# Regularized discrete probability distribution with dependence tree

A. Mkhadri, S. Bochi

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Regularized Discrete Probability Distribution with Dependence Tree*

Abdallah MKHADRI
Sami BOCHI

**N° 2210**
Mars 1994

PROGRAMME 5

Traitement du signal,
automatique et
productique

# Regularized discrete probability distribution with dependence tree

# Régularisation en discrimination qualitative par les modèles graphiques décomposables

Abdallah Mkhadri(*) and Sami Bochi(**)

(*) Département de Maths, FSS, B.P.:S15, Marrakech, Maroc

(**) Inria - Rocquencourt, B. P. 105, 78153 Le Chesnay, France

**Abstract** - In the context of discrete discriminant analysis, Chow and Liu introduced the notion of dependenc *tree* to approximate a *n*th order discrete probability distribution. More recently, Wong and Wang proposed a different product approximation. These two procedures have some resemblance with the problem of the choice between the classical linear and quadratic discriminant analysis. We propose an alternative discrete regularized method which is intermediate between the conditional independence model, the Chow and Liu's method and the Wong and Wang's method. Our method is characterized by two regularization parameters. The choice of the optimal regularization parameters can be computed explicitly by minimizing the cross-validated misclassification risk. The method is illustrated through application to real and simulated data.

**Key-words**: *classification, entropy, dependence tree, regularization.*

**Résumé :** Cet article traite de la discrimination par les modèles graphiques décomposables sur variables qualitatives. Ces modèles fournissent une estimation la plus proche de la densité de probabilité conditionnelle en tenant compte de certaines relations de dépendance conditionnelle, définies par un graphe, entre les variables explicatives. Nous proposons une méthode de régularisation, concernant les petits échantillons, à la place du choix entre un modèle graphique différent pour chaque groupe et un seul modèle graphique identique pour tous les groupes. Notre méthode est un compromis entre trois méthodes : ces deux modèles et le modèle d'indépendance conditionnelle. Elle utilise deux paramètres de régularisation qui sont déterminés par minimisation du taux d'erreur évalué par validation croisée. Deux applications sur données réelles et simulées sont présentées pour illustrer ses qualités par rapport à d'autres méthodes classiques.

**Mots-clés** : *discrimination, entropie, graphe de dépendance, régularisation.*

# 1 Introduction

The design of intelligent information system such as pattern recognition, inductive learning and expert systems is concerned with the problem of discrete *classification*. In many of these applications, the central task is to estimate the underlying $n$-dimensional discrete probability distributions from a finite number of samples. Because of the curse of dimensionality, the probability function is often approximated by some simplifying assumptions, such as statistical independence. This hypothesis is simple but may be unrealistic in certain applications. Lewis [1] and Brown [2] are the first who considered the problem of approximating an $n$th-order binary probability distribution by a product of its component distributions of lower order. Lewis [1] showed, under suitably restricted conditions, that the optimal product approximation can be obtained by minimizing a divergence measure between the true and approximate distribution. However, the simple and practical solution to the problem of selecting a set of component distributions, of given complexity to compose the best approximation, was proposed by Chow & Liu [3]. They introduced the notion of *tree* dependence, called 1-*tree* method, to approximate a $n$th order discrete probability distribution by a product of a $n$ - 1 second-order component distributions. One can then reduce the problem to find a dependence *tree* with maximum total branch weight of mutual information. Wong & Wang [4] suggested another product approximation, called 2-*tree* method, by minimizing an upper bound of the Bayes error rate. Wong & Poon [5] showed that the later procedure is a special case of a such minimization procedure of Chow & Liu. The important point is that 1-*tree* procedure use one tree structure for *each* individual class, while 2-*tree* procedure is obtained by using one tree structure for *all* classes. Based on simulation study, Wong & Poon [5] conclued that *1-tree* method is more restricted than the *2-tree* method. In fact, the *2-tree* method has the advantage of being computationally more efficient, especially when the number of features is very large. However, if accuracy is the predominant factor in a particular application, *1-tree* method is prefered. So, there is some need to a compromise between these methods. This problem is similar to the choice between Linear and Quadratic discrimminant analysis (denoted LDA and QDA hereafter) for continuous data, in small sample high-dimensional setting. Friedman [6] has proposed an alternative method, called Regularized Discriminant Analysis (RDA hereafter). RDA has a median position between LDA and QDA. On the other hand, Celeux & Mkhadri [7] proposed an alternative method for addressing the problem of discrete discriminant analysis in small sample setting. Celeux & Mkhadri's method is intermediate between: the full multinomial model (FMM), the conditional independence model and the kernel discrimant analysis of Aitchison & Aitken [8]. The main aim of this method is to regularize the FMM.

The objectif of this note is to suggest an alternative approximation of discrete probability distribution which is intermediate between conditional independence model, 1-*tree* method and 2-*tree* method. Our Regularized approximating discrete probability distributions, denoted RADP hereafter, is characterized by two parameters. The selection of the optimal values of the parameters is based on the cross-validated misclassification risk. We show that these optimal parameter values can be computed explicitly.
In section 2, the approximating discrete probability distributions are sketched in the context of discriminant analysis. Section 3 contains a detailed description of the 1-*tree* method and 2-*tree* method. In section 4, we detail our method of regularization along with the discussion of the method of the choice of regularization parameters. In section 5, the performance of four approaches are invistigated through applications to real and simulated data.

## 2 Approximating discrete probability distributions

In this section, we present three approximations of discrete probability distributions. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ denotes a $n$-dimensional random vector. The component $X_j$ of $\mathbf{X}$ represents the $j$th discrete-valued feature. Let $\mathbf{W}$ be a random variable whose values are used to label the classes. We denote by P(x;w) the joint discrete probability distribution for $\mathbf{X} = \mathbf{x} = (x_1, x_2, ..., x_n)$ and $\mathbf{W} = w$, where $\mathbf{x}$ is a value of the ramdom vector $\mathbf{X}$. Let $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N$ denotes the training sample of size $N$ and suppose that several classes $G_1, G_2, ..., G_K$ occur with prior probability $\delta_1, \delta_2, ..., \delta_K$ ($\Sigma_j \delta_j = 1$). The Bayes classification rule classifies an individual vector $\mathbf{x}$ into $G_k$ if

$$\delta_k P(\mathbf{x}; k) \geq \delta_h P(\mathbf{x}; h) \tag{1}$$

for h = 1,..., K, $h \neq k$.
For discrete data, the most natural model is to assume that the conditional probability P($\mathbf{x}$,$k$), where $\mathbf{x}$ is the $n$-dimensional vector of discrete components and k = 1,..., K, are multinomial probabilities. This model involves $2^n - 1$ parameters in each class for binary data. Hence, in practice and even for moderate $n$, not all of the parameters are identifiable. One way to deal with this high-dimensional problem consists in reducing the number of parametres needed to be estimated. The conditional independence model (CIM) assumes that the $n$ variables are independent in each class $k$, $k = 1,..., K$. Then, the probability distribution can be written as

$$P_I(\mathbf{x}; k) = \prod_{j=1}^{n} P(x_j; k) \tag{2}$$

The estimated probability distribution, by the maximum likelihood, is

$$\hat{P}_I(\mathbf{x}; k) = \prod_{j=1}^{n} N(x_j; k)/N_k \tag{3}$$

where $N(x_j; k) = \#\{\mathbf{y} \in G_k \mid \mathbf{y}_j = \mathbf{x}_j\}$ and $N_k = \#G_k$.
It follows that the number of parameters to be estimated for each class is reduced from $2^n - 1$ to $n$. While this method of conditional independence is simple but may be unrealistic in certain applications. An alternative method based on the product approximation of conditional dependence was suggested by Chow & Liu [3]. These autors introduced the notion of *tree* dependence to approximate a $n$th-order discrete probability distribution by a product of ($n$ - 1) second-order component distributions. The probabilities that are permissible as approximations can be written as

$$P(\mathbf{x}; k) = \prod_{j}^{n} P(x_{m_j}; k \mid x_{m_{i(j)}}) \tag{4}$$

where $0 \leq i(j) \leq j, (m_1, m_2, ..., m_K)$ is an unknown permutation of the integers 1, 2, ..., $n$, $P(x_{m_j}; k \mid x_{m_{i(j)}})$ is the joint probability of $x_{m_j}$ and k condional on the variable $x_{m_{i(j)}}$, and $P(x_j \mid x_0, k)$ by definition equal to $P(x_j; k)$. The estimates of $P(x_{m_j}; k \mid x_{m_{i(j)}})$ are based on the classical maximum likelihood (cf. [3]).
Wong & Liu [9] proposed another decision-derived approach in which the distribution estimation adopted is modified from the dependence tree. The modified dependence tree estimate is

$$P_M(\mathbf{x}; k) = \prod_{j=1}^{n'} P(x_{m_j}; k \mid x_{m_{i'(j)}}) \prod_{j=n'+1}^{n} P(x_{m_j}; k)$$

where for each $j$ ($j = 1, \ldots, n'$), and $(x_{m_j}, x_{m'_i(j)})$ satisfies the significance test. The associated probability distribution employs $n$' second-order marginals and $(n - n')$ first-order marginals where $n' \leq n$. So, this modified dependence tree approximation estimates fewer numbers of parameter than 1-tree method. Hence, this modified procedure with statistically insignificant dependence branches excluded yields better results in certain cases where only a limited number of samples are available (cf. [9]).
For notation convenience, we will drop the subscript $m$ and denote for example, $x_{m_j}$ by $x_j$ in subsequent discussions.

## 3 Tree dependence approximation

Lewis [1] and Brown [2] are the first who considered the problem of approximation of the $n$th-order binary distribution by a product of several of its components dis-

tributions of lower order. They showed, under suitably conditions, that the product approximation has the property of minimum information. However, Chow & Liu [3] are the first who have developped a method to best approximate a $n$th-order distribution by a product of $n$ - 1 second-order component distributions. This method is based on the minimization of the Kullback-Leibler distance between the true distribution P and it approximation $\hat{P}$. For instance, let the $n$th-order probability distribution $P(x_1, ..., x_n; k)$, $x_j$ being discrete, we wish to find a distribution of *tree* dependence $P_\tau(x_1, ..., x_n; k)$, such that $I_k(P, P_\tau) \leq I_k(P, P_t)$ for all $t \in T_n$, where $T_n$ is a set of all possible first-order dependence tree and

$$I_k(P, P_t) = \sum_{\mathbf{x}} P(\mathbf{x}; k) \ln\{P(\mathbf{x}; k)/P_t(\mathbf{x}; k)\}. \tag{5}$$

The solution $\tau$ is called, by Chow & Liu, the optimal first-order dependence tree. Since they are $n^{n-2}$ trees with $n$ vertices, the number of dependence trees in $T_n$ for any moderate value $n$ is so large as to exlude any approach of exhaustive search. So, Chow & Liu [3] showed that: the probability distribution of *tree* dependence $P_t(\mathbf{x})$ is an optimum approximation to $P_\mathbf{x}$ if and only if its dependence tree t has maximum weight. Indeed, $I_k(P, P_t)$ can written as

$$I_k(P, P_t) = -\sum_{j}^{n} I_k(X_j, X_{i(j)}) + \sum_{j}^{n} H_k(X_j) - H_k(\mathbf{X}) \tag{6}$$

where $H_k(\mathbf{X}) = \sum_\mathbf{x} P(\mathbf{x}; k) \ln P(\mathbf{x}; k)$, $H_k(X_j) = \sum_j^n P(x_j; k) \ln P(x_j; k)$ and

$$I_k(X_j, X_i) = \sum_{x_i, x_j} P(x_i, x_j; k) \ln \frac{P(x_i, x_j; k)}{P(x_i; k)P(x_j; k)},$$

is the mutual information between two variables $x_i$ and $x_j$ for $k$ = 1,..., $K$. Since $H_k(\mathbf{X})$ and $H_k(X_j)$, for all $j$, are independent of the dependence *tree* and $I_k(P, P_t)$ is non-negative, then minimizing the closeness measure $I_k(P, P_t)$ is equivalent to maximizing the total branch weight

$$\sum_{j}^{n} I_k(X_j, X_{i(j)}).$$

To relay the dependence *tree* selection criterion to the Bayes error rate, Wong & Wang [4] suggested an other product approximation by minimizing an upper bound of the Bayes error rate. In effect, let $\sigma_e$ denotes the Bayes error rate. It was proved by Hellman & Raviv [10] that

$$\sigma_e \leq \frac{1}{2} H(k \mid \mathbf{X})$$

where the entropy function $H(k \mid \mathbf{X})$ is defined by

$$H(k \mid \mathbf{X}) = -\sum_{\mathbf{x}} P(\mathbf{x}) \sum_{k} P(k \mid \mathbf{x}) \ln P(k \mid \mathbf{x}) \tag{7}$$

where $k = 1,..., K$. Let $\hat{H}(k \mid \mathbf{X})$ be the estimator of $H(k \mid \mathbf{X})$ in which $P(k \mid \mathbf{x})$ is replaced by the equation (4). Then, Wong & Poon [5] showed that if $H_k(\mathbf{X})$ is independent of the dependence tree chosen for *each* individual class, by minimizing $\hat{H}(k, \mathbf{X})$ it follows that

$$min\hat{H}(k \mid \mathbf{X}) = max \sum_{k} \sum_{j}^{n} I_k(X_j, X_{i(j)}), \tag{8}$$

which is the result obtained by Chow & Liu. Kruskal's algorithm [11] can be easily applied to finding a tree with maximum branch weight

$$B_k = \sum_{j=1}^{n} I_k(X_j, X_{i(j)})$$

for each individual class, where $P(x_i; k)$ and $P(x_i, x_j; k)$ are estimated by the maximum likelihood ([3]). On the other hand, as suggested by Wong & Wang [4], we may assume that the probability distribution for *all* the classes can be approximated by the same dependence tree. In this case, for $0 \leq i(j) \leq j$, the minimization problem becomes

$$min\hat{H}(k, \mathbf{X}) = max \sum_{j}^{n} [\sum_{k} P(k) I_k(X_j, X_{i(j)}) - I(X_j, X_{i(j)})], \tag{9}$$

where $I(X_j, X_{i(j)})$ has the same form as $I_k(X_j, X_{i(j)})$, but the conditional probabilities are estimated from the total sample.

The important point is that Chow & Liu's method uses one tree structure for *each* individual class. In contrast, by adopting the same minimization procedure, the result of Wong & Wang is obtained by using one tree structure for *all* classes. Obviously, Wong & Wang's method has the advantage of being computationally more efficient, especially when the number of features is very large. However, if accuracy is the predominant factor in a particular application, Chow & Liu's method is preferred as showing by the simulation results of Wong & Poon [5]. Thus, this problem is similar to the choice between the Quadratic discriminant analysis (different covariance matrix for *each* individual class) and the Linear discriminant analysis (the same covariance matrix for *all* classes) for Gaussian distributioins. So, there is some need to a compromise between these methods of approximation of discrete probability distributions.

# 4 Regularization and Shrinkage

Friedman [6] proposed a regularized discriminant analysis (RDA) conceived in a Gaussian framework. RDA has a median position between LDA and QDA. In the same line, Celeux & Mkhadri [7] proposed an alternative discrete version of RDA, called DRDA, which has an intermediate position between the FMM, the CIM and the kernel discrimination of Aitchison & Aitken [8]. The performance of this method with an other bayesian procedure is discussed in Mkhadri [11]. RDA and DRDA are controled by two regularization parameters which are selected by the minimization of the cross-validated misclassification risk. Alternative regularized approximation of discrete probability distributions (denoted RADP hereafter) are proposed in the following section.

## 4.1 Regularization scheme

Recall that $\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^N$ denotes the discrete training sample of size $N$. Let $\hat{P}_C(\mathbf{x}; k)$ (resp. $\hat{P}_W(\mathbf{x}; k)$) denotes the approximation probability distribution based on the 1-*tree* dependence of Chow & Liu (resp. on the 2-*tree* dependence of Wong & Wang). As mentioned above, Chow & Liu's method uses one tree structure for *each* class. Hence, this method is clearly ill-posed if $N_k = \#G_k \leq n$ for any class $k$, and poorly ill-posed whenever $N_k$ is not considerably large than $n$. One method of regularization is to use one tree structure for *all* classes as suggested by Wong & Wang. This applies a considerable degree of regularization by substancially reducing the number of parameters to be estimated. The choice between 1-*tree* method and 2-*tree* method represents a fairly predective set of regularization alternatives. A less limited set of alternatives is represented by

$$\hat{P}_\alpha(\mathbf{x}, k) = (1 - \alpha)\hat{P}_C(\mathbf{x}; k) + \alpha\hat{P}_W(\mathbf{x}; k) \tag{10}$$

where $\hat{P}_\alpha(\mathbf{x}; k)$ denotes RADP estimates of the group conditional probability for any discrete vector $\mathbf{x}$, with $\alpha$ $(0 \leq \alpha \leq 1)$ denoting the regularization parameter. It controls the degree of shrinkage of the individual class probability distribution estimates based on 1-*tree* method toward the pooled estimate (i.e. estimates based on 2-*tree* method). So, the value $\alpha = 0$ gives rise to 1-*tree* method, whereas $\alpha = 0$ yields 2-*tree* method. Values between these limits represent degree of regularization less severe than 2-*tree* method. But, this regularization is still fairly limited and is not the only natural way to regularize the 1-*tree* method. For instance, if the total sample size $N$ is less than or comparable to $n$, then even 2-*tree* method is ill- or poorly-posed. To this end we further regularize the RADP estimates of the group

conditional probability defined by Equation (10) through

$$\hat{P}_{\alpha,\gamma}(\mathbf{x};k) = (1-\gamma)\hat{P}_{\alpha}(\mathbf{x};k) + \gamma\hat{P}_{I}(\mathbf{x};k) \tag{11}$$

where $\hat{P}_{\alpha}(\mathbf{x};k)$ is given by Equation (10) and $\hat{P}_{I}(\mathbf{x};k)$ is the estimates of CIM given by Equation (3). For a given value of $\alpha$, the additional regularization parameter $\gamma$ ($0 \leq \gamma \leq 1$) controls shrinkage toward a conditional independence model (CIM). Some experiments show that CIM performs well, for small or moderate sample size, relative to other classical methods of discrete discriminant analysis ([13]).
Now, our RADP is defined by Equations (10) and (11) using two regularization parameters, ($0 \leq \alpha \leq 1$) and ($0 \leq \gamma \leq 1$). The three corners defining the extremes of the $\alpha, \gamma$ plane represent fairly well-known classification procedures. The 1-*tree* method corresponds to the case $\alpha = 0$ and $\gamma = 0$. The 2-*tree* method corresponds to the case $\alpha = 1$ and $\gamma = 0$. The CIM corresponds to the case $\gamma = 1$. Holding $\gamma$ fixed at 0 and varying $\alpha$ produces methods between 1-*tree* method and 2-*tree* method. Holding $\alpha$ fixed at 0 (resp. at 1) produces methods between 1-*tree* method (resp. 2-*tree* method) and CIM.

## 4.2 Model selection

Now, the problem is to select the best values of $\alpha$ and $\gamma$. As in [6], we can choose a grid of points on the $(\alpha, \gamma)$ -plane $(0 \leq \alpha \leq 1), (0 \leq \gamma \leq 1)$, evaluate the cross-validated estimate of future misclassification risk at each prescribed point on the grid, and then choose the point with the smallest estimated risk as its estimates for the optimal regularization parameter values $\hat{\alpha}$ and $\hat{\gamma}$. While this approach has the advantage of selecting regularization parameters on the basis of the actual misclassification rate, it can partially ignore information from a substantial portion of the data ([14]). However, by using the method of Celeux & Mkhadri [7], it is possible to select the regularization parameters explicitly and in a nearly optimal fashion. Holding $\gamma$ fixed, we can find in closed form the *complexity parameter* $\alpha^*$ which minimizes the cross-validated misclassification risk; holding $\alpha^*$ fixed, we can find the *shrinkage parameter* $\gamma^*$ which minimizes the cross-validated misclassification risk. We opted for this strategy which is saving a substantial amount of computation. We restricted our attention to two groups case (K = 2) and for the general case ([7]). This strategy is defined as follows:
**Step 1**: $\gamma$ is fixed and is assumed to be 0. Since RADP is essentially a variation arownd 1-*tree* and 2-*tree* dependence methods, it is natural to choose $\gamma = 0$ when deriving the optimal *complexity parameter* $\alpha^*$. Then it is easy to show that:

**Proposition 1**: The optimal *complexity parameter* which minimizes the cross-validated misclassification rule is either 0, 1 or takes the form

$$\frac{B_C(\mathbf{x}^i)}{B_C(\mathbf{x}^i) - B_W(\mathbf{x}^i)} \tag{12}$$

for each $\mathbf{x}^i \in \{\mathbf{x}^1, ..., \mathbf{x}^N\}$, where

$$B_C(\mathbf{x}^i) = \delta_1 \hat{P}_C^{(i)}(\mathbf{x}^i; 1) - \delta_2 \hat{P}_C^{(i)}(\mathbf{x}^i; 2),$$

$$B_W(\mathbf{x}^i) = \delta_1 \hat{P}_W^{(i)}(\mathbf{x}^i; 1) - \delta_2 \hat{P}_W^{(i)}(\mathbf{x}^i; 2),$$

and $\hat{P}_C^{(i)}(\mathbf{x}^i; j)$ (resp. $\hat{P}_W^{(i)}(\mathbf{x}^i; j)$) denotes the estimate conditional probability $\hat{P}_C(\mathbf{x}^i; j)$ (resp. $\hat{P}_W(\mathbf{x}^i; j)$, j = 1, 2) where $\boldsymbol{x}^i$ is removing from the sample.

**Proof**: It is similar to the proof of proposition 1 in [7]. $\Delta$
*Remark*: Recall that the cross-validated classification rule for any $\mathbf{x}^i$ $(1 \leq i \leq N)$ is: $\mathbf{x}^i$ is assigned to class 1 if and only if $C(\mathbf{x}^i,\alpha) \geq 0$, where

$$C(\mathbf{x}^i, \alpha) = (1 - \alpha)B_C(\mathbf{x}^i) + \alpha B_W(\mathbf{x}^i)$$

So, in practical situations, the number of sample points $\mathbf{x}^i$ $(1 \leq i \leq N)$ for which the linear equation $C(\mathbf{x}^i, \alpha) = 0$ has a solution $\alpha$ in (0,1) is very small. This number represents the number of points for which both models ($\hat{P}_C$ and $\hat{P}_W$) provide different assignments.

**Step 2**: Now, holding fixed the optimal *complexity parameter* $\alpha^*$, we proposed to choose the *shrinkage parameter* $\gamma^*$ which minimizes the cross-validated misclassification risk. In the same manner as above, it is easy to show that:

**Proposition 2**: the optimal *shrinkage parameter* $\gamma^*$ which minimizes the cross-validated misclassification rule is either 0, 1 or takes the form

$$\frac{B_{\alpha^*}(\mathbf{x}^i)}{B_{\alpha^*}(\mathbf{x}^i) - B_I(\mathbf{x}^i)} \tag{13}$$

for each $\mathbf{x}^i \in \{\mathbf{x}^1, ..., \mathbf{x}^N\}$, where

$$B_I(\mathbf{x}^i) = \delta_1 \hat{P}_I^{(i)}(\mathbf{x}^i; 1) - \delta_2 \hat{P}_I^{(i)}(\mathbf{x}^i; 2),$$

$$B_{\alpha^*}(\mathbf{x}^i) = \delta_1 \hat{P}_{\alpha^*}^{(i)}(\mathbf{x}^i; 1) - \delta_2 \hat{P}_{\alpha^*}^{(i)}(\mathbf{x}^i; 2),$$

and $\hat{P}_{\alpha^*}^{(i)}(\mathbf{x}^i;k)$ (resp. $\hat{P}_I^{(i)}(\mathbf{x}^i;k)$) denotes the estimate conditional probability $\hat{P}_{\alpha^*}(\mathbf{x}^i;k)$, defined by equation (11) (resp. $\hat{P}_I(\mathbf{x}^i;k)$ defined by equation (3)), where $\mathbf{x}^i$ is removing from the sample.

*Remark*: Now, the general case (K $\geq$ 3) can be takled, for the regularization parameters $\alpha$ and $\gamma$, in the same way as in [7]. It is worth noting that, for each $\mathbf{x}^i$ $(1 \leq i \leq N)$, there are, generally, at most two possible optimal values for each parameter.

### 4.3 Alernative choices for the regularization parameters

Our strategy of selecting the regularization parameters is not optimal as shown in example 1. An alternative strategy is to choose a grid of points on the $(\alpha,\gamma)$-plane $(0 \leq \alpha \leq 1, 0 \leq \gamma \leq 1)$ as Friedman did [6]. But this solution is expensive. Other more intersting approaches consist in using our strategy, described in Section 4, permeting the role of $\alpha$ and $\gamma$. More precisely, holding $\alpha = 0$, we derive the optimal shrinkage parameter $\gamma_0^*$. In this case, the formula (12) becomes, for each $\mathbf{x}^i \in \{\mathbf{x}^1,\ldots,\mathbf{x}^N\}$,

$$\frac{B_C(\mathbf{x}^i)}{B_C(\mathbf{x}^i) - B_I(\mathbf{x}^i)} \tag{14}$$

where $B_I(\mathbf{x}^i) = \delta_1 \hat{P}_I^{(i)}(\mathbf{x}^i;1) - \delta_2 \hat{P}_I^{(i)}(\mathbf{x}^i;2)$. Then holding $\gamma = \gamma_0^*$, we derive the optimal complexity parameter $\alpha^*$, which in this case takes the form

$$\frac{B_C^*(\mathbf{x}^i) + \overline{B}_I(\mathbf{x}^i)}{B_C^*(\mathbf{x}^i) - B_W^*(\mathbf{x}^i)} \tag{15}$$

where

$$\begin{aligned}
B_C^*(\mathbf{x}^i) &= (1-\gamma^*)\delta_1 \hat{P}_C^{(i)}(\mathbf{x}^i;1) - (1-\gamma^*)\delta_2 \hat{P}_C^{(i)}(\mathbf{x}^i;2),\\
B_W^*(\mathbf{x}^i) &= (1-\gamma^*)\delta_1 \hat{P}_W^{(i)}(\mathbf{x}^i;1) - (1-\gamma^*)\delta_2 \hat{P}_W^{(i)}(\mathbf{x}^i;2),\\
\overline{B}_I(\mathbf{x}^i) &= \gamma^*\delta_1 \hat{P}_I^{(i)}(\mathbf{x}^i;1) - \gamma^*\delta_2 \hat{P}_I^{(i)}(\mathbf{x}^i;2).
\end{aligned}$$

RAPD1 denotes this modified version of RAPD.
An other modified version of RAPD, which we call RAPD2, is to hold $\alpha = 1$ and we derive the shrinkage parameter $\gamma^*$. It is similar to RAPD1 in which the role of $P_C$ and $P_W$ are permited. In this case, the shrinkage parameter $\gamma_1^*$ is obtained by the equation (14) in which we replace $B_C(.)$ by $B_W(.)$. Holding $\gamma = \gamma_1^*$, we derive

the optimale complexite parameter $\alpha^*$ from the equation (15) in which $B_W^*$ takes the role of $B_C^*$.
Our examples illustrate that these modified versions of RAPD can perform better than RAPD in certain situations.

# 5 Exprimental results

In the following section, the performance of the four procedures (CIM, 1-tree, 2-tree and RAPD) is examined through application to real and simulated data sets.

## 5.1 Example 1

The data consists of 241 patients suffering from arthose disease. The whole sample was divided into two groups. The first group contained patients for which an aggravation of disease has been discovered from a radiology examination and the second contained the other patients. For each patient, the values of 10 binary variables were available ([7] ). For our illustrative experiment, we drew at radom a training sample of 141 patients and the rest constitutes the test sample. Table1 summarizes the results of four classification methods for this data set.

**Table 1:** Misclassification risk and regularization parameter values for arthose data set

| Methods | test | $\alpha$ | $\gamma$ |
|---|---|---|---|
| CIM | 42 | | |
| 1-Tree | 41 | | |
| 2-Tree | 46 | | |
| RAPD | 44 | .85 | .0 |
| RAPD1 | 40 | .5 | .47 |
| RAPD2 | 40 | .0 | .33 |

For each method, we give the misclassification estimated on the test sample. Also shown, are the selected regularizations parameters $(\alpha, \gamma)$ for RAPD rule. The prior probabilities were taken to be equal, $\delta_k = 1/2$ $(k = 1, 2)$, for each group.

According to 1-tree method, the following tree structures were selected.

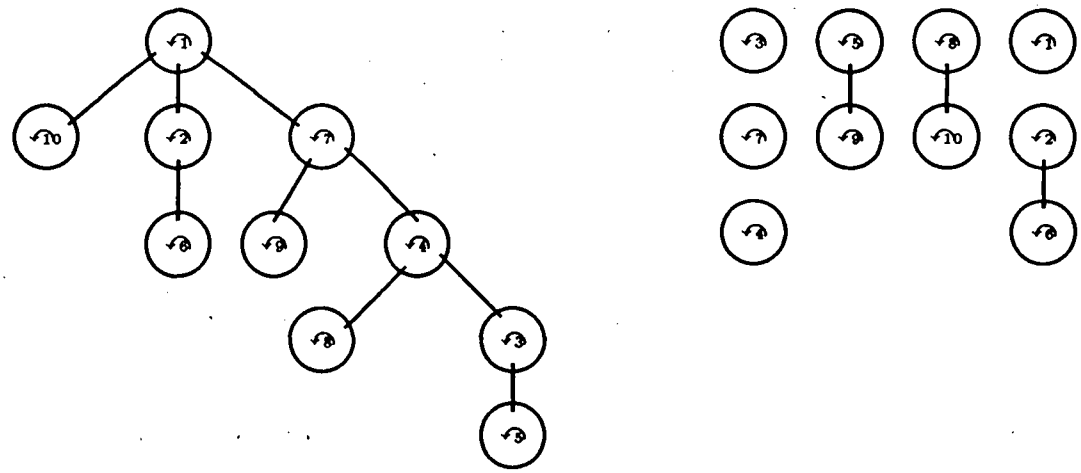

Fig. 1 Dependence trees of group 1 and group 2 for data set arthose

For the same sample, the tree structure for both groups selected by 2-tree method is shown below

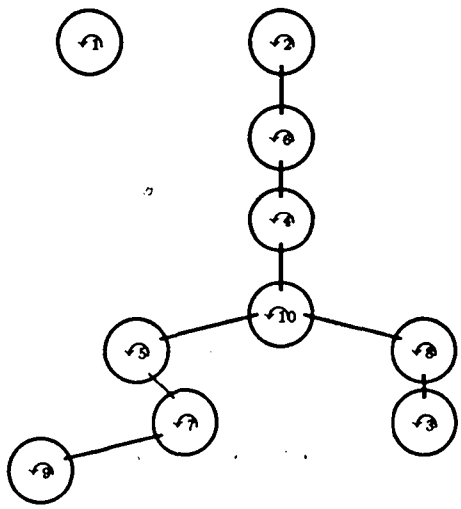

Fig. 2 Dependence tree of all groups for data set arthose

From Table 1 it can be seen that 1-tree performs better than CIM and CIM gives best result than 2-tree. Since RAPD is essentially a variation of 1-tree and 2-

tree, it is natural that it works worse than 1-tree. While the modified versions of RAPD, RAPD1 and RAPD2, provided better misclassification risk because there are essentially a variation of 1-tree (or 2-tree) and CIM. The same result of RAPD1 and RAPD2 was obtained by DRDA in [7].

### 5.2 Example 2

The performance of these methods is evaluated through one additional Monte-Carlo sampling experiment implemented from the Bahadur model as discussed in [7]. The training data set consists of 50 points in $\{0,1\}^6$ randomly generated as described in [7] with different correlation matrices for each group. There are two groups with equal size $n_1 = n_2 = 25$. An additionnal test data of the same size was randomly generated with the population structure and classified with the 6 rules derived from the training data set. The prior probabilities were taken to be equal.
In the following, the tree structures selected by 1-tree and 2-tree methods are shwon below.

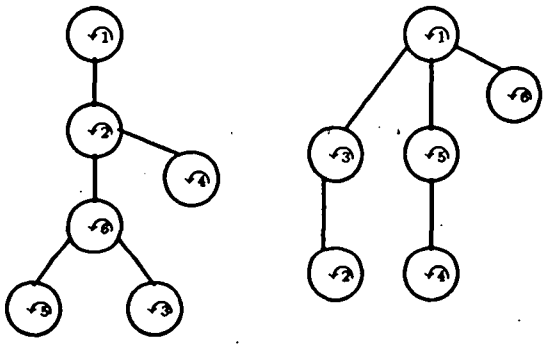

Fig. 3 Trees for group 1 and group 2

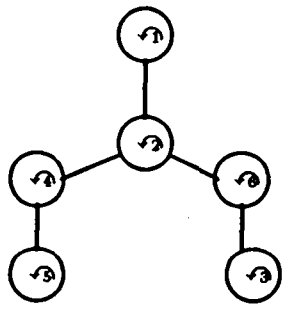

Fig. 4 tree for all groups

**Table 2:** Misclassification risk and regularization parameter values for simulated data set

| Methods | test | $\alpha$ | $\gamma$ |
|---|---|---|---|
| CIM | 32 | | |
| 1-Tree | 20 | | |
| 2-Tree | 28 | | |
| RAPD | 18 | .43 | .0 |
| RAPD1 | 20 | .0 | .0 |
| RAPD2 | 28 | .0 | .99 |

Table 2 summarizes the results of the six methods. Here, RAPD was the best method. As in the example 1, 1-tree method gave the best result than 2-tree method which in turn performs better than CIM. Hence, it is natural that the modified versions of RAPD did not perform better than RAPD in this situation. Also, this example showed that the model selection of the regularization parameters depend on the structure of the training sample. Thus, there is no unique solution to the optimal regularization parametrs. So, if the optimal regularization parameters exist then the choice of one verion of RAPD will be based on the result of the three classical method (1-tree, 2-tree and CIM) as example 1 and 2 showed.

# 6 Conclusion

The numerical experiments showed that good performances can be expected from RAPD. However, unlike RDA in the Gaussian framework [6], we have not yet exhibited situations where RAPD improved substancially on both 1-tree and 2-tree (or CIM). Roughly speaking, in our experiments, RAPD is related to both 1-tree and 2-tree. While the modified versions of RAPD are related to both 1-tree (or 2-tree) and CIM. Nevertheless, if the difference in misclassification risk between two classical methods (1-tree, 2-tree and CIM) is more important, then RAPD could be expected to dominate these three methods. While, if this difference is less important (for example less than 4), then RAPD will not be really useful.
Despite these restrictions, we think that RAPD can be quite beneficial for discrete discriminant analysis in a setting for which sample sizes are small and the groups not well separated.

## References


[1] P. M. Lewis," Approximating probability distributions to reduce storage requirement," *Inform. and Contr.*, **vol. 2**, 1959, pp. 214-225.

[2] D. T. Brown," A note on approximations to discrete probability distributions," *Inform. and Contr.*, **vol. 2**, 1959, pp. 386-392.

[3] C. K. Chow & C. N. Liu," Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, **vol. IT-14**, 1968, pp. 462-467.

[4] A. C. K. Wong & C. C. Wang," Classification of discrete biomedical data with error probability minimax," *In Proc. Seventh Int. Conf. Cybern. Soc.*, Washington, DC, 1979, pp. 19-21.

[5] S. K. M. Wong & F. C. S. Poon," Comments on approximating discrete probability distributions with dependence trees," *IEEE Trans. on Patern Anal. and Mach. Intel.*, **vol 11**, No 3, 1989, pp. 333-335.

[6] J. H. Friedman," Regularized discriminant analysis," *Journal of the Americain Statistical Association* **84**, 1989, pp. 165-175.

[7] G. Celeux & A. Mkhadri," Discrete regularized discriminant analysis," *Statistics & Computing*, **vol 2**, p. 143-151, 1992.

[8] C. G. G. Aitchison & J. Aitken," Multivariate binary discrimination by the kernel method," *Biometrika* **63**, 1976, pp. 413-20.

[9] A. K. C. Wong & T. S. Liu," A decision-directed clustering Algorithm, for discrete data," *IEEE Trans. Comput.*, **vol. C-26**, pp. 75-82, 1977.

[10] M. E. Hellman & J. Raviv," Probabilaty of error, equivocation, and the Chernoff bound," *IEEE Trans. Inform. Theory*, **vol. IT-16**, 1970, pp. 368-372.

[11] J. B. Kruskal," On the shortest spanning subtree of a graph and traveling salesman problem," *Proc. Amer. Math. Soc.*, **vol. 7**, 1956, pp. 48-50.

[12] A. Mkhadri," A comparative study of two methods of regularized discrete discriminant analysis," *COMPSTAT 92*, p. 185-90, Eds. Y. Dodge & J. Whittaker, Springler-Verlag, 1992.

[13] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, J. D. F. Habbema & G. J. Gelpke," Comparative of discrimination techniques applied to a computer data set of head injured patients," *Journal of the Royal Statistical Society* **A 144**, 1981, 145-175.

[14] W. Rayens & T. Greene," Covariance pooling and stabilization for classification," *Comput. Stat. & Data Analysis*, **11**, 1991, pp. 17-42.

[15] D. C. Martin & R. A. Bradley," Probability models, estimation, and classification for multivariate dichotomous populations," *Biometrics* **28**, 1972, pp. 203-222.

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399