Localised Mixtures of Experts for Mixture of Regressions

Guillaume Bouchard

Institut National de Recherche en Informatique et en Automatique,
Projet IS2 – ZIRST – 655 avenue de l'Europe
38330 Montbonnot Saint-Martin, France

Abstract. In this paper, an alternative to Mixture of Experts (ME) called localised mixture of experts is studied. It corresponds to ME where the experts are linear regressions and the gating network is a Gaussian classifier. The underlying regressors distribution can be considered to be Gaussian, so that the joint distribution is a Gaussian mixture. This provides a powerful speed-up of the EM algorithm for localised ME. Conversely, when studying Gaussian mixtures with specific constraints, one can use the standard EM algorithm for mixture of experts to carry out maximum likelihood estimation. Some constrained models should be very useful, and the corresponding modifications to apply to the EM algorithm are described.

1 Introduction

Let consider a regression model, where the dependent variable Y can be fully explained with a given set of variables $X_1, \cdots, X_{d+1}$. Assume that X_{d+1} is a not observed discrete variable. This regressor is called a latent variable. A natural way of carrying out regression is to explain Y with the d remaining regressors. But the latent variable can carry much information and it could be important to try to recover it. For example, for each value of the latent variable X_{d+1}, the conditional model is completely different. The missing information can be estimated in a mixture of regressions model (Quandt and Ramsey (1978)). Distinction between clusterwise regression models was performed by Hennig (1999).

Switching regression is well known in econometrics literature ; it is a special case of mixture of linear regressions, assuming that the mixture proportions do not depend on the regressors. This model was first examined in Quandt (1972) and Kiefer (1978) gave consistency proof of maximum likelihood (ML) estimators. See Hurn et al. (2001) for Bayesian analysis of this model.

In a general framework, mixtures of regressions are often referred as *Mixtures of Experts* (ME), due to their first introduction in the machine learning community (Jacobs et al. (1991)). ME considers a gating network which is the conditional distribution of the hidden variable given the regressors. These models are therefore called *conditional mixture models*. Some useful results have been established, regarding the convergence rate of EM algorithm (Jordan and Xu (1995)) or identifiability (Jiang and Tanner (1999)). Direct extension is Hierarchical ME where the gating network has a tree structure (Jordan and Jacobs, 1994).

The motivation of this paper is to study mixtures of regressions where we assume a Gaussian distribution for the regressors. It leads to the so-called *localized mixture of experts* (Moerland (1999)) first introduced by Xu et al. (1995). It is also referred as normalized Gaussian networks by Sato and Ishii (2000). Our approch was to work on the joint distribution of the observations. It enables us to link localized mixture of experts with mixture models in their standard form, and thus to take profit of well established theoretical results (McLachlan and Peel (2000)). In this way, we provide a version of the EM algorithm that dramatically decreases the computing time. Conversely, mixtures of experts can be used in the mixture model context to estimate models with specific constraints on parameters. Detailed formula of the EM algorithm for such models are given in this paper.

2 The Model

We consider relashionships between three variables X, Y and H:

- X in $\mathbb{R}^d$ is a vector of d real regressors,
- Y in $\mathbb{R}$ is the dependent variable,
- H in $\{1, \cdots, K\}$ is the latent or hidden discrete variable.

Let $(x,y) = \{(x_i, y_i)_{i=1,\cdots,n}\}$ be iid observations of the couple (X,Y). Since H is not observed, the density of (X,Y) is obtained by marginalization:

$$p(X,Y) = \sum_{k=1}^{K} p(X,Y,H=k). \tag{1}$$

Applying the Bayes rule on $p(X,Y,H)$, we derive two useful expressions of the joint probability:

$$p(X,Y) = \sum_{k=1}^{K} p(X)p(H=k|X)p(Y|X,H=k) \tag{2}$$

$$p(X,Y) = \sum_{k=1}^{K} p(H=k)p(X|H=k)p(Y|X,H=k). \tag{3}$$

For these two parametrizations, the distribution of Y conditionally on $H=k$ and $X=x$ is, as usual in linear regression, an univariate Gaussian with mean $\beta_k' x + \alpha_k$ and variance τ_k^2:

$$Y|X=x, H=k \sim \mathcal{N}(\beta_k' x + \alpha_k, \tau_k^2). \tag{4}$$

. We now present models that find estimators of β_k,α_k and τ_k.

2.1 Standard mixture of experts

The expression (2) corresponds to the conditional mixture model, since maximizing its log-likelihood does not require knowledge of the distribution of X. It is equivalent to work with the conditional probability of Y given X:

$$p(Y|X) = \sum_{k=1}^{K} \underbrace{p(H=k|X)}_{gating\ network} \underbrace{p(Y|X, H=k)}_{expert} \tag{5}$$

In this case, the gating network classifier $p(H|X)$ has to be specified. The multinomial logit model is usually choosed. It is a generalised linear model with conditional density

$$p(H=k|X) = \frac{p_k e^{v_k x}}{\sum_{l=1}^{K} p_l e^{v_l x}}, \quad k = 1, \cdots, K, \tag{6}$$

where vectors v_k and proportions p_k are parameters such that $v_K = 0$ and $\sum_{k=1}^{K} p_k = 1$.

2.2 Localized mixture of experts

In the sequel, we opt for parametrization (3), which corresponds to a standard *mixture model* where each component has density $p(X|H=k)p(Y|X,H=k)$. A multinomial distribution is assumed for H:

$$H \sim \mathcal{M}(1, p), \tag{7}$$

where $p = (p_1, \cdots, p_K)'$ is a vector of component proportions such that $\sum_{k=1}^{K} p_k = 1$. Given mixture component H, the regressors X are assumed to arise from a multivariate Gaussian distribution:

$$X|H=k \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{8}$$

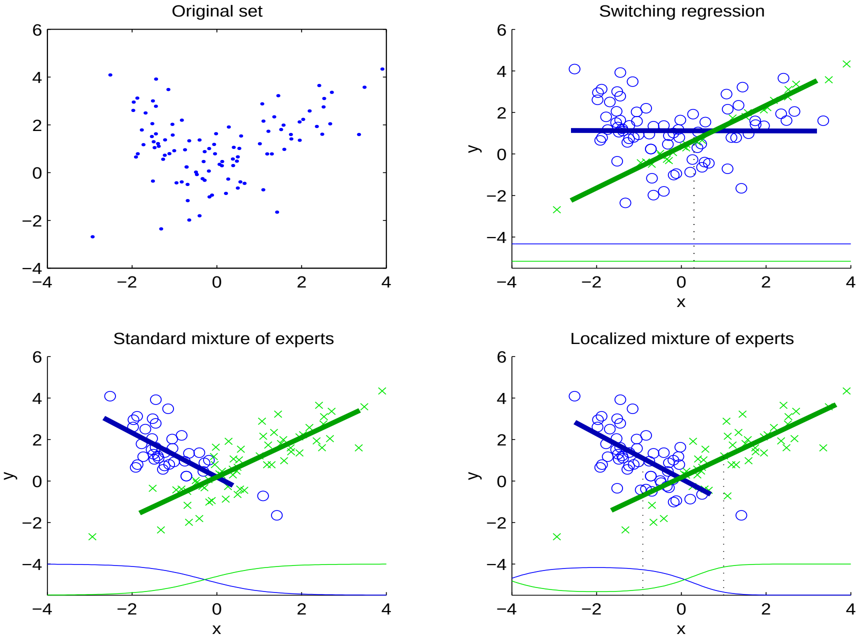
With the Gaussian parametrization, components can be interpreted in a more natural way than the standard ME, since the means μ_k summarize the regressors. The corresponding gating network classifier can be obtained by direct application of the Bayes rule:

$$p(H=k|X=x) = \frac{p(H=k)p(X=x|H=k)}{p(X=x)} \tag{9}$$

$$= \frac{p_k |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-m_k)'\Sigma_k^{-1}(x-m_k)}}{\sum_{l=1}^{K} p_l |\Sigma_l|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-m_l)'\Sigma_l^{-1}(x-m_l)}}. \tag{10}$$

This is exactly the gating network with Gaussian kernel proposed by Xu et al. (1995). This parametrization differs from the usual *softmax* fonction primarily by the quadratic form of the canonical link. We refer this model as

Fig. 1. Illustration of a mixture of two regression: from a dataset where a simple linear regression is not suitable (top left), switching regression (top right) finds two optimal regression lines equally distributed on the regressors. Standard mixture of experts (bottom left) expresses the proportions from a linear logistic model on the regressors. Localized ME (bottom right) assume that the distribution of the regressors is normal with parameters depending on the hidden variable. We see that the last two models give similar results. Proportions are represented at the bottom of the graphs.



localized mixture of experts, following Moerland (1999), who compared Gaussian and standard gating network for classification and noted a slight superiority of standard ME. Localized ME was successfully exploited in Fritsch (1996) and Fritsch, Finke and Waibel (1997) for speech recognition. They show that such kernels can reduce significantly the time of convergence of the EM algorithm for large databases. They obtain near optimal initial parameter values μ_k and Σ_k by an unsupervised learning applied on regressors only. We should stress that originally, localized ME introduction was for convenience, but we give a natural justification of their use in terms of probability assumption.

It can easily be proved that the joint distribution of the observations (X', Y') is a mixture of $(d+1)$-dimensional Gaussian distributions. The proportions are the p_k, $k = 1, \cdots, K$ defined above, the mean and covariance matrix of the kth component are

$$m_k = \begin{pmatrix} \mu_k \\ \mu'_k\beta_k + \alpha_k \end{pmatrix}, \quad \varGamma_k = \begin{bmatrix} \varSigma_k & \varSigma_k\beta_k \\ \beta'_k\varSigma_k & \tau_k^2 + \beta'_k\varSigma_k\beta_k \end{bmatrix}. \tag{11}$$

Then, localised ME is just a Gaussian mixture with a specific parametrization.

2.3 Adding constraints

The basic model described above has $(\frac{d^2}{2} + \frac{5}{2}d + 3)K - 1$ parameters. It can be excessive since the number of parameters grows quadraticaly with the dimension d of the data. To avoid overfitting, we can derive more parsimonious models by adding constraints on parameter values. The first assumption would be to constraint the component covariance matrix $\varSigma_k$ to be diagonal. This assumption corresponds to the conditional independence of regressors given the component, and we claim that it may not be very severe since regression focus essentially on coefficients β_k. We then obtain a particular covariance matrices $\varGamma_k$ of the joint distribution for each component k:

$$\varGamma_k = \begin{bmatrix} \sigma_{k1}^2 & 0 & \dots & \sigma_{k1}^2\beta_{k1} \\ \vdots & \ddots & 0 & \vdots \\ 0 & \dots & \sigma_{kd}^2 & \sigma_{kd}^2\beta_{kd} \\ \sigma_{k1}^2\beta_{k1} & \dots & \sigma_{kd}^2\beta_{kd} & \tau_k^2 + \prod_{i=1}^d \beta_{ki}^2\sigma_{ki}^2 \end{bmatrix}. \tag{12}$$

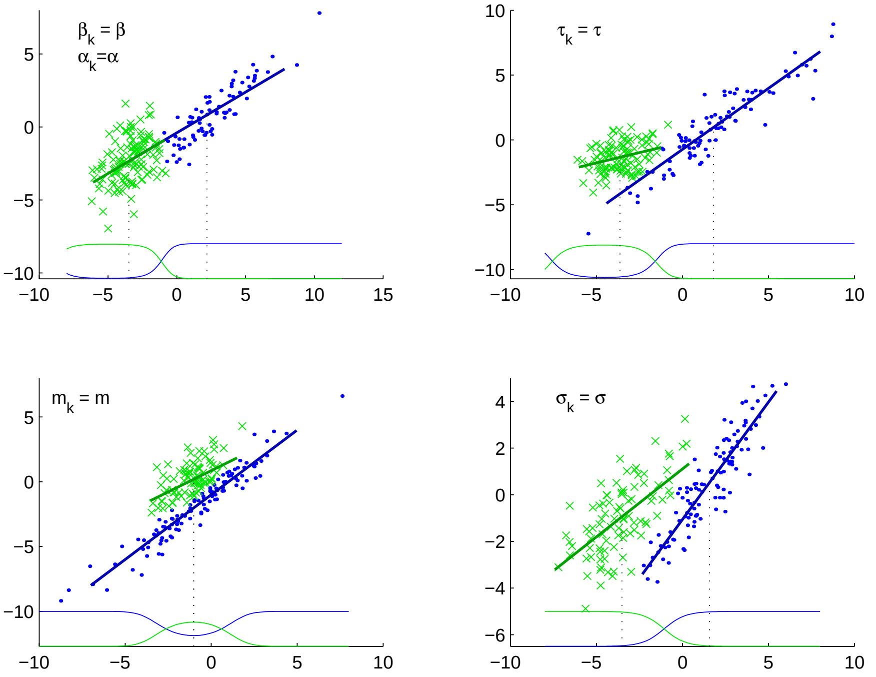
To our knowledge, this type of covariance matrix was never mentioned in the literature on Gaussian mixtures. It can yet be useful for specific problems. This model has now $K(2d+3) - 1$ parameters which is linear in d, so that it can be more suitable in high dimension problems.

Another class of models can be obtained by constraining some parameters to be equal between groups:

1. $p_k = p$: components proportions are equal. This assumption can be regarded as unrealistic. However, maximum likelihood estimator can be expected to be more stable since the number of local maxima of likelihood dramatically decreases when proportions are equal.
2. $\beta_k = \beta$: common slope between components. The model is then a linear regression for which the error term can be dependent on regressors values.
3. $\tau_k = \tau$: common error term between regression. This constraint forces the model to have a constant error term in each component.
4. $\varSigma_k = \varSigma$: common regressors covariance matrix. This assumption is useful when we want a linear separations between groups instead of a quadratic one. This is illustrated on Figure 2 with $\sigma_k = \sigma$. The probabilities of each component are split between left and right, contrary to other models.

Some other contraints would be to set $\alpha_k = \alpha$ or $\mu_k = \mu$, i.e. assuming that components have a common intercept or a common mean. They correspond to very specific models, as it can be viewed on Figure 2. Combinations of these constraints leads to a large variety of different models. Note that some of these constraints may be very severe and only applicable to specific distributions.

Fig. 2. Some illustrations of data sets in accordance with constrained models.



3 Maximum likelihood estimation

Before carrying out ML estimation, we must ensure that the model is identifiable. Hennig (2000) gives necessary conditions for the existence of consistent estimators in mixture of regressions with random regressors: the regressor distribution must not give positive probability to any $(d-1)$-dimensional hyperplanes. Of course, this does not occur as soon as the Σ_k are not singular, which is not a mild condition. Following Dempster et al. (1977), we now describe the EM algorithm for the Gaussian mixture of experts. We write $\tilde{\beta}_k = (\alpha_k, \beta'_k)'$ for $k = 1, \cdots, K$. Let θ be the vector of parameters containing p_k, $\tilde{\beta}$, τ_k, μ_k or Σ_k for $k = 1, \cdots, K$.

E step. The expectation step require the computation of the conditional expectation of the complete log likelihood

$$Q(\theta|\theta^{(t)}) = E\{L_c(\theta; x, y)|x, y, \theta^{(t)}\} \tag{13}$$

where $\theta^{(t)}$ is the value of parameter vector at iteration t and $L_c(\theta; x, y)$ is the complete log likelihood of the model. Denoting h the density of $X|H$ and g the density of $Y|X, H$, we have

$$L_c(\theta; x, y) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log\left(p_k h(x_i; \mu_k, \Sigma_k) g(y_i; x_i, \tilde{\beta}_k, \tau_k)\right). \tag{14}$$

Here, c_{ik} equals to 1 if data i comes from component k, and 0 otherwise. Its expectation conditionally on parameters $\theta^{(t)}$ is

$$w_{ik}^{(t)} = \frac{p_k^{(t)} h(x_i; \mu_k^{(t)}, \Sigma_k^{(t)}) g(y_i; x_i, \tilde{\beta}_k^{(t)}, \tau_k^{(t)})}{\sum_{l=1}^{K} p_l^{(t)} h(x_i; \mu_l^{(t)}, \Sigma_l^{(t)}) g(y_i; x_i, \tilde{\beta}_l^{(t)}, \tau_l^{(t)})}. \tag{15}$$

M step. The maximization step is computing

$$\theta^{(t+1)} = \underset{\theta}{argmax}\, Q(\theta|\theta^{(t)}). \tag{16}$$

From equations (13) and (14) we get

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t)} \log[p_k^{(t)} h(x_i; \mu_k^{(t)}, \Sigma_k^{(t)}) g(y_i; x_i, \tilde{\beta}_k^{(t)}, \tau_k^{(t)})] \tag{17}$$

Let $X = [x_1, \cdots, x_n]'$ be the matrix of regressors and $\tilde{X} = [\mathbb{I}\ X]$ where $\mathbb{I}$ is a $n \times 1$ vector of ones. Y is the vector of y_i and $W_k^{(t)}$ are $n \times n$ diagonal matrices with $w_{ik}^{(t)}$ on their diagonal. Expression (17) is maximised by setting its partial on p_k, $\tilde{\beta}$, τ_k, μ_k and Σ_k to zero. We obtained closed form solutions:

$$p_k^{(t+1)} = \frac{1}{n} tr W_k^{(t)}, \tag{18}$$

$$\tilde{\beta}_k^{(t+1)} = (\tilde{X}' W_k^{(t)} \tilde{X})^{-1} \tilde{X} W_k^{(t)} Y, \tag{19}$$

$$\tau_k^{2\,(t+1)} = \frac{1}{tr W_k^{(t)}} (Y - \tilde{X} \tilde{\beta}_k^{(t+1)})' W_k^{(t)} (Y - \tilde{X} \tilde{\beta}_k^{(t+1)}), \tag{20}$$

$$\mu_k^{(t+1)} = \frac{1}{tr W_k^{(t)}} X' W_k^{(t)} \mathbb{I}', \tag{21}$$

$$\Sigma_k^{(t+1)} = \frac{1}{tr W_k^{(t)}} (X - \mathbb{I} \mu_k^{'(t)})' W_k^{(t)} (X - \mathbb{I} \mu_k^{'(t)}). \tag{22}$$

We can notice that equations (19) and (20) correspond to a weighted least square fit and equations (21) and (22) give weighted mean and variance.

3.1 ML estimation of constrained models

To adapt the previous algorithm to the constraint models defined above, the procedure is the same. Derivatives of expression (17) are slightly different, but except for $\tilde{\beta}_k^{(t+1)}$, the modifications are straightforward and no included here, except for the constraint $\tilde{\beta}_k = \tilde{\beta}$: we have to solve the nonlinear system

$$\begin{cases} \tau_k^{2\,(t+1)} = \frac{1}{trW_k^{(t)}}(Y - \tilde{X}\tilde{\beta}_k^{(t+1)})'W_k^{(t)}(Y - \tilde{X}\tilde{\beta}_k^{(t+1)}), \\ \tilde{\beta}_k^{(t+1)} = \left(\tilde{X}'(\sum_{k=1}^{K} \frac{1}{\tau_k^{2\,(t+1)}} W_k^{(t)})\tilde{X}\right)^{-1} \tilde{X}(\sum_{k=1}^{K} \frac{1}{\tau_k^{2\,(t+1)}} W_k^{(t)})Y. \end{cases} \tag{23}$$

Since closed form solution of this system is not available, we simply replace the term $\tau_k^{2(t+1)}$ by its previous value $\tau_k^{2(t)}$ in the expression of $\tilde{\beta}_k^{(t+1)}$. It can be proved that acting in such a way, the modified M step still increases the likelihood. Thus, we have defined a Generalised EM algorithm which have the same properties than EM (see Dempster et a. (1977)).

3.2 Reducing the computing time

If no constraint on component parameters is applied, ML estimation is straightforward since the model is a mixture of normal distributions. This approch differs from Fritsch's one in the sense that we directly obtain ML estimators by an unsupervised learning on the joint distribution $p(X, Y)$ instead of $p(X)$. There exists various effective ways of finding ML estimators of μ_k and Γ_k, but the most used is the EM algorithm (McLachlan and Peel (2000)). Once we get the ML estimators $\hat{\mu}_k$ and $\hat{\Gamma}_k$, we write $\hat{\mu}_k = \begin{bmatrix} e_k \\ f_k \end{bmatrix}$ and $\hat{\Gamma}_k = \begin{bmatrix} A_k & b_k \\ b_k' & c_k \end{bmatrix}$, A_k being a $d \times d$ matrix, b_k and e_k vectors of $\mathbb{R}^d$ and f_k and e_k real values. We solve equations (11), getting $\hat{\mu}_k = e_k$, $\hat{\Sigma}_k = A_k$, $\hat{\beta}_k = \hat{\Sigma}_k^{-1} b_k$, $\hat{\tau}_k = c_k - \hat{\beta}_k' \hat{\Sigma}_k \hat{\beta}_k$, and finally $\alpha_k = f_k - \hat{\mu}_k' \hat{\beta}_k$. Not only this estimation is simple, but it is faster than the previous EM algorithm. Namely, in the first algorithm, two inversions of matrices are needed for each EM step and each component (one in the evaluation of the density $h((x_i; \mu_k^{(t)}, \Sigma_k^{(t)})$ in the E step and one for the weighted least square fit (19) in the M step). By constrast, in this new version, each step requires one matrix inversion (evaluation of the Gaussian density in the E step), thus dividing the computing time by 2 when the dimension d is large. It proves that this specific parametrization can appreciably simplify estimation.

4 Discussion

We studied a mixture of Gaussian distributions used in a regression purpose, and showed that it can be viewed as a mixture of experts with Gaussian gating network. The specific parametrization provide a natural interpretation of clusters and enable us to add constraints that were never mentioned in Gaussian mixture literature. A wide variety of clusterwise regression models is therefore available, that can be used for multiple purposes. Robust linear regression is possible, so that non-Gaussian error terms can be handle easily.

Some further work is needed to provide results on this topic. Independence constraints also permit to reduce significantly the number of parameters, and this can be particulary desirable in high dimension. Finally, on the model without constraint on component parameters, we gave a powerful way of estimating ML parameters.

References

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

J. FRITSCH. (1996): Modular neural networks for speech recognition. Master's thesis, Carnegie Mellon University & University of Karlsruhe. `ftp://reports.adm.cs.cmu.edu/usr/anon/1996/CMU-CS-96-203.ps.gz`.

FRITSCH, J., FINKE, M. and WAIBEL, A. (1997): Adaptively growing hierarchical mixtures of experts. In M. C. Mozer, M. I. Jordan and T. Petsche (Eds.), *Advances in Neural Informations Processing Systems*, 9. MIT Press.

HENNIG, C. (1999): Models and Methods for Clusterwise Linear Regression. Gaul, W. and Locarek-Junge, H. (Eds): *Classification in the Information Age.* Springer, Berlin, 179–187.

HENNIG, C. (2000): Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification*, 17, 273–296.

HURN, M. A., JUSTEL, A. and C. P., ROBERT. (2000): Estimating mixtures of regressions. Technical report, CREST, France.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991): Adaptive mixture of local experts. *Neural Computation*, 3(1), 79–87.

JIANG, W. and TANNER, M.A. (1999): Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Ann. Statistics*, 27, 987–1011.

JORDAN, M. I. and JACOBS, R. A. (1994): Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6, 181–214.

QUANDT, R. E.(1972): A new Approach to Estimating Switching Regressions *Journal of the American Statistical Association*, 67, 306–310.

KIEFER, N. M.(1978): Discrete Parameter Variation : Efficient Estimation of a Switching Regression Model *Econometrica*, 46, 427–434.

QUANDT, R. E. and RAMSEY, J. B. (1978): Estimating mixtures of normal distributions and switching regressions, *JASA*, 73, 730–752.

McLACHLAN, G. J. and PEEL., D. (2000): *Finite Mixture Models*, Wiley.

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

MOERLAND, P. (1999) Classification using localized mixtures of experts. In proc. of the *International Conference on Artificial Neural Networks.*

SATO, M. and ISHII, S. (2000): On-line EM algorithm for the normalized gaussian network. *Neural Computation*, 12(2), 407–432.

XU, L. and JORDAN, M.I. (1995): On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1), Jan.

XU, L., HINTON, G. and JORDAN, M. I. (1995): An alternative model for mixtures of experts. In G. Tesauro et al. edts., *Advances in Neural Information Processing Systems*, 7, 633–640, Cambridge MA, MIT Press.