only cope with standard noise assumptions:

- The measurements are likely to be influenced by complex noise patterns, due to such properties as movements of the body, specific preprocessing of the data and specific features of the different measurement devices involved.
- The data is non-stationary, the three main dynamical patterns are identified corresponding to the apnea, the normal and an intermediate sleeping phase of the patient.
- There are artifacts in the data, some of which can easily be identified by physicians, however, others might well have not been recognized. We treat these artifacts as missing data.

In our experiments (as in [18]) we focus on the prediction of the heart rate five steps ahead. As

[1] Sleep apnea is medically important because it can lead to several serious heart conditions and occasionally to death. For the physician it would therefore be desirable to predict the onset and the duration of apnea. It would be beneficial for the patients to quickly wake up and fall asleep again.

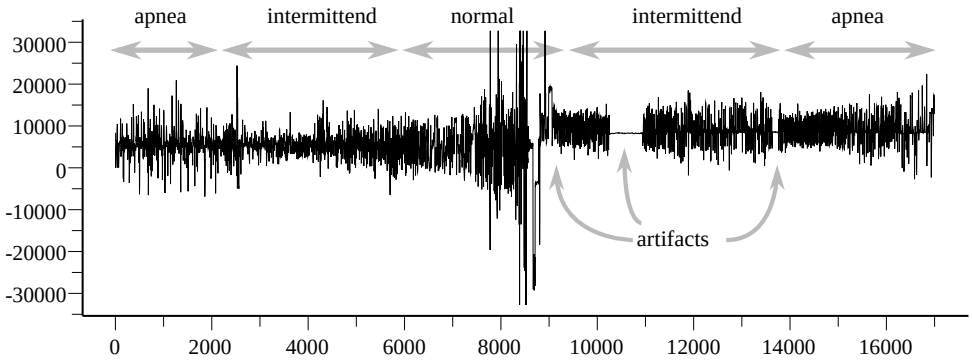

Figure 4: Time series of the respiration rate together with a rough indication of the different sleeping periods and of the erroneous measurement segments.

input variables we choose the heart rate, the chest volume, and the blood oxygen concentration at the current time step (setting 1). For comparison, we used a second experimental setting with the same input variables at the current time step and, additionally, five time steps ago (setting 2). In both cases the complete data set was divided into a training set (first half of the data set) and an independent test set (second half). We compared our method of structure optimization of mixture models with evolutionary algorithms with standard RBF networks, which are known to work particularly well for forecasting tasks. The results are presented in Table 1.

The prediction errors of the RBFN show that for no number of basis functions the error $\mathcal{E}_{\text{test}}$ is below one[2]. With setting 2 the errors are somewhat lower. In both cases, however, the errors increase with larger numbers of basis functions which corresponds to an over-fitting of the data. The results indicate that the RBFN is not able to capture any regularities of the data which is necessary in order to generalize well. The evolutionary optimized mixture model generalizes well for both settings. It is able to cope with the properties of the physiological time series, which makes it a particular complex prediction task, far better than the RBFN.

Table 1: Results for the five-step prediction of the heart rate. The RBFN (a) was trained with the BFGS algorithm until convergence ($\approx$ 3000 steps), the mixture model (b) was optimized with the evolutionary algorithm over 500 generations using a (4,20) selection strategy with one elitist and a regularization parameter of 10^{-20}. All results were averaged over 50 independent runs. The first column shows the number of kernels in the model (which varies in the case of evolutionary optimized models), the next columns show the root of the normalized mean squared errors for training and test and both settings.

#	setting 1		setting 2	
kernels	$\mathcal{E}_{\text{train}}$	$\mathcal{E}_{\text{test}}$	$\mathcal{E}_{\text{train}}$	$\mathcal{E}_{\text{test}}$
(a) RBFN				
10	0.8844	1.2451	0.8867	1.1588
20	0.8645	1.1312	0.8719	1.1651
50	0.8419	1.5403	0.8260	1.0445
100	0.8272	3.0368	0.8021	1.3727
150	0.8118	3.6850	0.7949	2.5839
200	0.7943	6.0067	0.7712	3.3723
(b) mixture of densities				
112 – 144	0.5815	0.4193	0.5603	0.4451

5 Conclusion

We considered the evolutionary structure optimization of nonlinear statistical models, mixtures of densities, combined with learning of the unknown parameters of these models.

Furthermore, we compared different regularization methods in the context of structure and parameter optimization. Most of them depend on additional parameters which have a strong impact on the performance of the method. In order to avoid the choice of a regularization parameter 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.

The interaction between evolution of the structure and learning of the parameters by means of local adaptation methods offers the possibility to combine computationally efficient adaptation methods with complex objective functions.

The application of the proposed method to the task of forecasting of a real-world system and the comparison to RBFNs gives experimental evidence for superiority of the chosen approach.

[2] $\mathcal{E}_{\text{test}} = 1$ is attained by a prediction that uses the mean of the time series. Rigney et al. [18] trained a RBFN with $\mathcal{E}_{\text{test}} = 0.98$.

References

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

[2] C. M. Bishop. Curvature-driven smoothing: A learning algorithm for feed-forward networks. *IEEE Trans. Neural Networks*, 4(5):882–884, 1993.

[3] B. Burdsall and C. Giraud-Carrier. GA-RBFN: A self-optimising RBF network. In G. D. Smith et al., editors, *Artificial Neural Nets and Genetic Algorithms*, pages 346–349. Springer, 1998.

[4] 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 et al., editors, *Parallel Problem Solving from Nature IV*, pages 710–719. Springer, 1996.

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

[6] 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.

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

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

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

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

[11] C. Igel and M. Kreutz. Using fitness distributions to improve the evolution of learning structures. In *Congress on Evolutionary Computation (CEC99)*, volume 3, pages 1902–1909. IEEE Press, 1999.

[12] 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 et al., editors, *Parallel Problem Solving from Nature V*, pages 998–1007. Springer, 1998.

[13] M. Kreutz, A. M. Reimetz, B. A. Sendhoff, C. Weihs, and W. von Seelen. Regularization and model selection in the context of density estimation. Technical Report 27/1999, Universität Dortmund, Germany, 1999.

[14] M. Kreutz, A. M. Reimetz, B. A. 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.

[15] R. Neruda. Canonical genetic learning of RBF networks is faster. In G. D. Smith et al., editors, *Artificial Neural Nets and Genetic Algorithms*, pages 350–353. Springer, 1998.

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

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

[18] D. R. Rigney et al. Multi-channel physiological data: Description and analysis (data set B). In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, Santa Fe Institute Studies in the Science of Complexity, vol. XV, pages 105–129. Addison Wesley, 1994.

[19] 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 *Combinations of Genetic Algorithms & Neural Networks (COGANN92)*, pages 1–37, 1992.

[20] 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.

[21] B. A. Sendhoff. *Evolution of Structures: Optimization of Artificial Neural Structures for Information Processing*. Berichte aus der Physik. Shaker Verlag, 1998.

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

[23] J. E. Smith and T. C. Fogarty. Operator and parameter adaptation in genetic algorithms. *Soft Computing*, 1(2):81–87, 1997.

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

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