Author(s): Dybowski, Richard.
Article Title: Classification of incomplete feature vectors by radial basis function networks
Year of publication: 1998
Citation: Dybowski R. (1998) "Classification of incomplete feature vectors by radial basis function networks". Pattern Recognition Letters, 19 (14) 1257-1264
Link to published version: http://dx.doi.org/10.1016/S0167-8655(98)00096-8
DOI: 10.1016/S0167-8655(98)00096-8
ISSN: 0167-8655

Submitted for publication in *Pattern Recognition Letters* [Submission date 27 January 1998]

Classification of Incomplete Feature Vectors by Radial Basis Function Networks

Richard Dybowski

Intensive Care Unit (Division of Medicine), UMDS (St Thomas' Hospital), London, UK

Abstract
The paper describes the use of radial basis function neural networks with Gaussian basis functions to classify incomplete feature vectors. The method exploits the fact that any marginal distribution of a defined Gaussian joint distribution can be determined from the mean vector and covariance matrix of the joint distribution. The method is discussed in the context of complete and incomplete training sets.

Key words: Incomplete data; Gaussian mixture models; Radial basis functions; Imputation; EM algorithm

1. Introduction

Highly flexible discriminant functions such as projection pursuit regression models (Friedman, 1981) and multilayer perceptrons (Bishop, 1995) can be used to estimate the probability $P(C_k|\mathbf{x})$ of a class of interest C_k conditional on a given vector $\mathbf{x}$ of measurements (a *feature vector*) taken from an object, thereby enabling the object to be classified.

In deriving a function which discriminates between classes of interest by estimating $P(C_k|\mathbf{x})$, it is assumed that each constituent variable of $\mathbf{x}$ has a value, but what if at least one of these variables has not been assigned a value?

1.1 Imputation of an incomplete feature vector

A natural response to the problem of classifying an incomplete feature vector $\mathbf{x}$ is to make up the deficit in $\mathbf{x}$ by supplying values for the missing data (*imputation*). If $\mathbf{x}_{obs}$ is the part of $\mathbf{x}$ that is observed and $\mathbf{x}_{mis}$ is the subset of $\mathbf{x}$ that is missing (*i.e.* $\mathbf{x} = \mathbf{x}_{obs} \cup \mathbf{x}_{mis}$), an intuitively reasonable approach is to impute $\mathbf{x}$ with the expected values for $\mathbf{x}_{mis}$ given the observed values $\mathbf{x}_{obs}$. In other words, in the absence of $P(C_k|\mathbf{x}_{obs}, \mathbf{x}_{mis})$, use $P(C_k|\mathbf{x}_{obs}, E[\mathbf{x}_{mis}|\mathbf{x}_{obs}])$ in place of $P(C_k|\mathbf{x}_{obs})$. But a problem with this approach is that $P(C_k|\mathbf{x}_{obs}, E[\mathbf{x}_{mis}|\mathbf{x}_{obs}])$ does not necessarily equal $P(C_k|\mathbf{x}_{obs})$ since, from Bayes' theorem,

$$P\left(C_k \middle| \mathbf{x}_{obs}, \mathbf{x}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}]\right) = P(C_k|\mathbf{x}_{obs}) \frac{f\left(\mathbf{x}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}] \middle| \mathbf{x}_{obs}, C_k\right)}{f\left(\mathbf{x}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}] \middle| \mathbf{x}_{obs}\right)},$$

and $f\left(\mathbf{x}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}] \middle| \mathbf{x}_{obs}, C_k\right)$ is not necessarily equal to $f\left(\mathbf{x}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}] \middle| \mathbf{x}_{obs}\right)$, where f denotes a probability density function (*pdf*). However, an approximation between $P(C_k|\mathbf{x}_{obs})$ and $P(C_k|\, \mathbf{x}_{obs}, E[\mathbf{x}_{mis}|\mathbf{x}_{obs}])$ can be made as follows.

Theorem 1. *$P(A|B) = E[P(A|B,\mathbf{x})|B]$, where A and B are any two statements, and the expectation is taken over a vector of values* $\mathbf{x}$.

Proof. If $P(A|B,\mathbf{x})$ is regarded as a function of $\mathbf{x}$ then

$$E[P(A|B,\mathbf{x})|B] = \int_{\Re^{\|\mathbf{x}\|}} P(A|B,\mathbf{x}) f(\mathbf{x}|B) d\mathbf{x}, \qquad (1)$$

where $f(\mathbf{x}|B)$ is the pdf for $\mathbf{x}$ conditioned on B. Upon unfactorizing the integrand of Eq. (1) we have

$$E[P(A|B,\mathbf{x})|B] = \int_{\Re^{\|\mathbf{x}\|}} f(A,\mathbf{x}|B)d\mathbf{x}$$

$$= P(A|B). \quad \square$$

Let $\hat{\mathbf{x}}_{mis} = E[\mathbf{x}_{mis}|\mathbf{x}_{obs}]$. By regarding $P(C_k|\mathbf{x}_{obs}, \mathbf{x}_{mis})$ as a function of $\mathbf{x}_{mis}$, the Taylor series gives

$$P(C_k|\mathbf{x}_{obs},\mathbf{x}_{mis}) \approx P(C_k|\mathbf{x}_{obs},\hat{\mathbf{x}}_{mis}) + (\mathbf{x}_{mis} - \hat{\mathbf{x}}_{mis})\left[\nabla_{\mathbf{x}_{mis}} P(C_k|\mathbf{x}_{obs},\mathbf{x}_{mis})\right]^{\mathrm{T}}_{\hat{\mathbf{x}}_{mis}} \tag{2}$$

if $E[\mathbf{x}_{mis}|\mathbf{x}_{obs}]$ is sufficiently close to $\mathbf{x}_{mis}$. Taking expectations on both sides of Eq. (2) over $\mathbf{x}_{mis}$ for a given $\mathbf{x}_{obs}$ results in

$$E[P(C_k|\mathbf{x}_{obs},\mathbf{x}_{mis})|\mathbf{x}_{obs}] \approx P(C_k|\mathbf{x}_{obs},\hat{\mathbf{x}}_{mis})$$

since $E[\mathbf{x}_{mis} - \hat{\mathbf{x}}_{mis}]=0$, therefore, from Theorem 1,

$$P(C_k|\mathbf{x}_{obs}) \approx P(C_k|\mathbf{x}_{obs},\hat{\mathbf{x}}_{mis}) . \tag{3}$$

Note that Theorem 1 implies

$$P(C_k|\mathbf{x}_{obs}) = E[P(C_k|\mathbf{x}_{obs},\mathbf{x}_{mis})|\mathbf{x}_{obs}] = \int_{\Re^{\|\mathbf{x}_{mis}\|}} P(C_k|\mathbf{x}_{obs},\mathbf{x}_{mis}) f(\mathbf{x}_{mis}|\mathbf{x}_{obs})d\mathbf{x}_{mis} ,$$

but solving the integration may be far from straightforward for a chosen discriminant function.

In this paper we will tackle the problem of an incomplete feature vector by reformulating the desired conditional probability using Bayes' theorem and expressing class-conditional probability densities as Gaussian mixture models.

2. Gaussian mixture models

In kernel density estimation (Silverman, 1986), the pdf $f(\mathbf{x})$ from which a set of N data points $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}\}$ was randomly sampled is modelled by a linear combination of N kernel functions $K((\mathbf{x}-\mathbf{x}^{(n)})/h)$, a kernel function being centred on each data point:

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^{\|\mathbf{x}\|}} \sum_{n=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}^{(n)}}{h}\right),$$

where h is a smoothing parameter. Usually, K is a radially symmetrical unimodal pdf such as a multivariate Gaussian distribution.

Kernel density estimators satisfy the asymptotic requirement that $\hat{f}(\mathbf{x}) \to f(\mathbf{x})$ as $N \to \infty$ (*e.g.* Parzen (1962)) and motivate the concept of a Gaussian *mixture model* (Everitt and Hand, 1981)

$$\hat{f}(\mathbf{x}|C_k) = \sum_{j=1}^{M_k} p_{kj} G(\mathbf{x}|\boldsymbol{\mu}^{[kj]}, \boldsymbol{\Sigma}^{[kj]}) \tag{4}$$

in which M_k (< N) multivariate Gaussian pdfs G (Gaussian *basis functions*) are employed, each basis function being defined by a mean vector $\boldsymbol{\mu}^{[kj]}$ and covariance matrix $\boldsymbol{\Sigma}^{[kj]}$. *Mixing coefficient* p_{kj} is equal to $P(j|C_k)$, which is interpreted as the probability that, for class C_k, a data point will originate from the j-th basis function associated with that class.

Before continuing, we state the following theorem, which allows us to determine the marginal distribution of any subset of **x** when **x** has a defined Gaussian distribution.

Theorem 2. (*e.g.* Krzanowski (1988)) *Let* **x** *be a* $p\times1$ *vector with multivariate normal distribution* $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *If* **x** *is partitioned as* $\mathbf{x}=\left(\mathbf{x}_1,\mathbf{x}_2\right)^{\mathrm{T}}$, *where* $\mathbf{x}_i$ *is a* $p_i\times1$ *vector, and the corresponding partitions of* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *are*

$$\boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\quad \boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix},$$

where $\boldsymbol{\mu}_i$ *is a* $p_i\times1$ *vector and* $\boldsymbol{\Sigma}_{ii}$ *a* $p_i\times p_i$ *matrix, then* $\mathbf{x}_i$ *has the distribution* $\mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$.

The conditional probability $P(C_k|\mathbf{x}_{obs})$ presented in Section 1 can be reformulated using Bayes' theorem,

$$P(C_k|\mathbf{x}_{obs})=\frac{P(C_k)\,f(\mathbf{x}_{obs}|C_k)}{\sum_{k'}P(C_{k'})\,f(\mathbf{x}_{obs}|C_{k'})}, \tag{5}$$

and distribution $f(\mathbf{x}_{obs}|C_k)$ within Eq. (5) can be estimated by applying Theorem 2 to a collection of M class-conditional Gaussian mixture models:

$$\begin{aligned} f(\mathbf{x}_{obs}|C_k) &= \int_{\Re^{\|\mathbf{x}_{mis}\|}} f(\mathbf{x}_{obs},\mathbf{x}_{mis}|C_k)\,d\mathbf{x}_{mis}\\ &\approx \sum_{j=1}^{M} p_{kj} \int_{\Re^{\|\mathbf{x}_{mis}\|}} G(\mathbf{x}_{obs},\mathbf{x}_{mis}|\boldsymbol{\mu}^{[kj]},\boldsymbol{\Sigma}^{[kj]})\,d\mathbf{x}_{mis} \quad \text{(from Eq. (4))}\\ &= \sum_{j=1}^{M} p_{kj}\,G(\mathbf{x}_{obs}|\boldsymbol{\mu}_1^{[kj]},\boldsymbol{\Sigma}_{11}^{[kj]}) \quad \text{(from Theorem 2),} \end{aligned} \tag{6}$$

where $\boldsymbol{\mu}_1^{[kj]}$ and $\boldsymbol{\Sigma}_{11}^{[kj]}$ are obtained from partitions

$$\boldsymbol{\mu}^{[kj]}=\begin{pmatrix}\boldsymbol{\mu}_1^{[kj]}\\ \boldsymbol{\mu}_2^{[kj]}\end{pmatrix},\quad \boldsymbol{\Sigma}^{[kj]}=\begin{pmatrix}\boldsymbol{\Sigma}_{11}^{[kj]} & \boldsymbol{\Sigma}_{12}^{[kj]}\\ \boldsymbol{\Sigma}_{21}^{[kj]} & \boldsymbol{\Sigma}_{22}^{[kj]}\end{pmatrix}, \tag{7}$$

corresponding to partition $\mathbf{x}=\left(\mathbf{x}_1,\mathbf{x}_2\right)^{\mathrm{T}}$, where $\mathbf{x}_1=\mathbf{x}_{obs}$ and $\mathbf{x}_2=\mathbf{x}_{mis}$. Determination of the model parameters via the EM algorithm is described in Appendix A.

The ability of Eq. (5) in conjunction with Eq. (6) to provide an estimate for $P(C_k|\mathbf{x}_{obs})$ suggests that the classification of a feature vector, whether complete or incomplete, can be achieved from the parameters of a trained radial basis function neural network, to be described in the next section.

3. Radial basis function networks

Mixture model (4) can be rewritten as

$$\hat{f}(\mathbf{x}|C_k) = \sum_{j=1}^{M_k} P(j|C_k) f(\mathbf{x}|j, C_k). \tag{8}$$

Implementation of this model requires $P(j|C_k)$ and $f(\mathbf{x}|j, C_k)$ for each class C_k, but an alternative to Eq. (8) is

$$\hat{f}(\mathbf{x}|C_k) = \sum_{j=1}^{M} P(j|C_k) f(\mathbf{x}|j) \tag{9}$$

in which, through the use of a common pool of M basis functions, only $P(j|C_k)$ is required for each class (Bishop, 1995, p.180). Although one would expect Eq. (8) to provide a better representation of each class-conditional pdf of interest, a disadvantage is that the total number of adjustable parameters (summed over each class) may be greater than that for Eq. (9), whereas the modelling accuracy provided by Eq. (9) may suffice.

Substituting Eq. (9) into the Bayesian reformulation of $P(C_k|\mathbf{x})$ gives

$$\hat{P}(C_k|\mathbf{x}) = \sum_{j=1}^{M} w_{kj} f_j(\mathbf{x}), \tag{10}$$

where $w_{kj} = \hat{P}(C_k|j)$ and

$$f_j(\mathbf{x}) = \frac{f(\mathbf{x}|j)P(j)}{\sum_{j'=1}^{M} f(\mathbf{x}|j')P(j')}.$$

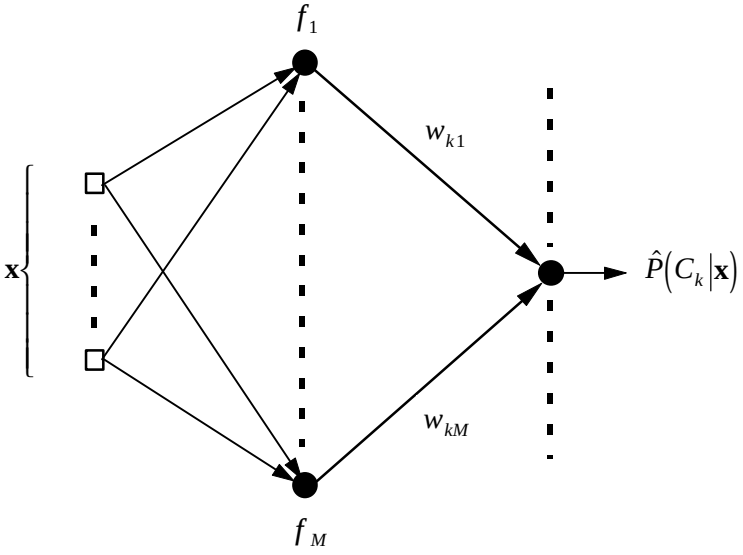

Fig. 1. The architecture of a radial basis function network

When Eq. (10) is rendered as the network shown in Fig. 1, and $f(\mathbf{x}|j)$ is modelled by a radially symmetrical function, the structure is referred to as a *radial basis function* (*RBF*) *network* in which $f(\mathbf{x}|j)$ are the RBFs, $f_j(\mathbf{x})$ are the normalised RBFs, and w_{kj} are the weights. Only one layer of RBFs in Fig. 1 is required to approximate any continuous function to arbitrary accuracy (Hartmen *et al*, 1990).

This architecture has been proposed by a number of workers, including Broomhead and Lowe (1988) and Poggio and Girosi (1990). Whereas multilayer perceptrons partition feature space with hyperplanes, RBF networks perform the partitioning with hyperellipsoids. An advantage of RBF networks over multilayer perceptrons is that, once the parameters of the normalised RBFs have been decided (see below), the weights can be easily determined by linear optimisation since the RBFs have mapped the inputs into a higher-dimensional space where they are linearly separable (Moody and Darken, 1989).

The parameters for $f_j(\mathbf{x})$ in Eq. (10) consist of $P(j)$ and the parameters for $f(\mathbf{x}|j)$ (j=1,…,M). These can be determined by modelling the density of $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}\}$ by the Gaussian mixture model

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{M} P(j) f(\mathbf{x}|j) \tag{11}$$

and applying the EM algorithm described in Appendix A to the entire dataset rather than to a class-wise subset of it. Once these parameters have been acquired, they are held constant whilst linear optimisation is used to provide the weights w_{kj} for Eq. (10) through supervised learning.

3.1 Classification of an incomplete feature vector

If we repeat the derivation of Eq. (6), replacing Eq. (4) with Eq. (9) and letting $f(\mathbf{x}|j) = G(\mathbf{x}|\boldsymbol{\mu}^{[j]}, \boldsymbol{\Sigma}^{[j]})$, then

$$f(\mathbf{x}_{obs}|C_k) \approx \sum_{j=1}^{M} p_{kj} G(\mathbf{x}_{obs}|\boldsymbol{\mu}_1^{[j]}, \boldsymbol{\Sigma}_{11}^{[j]}), \tag{12}$$

where $\boldsymbol{\mu}_1^{[j]}$ and $\boldsymbol{\Sigma}_{11}^{[j]}$ are analogous to subvector $\boldsymbol{\mu}_1^{[kj]}$ and submatrix $\boldsymbol{\Sigma}_{11}^{[kj]}$ in Eq. (7) and are thus contained within $\boldsymbol{\mu}^{[j]}$ and $\boldsymbol{\Sigma}^{[j]}$, respectively. Substituting Eq. (12) into Eq. (5) gives the required expression:

$$\hat{P}(C_k|\mathbf{x}_{obs}) = \frac{\sum_{j=1}^{M} w_{kj} G(\mathbf{x}_{obs}|\boldsymbol{\mu}_1^{[j]}, \boldsymbol{\Sigma}_{11}^{[j]}) P(j)}{\sum_{j=1}^{M} G(\mathbf{x}_{obs}|\boldsymbol{\mu}_1^{[j]}, \boldsymbol{\Sigma}_{11}^{[j]}) P(j)}$$

since $\pi_{kj} = P(j|C_k) = P(j)w_{kj}/P(C_k)$ and $\sum_{k'} w_{k'j} = 1$.

In sum, if

$$\hat{P}(C_k|\mathbf{x}) = \frac{\sum_{j=1}^{M} w_{kj} G(\mathbf{x}|\boldsymbol{\mu}^{[j]}, \boldsymbol{\Sigma}^{[j]}) P(j)}{\sum_{j=1}^{M} G(\mathbf{x}|\boldsymbol{\mu}^{[j]}, \boldsymbol{\Sigma}^{[j]}) P(j)}$$

is a trained RBF network with Gaussian radial basis functions $G(\mathbf{x}|\boldsymbol{\mu}^{[j]}, \boldsymbol{\Sigma}^{[j]})$, then $\hat{P}(C_k|\mathbf{x}_{obs})$ can be determined for any $\mathbf{x}_{obs} \subseteq \mathbf{x}$ using the existing set of RBF-network parameters $\{w_{kj}, \boldsymbol{\mu}^{[j]}, \boldsymbol{\Sigma}^{[j]}, P(j)\}$ and class priors $P(C_k)$.

4. Incomplete training sets

The presence of an incomplete feature vector $\mathbf{x}_{obs}$ could be symptomatic of an underlying missing-data mechanism which has also caused previous feature vectors to be incomplete. If these previous vectors are collected together as tuples $\langle \mathbf{x}_{obs}^{\mathrm{T}}, C_k \rangle$, where C_k is the class assigned to $\mathbf{x}_{obs}$, the resulting dataset will be an incomplete training set.

The reality of incomplete datasets has been a long-standing problem in statistics and a number of solutions have been proposed. A standard technique is to assign values to the missing data (imputation) and perform the required analysis or modelling on the resultant complete dataset. The choice of method for conducting imputation is dependent upon the pattern of missingness within the dataset, on whether the precise nature of the missing-data mechanism can be ignored, and on the type of data involved (*i.e.* whether it is continuous, categorical, or a mixture of the two). Little and Rubin (1987) provide a comprehensive discussion of this topic, a summary of which is available (Little and Rubin, 1990).

In this paper we are focusing on real-valued variables, and a sound approach for the imputation of such variables in a dataset is to use the EM algorithm of Appendix B. This imputation determines the underlying model most likely to have given rise to the observed data. When this model is known, the values $\mathbf{x}_{mis}$ missing from tuple $\langle \mathbf{x}^{\mathrm{T}}, C_k \rangle$ can be estimated by using the parameters of the model.

Once a dataset has been imputed by the EM algorithm, several routes can be taken to the classification of a new but incomplete feature vector:

(a) Proceed with the derivation of an RBF network as described in Section 3 for the case when a complete dataset is available and use the method of Section 3.1.

(b) When a dataset has been imputed by the method described in Appendix B, the resulting Gaussian mixture model (14) is equivalent to Eq. (4), therefore, the model can be used to classify incomplete feature vectors, as described in Section 2. There is no gain in supplementing this mixture model with an RBF network.

(c) If the condition for approximation (3) holds then Eq. (16) can be used to determine the expectation in $P(C_k|\mathbf{x}_{obs}, E[\mathbf{x}_{mis}|\mathbf{x}_{obs}])$.

The EM algorithm of Appendix B imputes each of the class-wise partitions of the dataset separately, but if the class labels of the dataset are momentarily ignored and the algorithm is applied to the incomplete dataset as a whole, then mixture model (11) is obtained in place of model (14). Once the parameters of the basis functions of Eq. (11) have been obtained on conclusion of the imputation, the class labels can be re-instated to allow the weights of Eq. (10) to be determined. Hence, an RBF network can be obtained directly from an incomplete dataset.

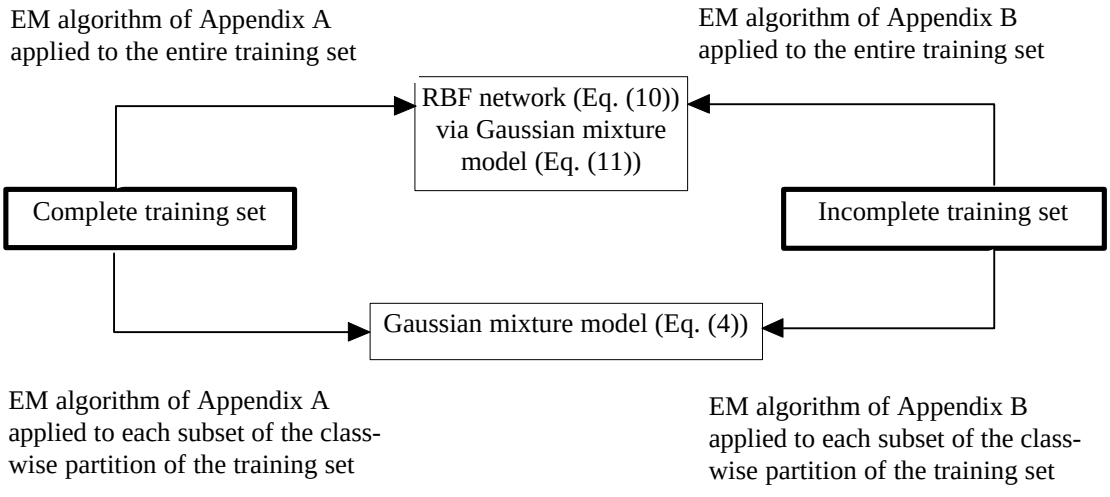

Fig. 2. EM algorithms for the derivation of Gaussian mixture models

5. Conclusion

The derivation of RBF networks from complete and incomplete datasets is summarised in Fig. 2. Whichever route is taken to produce an RBF network with Gaussian basis functions, an incomplete feature vector can be classified by the network as detailed in Section 3.1.

Acknowledgement
The author thanks the Special Trustees for Guy's and St Thomas' Hospitals for their financial support of this work.

Appendix A. The EM algorithm for the parameters of Gaussian mixture models

The standard approach for determining the parameters $\{p_{kj}, \boldsymbol{\mu}^{[kj]}, \boldsymbol{\Sigma}^{[kj]}\}$ of a Gaussian mixture model from a given dataset is to use maximum-likelihood estimation. Various procedures have been developed for this purpose (Redner and Walker, 1984), including the EM algorithm.

The *expectation-maximisation* (*EM*) *algorithm* (Dempster et al., 1977) is a general iterative technique for computing maximum-likelihood estimates when the observed data can be regarded as incomplete. Each iteration of the EM algorithm is composed of two steps: an expectation (*E*) step and a maximisation (*M*) step. Let **v** be observed data, **y** unobserved data, and let $\boldsymbol{\theta}^{cur}$ denote the current value of a parameter vector $\boldsymbol{\theta}$. After assigning an initial set of values to $\boldsymbol{\theta}^{cur}$, the EM algorithm proceeds via the following cycle:

E-step: Let $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{cur}) = E\left[\sum_{n=1}^{N} \ln f(\mathbf{v}, \mathbf{y}|\boldsymbol{\theta}) \middle| \{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(N)}\}, \boldsymbol{\theta}^{cur}\right]$.

M-step: Determine $\boldsymbol{\theta}^{*} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{cur})$, make the assignment $\boldsymbol{\theta}^{cur} \leftarrow \boldsymbol{\theta}^{*}$, and return to the E-step.

Let $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_k)}\}$ be a dataset of complete tuples $\mathbf{x}^{\mathrm{T}}$ representing class C_k. If we regard the identity j of the kernel function which generated **x** as unobserved data, the E-step for mixture model (4) can be written as (Bishop, 1995, pp.69-72]

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{cur}) = \sum_{n=1}^{N_k} \sum_{j=1}^{M} P(j|\mathbf{x}^{(n)}, C_k, \boldsymbol{\theta}^{cur}) \ln\left[P(j|C_k, \boldsymbol{\theta}) f(\mathbf{x}^{(n)}|j, C_k, \boldsymbol{\theta})\right],$$

where

$$P(j|\mathbf{x}^{(n)}, C_k, \boldsymbol{\theta}^{cur}) = \frac{p_{kj}^{cur} f(\mathbf{x}^{(n)}|j, C_k, \boldsymbol{\theta}^{cur})}{\sum_{j'=1}^{M} p_{kj'}^{cur} f(\mathbf{x}^{(n)}|j', C_k, \boldsymbol{\theta}^{cur})}$$

and $\boldsymbol{\theta}$ is the vector of parameters p_{kj}, $\boldsymbol{\mu}^{[kj]}$, and $\boldsymbol{\Sigma}^{[kj]}$.

For the M-step, $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{cur})$ is maximised with respect to mixing coefficient π_{kj}, and with respect to the Gaussian parameters $\boldsymbol{\mu}^{[kj]}$ and $\boldsymbol{\Sigma}^{[kj]}$. Equations for the maximisation of $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{cur})$ are as follows [Everitt and Hand, 1981, p.37]:

$$\left.\begin{aligned} p_{kj}^{*} &= \frac{1}{N_k}\sum_{n=1}^{N_k} P(j|\mathbf{x}^{(n)}, C_k, \boldsymbol{\theta}^{cur}) \\ \left(\boldsymbol{\mu}^{[kj]}\right)^{*} &= \frac{1}{N_k p_{kj}^{*}}\sum_{n=1}^{N_k} P(j|\mathbf{x}^{(n)}, C_k, \boldsymbol{\theta}^{cur})\mathbf{x}^{(n)} \\ \left(\boldsymbol{\Sigma}^{[kj]}\right)^{*} &= \frac{1}{N_k p_{kj}^{*}}\sum_{n=1}^{N_k} P(j|\mathbf{x}^{(n)}, C_k, \boldsymbol{\theta}^{cur})\left(\mathbf{x}^{(n)} - \left(\boldsymbol{\mu}^{[kj]}\right)^{*}\right)\left(\mathbf{x}^{(n)} - \left(\boldsymbol{\mu}^{[kj]}\right)^{*}\right)^{\mathrm{T}} \end{aligned}\right\} \quad (13)$$

For the covariance matrix, a simplification can be made by assuming that the Gaussian distributions are symmetrical with variance s_{kj}^{2}. Even with this simplification, the mixture model can approximate any given density function to arbitrary accuracy, provided that an appropriate choice has been made of the model parameters (McLachlan and Basford, 1988).

Appendix B. The EM algorithm for the imputation of real-valued datasets

Let $\{\mathbf{x}_{obs}^{(1)},\ldots,\mathbf{x}_{obs}^{(N_k)}\}$ be the subset of all the rows of an incomplete real-valued dataset that are labelled as belonging to class C_k. The aim is to obtain an imputed dataset $\{\hat{\mathbf{x}}^{(1)},\ldots,\hat{\mathbf{x}}^{(N_k)}\}$, where $\hat{\mathbf{x}}^{(n)} = \mathbf{x}_{obs}^{(n)} \cup \hat{\mathbf{x}}_{mis}^{(n)}$ and $\hat{\mathbf{x}}_{mis}^{(n)}$ is assumed to be unobserved data associated with $\mathbf{x}_{obs}^{(n)}$. The basic EM algorithm of Appendix A is adapted for this imputation as follows:

E-step: For n-1,…,N_k, let $\hat{\mathbf{x}}_{mis}^{(n)} = E[\mathbf{x}_{mis}^{(n)}|\mathbf{x}_{obs}^{(n)}, \boldsymbol{\theta}^{cur}]$

M-step: Obtain the maximum-likelihood estimate for $\boldsymbol{\theta}$ from the current estimate of the completed dataset $\{\hat{\mathbf{x}}^{(1)},\ldots,\hat{\mathbf{x}}^{(N_k)}\}$, let this be the new value for $\boldsymbol{\theta}^{cur}$, and return to the E-step.

These two steps will now be detailed.

Let $\mathbf{x}$ in the completed dataset $\{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N_k)}\}$, where $\mathbf{x} = \mathbf{x}_{obs} \cup \mathbf{x}_{mis}$, be regarded as a random sample drawn from the pdf $f(\mathbf{x}|C_k,\boldsymbol{\theta})$ independently of the other samples, and let the pdf be expressed by the finite mixture model

$$\hat{f}(\mathbf{x}|C_k,\boldsymbol{\theta}) = \sum_{j=1}^{M_k} p_{kj} G(\mathbf{x}|j, C_k, \boldsymbol{\theta}). \quad (14)$$

For the E-step we have that

$$\begin{aligned} E[\mathbf{x}_{mis}^{(n)}|\mathbf{x}_{obs}^{(n)}, C_k, \boldsymbol{\theta}] &= \int_{\Re^{\|\mathbf{x}_{mis}^{(n)}\|}} \mathbf{x}_{mis}^{(n)} f(\mathbf{x}_{mis}^{(n)}|\mathbf{x}_{obs}^{(n)}, C_k, \boldsymbol{\theta}) d\mathbf{x}_{mis}^{(n)} \\ &= \frac{\sum_{j=1}^{M_k} p_{kj} G(\mathbf{x}_{obs}^{(n)}|C_k, \boldsymbol{\theta}_j) E[\mathbf{x}_{mis}^{(n)}|\mathbf{x}_{obs}^{(n)}, C_k, \boldsymbol{\theta}_j]}{\sum_{j=1}^{M_k} p_{kj} G(\mathbf{x}_{obs}^{(n)}|C_k, \boldsymbol{\theta}_j)}, \end{aligned} \quad (15)$$

where $\boldsymbol{\theta}_j$ is the subset of $\boldsymbol{\theta}$ pertaining to the j-th component of the mixture model. If the partitions of $\boldsymbol{\mu}^{[kj]}$ and $\boldsymbol{\Sigma}^{[kj]}$ corresponding to partition $\mathbf{x} = \left(\mathbf{x}_{obs}^{(n)}, \mathbf{x}_{mis}^{(n)}\right)^{\mathrm{T}}$ are

$$\boldsymbol{\mu}^{[kj]} = \begin{pmatrix} \boldsymbol{\mu}_1^{[kjn]} \\ \boldsymbol{\mu}_2^{[kjn]} \end{pmatrix}, \quad \boldsymbol{\Sigma}^{[kj]} = \begin{pmatrix} \boldsymbol{\Sigma}_{11}^{[kjn]} & \boldsymbol{\Sigma}_{12}^{[kjn]} \\ \boldsymbol{\Sigma}_{21}^{[kjn]} & \boldsymbol{\Sigma}_{22}^{[kjn]} \end{pmatrix},$$

then, in accordance with Theorem 2, the mean vector and covariance matrix for $G(\mathbf{x}_{obs}^{(n)}|C_k, \boldsymbol{\theta}_j)$ are $\boldsymbol{\mu}_1^{[kjn]}$ and $\boldsymbol{\Sigma}_{11}^{[kjn]}$, respectively. For the expectation in Eq. (15), we use the relationship

$$E[\mathbf{x}_{mis}^{(n)}|\mathbf{x}_{obs}^{(n)}, C_k, \boldsymbol{\theta}_j] = \boldsymbol{\mu}_1^{[kjn]} + \boldsymbol{\Sigma}_{12}^{[kjn]}\left(\boldsymbol{\Sigma}_{22}^{[kjn]}\right)^{-1}\left(\mathbf{x}_{obs}^{(n)} - \boldsymbol{\mu}_2^{[kjn]}\right). \tag{16}$$

The M-step is achieved by equations (13) using $\hat{\mathbf{x}}^{(n)}$ in place of $\mathbf{x}^{(n)}$.

References

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Broomhead, D.S. and D. Lowe (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems* 2, 321-355.

Everitt, B.S. and D.J. Hand (1981). *Finite Mixture Distributions*. Chapman & Hall, London.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* B39, 1-38.

Friedman, J.H. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association* 76, 817-823.

Hartman, E.J., J.D. Keeler and J.M. Kowalski (1990). Layered neural networks with Gaussian hidden units as universal approximators. *Neural Computation* 2, 210-215.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis*. Oxford University Press, Oxford, p.206.

Little, R.J.A. and D.B. Rubin (1987). *Statistical Analysis with Missing Data*. John Wiley, New York.

Little, R.J.A. and D.B. Rubin (1990). The analysis of social science data with missing values. In: J. Fox and J.S. Long , Eds., *Modern Methods of Data Analysis*. Sage Publications, Newbury Park, CA, pp. 374-409.

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

Moody, J. and C.J. Darken (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation* 1, 281-294.

Parzen, E. (1961). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

Poggio, T. and F. Girosi (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science* 247, 978-982.

Redner, R.A. and H.F. Walker (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26, 195-239.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.