

Cluster Weighted Modeling with Student- t components

Salvatore Ingrassia, Simona C. Minotti, and Giorgio Vittadini

Abstract We consider Cluster-Weighted Modeling in order to model functional dependence between some input and output variables based on data coming from an heterogeneous population. While such approach is usually based on Gaussian assumptions, here we extend this framework to Student- t distributions which provide both more realistic tails for real-world data and robust parametric extension to the fitting of data with respect to the alternative Gaussian models. Theoretical results are illustrated on the ground of some numerical simulations.

Key words: Cluster-weighted modeling, Classification and pattern recognition, t -distribution.

1 Introduction

The functional dependence between some input and output variables $(\mathbf{X}, Y)$ based on data coming from an heterogeneous population Ω constituted by G homogeneous subpopulations $\Omega_1, \dots, \Omega_G$, is often estimated using methods able to model local behaviour, like Finite Mixtures of Regression (FMR) and Finite Mixtures of

Salvatore Ingrassia
Dipartimento di Impresa, Culture e Società
Università di Catania
Corso Italia 55, - Catania (Italy). e-mail: s.ingrassia@unict.it

Simona C. Minotti
Dipartimento di Statistica
Università di Milano-Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano (Italy). e-mail: simona.minotti@unimib.it

Giorgio Vittadini
Dipartimento di Metodi Quantitativi per l'Economia e le Scienze Aziendali
Università di Milano-Bicocca
Via Bicocca degli Arcimboldi 8 - 20126 Milano (Italy). e-mail: giorgio.vittadini@unimib.it

Regression with Concomitant variables (FMRC), see e.g. [4], [1]. As a matter of fact, the purpose of these models is to identify groups by taking into account the local relationships between some response variable Y and some d -dimensional explanatory variables $\mathbf{X} = (X_1, \dots, X_d)$.

Here we focus on a different approach called *Cluster-Weighted Modeling* (CWM), proposed first in [5]. While mixtures of regressions consider the conditional probability density $p(y|\mathbf{x})$, the CWM approach models the joint probability density $p(\mathbf{x}, y)$ as a weighted sum over G clusters, where each cluster is modeled by the product of an input distribution $p(\mathbf{x}|\Omega_g)$ (i.e. a local model for the input variable $\mathbf{X}$) and an output distribution $p(y|\mathbf{x}, \Omega_g)$ (i.e. a local model for the dependence between $\mathbf{X}$ and Y). In the literature, the local densities $p(\mathbf{x}|\Omega_g)$ and $p(y|\mathbf{x}, \Omega_g)$ are usually Gaussian distributed. In this paper we consider cluster weighted modeling based on Student- t distributions which are becoming popular and popular in multivariate statistics because they provide more realistic tails for real-world data with respect to the alternative Gaussian models, see e.g. [10], [11]. Moreover models based on t -distributions provide a robust parametric extension to the fitting of data with respect to normal mixtures. Indeed, in presence of some noise component, potential outliers or data with longer-than-normal tails, one way to broaden the model can be provided by considering t distributions, see e.g. [13].

The rest of the paper is organized as follows. In Section 2 Cluster-Weighted Modeling is introduced; in Section 3 we introduce the Cluster-Weighted Modeling based on Student- t distributions; in Section 4 some numerical studies are outlined. Finally, in Section 5 we provide some conclusions and remarks for further research.

2 The Cluster-Weighted Modeling

Let $(\mathbf{X}, Y)$ be a pair of a random vector $\mathbf{X}$ and a random variable Y defined on Ω with joint probability distribution $p(\mathbf{x}, y)$, where $\mathbf{X}$ is the d -dimensional input vector with values in some space $\mathcal{X} \subseteq \mathbb{R}^d$ and Y is a response variable having values in $\mathcal{Y} \subseteq \mathbb{R}$. Thus $(\mathbf{x}, y) \in \mathbb{R}^{d+1}$. Assume that Ω can be partitioned into G disjoint groups, say $\Omega_1, \dots, \Omega_G$, that is $\Omega = \Omega_1 \cup \dots \cup \Omega_G$. *Cluster-Weighted Modeling* (CWM) decomposes the joint probability $p(\mathbf{x}, y)$ as:

$$p(\mathbf{x}, y) = \sum_{g=1}^G p(y|\mathbf{x}, \Omega_g) p(\mathbf{x}|\Omega_g) \pi_g, \quad (1)$$

where $\pi_g = p(\Omega_g)$ is the mixing weight of group Ω_g , $p(\mathbf{x}|\Omega_g)$ is the probability density of $\mathbf{x}$ given Ω_g and $p(y|\mathbf{x}, \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\mathbf{x}$ and the group Ω_g , $g = 1, \dots, G$, see [5]. Throughout this paper we assume that the input-output relation can be written as $Y = \mu(\mathbf{x}; \beta) + \varepsilon$, where ε is a random variable with zero mean and finite variance and β denotes the set of the parameters of $\mu(\cdot)$. Sometimes in (1) we shall write $p(y|\mathbf{x}, \Omega_g; \mu)$ rather than $p(y|\mathbf{x}, \Omega_g)$.

Formula (1) highlights that the joint density of $(\mathbf{X}, Y)$ can be viewed as a mixture of local models $p(y|\mathbf{x}, \Omega_g)$ weighted (in a broader sense) on both the local densities $p(\mathbf{x}|\Omega_g)$ and the mixing weights π_g . In the model based approach to clustering, the posterior probability $p(\Omega_g|\mathbf{x}, y)$ of the g -th group ($g = 1, \dots, G$) is given by:

$$p(\Omega_g|\mathbf{x}, y) = \frac{p(\mathbf{x}, y, \Omega_g)}{p(\mathbf{x}, y)} = \frac{p(y|\mathbf{x}, \Omega_g)p(\mathbf{x}|\Omega_g)\pi_g}{\sum_{j=1}^G p(y|\mathbf{x}, \Omega_j)p(\mathbf{x}|\Omega_j)\pi_j}. \quad (2)$$

Usually, in the literature about CWM, the local densities $p(\mathbf{x}|\Omega_g)$ are assumed to be multivariate Gaussian with parameters (μ_g, Σ_g) , that is $\mathbf{X}|\Omega_g \sim N_d(\mu_g, \Sigma_g)$; moreover also the conditional density $p(y|\mathbf{x}, \Omega_g)$ is often modeled by a Gaussian distribution with variance $\sigma_{\varepsilon, g}^2$ around some deterministic function of $\mathbf{x}$, say $\mu_g(\mathbf{x}; \beta)$. Thus $p(\mathbf{x}|\Omega_g) = \phi_d(\mathbf{x}; \mu_g, \Sigma_g)$ and $p(y|\mathbf{x}, \Omega_g) = \phi(y; \mu(\mathbf{x}; \beta_g), \sigma_{\varepsilon, g}^2)$ ($g = 1, \dots, G$), where $\phi_d(\mathbf{x}; \mu_g, \Sigma_g)$ denotes the probability density of a d -dimensional multivariate Gaussian. This implies that the response variable Y in the g -th group is given by:

$$Y|\mathbf{x}, \Omega_g = \mu(\mathbf{x}; \beta_g) + \varepsilon_g \quad g = 1, \dots, G,$$

where $\varepsilon_g \sim N(0, \sigma_{\varepsilon, g}^2)$. Thus (1) can be rewritten as:

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G \phi(y; \mu(\mathbf{x}; \beta_g), \sigma_{\varepsilon, g}^2) \phi_d(\mathbf{x}; \mu_g, \Sigma_g) \pi_g, \quad (3)$$

where $\theta = (\beta_1, \dots, \beta_G, \mu_1, \dots, \mu_G, \Sigma_1, \dots, \Sigma_G, \sigma_{\varepsilon, 1}^2, \dots, \sigma_{\varepsilon, G}^2, \pi_1, \dots, \pi_G)$ summarizes all unknown parameters of the model. In the following such model will be referred to as the *Gaussian CWM*.

In the simplest case, the conditional densities (3) are based on linear mappings $\mu(\mathbf{x}; \beta_g) = \mathbf{b}'_g \mathbf{x} + b_{g0}$, with $\beta = (\mathbf{b}'_g, b_{g0})'$, where $\mathbf{b} \in \mathbb{R}^d$ and $b_{g0} \in \mathbb{R}$ yielding the linear Gaussian CWM:

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G \phi(y; \mathbf{b}'_g \mathbf{x} + b_{g0}, \sigma_{\varepsilon, g}^2) \phi_d(\mathbf{x}; \mu_g, \Sigma_g) \pi_g. \quad (4)$$

Under suitable assumptions, one can show that the CWM (3) includes as special cases finite mixtures of Gaussian distributions (FMG), finite mixture of regression (FMR) given by

$$f(y|\mathbf{x}; \psi) = \sum_{g=1}^G p(y|\mathbf{x}, \Omega_g) \pi_g = \sum_{g=1}^G \phi(y; \mu(\mathbf{x}; \beta_g), \sigma_{\varepsilon, g}^2) \pi_g,$$

where ψ denotes the overall parameters of the model, and finite mixture of regression with concomitant variables (FMRC), given by

$$f^*(y|\mathbf{x}; \psi^*) = \sum_{g=1}^G \phi(y; \mu(\mathbf{x}; \beta_g), \sigma_{\varepsilon, g}^2) p(\Omega_g|\mathbf{x}, \xi), \quad (5)$$

where the mixing weight $p(\Omega_g|\mathbf{x}, \xi)$ now is a function depending on $\mathbf{x}$ through some ξ , which denotes the parameters of the weight function π_g , and ψ^* is the augmented set of all parameters to be estimated in the model. Table 1 summarizes the relationships between CWM and FMG, FMR, FMRC. Moreover the CWM is a Ω_g -to- $\mathbf{x}$ model, while the FMRC is a $\mathbf{x}$ -to- Ω_g model; according to [8], in the framework of neural networks they are called the *generative direction* model and the *diagnostic direction* model respectively. Finally, an important aspect concerns

	$p(\mathbf{x} \Omega_g)$	$p(y \mathbf{x}, \Omega_g)$	assumption
FMG	Gaussian	Gaussian	linear relationship between $\mathbf{X}$ and Y
FMR		Gaussian	$\mathbf{X} \Omega_g \sim N_d(\mu, \Sigma)$ not depending on g
FMRC		Gaussian	$G = 2$ and $\mathbf{X}$ is a Gaussian mixture with $\Sigma_1 = \Sigma_2$

Table 1 Some special cases of CWM based on Gaussian distributions.

the number of parameters to be estimated. Obviously CWM has a larger number of parameters than FMR and FMRC and thus the estimation of the parameters in CWM requires a quite larger amount of data than the other two models; the counterpart is that this leads to a more flexible approach. Some simulation studies showed that there are situations in which we need a less parsimonious model in order to identify input-output relationships which cannot be captured by either FMR or FMRC.

3 Student- t based CWM

In this section we consider cluster-weighted modeling based on Student- t distributions which provide more realistic tails for real-world data with respect to the Gaussian models. Moreover models based on t -distributions provide a robust parametric extension to the fitting of data with respect to normal mixtures. Indeed, in presence of some noise component, potential outliers or data with longer-than-normal tails, one way to broaden the model can be provided by considering t distributions.

Let us assume that in model (1) both $p(\mathbf{x}|\Omega_g)$ and $p(y|\mathbf{x}, \Omega_g)$ are Student- t densities. In particular we assume that $\mathbf{X}|\Omega_g$ has a multivariate t distribution with location parameter μ_g , inner product matrix Σ_g and degrees of freedom ν_g , that is $\mathbf{X}|\Omega_g \sim t_d(\mu_g, \Sigma_g, \nu_g)$, and $Y|\mathbf{x}, \Omega_g$ has a t distribution with location parameter $\mu(\mathbf{x}; \beta_g)$, scale parameter σ_g^2 and degrees of freedom ζ_g , that is $Y|\mathbf{x}, \Omega_g \sim t(\mu(\mathbf{x}; \beta_g), \sigma_g^2, \zeta_g)$, so that

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G p_Y(y; \mu(\mathbf{x}; \beta_g), \sigma_g^2, \zeta_g) p_{\mathbf{X}}(\mathbf{x}; \mu_g, \Sigma_g, \nu_g) \pi_g. \quad (6)$$

This implies that

$$\begin{aligned} \mathbf{X}|\mu_g, \Sigma_g, \nu_g, U_g &\sim N_d\left(\mu_g, \frac{\Sigma_g}{U_g}\right) \quad g = 1, \dots, G \\ Y|\mu(\mathbf{x}, \beta_g), \sigma_g, \zeta_g, W_g &\sim N\left(\mu(\mathbf{x}, \beta_g), \frac{\sigma_g}{W_g}\right) \quad g = 1, \dots, G \end{aligned}$$

where U_g and W_g are independent random variables such that

$$U_g|\mu_g, \Sigma_g, \nu_g \sim \Gamma\left(\frac{\nu_g}{2}, \frac{\nu_g}{2}\right) \quad \text{and} \quad W_g|\mu(\mathbf{x}, \beta_g), \sigma_g, \zeta_g \sim \Gamma\left(\frac{\zeta_g}{2}, \frac{\zeta_g}{2}\right),$$

for $g = 1, \dots, G$. The model (6) will be referred to as the t -CWM; the special case in which $\mu(\mathbf{x}; \beta_g)$ is some linear mapping will be called the *linear t -CWM*.

Let us consider the following Lemma, see e.g. [12] for details.

Lemma 1 Let $\mathbf{Z}$ be a q -variate random vector having multivariate t distribution with degrees of freedom $\nu \in (0, \infty)$, location parameter μ and positive definite inner product matrix Σ . Assume that $\mathbf{Z}$ is partitioned as $\mathbf{Z} = (\mathbf{Z}'_1, \mathbf{Z}'_2)'$, where $\mathbf{Z}_1$ takes values in $\mathbb{R}^{q_1}$ and $\mathbf{Z}_2$ in $\mathbb{R}^{q_2} \subset \mathbb{R}^{q-q_1}$ so that the parameters of $\mathbf{Z}$ can be written as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then

$$\mathbf{Z}_1 \sim t_{q_1}(\mu_1, \Sigma_{11}, \nu) \quad \text{and} \quad \mathbf{Z}_2|\mathbf{z}_1 \sim t_{q_2}(\mu_{2|1}, \Sigma_{2|1}^*, \nu + q_1) \quad (7)$$

where

$$\begin{aligned} \mu_{2|1} &= \mu_{2|1}(\mathbf{z}_1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 - \mu_1) \\ \Sigma_{2|1}^* &= \Sigma_{2|1}^*(\mathbf{z}_1) = \frac{\nu + \delta(\mathbf{z}_1; \mu_1, \Sigma_{11})}{\nu + q_1} \Sigma_{2|1} \end{aligned} \quad (8)$$

with $\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ and $\delta(\mathbf{z}_1; \mu_1, \Sigma_{11}) = (\mathbf{z}_1 - \mu_1)' \Sigma_{11}^{-1}(\mathbf{z}_1 - \mu_1)$. $\square$

In particular we point out that the covariance matrix of $\mathbf{Z}_2|\mathbf{z}_1$ depends on $\mathbf{z}_1$, basis on (8).

Now let us set $\mathbf{Z} = (\mathbf{X}', Y)'$ where $\mathbf{X}$ is a d -dimensional input vector and Y is a random variable defined on Ω , thus $\mathbf{Z}$ is a random vector with values in $\mathbb{R}^{d+1}$. Assume that the joint density of $\mathbf{Z}$ can be decomposed as a finite mixture of G multivariate t distributions (FMT) with parameters $(\mu_g^*, \Sigma_g^*, \nu_g)$, $g = 1, \dots, G$:

$$p(\mathbf{z}) = \sum_{g=1}^G p(\mathbf{z}; \mu_g^*, \Sigma_g^*, \nu_g) \pi_g.$$

Based on the above lemma, we have the following result.

Proposition 2 Let us consider the linear t -CWM

$$p(\mathbf{x}, y; \theta) = \sum_{g=1}^G p_Y(y; \mathbf{b}_g' \mathbf{x} + b_{g0}, \sigma_g^{*2}, \zeta_g) p_{\mathbf{X}}(\mathbf{x}; \mu_g, \Sigma_g, v_g) \pi_g. \quad (9)$$

If $\zeta_g = v_g + d$ and $\sigma_g^{*2} = \sigma_g^2[v_g + \delta(\mathbf{x}; \mu_g, \Sigma_g)]/(v_g + d)$ then the model (9) coincides with a mixture of multivariate t distribution for suitable parameters $\mathbf{b}_g, b_{g0}$ and σ_g^2 , $g = 1, \dots, G$. $\square$

This result show that, in general, the t -CWM is not a mixture of multivariate t -distributions.

Modeling based on the t -distribution is often motivated by robust clustering of multivariate normal data with noise. Let $\mathcal{D} = \{(x_n, y_n)\}_{n=1, \dots, N}$ be a sample dataset coming from a population $\Omega = \Omega_1 \cup \dots \cup \Omega_G$, including also some noisy data. Parameters in cluster weighted models are usually estimated according to the likelihood approach using the EM algorithm. In such cases, once the maximum likelihood estimates of the parameters $\hat{\pi}_g, \hat{\mu}_g, \hat{\Sigma}_g, \hat{v}_g$ ($g = 1, \dots, G$), have been obtained, outliers are labeled according to the following criterion, see e.g. [13]. Let us set $\delta(\mathbf{x}_n; \hat{\mu}_g, \hat{\Sigma}_g) = (\mathbf{x}_n - \hat{\mu}_g)' \hat{\Sigma}_g^{-1} (\mathbf{x}_n - \hat{\mu}_g)$ and assume that the distribution of $\delta(\mathbf{X}_n; \hat{\mu}_g, \hat{\Sigma}_g)$ can be approximated by a chi-square distribution with q degrees of freedom. This assumption is based on recalling that the distribution of $\delta(\mathbf{X}; \mu, \Sigma)$ is exactly $\chi^2(q)$ when $\mathbf{X}$ is a random vector having a multivariate normal distribution. Then an observation $\mathbf{x}_n$ is treated as an outlier (and thus it will be classified as a noise data) if

$$\sum_{g=1}^G \hat{z}_{jn} \delta(\mathbf{x}_n; \hat{\mu}_g, \hat{\Sigma}_g) > \chi_{1-\alpha}^2(q) \quad (10)$$

where $\hat{z}_{jn} = 1$ if $\mathbf{x}_n$ unit is classified in the j -th group according to maximum posterior probability and 0 otherwise, while $\chi_{1-\alpha}^2(q)$ denotes the quantile of order $(1 - \alpha)$ of the chi-squared distribution with q degrees of freedom.

4 Numerical studies

The statistical models introduced before have been evaluated on the grounds of many numerical studies concerning robust classification of noisy data, where the noise has been simulated by adding units generated with uniform distribution. We considered both simulated and real data sets. The real data sets concern environmental data, see [7]; as for as simulated data are concerned, we considered mixtures of three regression lines ($d = 1$) and mixtures of two bivariave regressions ($d = 2$). Two examples of datasets are shown in Figure 1.

Data have been fitted first by a Student model (6) with linear relationships; afterwards each unit (x_n, y_n) ($n = 1, \dots, N$) satisfying the relation (10) with $\alpha = 0.005$, at least for either x_n or y_n , has been marked as an outlier. Then we run the Gaussian CWM on the reduced data set $\mathcal{D}' = \mathcal{D} \setminus \mathcal{O}$, where $\mathcal{O}$ denotes the set containing the N_o units marked as outliers (thus $\mathcal{D}'$ contains $N' = N - N_o$ units), in order to estimate

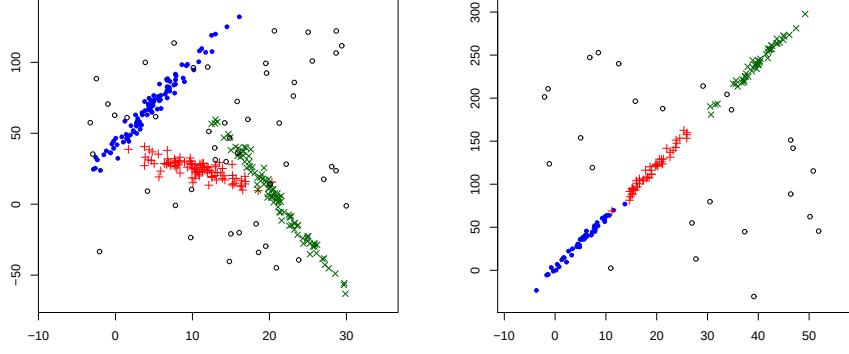


Fig. 1 Two datasets (circles represent noise)

the parameters of the model. Finally we performed the classification of the whole dataset $\mathcal{D}$ (containing N units) according to the obtained model. A similar strategy has been first proposed in [6]. The procedure can be summarized as follows:

1. Run the Student CWM on the data set $\mathcal{D}$;
2. Remove from the dataset the N_o units of $\mathcal{O}$ marked as outlier;
3. Run the Gaussian CWM on the reduced dataset $\mathcal{D}' = \mathcal{D} \setminus \mathcal{O}$;
4. Classify the N units of $\mathcal{D}$ according to the model computed at the previous step.

In the following, the above procedure will be referred to as the *Student-Gaussian CWM* (tG -CWM). In this procedure the step 3 can be also replaced with

- 3'. Run the Student CWM on the reduced dataset containing $N' = N - N_o$ units;

and this will be referred to as the *Student-Student CWM* (tt -CWM).

In general, neither tG -CWM nor tt -CWM outperforms the other one in terms of the misclassification error. Thus the best model has been selected according to smallest value of the mean squared error

$$\mathcal{E} = \left(\frac{1}{N} \sum_{n=1}^N \left[y_n - \left(\sum_{g=1}^G \mu(\mathbf{x}_n; \beta_g) p(\Omega_g | \mathbf{x}_n, y_n) \right) \right]^2 \right)^{1/2}. \quad (11)$$

We remark that, in general, this criterion leads to the model having the smallest mean squared error $\mathcal{E}$.

5 Concluding remarks

We presented a statistical analysis of Cluster-Weighted Modeling (CWM) based on Student- t distributions. In the robust fitting of noisy data, we proposed a double step procedure for removing noise and then estimate the parameters of the model. In particular the first step of the procedure is performed according to a Student based CWM while the second step can be performed using either a Gaussian model (t G-CWM) or a Student model (tt -CWM). Then we selected the model with the smallest mean squared error $\mathcal{E}$.

An important issue concerns some computational aspects. Model parameters have been estimated according to the maximum likelihood approach by means of the EM algorithm. However the simulations we carried out confirmed the conclusion of [3] in the area of mixture of regression, in particular the initialization of the algorithms is crucial. In our simulations the initial guess has been chosen according to a preliminary clustering of data using a k -means algorithm.

References

1. Dayton, C.M., Macready, G.B. (1988). Concomitant-Variable Latent-Class Models, *Journal of the American Statistical Association*, **83**, 173-178.
2. Engster, D., Parlitz, U. (2006). Local and Cluster Weighted Modeling for Time Series Prediction. In: Schelter, B., Winterhalder, M., Timmer, J. (Eds.), *Handbook of Time Series Analysis. Recent theoretical developments and applications*. Wiley, Weinheim, 39-65.
3. Faria, S., Soromenho, G. (2010). Fitting mixtures of linear regressions, *Journal of Statistical Computation and Simulation*, **80**, 201225.
4. Frühwirth-Schnatter, S. (2005). *Finite Mixture and Markov Switching Models*. Springer, Heidelberg.
5. Gershenfeld, N., Schoner, B., Metois, E. (1999). Cluster-weighted modelling for time-series analysis. *Nature*, **397**, 329-332.
6. Greselin, F., Ingrassia, S. (2010). Constrained monotone EM algorithms of multivariate t distributions, *Statistics & Computing*, **20**, 9-22.
7. Hurn, M., Justel, A. and Robert C.P. (2003), Estimating Mixtures of Regressions *Journal of Computational and Graphical Statistics*, **12**, 55-79.
8. Jordan, M.I. (1995), Why the logistic function? A tutorial discussion on probabilities and neural networks, MIT Computational Cognitive Science Report 9503.
9. Jordan, M.I. and Jacobs, R.A. (1994), Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**, 181-224.
10. Kotz, S., Nadarajah, S. (2004), *Multivariate t Distributions and Their Applications*, Cambridge University Press, New York.
11. Lange, K.L., Little, R.J.A., Taylor, J.M.G. (1989). Robust Statistical Modeling Using the t Distribution, *Journal of the American Statistical Society*, **84**, n. 408, 881-896.
12. Liu, C., Rubin D.M. (1995). ML Estimation of the t Distribution using EM and its Extensions, ECM and ECME, *Statistica Sinica*, **5**, 19-39.
13. Peel, D., McLachlan, G.J. (2000). Robust mixture modelling using the t distribution, *Statistics & Computing*, **10**, 339-348.