Communicated by Steven Nowlan

Modeling Conditional Probability Distributions for Periodic Variables

Christopher M. Bishop
Ian T. Nabney
Neural Computing Research Group, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, B4 7ET, U.K.

Most conventional techniques for estimating conditional probability densities are inappropriate for applications involving periodic variables. In this paper we introduce three related techniques for tackling such problems, and investigate their performance using synthetic data. We then apply these techniques to the problem of extracting the distribution of wind vector directions from radar scatterometer data gathered by a remote-sensing satellite.

1 Introduction

Many applications of neural networks can be formulated in terms of a multivariate nonlinear mapping from an input vector $\mathbf{x}$ to a target vector $\mathbf{t}$. A conventional neural network approach, based on least squares, for example, leads to a network mapping that approximates the regression (i.e., the conditional average) of $\mathbf{t}$ given $\mathbf{x}$. For mappings that are multivalued, however, this approach breaks down, since the average of two solutions is not necessarily a valid solution. This problem can be resolved by recognizing that the conditional mean is just one aspect of a more complete description of the relationship between input and target, obtained by estimating the conditional probability density of $\mathbf{t}$ conditioned on $\mathbf{x}$, written as $p(\mathbf{t} \mid \mathbf{x})$. The least-squares approach then corresponds to maximum likelihood for the special case in which $p(\mathbf{t} \mid \mathbf{x})$ is modeled by a gaussian distribution which is spherically symmetric in $\mathbf{t}$-space and which has an $\mathbf{x}$-dependent mean.

Although techniques exist for modeling general conditional densities when the target vectors lie in Euclidean space, they are not appropriate when the targets are periodic. Direction and (calendar) time are two quantities that are periodic and that occur frequently in applications.

In this paper, we introduce three general techniques for modeling the conditional distribution of a periodic variable. We then investigate and compare their performance using synthetic data, as well as data collected from the ERS-1 remote sensing satellite.

Neural Computation **8, 1123–1133** (1996)

2 Density Estimation for Periodic Variables

A commonly used technique for *unconditional* density estimation is based on mixture models of the form

$$p(\mathbf{t}) = \sum_{i=1}^{m} \alpha_i \phi_i(\mathbf{t}) \tag{2.1}$$

where α_i are called mixing coefficients, and the component functions, or kernels, $\phi_i(\mathbf{t})$ are typically chosen to be Gaussians (McLachlan and Basford 1988; Titterington *et al.* 1985). To turn this into a model for conditional density estimation, we simply make the mixing coefficients, as well as any adaptive parameters in the component densities, into functions of the input vector $\mathbf{x}$. To achieve this we set the mixing coefficients and parameters from the outputs of a neural network that takes $\mathbf{x}$ as input. This approach underlies the "mixture of experts" model (Jacobs *et al.* 1991) and has also been considered by a number of other authors (Bishop 1994; Liu 1994).

In this section we extend this technique to provide three distinct methods for modeling the conditional density $p(\theta \mid \mathbf{x})$ of a periodic variable θ conditioned on an input vector $\mathbf{x}$. We also compare these methods with earlier approaches for treating periodic variables.

2.1 Mixtures of Wrapped Normal Densities. The first technique that we consider involves a transformation from a Euclidean variable $\chi \in (-\infty, \infty)$ to the periodic variable $\theta \in [0, 2\pi)$ of the form $\theta = \chi \bmod 2\pi$. This can be visualized as wrapping the infinite real line around a circle of unit radius. It induces a transformation that maps density functions p with domain $\mathbb{R}$ into density functions $\tilde{p}$ with domain $[0, 2\pi)$ as follows:

$$\tilde{p}(\theta \mid x) = \sum_{N=-\infty}^{\infty} p(\theta + N2\pi \mid \mathbf{x}) \tag{2.2}$$

It is clear by construction that the function $\tilde{p}$ satisfies the periodicity requirement $\tilde{p}(\theta + 2\pi \mid \mathbf{x}) = \tilde{p}(\theta \mid \mathbf{x})$. It is also normalized on the interval $[0, 2\pi)$, provided $p(\chi \mid \mathbf{x})$ is normalized on $\mathbb{R}$, since

$$\begin{aligned} \int_0^{2\pi} \tilde{p}(\theta \mid \mathbf{x})\, d\theta &= \sum_{N=-\infty}^{\infty} \int_0^{2\pi} p(\theta + N2\pi \mid \mathbf{x})\, d\theta \\ &= \sum_{N=-\infty}^{\infty} \int_{2N\pi}^{2(N+1)\pi} p(\theta \mid \mathbf{x})\, d\theta \\ &= \int_{-\infty}^{\infty} p(\chi \mid \mathbf{x})\, d\chi = 1 \end{aligned} \tag{2.3}$$

Various choices for the component density functions that make up the mixture $\tilde{p}(\chi \mid \mathbf{x})$ are possible, but here we shall restrict attention to

functions that are Gaussian of the form

$$\phi_i(\chi \mid \mathbf{x}) = \frac{1}{(2\pi)^{1/2}\sigma_i(\mathbf{x})} \exp\left\{-\frac{(\chi - \mu_i(\mathbf{x}))^2}{2\sigma_i(\mathbf{x})^2}\right\} \tag{2.4}$$

where $t \in \mathbb{R}$. The transformed density function $\tilde{\phi}_i$ is known as the "wrapped normal" distribution (Kotz and Johnson 1992).

The density function $p(\chi \mid \mathbf{x})$ is modeled using a combination of a neural network and a mixture model as described above. In this paper we use a standard multilayer perceptron with a single hidden layer of sigmoidal units and an output layer of linear units. To ensure that the mixture model in equation 2.1 is a density function, it is necessary that the mixing coefficients $\alpha_i(\mathbf{x})$ satisfy the constraints

$$\sum_{i=1}^{m} \alpha_i(\mathbf{x}) = 1, \qquad 0 \leq \alpha_i(\mathbf{x}) \leq 1 \tag{2.5}$$

for all $\mathbf{x}$. This can be achieved by choosing the $\alpha_i(\mathbf{x})$ to be related to the network outputs by a normalized exponential, or *softmax* function (Jacobs *et al.* 1991)

$$\alpha_i = \frac{\exp(z_i^\alpha)}{\sum_{j=1}^{M} \exp\left(z_j^\alpha\right)} \tag{2.6}$$

where z_j^α represents the corresponding network outputs. The centers μ_i of the kernel functions are represented directly by the network outputs; this is motivated by the corresponding choice of an uninformative Bayesian prior, assuming that the relevant network outputs have uniform probability distributions (Berger 1985; Jacobs *et al.* 1991). The standard deviations $\sigma_i(\mathbf{x})$ of the kernel functions represent *scale* parameters and so it is convenient to represent them in terms of the exponentials of the corresponding network outputs. This ensures that $\sigma_i(\mathbf{x}) > 0$ and discourages $\sigma_i(\mathbf{x})$ from tending to 0. Again, it corresponds to an uninformative prior in the Bayesian framework.

The adaptive parameters of the model (the weights and biases in the network) are optimized by maximum likelihood. In practice it is convenient to minimize an error function E given by the negative logarithm of the likelihood function. Derivatives of E with respect to the network weights can be computed using the rules of calculus (Bishop 1994), and these derivatives can then be used with standard optimization procedures to find a minimum of the error function. The results presented in this paper were obtained using a conjugate gradient algorithm.

One limitation of the maximum likelihood approach is that it leads to *biased* solutions since it underestimates the variance of a distribution in regions of low data density (Bishop 1995). An extreme example occurs if a component density function collapses onto one of the data points, giving zero variance and an infinite likelihood. For the applications reported in this paper, this effect will be small since the number of data points

is large and we are dealing with a one-dimensional target space. The use of an exponential relationship between the variance and the network output, discussed above, also helps to avoid pathological solutions.

In a practical implementation, it is necessary to restrict the value of N in the summation. We have taken the summation over 7 complete periods of 2π spanning the range $(-7\pi, 7\pi)$. Since the component Gaussians have exponentially decaying tails and the standard deviations are typically $\lesssim 2\pi$, this truncation introduces negligible error provided that care is taken in initializing the network weights so that the kernels have their means in the central interval $(-\pi, \pi)$.

2.2 Mixtures of Circular Normal Densities. The second approach to periodic conditional density estimation is also based on a mixture of kernel functions, but in this case the kernels themselves are periodic, thereby ensuring that the overall conditional density function will be periodic. The particular form of kernel function that we use can be motivated by considering a vector $\mathbf{v}$ in two-dimensional Euclidean space for which the probability distribution $p(\mathbf{v})$ is a symmetric Gaussian. By using the transformation $v_x = \|\mathbf{v}\| \cos\theta$, $v_y = \|\mathbf{v}\| \sin\theta$, we can show that the conditional distribution of the direction θ, given the vector magnitude $\|\mathbf{v}\|$, is given by

$$p(\theta) = \frac{1}{2\pi I_0(m)} \exp\{m \cos(\theta - c)\} \tag{2.7}$$

which is known as a *circular normal* or *von Mises* distribution (Mardia 1972). The normalization coefficient is expressed in terms of the zeroth-order modified Bessel function of the first kind, $I_0(m)$, and the parameter m (which depends on $\|\mathbf{v}\|$ in this derivation) is analogous to the inverse variance parameter in a conventional normal distribution. The parameter c corresponds to the mean of the density function.

Again the parameters $\alpha_i(\mathbf{x})$, $c_i(\mathbf{x})$, and $m_i(\mathbf{x})$ in the corresponding mixture model are determined by the outputs of a neural network taking $\mathbf{x}$ as input, and the network weights are determined by minimizing the negative log likelihood defined with respect to the training data. Because $I_0(m)$ is asymptotically an exponential function of m, some care must be taken in the implementation of the error function and its derivatives to avoid overflow in the results of intermediate calculations.

2.3 Expansion in Fixed Kernels. The third and final technique introduced in this paper involves a conditional density model consisting of a fixed set of periodic kernels, again given by circular normal functions as in equation 2.7. In this case the mixing proportions alone are determined by the outputs of a neural network (through a softmax activation function equation 2.6) and the centers c_i and width parameters m_i are fixed. We have selected a uniform distribution of centers, and set $m_i = m$ for

each kernel, where the value for m was chosen to give moderate overlap between the component functions.

Clearly a major drawback of fixed-kernel methods is that the number of kernels must grow exponentially with the number of output-space variables. This is an example of the "curse of dimensionality" (Bellman 1961; Bishop 1995). For the single output variable considered here, however, the number of kernel functions that is required is small, and the technique can be regarded as practical.

2.4 Related Work. The problem of modeling periodic variables has been well studied. In Mardia (1972) there is an introduction to conventional statistical approaches including the modeling of simple distributions. An approach to the problem of modeling more complex distributions involving multiple variables is that of Directional Unit Boltzmann Machines (DUBM) contained in Zemel *et al.* (1995). In this paper the theory of a Boltzmann machine whose units have associated densities that are circular normal distributions is developed. However, their approach is not suitable for the applications considered here for two reasons. First, the applications we consider have real-valued inputs, but in the DUBM all units must be directional. Second, the applications have a multimodal conditional distribution of the target variable. However, the deterministic version of the DUBM models the output density with a single circular normal, which is adequate only for unimodal distributions. The stochastic version does not suffer from this restriction, but requires extremely long training times.

3 Application to Synthetic Data

To test and compare the methods introduced in Section 2, we first consider a simple problem involving synthetic data, for which the true underlying distribution function is known. This data set is intended to mimic the salient features of the real data to be discussed in the next section. It is generated from a mixture of two triangular distributions where the centers and widths are taken to be linear functions of a single input variable x, and the mixing coefficients are fixed at 0.6 and 0.4. Any values of θ that fall outside $(-\pi, \pi)$ are mapped back into this range by shifting in multiples of 2π to give a distribution which is periodic. An example data set generated in this way is shown in Figure 1.

Three independent data sets (for training, validation, and testing) were generated from this distribution, each containing 1000 data points. For each technique, training runs were carried out in which the number of hidden units, as well as the number of kernels in the mixture model, were varied systematically to determine good values by minimizing the error obtained on the validation set. Table 1 gives a summary of the best

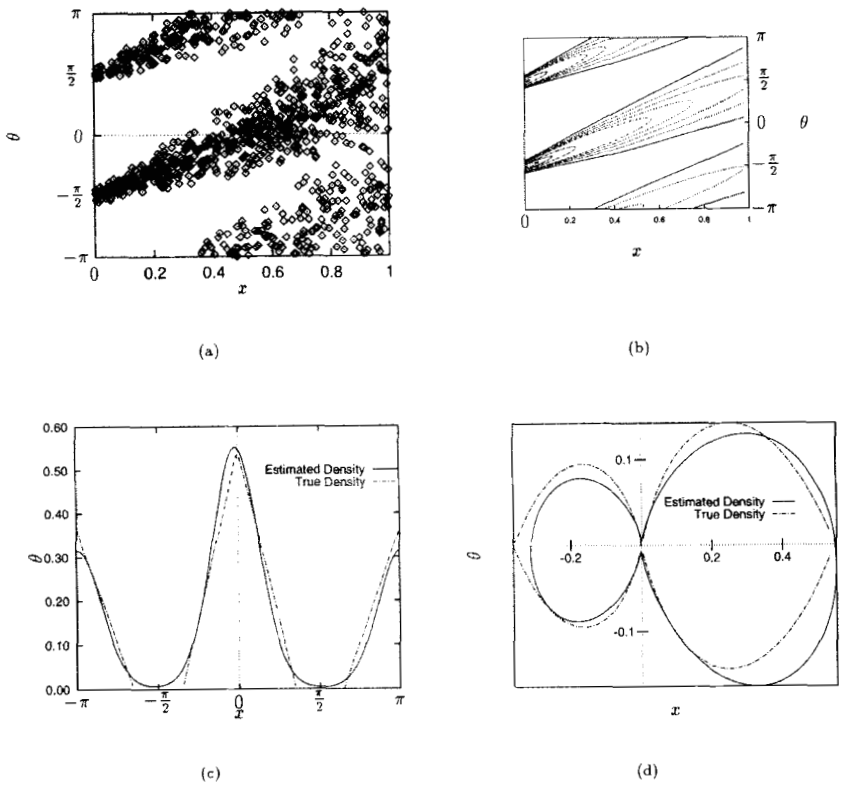

Figure 1: (a) Scatter plot of the synthetic training data. (b) Contours of the conditional density $p(\theta \mid x)$ obtained from a mixture of adaptive circular normal functions as described in Section 2.2. (c) The distribution $p(\theta \mid x)$ for $x = 0.5$ (solid curve) from the adaptive circular normal model, compared with the true distribution (dashed curve) from which the data were generated. (d) The same data as in (c) shown as a polar plot.

results from each of the three methods. We see that, for this data set, the best results, as determined from the test set, were obtained using the mixture of adaptive circular normal functions. Plots of the corresponding distributions are shown in Figure 1.

Table 1: Results Obtained Using Synthetic Data[a]

Method	Centers	Hidden units	Validation error	Test error
1	6	7	1177.1	1184.4
2	6	8	1109.5	1133.9
3	36	7	1184.6	1223.5

[a]Method 1: Mixture of wrapped normal functions.
Method 2: Mixture of adaptive circular normal functions.
Method 3: Mixture of fixed kernel functions.

4 Application to Radar Scatterometer Data

One of the original motivations for developing the techniques described in this paper was to provide an effective, principled approach to the analysis of radar scatterometer data from satellites such as the European Remote Sensing satellite ERS-1 (Thiria *et al.* 1993; Bishop and Legleye 1995). The ERS-1 satellite is equipped with three C-band radar antennae that measure the total backscattered power (written as σ_0) along three directions relative to the satellite track, as shown in Figure 2. When the satellite passes over the ocean, the strengths of the backscattered signals are related to the surface ripples of the water (on a scale of a few centimeters), which in turn are determined by the low-level winds.

Extraction of the wind vector from the radar signals represents an inverse problem that is typically multivalued. Although determining the wind speed is relatively straightforward, the data give rise to *aliases* for the wind direction. For example, a wind direction of θ will sometimes give rise to similar radar signals to a wind direction of $\theta + \pi$, and there may be further aliases at other angles. A conventional neural network approach to this problem, based on a least-squares estimate of θ, would predict directions that were given by conditional averages. Since the average of several valid wind directions is not itself a valid direction, such an approach would clearly fail. In this paper we show how such problems can be avoided by extracting a complete distribution function of wind directions, conditioned on the satellite measurements.

For this application, the modeling of the conditional distribution of wind direction provides the most complete information for the next stage of processing, which is to "dealias" the wind directions by combining information from groups of wind-field cells, together with prior knowledge, to determine the most probable overall wind field.

A large data set of ERS-1 measurements, spanning a wide range of meteorological conditions, has been assembled by the European Space Agency in collaboration with the UK Meteorological Office. Labeling of the data set was performed using wind vectors from the Meteorological

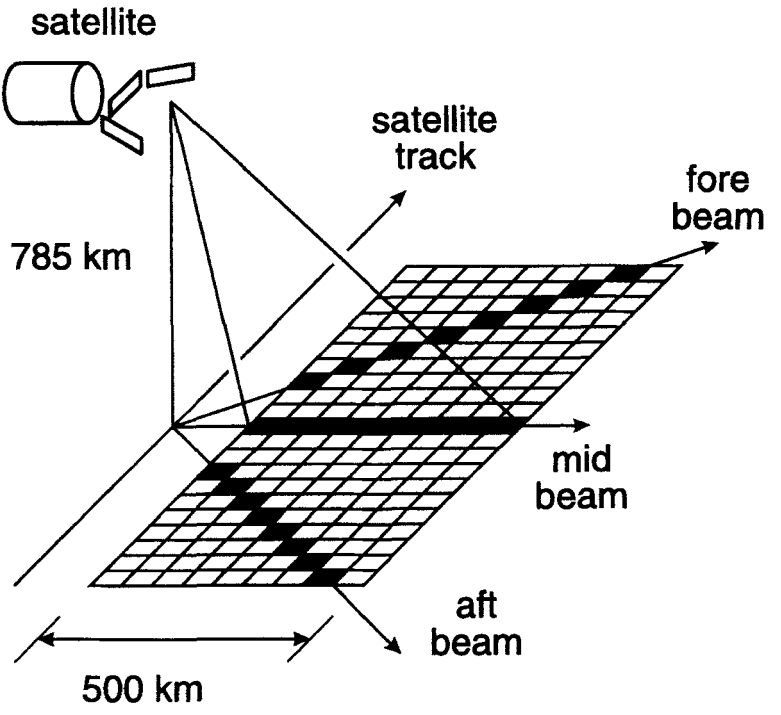

Figure 2: Schematic illustration of the ERS-1 satellite showing the footprints of the three radar scatterometers.

Office Numerical Weather Prediction model. These values were interpolated from the inherently coarse-grained model to regions coincident with the scatterometer cells.

The data that were selected for the experiments reported in this paper were collected from low-pressure (cyclonic) and high-pressure (anticyclonic) circulations. These conditions, rather than cases that were homogeneous or with a simple gradient in speed or direction, were chosen to provide a more challenging task to test the modeling techniques. Ten wind fields from each of the two categories were used: each wind field contains $19 \times 19 = 361$ cells, each of which represents an area of approximately 26×26 km. After removal of completed data, this resulted in training, validation, and test sets each containing 1963 patterns.

The inputs used for modeling the data were the three values of σ_0 for the aft-beam, mid-beam, and fore-beam, together with the sine of the incidence angle of the mid-beam, since this angle strongly influences the

Table 2: Results on Satellite Data[a]

Method	Centers	Hidden units	Validation error	Test error
1	4	20	2474.6	2446.2
2	6	20	2308.0	2337.9
3	36	24	2028.9	1908.9

[a]Method 1: Mixture of wrapped normal functions.
Method 2: Mixture of adaptive circular normal functions.
Method 3: Mixture of fixed kernel functions.

reflected signal received by the scatterometer. Each σ_0 input was scaled to have zero mean and unit variance, while the fourth input value was passed to the network unchanged. The target value was expressed as an angle clockwise from the satellite's forward path and converted to radians. Again, a conjugate gradient algorithm was used to optimize the network weights.

Table 2 gives a summary of the preliminary results obtained with each of the three methods. As expected, the fact that this is a more complex domain than the synthetic problem meant that there were more difficulties with local optima. In fact, over 75% of the training runs ended with the network trapped in a poor minimum of the error function. This problem was considerably reduced (to about 25% of the runs by initializing the network weights so that the initial centers of the kernel functions were approximately uniformly spaced in $[0, 2\pi)$. Of the adaptive-center models, the one with six centers has the lowest error on the validation data; however, fewer centers are actually required to model the conditional density function reasonably well. This can also be seen in Figure 3, which shows the conditional distribution of wind directions at a typical data point from the test set, and which clearly has fewer than eight peaks.

5 Discussion

In this paper we have introduced three distinct but related techniques for modeling the conditional probability distribution of a periodic variable, and we have illustrated the use of these techniques in a simple synthetic problem, and on radar scatterometer data from a remote sensing satellite. All three methods give reasonable results, with the adaptive-kernel approaches somewhat outperforming the fixed-kernel technique on synthetic data, and vice versa on the scatterometer data. A conventional network approach, involving the minimization of a sum-of-squares error function or the use of a DUBM, would perform very poorly on these problems since the required mapping is multivalued.

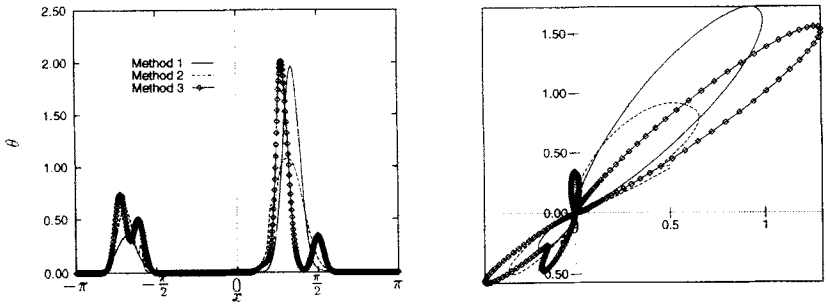

Figure 3: Linear and polar plots of the conditional distribution $p(\theta \mid \mathbf{x})$ for a sample input vector from the test set. The dominant alias at π is evident. In both plots, the solid curve represents method 1, the dashed curve represents method 2, and the curve with diamonds represents method 3.

The two fully adaptive methods (methods 1 and 2) give largely similar results. This is not unexpected, since the kernels used are similar functions. Generalizing the approach, there is a range of possible models with different parameters fixed: the third method is an extreme case where the only adaptive parameters were the mixing coefficients.

One aspect of these algorithms that is more complex than conventional network optimization is the problem of model order selection. The incorporation of a mixture model means that there are two structural parameters to select: the number of hidden units in the network and the number of components in the mixture model. Changes to either of these parameters will change the number of adaptive weights in the network, and so the two parameters are closely coupled. In this paper we have varied both of these structural parameters in a systematic way and sought the optimum network by measuring performance on an independent validation set. It is likely that the use of a larger, fixed network structure, together with regularization to control the effective model complexity, will significantly simplify the process of model order selection.

Acknowledgments

We are grateful to the European Space Agency and the UK Meteorological Office for making available the ERS-1 data. The contributions of Claire Legleye in the early stages of this project are also gratefully acknowledged. We would also like to thank Iain Strachan and Ian Kirk

of AEA Technology for a number of useful discussions relating to the interpretation of this data.

References

Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, Princeton, NJ.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis,* 2nd ed. Springer, New York.

Bishop, C. M. 1994. *Mixture density networks.* Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University, U.K.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

Bishop, C. M., and Legleye, C. 1995. Estimating conditional probability distributions for periodic variables. In *Advances in Neural Information Processing Systems,* D. S. Touretzky, G. Tesauro, and T. K. Leen, eds., Vol. 7, pp. 641–648. MIT Press, Cambridge, MA.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Comp.* **3**, 79–87.

Kotz, S., and Johnson, N. L. eds. 1992. *Encyclopedia of Statistical Sciences,* pp. 381–386. John Wiley, New York.

Liu, Y. 1994. Robust neural network parameter estimation and model selection for regression. In *Advances in Neural Information Processing Systems,* Vol. 6, pp. 192–199. Morgan Kaufmann, San Mateo, CA.

Mardia, K. V. 1972. *Statistics of Directional Data.* Academic Press, London.

McLachlan, G. J., and Basford, K. E. 1988. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Thiria, S., Mejia C., Badran, F., and Crepon, M. 1993. A neural network approach for modeling nonlinear transfer functions: Application for wind retrieval from spaceborne scatterometer data. *J. Geophys. Res.* **98**(C12), 22827–22841.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. 1985. *Statistical Analysis of Finite Mixture Distributions.* John Wiley, New York.

Zemel, R. S., Williams, C. K. I., and Mozer, M. C. 1995. Lending direction to neural networks. *Neural Networks* **8**, 503–512.

Received June 9, 1995; accepted December 19, 1995.