On the Optimal Weighting of High-Dimensional Bayesian Networks

Tatjana Pavlenko and Dietrich von Rosen

Research Report
Centre of Biostochastics

Swedish University of
Agricultural Sciences

Report 2003:6
ISSN 1651-8543

On the Optimal Weighting of High-Dimensional Bayesian Networks

Tatjana Pavlenko[1]

Department of Engineering, Physics and Mathematics
Mid Sweden University, SE-85170, Sundsvall, Sweden

Dietrich von Rosen

Centre of Biostochastics
Swedish University of Agricultural Sciences
Box 7013, SE-75007, Uppsala, Sweden

Abstract

For an augmented Bayesian network classifier we propose a method of scoring a set of feature nodes for the separation strength, wherein we have combined a weighting technique and growing dimension asymptotics in a single framework. We show that the distribution of the weighted classifier is asymptotically Gaussian and establish the weight-function which is optimal in a sense of minimum misclassification probability.

Keywords: Bayesian network, augmenting, separation strength, growing dimension asymptotic, weighted classifier, limiting error probability.

AMS 1991 subject classifications: Primary 62H30; Secondary 62F12

[1]E-mail address all correspondence and requests for reprints to: tatjana@fmi.mh.se

1 Introduction

Bayesian network (BN) models have an increasing number of applications in decision analysis and artificial intelligence (Korb & Nicholson, 2003) as well as in statistics (see e.g. Cowell et al., 1999). A BN model $\mathcal{M} = \langle G, \mathcal{F}_G \rangle$ for a set of random variables $\mathbf{x} = \{x_1, \ldots, x_p\}$ is a set of joint probability distributions, specified via two components: a structure G and a set of local distribution families $\mathcal{F}_G$. The structure G for $\mathbf{x}$ is a directed acyclic graph having for every variable x_i in $\mathbf{x}$ a node labeled by x_i with parents labeled by $Pa_i^{\mathcal{M}}$. The structure G represents the set of conditional independence assertions which are implied by a factorization of a joint distribution for $\mathbf{x}$ given by $F(\mathbf{x}) = \prod_{i=1}^{p} F(x_i|Pa_i^{\mathcal{M}})$. The local distributions $F(x_i|Pa_i^{\mathcal{M}})$ are the p conditional and marginal probability distributions that constitute the factorization of $F(\mathbf{x})$. Each such distribution belongs to the specific family of allowable probability distributions $\mathcal{F}_G$.

We assume that $\mathbf{x}$ consists of continuous random variables and each local probability distribution is selected from a family $\mathcal{F}_G$ which depends on a finite set of parameters $\theta \in \Theta$. The parameters for a local distribution are a set of real numbers that completely determine the functional form of $\mathcal{F}_\kappa(x_i|Pa_i^{\mathcal{M}})$, given the network structure. Consequently, the joint probability density for a BN model is given by

$$f(x_1, \ldots, x_p; \theta) = \prod_{i=1}^{p} f(x_i; \theta_i|Pa_i^{\mathcal{M}}),$$

where $\theta_1, \ldots, \theta_p$ are subsets of θ and $f(\mathbf{x}; \theta) = F'(\mathbf{x}; \theta)$. Whereas in a general formulation of BN models, the subsets $\{\theta_i\}_{i=1}^{p}$ could overlap allowing several local distributions to have common parameters, here we shall exclude this possibility (see subsection 2.1).

In the current study, BN models will be considered in the *classification framework* where the outcome of interests, $\mathcal{C}$, falls into ν unordered classes, which for convenience we denote by the set $\{1, 2, \ldots, \nu\}$. The goal is to build a rule for assessing the class membership of an item based on p feature variables $\mathbf{x} \in R^p$, whose joint conditional probability density in each class is represented by a BN model, $\mathcal{M}$, having its own set of parameters, but sharing a common structure. Using Bayes' theorem and flipping the densities into class posterior probabilities $\mathsf{Pr}(\mathcal{C}|\mathbf{x})$ we construct the classification rule

$$\mathcal{C} = j \quad \text{if} \quad \mathsf{Pr}(\mathcal{C} = j|\mathbf{x}) = \max_k \mathsf{Pr}(\mathcal{C} = k|\mathbf{x}), \tag{1.1}$$

where $\mathsf{Pr}(\mathcal{C} = j|\mathbf{x}) \propto \pi_j f(\mathbf{x};\theta^j)$, $\mathsf{Pr}(\mathcal{C} = j) = \pi_j$ are class prior probabilities, $j = 1,\ldots,\nu$ and $\propto$ denotes proportionality. This is in fact the definition of a *general Bayesian network classifier* (BN classifier) commonly found in the literature; see e.g. Cowell et al. (1999). A well known example of BN classifiers is the *naive Bayesian classifier* which is represented by the network structure $\mathcal{G}$ requiring for the set $\{x_1,\ldots,x_p,\mathcal{C}\}$ that the class variable $\mathcal{C}$ is the *only* parent for each node variable x_i, i.e. $Pa_i^{\mathcal{M}} = \mathcal{C}$ for all $i = 1,\ldots,p$, and no other connection is allowed. This implies that the feature variables are independent given the class variable and G induces the following factorization of each class probability density $f(\mathbf{x},\theta^j) = \prod_{i=1}^p f(x_i,\theta_i^j)$. Using Bayes theorem we get the classifier of the form $\mathsf{Pr}(\mathcal{C} = j|\mathbf{x}) \propto \pi_j \prod_{i=1}^p f(x_i;\theta_i^j)$. Despite its naive assumption, a variety of empirical results shows surprisingly that the naive BN classifier outperformed many sophisticated classifiers even in the domains where clear feature dependence exists; see for instance Barash & Friedman (2002). Theoretical analysis is provided by e.g. Friedman (1997) and Zang & Ling (2001).

We examine approaches that maintain the basic structure of the naive BN classifier, however allowing its *augmenting* by adding arcs between feature nodes, when needed, thus dispensing with its strong assumptions about independence. Among these we single out a model of the form $f(\mathbf{x};\theta^j) = \prod_{i=1}^{\kappa} f_i(\mathbf{x}_i;\theta_i^j)$ where $\mathbf{x}_1,\ldots,\mathbf{x}_\kappa$ are pairwise disjoint subsets of feature variables and the correspondent augmented nodes in G are connected via κ fully connected subnetworks, assuming however *no arcs* between the subnetworks.

The focus of this paper is on the technique for evaluating a feature separation properties that uses the *weighting* of the augmented BN model as a means to improving the classification accuracy. We emphasize that the approach described herein is carried out jointly and discriminatively together with the estimation of the specific classifier and is an extension of the results by Pavlenko & von Rosen (2001) for augmented BN models.

Classification performance is analyzed in a high-dimensional framework, i.e. assuming that the size of the training sample is comparable to the number of feature nodes, which can severely hurt a BN classifier. The degradation effect is known as "curse of dimensionality" and an important goal of this study is to evaluate this effect when using a weighted form of the augmented BN model. In order to tackle this problem effectively, we employ a *growing dimension asymptotic* approach (see Girko, 1995), meaning that the relationship between dimensionality, p, and size of the data set for learning the network,

n, satisfies the condition: $\lim_{n\to\infty} \lambda(p,n) < \infty$, where $\lambda(p,n)$ is a positive function increasing along p and decreasing along n. Herein we assume that p and n grow somehow simultaneously so that the asymptotics we are going to exploit can be based on the ratio

$$\lim_{n\to\infty} \frac{p}{n} = c, \tag{1.2}$$

where $0 < c < \infty$ is a certain constant.

The contributions of this paper are as follows: a unified methodology that combines the technique for scoring a set of features for their separation strength and evaluating the high dimensionality effects (Section 2); *weighted* form of the augmented BN classifier and analysis of its performance accuracy using growing dimension asymptotics; a formula that computes the optimal in a sense of minimum misclassification probability type of the weight-function for different *a priori* assumptions about the feature separation strength (Section 3).

2 Passage to the augmented BN model via binary classification

In what follows we restrict ourselves to *binary classification*, the special (but common) case in which $\nu = 2$. Although most of the concepts generalize to the case $\nu \geq 3$, the derivations and underlying intuition are more straightforward for this special "two-class" case. Hastie et al. (2001) suggested the following *pairwise coupling* technique for the multi-class setting: Solve each of the two-class problems, and then for a test observation, combine all the pairwise decisions to form a ν-class decision. Observe that pairwise coupling combination rule is quite intuitive: Assign to the class that wins the most pairwise comparisons. For convenience in what follows, we will make use of the decision boundaries that are expressed in terms of a logarithmic difference between two densities, i.e. the *discriminant score*,

$$\mathcal{D}(\mathbf{x};\theta^1,\theta^2) = \ell(\mathbf{x};\theta^1) - \ell(\mathbf{x};\theta^2),$$

where $\ell(\mathbf{x};\theta^j) := \ln f(\mathbf{x};\theta^j)$. To motivate why this representation of the classifier is attractive, note that discriminant preserves the ordering of the class posterior probabilities leading to the decision rule:

$$\mathcal{C}(\mathbf{x}) = \begin{cases} 1 & \text{whenever} \quad \mathcal{D}(\mathbf{x};\theta^1,\theta^2) > \ln\frac{\pi_2}{\pi_1}, \\ 2 & \text{otherwise.} \end{cases} \tag{2.1}$$

The main advantage of using the discriminative formulation is that the performance accuracy of $\mathcal{D}(\mathbf{x};\theta^1,\theta^2)$ can be measured by *misclassification probabilities* defined as follows:

$$\begin{aligned}\mathcal{E}_1 &= \mathsf{Pr}(\mathcal{D}(\mathbf{x};\theta^1,\theta^2) \leq \ln\frac{\pi_2}{\pi_1}|\mathcal{C}(\mathbf{x})=1),\\ \mathcal{E}_2 &= \mathsf{Pr}(\mathcal{D}(\mathbf{x};\theta^1,\theta^2) > \ln\frac{\pi_2}{\pi_1}|\mathcal{C}(\mathbf{x})=2).\end{aligned} \tag{2.2}$$

These can then form the *Bayes risk*, $\mathcal{R}_{\mathcal{D}(\mathbf{x};\theta^1,\theta^2)} = \pi_1\mathcal{E}_1 + \pi_2\mathcal{E}_2$, which in turn gives a straightforward way of judging the classification accuracy. Note also that in the symmetric case with equal prior probabilities both class-wise error rates are equal, and the minimum attainable Bayes risk is $\mathcal{R}_{\mathcal{D}(\mathbf{x};\theta^1,\theta^2)} = \frac{1}{2}(\mathcal{E}_1 + \mathcal{E}_2)$.

2.1 κ-blocking (augmenting) and estimation procedure

In our approach, the strong independence assumption of the naive BN model is relaxed by merging highly dependent feature variables together to subset of variables (blocks), or equivalently, by connecting the correspondent feature nodes in the network via a fully connected subnetworks, assuming *no arcs* between the subnetworks.

In the present study, we fix the subset size to some constant, m, and require the subsets to be *non-overlapping*, in which case the network structure forms a decomposition of both $\mathbf{x}$ and θ^j into κ pairwise disjoint, independent, m-dimensional subsets, so that $p = \kappa m$ and $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_\kappa)$, $\theta^j = (\theta^j_1,\ldots,\theta^j_\kappa)$ where $\mathbf{x}_i = (x_{i1},\ldots,x_{im})$, $\theta^j_i = (\theta^j_{i1},\ldots,\theta^j_{im})$ $i = 1,\ldots,\kappa$, $j = 1,2$. We call these structures *augmented naive Bayesian networks* (augmented BN) and the subsets *blocks*.

It is important to note that finding and adding the best set of augmenting arcs is generally an intractable problem, since it is equivalent to learning the best BN model among those in which C is the class variable. However, by restricting the network complexity first, it is possible to explore the best partitioning that can approximate the data set. For algorithms discovering the appropriate feature nodes decompositions in high dimensional classification problems see, e.g. Kusiak (2000) and Maimon & Rockach (2001) and references therein.

Augmenting the network followed by the block independence assertions implies that the true joint probability density of $\mathbf{x}$ for each class can be repre-

sented as a product of local interaction models, i.e. $f(\mathbf{x};\theta^j) = \prod_{i=1}^{\kappa} f_i(\mathbf{x}_i;\theta_i^j)$, where the local m-dimensional density $f_i(\mathbf{x}_i;\theta_i^j)$ belongs to a family $\mathcal{F}_{\theta^j}$ which depends on a finite set of parameters $\theta_i^j \in \Theta$, $i = 1,\dots,\kappa$, $j = 1,2$. We assume that the family $\mathcal{F}_{\theta^j}$ satisfies the following regularity conditions: for each $\mathbf{x}_i$, the function $\ell_i(\mathbf{x}_i;\theta_i^j) := \ln f_i(\mathbf{x}_i;\theta_i^j)$ is three times differentiable in the components of θ_i^j and all first-, second- and third- order derivatives with respect to θ_i^j of $\ell(\mathbf{x}_i;\theta_i^j)$ are integrable with respect to $f(\mathbf{x};\theta_j)\,d\mathbf{x}$, $j = 1,2$. Consequently for an augmented BN model the problem of learning the classifier (2.1) reduces to computing the appropriate estimates of the unknown parameter θ^j from the training set of data. To completely specify the learning method in a high-dimensional framework, we define the asymptotic properties of estimates $\hat{\theta}_i^j$ of the ith local model given data $\mathbf{x}_1^j,\dots,\mathbf{x}_n^j$, a random sample from $f(\mathbf{x};\theta^j)$, $j = 1,2$ and assuming the same rate of growing for both sample sizes so that $n_1 = n_2 = n$. We introduce the statistics $T_i^j = n^{1/2}(\hat{\theta}_i^j - \theta_i^j)' I^{1/2}(\theta_i^j)$, which for each i describes the standardized bias of the estimate $\hat{\theta}_i^j$, where

$$I^j = I(\theta^j)_{ik} = \int \frac{\partial \ell(\mathbf{x},\theta^j)}{\partial \theta_i^j} \frac{\partial \ell(\mathbf{x},\theta^j)}{\partial \theta_k^j} f(\mathbf{x},\theta^j)\,d\mathbf{x}$$

is the Fisher information matrix which is positively definite for all $\theta^j \in \Theta^j$. By the network structure, the matrices are of block-diagonal form with blocks $I_i^j = I(\theta_i^j)$ of dimension $m \times m, j = 1,2$. We assume that the estimate $\hat{\theta}_i^j$ is such that for each j uniformly in i:

1. $\lim_{n\to\infty} \max_i |\mathsf{E}[T_i^j]| = 0$, where $\mathsf{E}[\cdot]$ is the expectation operator.
2. All eigenvalues of the matrices $n\mathsf{E}[(\hat{\theta}_i^j - \theta_i^j)(\hat{\theta}_i^j - \theta_i^j)']$ are bounded from above so that

$$\lim_{n\to\infty} \max_i |n\mathsf{E}[(\hat{\theta}_i^j - \theta_i^j)' I(\theta_i^j)(\hat{\theta}_i^\nu - \theta_i^j)] - m| = \lim_{n\to\infty} \max_i |\mathsf{E}[\langle T_i^j, T_i^j\rangle] - m| = 0, \tag{2.3}$$

   where $\langle \bullet, \bullet \rangle$ denotes the scalar product.
3. $\max_i \mathsf{E}[|T_i^j|^3] = \mathcal{O}(\frac{1}{n^{3/2}})$.
4. The asymptotic distribution of T_i^j converges to $\mathcal{N}_m(0, I)$ as n approaches infinity.

These assumptions form the standard set of "good" asymptotic properties, of which the first three reflect unbiasedness, efficiency and boundness of the third absolute moment of $\hat{\theta}_i^j$, uniformly in i as $n \to \infty$, $i = 1,\dots,\kappa$.

Let us now in this framework have a look at the structure of the classifier $\mathcal{D}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2)$. By the network structure the block size is fixed to the constant m, which implies that the total number of blocks κ, must grow together with n according to (1.2) in such a way that

$$\lim_{n\to\infty} \frac{\kappa}{n} = \rho, \quad \text{where} \quad 0 < \rho < \infty \tag{2.4}$$

and $c = m\rho$. This assumption being designed for the special dependence structure among the feature nodes, is just a particular case of (1.2). Further, the classifier induced by augmenting the network is *log additive* in each block, i.e.

$$\mathcal{D}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2) = \sum_{i=1}^{\kappa} \mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2), \tag{2.5}$$

where $\mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2) = \ell_i(\mathbf{x}_i; \hat{\theta}_i^1) - \ell_i(\mathbf{x}_i; \hat{\theta}_i^2)$ and the corresponding classification procedure thus is within the frame of *Generalized Additive Models*; see Hastie et al. (2001). Observe that the naive BN model can be viewed as a particular case of the augmented one: if we assume that $m = 1$, (and so $\kappa = p$), then the resulting classifier $\mathcal{D}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2) = \sum_{i=1}^{p} \mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2)$ is additive in each of the features and corresponds to the usual naive BN.

The main advantage of the additive structure of the augmented classifier is that in the asymptotic framework specified by (2.4), $\mathcal{D}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2)$ can be viewed as a sum of a growing numbers (κ grows together with n) of independent random variables and, under the regularity conditions imposed on the family of local densities $\mathcal{F}_\theta$, we may state the convergence of this sum towards a Gaussian distribution. This methodology has been studied in details in Pavlenko & von Rosen (2001), where the asymptotic distribution of $\mathcal{D}(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2)$ was used to establish the minimum misclassification risk

$$\mathcal{R}_{\mathcal{D}(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)} = \Phi\Big(-\frac{\sqrt{\mathcal{J}}}{2} \frac{1}{\sqrt{1 + \frac{2m\rho}{\mathcal{J}}}} \Big), \tag{2.6}$$

where $\mathcal{J}$ denotes the cross-entropy distance between the classes defined by

$$\mathcal{J} = \int \ln \frac{f^1(\mathbf{x}; \theta^1)}{f^2(\mathbf{x}; \theta^2)} \Big(f^1(\mathbf{x}; \theta^1) - f^2(\mathbf{x}; \theta^2) \Big) d\mathbf{x}. \tag{2.7}$$

2.2 Blocks separation strength

In this section, we propose a distance-based measure by which a separation strength of a feature node, or a block of nodes, can be assessed. Since the

performance accuracy of a BN classifier is measured by the misclassification risk, $\mathcal{R}$, the latter seems to be a most appealing function for this assessment. However, as it is proved by (2.6), the misclassification risk is a monotone decreasing function of the cross-entropy distance between classes, which means that distance-based scoring measure induce over the set of all potential blocks, the same ranking as the one induced by $\mathcal{R}$. Details of this equivalence are given in Pavlenko (2003).

The product form of class densities implies that the cross-entropy distance $\mathcal{J} := \mathcal{J}(\kappa)$ defined by (2.7) is additive and decomposable as $\mathcal{J}(\kappa) = \sum_{i=1}^{\kappa} \mathcal{J}_i$, where

$$\mathcal{J}_i = \int \ln \frac{f_i(\mathbf{x}_i; \theta_i^1)}{f_i(\mathbf{x}_i; \theta_i^2)} (f(\mathbf{x}; \theta^1) - f(\mathbf{x}; \theta^2)) d\mathbf{x} \tag{2.8}$$

is the input of ith block into the distance $\mathcal{J}(\kappa)$. We define the *separation score* of the ith block by the value $\frac{n\mathcal{J}_i}{2}$, which is a normalized input of ith block towards the distance $\mathcal{J}(\kappa)$ and its sample based analogue $\frac{n\hat{\mathcal{J}}_i}{2}$. Normalization by n is to ensure that $0 < \frac{n\mathcal{J}_i}{2} < \infty$ as $n \to \infty$ according to (2.4).

In the asymptoitc framework specified by (2.4) it is worthwhile introducing a distribution function of the block scores as

$$H_\kappa(u) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \mathbf{1}_{\{\frac{n\mathcal{J}_i}{2}, \infty\}}(u),$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of the set A. We suppose also that the convergence $\lim_{\kappa \to \infty} H_\kappa(u) = H(u)$ takes place uniformly in u and $H(u)$ is a known distribution.

By the construction of $\hat{\mathcal{J}}_i$ it is clear that the sample based separation strength is affected by the high dimensionality. To give an impression about this effect we establish the asymptotic distribution of $\frac{n\hat{\mathcal{J}}_i}{2}$ and find the explicit expression of the bias induced by the sample based scoring technique. We start with the following auxiliary

Lemma 2.1 *Let* $\gamma_i^2 = \langle \gamma_i, \gamma_i \rangle = \frac{n}{2}(\theta_i^1 - \theta_i^2)' I(\theta_i)(\theta_i^1 - \theta_i^2)$, *where* $I(\theta_i)$ *is the* i*th block's information matrix, and* $\theta_i = \frac{\theta_i^1 + \theta_i^2}{2}$. *Then the true separation score admits the representation*

$$\frac{n\mathcal{J}_i}{2} = \gamma_i^2 + \mathcal{O}(n^{-1/2}),$$

and the values $\frac{n\hat{\mathcal{J}}_i}{2}$ *are uniformly bounded with respect to* $i = 1, \ldots, \kappa$ *as* $n \to \infty$.

A proof of this lemma is based on standard Taylor-expansion arguments and the regularity conditions imposed on the family $\mathcal{F}_{\theta^j}$. Details of the proof can be found in Pavlenko (2001).

Theorem 2.1 *Let* $g(u; m, \gamma^2)$ *be the probability density of a non-central* χ^2 *distribution* $\mathcal{G}(u; m, \gamma^2)$ *with* m *degrees of freedom and non-centrality parameter* γ^2. *Let also* $\mathcal{H}(u; \gamma_i^2)$ *be a distribution function of* $\frac{n\hat{J}_i(n)}{2}$, *where* γ_i^2 *is specified in Lemma 2.1. Then, uniformly in* i $|\mathcal{G}(u; m, \gamma^2) - \mathcal{H}(u; \gamma_i^2)| \longrightarrow 0$ *as* $n \to \infty$, $i = 1, \ldots, \kappa$.

PROOF: Observe that $\hat{J}_i(n)$ admits the representation

$$\hat{J}_i(n) = (\hat{\theta}_i^1 - \hat{\theta}_i^2)' I(\theta_i)(\hat{\theta}_i^1 - \hat{\theta}_i^2) + \mathcal{O}(n^{-3/2}), \tag{2.9}$$

where $I(\theta_i)$ is the ith block's information matrix, and $\theta_i = \frac{\theta_i^1 + \theta_i^2}{2}$. Furthermore,

$$(\hat{\theta}_i^2 - \hat{\theta}_i^1)' I(\theta_i)(\hat{\theta}_i^2 - \hat{\theta}_i^1) = n^{-1} \langle (\omega_i + T_i^2 - T_i^1), (\omega_i + T_i^2 - T_i^1) \rangle \tag{2.10}$$

where $\omega_i = \sqrt{n}[I^{1/2}(\theta_i)]'(\theta_i^1 - \theta_i^2), i = 1, \ldots, \kappa, \nu = 1, 2$. These results are proved in Pavlenko (2001). The distribution function of the random variable $\frac{T_i^2 - T_i^1}{\sqrt{2}}$ also approaches $\mathcal{N}_m(0, I)$ uniformly with respect to $i = 1, \ldots, \kappa$ since T_i^2 and T_i^1 are independent random vectors, whose distributions are, by assumptions (2.3), asymptotically Gaussian uniformly in i as $n \to \infty$. Therefore the distribution of $\langle \omega_i + T_i^2 - T_i^1, \omega_i + T_i^2 - T_i^1 \rangle$, as well as $\frac{n\hat{J}_i(n)}{2}$, approaches $\mathcal{G}(u; m, \gamma_i^2)$, where $\gamma_i = \omega_i/\sqrt{2}$, $i = 1, \ldots, \kappa$. □

Now, using properties of a χ^2 distribution (see Johnson et al. 1995, p.442), Lemma 2.1 and Theorem 2.1, we can under mild regularity conditions conclude that

$$\mathsf{E}\Big[\frac{n\hat{\mathcal{J}}_i}{2}\Big] = \frac{n\mathcal{J}_i}{2} + m + \mathcal{O}(n^{-3/2}),$$

which shows the effect of estimation: the true separation strength is *overestimated* and the bias term of each block- estimator is of order m, the block size. The accumulation of the bias over the increasing number of blocks in growing dimension asymptotics leads to that bias of the classifier (2.5) is of order $\mathcal{O}(\kappa/n)$ (curse of dimensionality factor). We collect these results in the following proposal:

(i) to utilize the block separation strength, thereby counteracting the equalizing of impacts of low- and highly- relevant blocks inherent in standard augmented BN classifier;

(ii) to account for (and bring down) the effect of a bias induced by high dimensionality when using estimates. This is the case where the results of Theorem 2.1 will be relevant.

3 Weighted BN classifier

We now elaborate the augmented BN classifier by a *weighting procedure* which takes into account the block separation strength. We specify the weight-function of the ith block by $w_i := w(\frac{n\hat{\mathcal{J}}_i}{2})$ where $w_i(u)$ is nonnegative and bounded for $u > 0$ and define the *weighted* BN classifier as

$$\mathcal{D}_w(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2) = \sum_{i=1}^{\kappa} w_i \mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2), \tag{3.1}$$

which provides us with the natural extension of the augmented BN model: each local classifier $\mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2)$ is weighted by the correspondent block separation score $\frac{n\hat{\mathcal{J}}_i}{2}$. One advantage of this approach is immediately clear: Applying such a weighting scheme gives the modified classifier of an additive form and we can, using the methodology proposed in Pavlenko & von Rosen (2001), prove that its distribution is asymptotically Gaussian. This in turn makes it possible to optimize the weight-function in a sense of minimum misclassification risk.

3.1 Asymptotic moments of the weighted BN classifier

To prove that the distribution of $\mathcal{D}_w(\mathbf{x}; \hat{\theta}^1, \hat{\theta}^2)$ is asymptotically Gaussian we need to specify the first three moments of $w_i \mathcal{D}(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2)$ and show that Liapunov conditions (see, e.g. Rao, 1973, p. 127) are applicable to the sum (3.1). To facilitate calculations we turn to integrated local classifier and use the following representations:

$$\begin{aligned} \int \mathcal{D}_i(\mathbf{x}_i; \hat{\theta}_i^1, \hat{\theta}_i^2) f(\mathbf{x}; \theta^1) d\mathbf{x} &= \frac{1}{2n} \langle (\sqrt{2}\gamma_i + T_i^2), (\sqrt{2}\gamma_i + T_i^2) \rangle - \frac{1}{2n} \langle T_i^1, T_i^1 \rangle \\ &\quad + \mathcal{O}(n^{-1/2}), \end{aligned} \tag{3.2}$$

$$\int [\mathcal{D}_i(\mathbf{x}_i;\hat{\theta}_i^1,\hat{\theta}_i^2)]^2 f(\mathbf{x};\theta^1)d\mathbf{x} = \frac{1}{n}\langle\sqrt{2}\gamma_i + T_i^2 - T_i^1, \sqrt{2}\gamma_i + T_i^2 - T_i^1\rangle + \mathcal{O}(n^{-1/2}), \quad (3.3)$$

$$\int [\mathcal{D}_i(\mathbf{x}_i;\hat{\theta}_i^1,\hat{\theta}_i^2)]^3 f(\mathbf{x};\theta^1)d\mathbf{x} = \mathcal{O}(n^{-1/2}), \quad (3.4)$$

where T_i^ν, $\nu = 1, 2$ is defined in subsection 2.1, $i = 1, \ldots, \kappa$. These representations have been obtained in Pavlenko (2003) using asymptotic expansions of ith local densities about θ_i^j. The advantage of considering the integrated $\mathcal{D}_i(\mathbf{x}_i;\hat{\theta}_i^1,\hat{\theta}_i^2)$ is that we can use asymptotic properties of T_i^ν (see (2.3)) when specifying the moments of the weighted BN classifier.

Lemma 3.1 *Under the regularity conditions for the family of the local densities, F_{θ^j}, and assuming that (2.3)-(2.4) hold, the moments of $\boldsymbol{\mathcal{D}}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ have the limits*

$$\mathsf{E}[\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)] \to E(w), \quad \mathsf{Var}[\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)] \to V(w),$$

as $n \to \infty$, where

$$E(w) = \rho\int \gamma^2[\int w(u)g(u;m+2,\gamma^2)du]dH(\gamma^2), \quad (3.5)$$

$$V(w) = 2\rho\int [\int uw^2(u)g(u;m,\gamma^2)du]dH(\gamma^2). \quad (3.6)$$

PROOF: See Appendix.

The results of Lemma 3.1 give us means to specify the asymptotic distribution of the weighted BN classifier. By (3.5)-(3.6) the sum $\sum_{i=1}^{\kappa} w_i\mathcal{D}_i(\mathbf{x}_i,\hat{\theta}_i^1,\hat{\theta}_i^2)$ satisfies the Liapunov conditions and consequently the distribution of the $\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ converges to the normal one as $n \to \infty$.

3.2 Misclassification risk and optimal choice of weight-function

Having established the asymptotic distribution of the weighted BN classifier we are now ready to compute the limiting error probabilities and analyze the classification performance.

Theorem 3.1 *Let $\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ be a weighted BN classifier, where the weighting is governed by the factor $w(\frac{n\hat{\mathcal{J}}_i}{2})$, $i = 1, \ldots, \kappa$. Then the misclassification*

probabilities of $\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)$ *given by*

$$\begin{aligned}\mathcal{E}_1(w) &= \mathsf{Pr}\Big(\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)\le\pi_0|\mathcal{C}(\mathbf{x})=1\Big),\\ \mathcal{E}_2(w) &= \mathsf{Pr}\Big(\mathcal{D}_w(\mathbf{x};\hat{\theta}^1,\hat{\theta}^2)>\pi_0|\mathcal{C}(\mathbf{x})=2\Big),\end{aligned}\tag{3.7}$$

where $\pi_0=\ln\frac{\pi_2}{\pi_1}$, *have the limits*

$$\mathcal{E}_1(w)\longrightarrow\Phi\Big(-\frac{E(w)-\pi_0}{\sqrt{V(w)}}\Big),\quad \mathcal{E}_2(w)\longrightarrow\Phi\Big(-\frac{E(w)+\pi_0}{\sqrt{V(w)}}\Big),\tag{3.8}$$

as $n\to\infty$. *Further, let* W *be a class of functions such that for all* $w(u)\in W$ *both* $E(w)$ *and* $V(w)$ *do not equal to zero, then assuming that* $\pi_1=\pi_2$, *and denoting* $w_0(u):=\arg\min_{w(u)\in W}\mathcal{R}(w)$, *we get*

$$w_0(u)=\frac{\int\gamma^2 g(u;m+2,\gamma^2)dH(\gamma^2)}{u\int g(u;m,\gamma^2)dH(\gamma^2)},\tag{3.9}$$

where $g(u;m,\gamma^2)$ *is the probability density of the non-central* χ^2 *distribution with* m *degrees of freedom and non-centrality parameter* γ^2.

PROOF: The proof of assertion (3.8) follows straightforwardly by using the asymptotic normality of $\mathcal{D}_w(\mathbf{x},\hat{\theta}^1,\hat{\theta}^2)$ and taking into account the limiting results from Lemma 3.1.

In order to prove (3.8) we notice that minimization of $\mathcal{R}=\mathcal{R}(w)$ is equivalent to maximization of

$$\frac{\int\gamma^2[\int w(u)g(u;m+2,\gamma^2)du]dH(\gamma^2)]^2}{\int[\int uw^2(u)g(u;m,\gamma^2)du]dH(\gamma^2)}\tag{3.10}$$

with respect to $w(u)$. By changing the order of integration in both numerator and denominator of (3.10) and then using the Cauchy-Schwartz inequality we obtain

$$\begin{aligned}&[\int w(u)\int\gamma^2 g(u;m+2,\gamma^2)dH(\gamma^2)du]^2 \qquad (3.11)\\ \le\ &\int uw^2(u)[\int g(u;m,\gamma^2)dH(\gamma^2)]du\int\frac{[\gamma^2 g(u;m+2,\gamma^2)dH(\gamma^2)]^2}{u\int g(u;m,\gamma^2)dH(\gamma^2)}du,\end{aligned}$$

with equality being attained if and only if

$$w(u)\sqrt{u\int g(u;m,\gamma^2)dH(\gamma^2)}\propto\frac{\int\gamma^2 g(u;m+2,\gamma^2)dH(\gamma^2)}{\sqrt{u\int g(u;m,\gamma^2)dH(\gamma^2)}},$$

from which (3.9) immediately follows. □

It is not difficult to show that $w_0(u)$ is bounded, continuous for $u > 0$ and $w_0(u) \in W$. The corresponding minimum value of $\mathcal{R}$ is

$$\mathcal{R}_0 = \Phi\Big(-\frac{1}{2}\sqrt{2\rho \int \frac{[\int \gamma^2 g(u; m+2, \gamma^2) dH(\gamma^2)]^2}{u \int g(u; m, \gamma^2) dH(\gamma^2)} du}\Big). \tag{3.12}$$

The practical implementation of the proposed weighting technique requires specification of the distribution $H(\gamma^2)$ of the block separation strength. To give an impression of how the weighting by $w_0(u)$ works, we consider one simple choice of $H(\gamma^2)$.

EXAMPLE: Distributions that can describe the *a priori* knowledge about the block separation strength include e.g. a *point mass distribution*, $dH(\gamma^2) = 1$ concentrated in a certain point, γ^2. Using this type of distribution means that the contributions of all blocks into the distance $\mathcal{J}(\kappa)$ are *identical* so that the separation strength of all blocks is assumed to be the same and equal to γ^2. Then in a view of (3.9) and given the point mass distribution $H(\gamma^2)$, the optimal weight-function, w_0, turns out to be

$$w_0(u) = \frac{\gamma^2 g(u; m+2, \gamma^2)}{u g(u; m, \gamma^2)},$$

which according to (3.12) gives the limiting risk

$$\mathcal{R}(w_0) = \Phi\Big(-\frac{1}{2}\sqrt{2\rho \int \frac{[\gamma^2 g(u; m+2, \gamma^2)]^2}{u g(u; m, \gamma^2)} du}\Big).$$

We now may understand the effect of weighting by comparing $\mathcal{R}(w_0)$ with $\mathcal{R}(1)$, i.e. with the misclassification risk when $w_0 = 1$ (no weighting is involved). For this case, we use (3.5)-(3.6) as well as properties of the non-central χ^2 distribution and find

$$\begin{aligned} E(1) &= \rho \int \gamma^2 dH(\gamma^2) = \rho\gamma^2, \\ V(1) &= \rho \int [\int u g(u; m, \gamma^2) du] dH(\gamma^2) = \rho(\gamma^2 + m), \end{aligned}$$

which in turn gives

$$\mathcal{R}(1) = \Phi\Big(-\frac{1}{2}\sqrt{2\rho \frac{\gamma^4}{\gamma^2 + m}}\Big).$$

Using standard arguments it is not difficult to show that

$$\begin{aligned}\int \frac{[g(u;m+2,\gamma^2)]^2}{ug(u;m,\gamma^2)}du &> \int \frac{g(u;m+2,\gamma^2)}{ug(u;m,\gamma^2)}du > \frac{\int g(u;m+2,\gamma^2)du}{\int ug(u;m+\gamma^2)du} \\ &= \frac{1}{\gamma^2+m}.\end{aligned}$$

Since $\Phi(y)$ is a decreasing function of y, we conclude that $\mathcal{R}(w_0) < \mathcal{R}(1)$.

Observe that the obtained result could be seen as somewhat counter intuitive: Assuming the true separation strength to be equal for all blocks and thereby giving them equal weights, should *not* effect classification accuracy. Our results however clearly indicate the decrease of misclassification risk when weighting by w_0. A clue to the decrease of $\mathcal{R}(w_0)$ is provided by the results of Theorem 2, where we have shown that when using sample based weight-function in a high dimensional setting, the block separation strength is heavily overestimated. Thus, even in a specific case of a point mass distribution $H(\gamma^2)$, the optimal weight-function w_0 established in Theorem 3 turns out to be sufficiently sensitive to the high dimensionality effects and provides the desirable down-weighting.

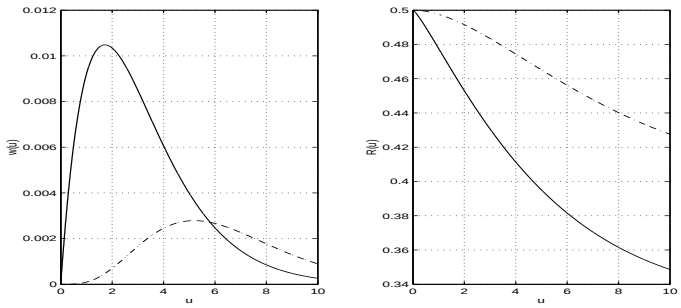

Figure 1: Optimal weight function $w_0(u)$ (left) and associated misclassification risk $\mathcal{R}(w_0)$ (right), given that $dH(\gamma^2) = 1$, $\gamma^2 = 1.8$ and $n = 36$. The behavior of $w_0(u)$ and $\mathcal{R}(w_0)$ with $m = 3$, $\kappa = 8$ and $\rho = 0.222$ (solid line); with $m = 6$, $\kappa = 4$ and $\rho = 0.111$ (dashed line).

We conclude with some graphical illustrations of the weighting technique, which is given by Figure 1 where w_0 and $\mathcal{R}(w_0)$ are plotted under different values of m, κ and ρ. As expected each weight-function places substantial part of its mass to the right tail (left panel) so that the block impacts with

high deviations of estimates are down-weighted. Observe also that the weight-function seems to be more "flat" as the block size m increases. The right panel shows the asymptotic misclassification risk $\mathcal{R}(w_0)$ when weighting by w_0. Not surprising it is seen to be slowly decreasing given the smaller number of independent blocks in the network, i.e. when $\kappa = 4$ (dashed line) whereas embedding more independence in the network structures, i.e. letting $\kappa = 8$ and reducing the block size, lead to a faster decrease (solid line). Roughly speaking, given that the structure of BN is correct, the corresponding additive classifier borrows strength from the local density which naturally results in a better classification accuracy. However, the design of these procedures should take into account block separation strength *combined* with the effect of high dimensionality induced by the sample based weight-function.

Appendix

PROOF OF LEMMA 3.1: To prove (3.5), we use the representation (3.2) for $\int \mathcal{D}_i(\mathbf{x}_i, \hat{\theta}_i^1, \hat{\theta}_i^2) f_i(\mathbf{x}_i, \theta_i^1) d\mathbf{x}_i$ and observe that

$$\begin{aligned} &\frac{1}{2n}\sum_{i=1}^{\kappa} \mathsf{E}[w_i \cdot (\langle \sqrt{2}\gamma_i + T_i^2, \sqrt{2}\gamma_i + T_i^2\rangle - \langle T_i^1, T_i^1\rangle)] \\ &= \quad \frac{1}{n}\sum_{i=1}^{\kappa} \mathsf{E}[w_i \cdot (\langle \gamma_i, \gamma_i\rangle + \sqrt{2}\langle \gamma_i, T_i^2\rangle)] \end{aligned} \tag{A.1}$$

since T_i^1 and T_i^2 are independent and identically distributed. By the distributional properties of $\frac{n\hat{J}_i(n)}{2}$ established in Theorem 2.1, the first summand of the right-hand side of (A.1) can be transformed as follows

$$\frac{1}{n}\sum_{i=1}^{\kappa} \mathsf{E}[w(\frac{n\hat{\mathcal{J}}_i}{2})\langle \gamma_i, \gamma_i\rangle] = \frac{\kappa}{n}\frac{1}{\kappa}\sum_{i=1}^{\kappa}\langle \gamma_i, \gamma_i\rangle \int w(u) g(u; m, \gamma_i^2) du + \mathcal{O}(n^{-1/2}).$$

Therefore, since $H_n(\cdot)$ converges towards $H(\cdot)$ as $n \to \infty$, we get by Helly-Bray theorem (see Rao, 1973)

$$\frac{1}{\kappa}\sum_{i=1}^{\kappa}\langle \gamma_i, \gamma_i\rangle \int w(u) g(u; m, \gamma_i^2) du \longrightarrow \int \gamma^2 [\int w(u) g(u; m, \gamma^2) du] dH(\gamma^2). \tag{A.2}$$

Now we take a closer look at the expectation

$$\frac{1}{n}\sum_{i=1}^{\kappa}\mathsf{E}\Big[w(\frac{n\hat{J}_i(n)}{2})\sqrt{2}\langle\gamma_i,T_i^2\rangle\Big],$$

and note that $\frac{n\hat{J}_i(n)}{2}$ and T_i^2 are statistically dependent. Therefore the main point is to evaluate $\mathsf{E}[w(u)y]$, where u tends to $\mathcal{G}(u;m,\gamma^2)$ and y tends to $N(0,1)$. First we perform the simple (univariate) case when $m=1$. As it will be explained below obvious changes will suffice to treat the vector case. We introduce the univariate analogies of T_i^1 and T_i^2 and define them by y_1 and y_2 respectively. Then the expectation to be evaluated is

$$\mathsf{E}[w((y_2-y_1+\gamma)^2)y_2].$$

Consider first only $\mathsf{E}[w((y_2-y_1+\gamma)^2)]$. Since y_1 and y_2 are independent and distributed as $N(0,1)$ we get

$$\mathsf{E}[w((y_2-y_1+\gamma)^2)]=\frac{1}{\sqrt{2\pi}}\iint w((y_2-y_1+\gamma)^2)e^{-y_1^2/2}e^{-y_2^2/2}dy_1dy_2,$$

and by changing variables

$$\mathsf{E}[w((z_2-z_1)^2)]=\frac{1}{\sqrt{2\pi}}\iint w((z_2-z_1)^2)e^{-z_1^2/2}e^{-(z_2-\gamma)^2/2}\,dz_1\,dz_2,$$

where $z_2=y_2+\gamma$ and $z_1=y_1$. It is further seen that z_2 under the integrals can be obtained by differentiating with respect to γ: Indeed

$$\frac{\partial}{\partial\gamma}\mathsf{E}[w((z_2-z_1)^2)]=\frac{1}{\sqrt{2\pi}}\iint(z_2-\gamma)w((z_2-z_1)^2)e^{-z_1^2/2}e^{-(z_2-\gamma)^2/2}dz_1dz_2,$$

or in terms of u and y_2

$$\frac{\partial}{\partial\gamma}\mathsf{E}[w(u)]=\mathsf{E}[y_2w(u)]. \tag{A.3}$$

Note that the expectation to be evaluated is given by the right-hand side of (A.3). Taking the same approach to the vector case (i.e. $m>1$) and coming back to T_i^2 instead of y_2 we establish the following

$$\langle\gamma_i,\mathsf{E}[T_i^2w(u)]\rangle=\langle\gamma_i,\frac{\partial}{\partial\gamma}\mathsf{E}[w(u)]\rangle. \tag{A.4}$$

Using the following recurrence relation for the non-central χ^2 distribution $\mathcal{G}(u; m, \gamma^2)$

$$\frac{\partial \mathcal{G}(u; m, \gamma^2)}{\partial \gamma} = \gamma[\mathcal{G}(u; m+2, \gamma^2) - \mathcal{G}(u; m, \gamma^2)], \tag{A.5}$$

see Johnson et al. (1988) and applying $(A.5)$ to $(A.4)$ we obtain

$$\begin{aligned} &\langle \gamma_i, \frac{\partial}{\partial \gamma} \mathsf{E}[w(u)] \rangle \\ &= \langle \gamma_i, \gamma_i \rangle \int w(u) d\mathcal{G}(u; m+2, \gamma^2) - \langle \gamma_i, \gamma_i \rangle \int w(u) d\mathcal{G}(u; m, \gamma^2). \end{aligned} \tag{A.6}$$

Combining $(A.4)$ with $(A.6)$ and using $(A.2)$ gives

$$\frac{1}{2n} \sum_{i=1}^{\kappa} \langle \gamma_i, \gamma_i \rangle \int w(u) d\mathcal{G}(u; m, \gamma_i^2),$$

from which we conclude that

$$\frac{1}{n} \sum_{i=1}^{\kappa} \gamma_i^2 \int w(u) d\mathcal{G}(u; m+2, \gamma_i^2) \longrightarrow \rho \int \gamma^2 [\int w(u) d\mathcal{G}(u; m+2, \gamma^2)] dH(\gamma^2)$$

as $n \to \infty$. Since the function under the integral is bounded, we can use the Helly-Bray theorem and turn to integration with respect to $H(u)$.

To evaluate the second moment we use the representation (3.3) and distributional properties of $\frac{n\hat{J}_i(n)}{2}$ from Theorem 2.1. These give

$$\begin{aligned} &\mathsf{E}\Big[\sum_{i=1}^{\kappa} w_i^2 \cdot \int [\mathcal{D}_i(\mathbf{x}_i, \hat{\theta}_i^1, \hat{\theta}_i^2)\Big]^2 f_i(\mathbf{x}_i, \theta_i^1) \mu(d\mathbf{x}_i) \\ &= \frac{1}{n} \sum_{i=1}^{\kappa} \mathsf{E}[w_i^2 \langle \sqrt{2}\gamma_i + T_i^2 - T_i^1, \sqrt{2}\gamma_i + T_i^2 - T_i^1 \rangle] + \mathcal{O}(n^{-1/2}). \end{aligned} \tag{A.7}$$

The right-hand side of $(A.7)$ yields

$$\frac{2}{n} \sum_{i=1}^{\kappa} \int u w^2(u) d\mathcal{G}(u; m, \gamma_i^2) \longrightarrow 2\rho \iint u w^2(u) d\mathcal{G}(u; m, \gamma^2) dH(\gamma^2),$$

by the same arguments that have been used for the first moment. This completes the proof.

Acknowledgments

Tatjana Pavlenko was supported in part by Västernorrland County Council under grant FoU: JA-2002-1001 RDC.

References

Barash, Y. & Friedman, N. (2002), Context-specific Bayesian clustering for gene expression data, *J. of Comput. Biology* **9**, 169-191.

Cowell, R., Dawid, A.P., Lauritzen, S.L & Spiegelhalter, D.J. (1999), *Probabilistic Networks and Expert Systems*. Springer, New York.

Friedman, J. (1997), On bias, variance, 0/1 - loss, and curse-of-dimensionality, *Data Mining and Knowledge Discovery* **1**, 55-77.

Girko, V.L. (1995), *Statistical Analysis of Observations of Increasing Dimension*. Kluwer, London.

Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2. Wiley, New York.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *Elements of Statistical Learning: Data Mining, Inference and Prediction*. (Springer, New York).

Korb, K.B. & Nicholson, A.E. (2003), *Bayesian Artificial Intelligence*. Chapman and Hall / CRC.

Kusiak, A. (2000), Decomposition in Data Mining: An Industrial Case Study, *IEEE transactions on Electronics Packaging Manufacturing* **23:4**, 345-353.

Maimon, O. & Rokach, L. (2001), Data mining by attribute decomposition with semiconductors manufacturing case study, in: D. Braha, eds. *Data Mining for Design and Manufacturing: Methods and Applications*. Kluwer, London. pp. 311-336.

Pavlenko, T. (2001), Feature informativeness, curse-of-dimensionality and error probability in discriminant analysis. PhD thesis, Lund University, Sweden.

Pavlenko, T. & von Rosen, D. (2001), Effect of dimensionality on discrimination, *Statistics* **35**, 191-213.

Pavlenko, T. (2003), On feature selection, curse-of-dimensionality and error probability in discriminant analysis. *J. of Stat. Planning and Inference* **115**, 565-584.

Rao, C.R. (1973), *Linear statistical inference and its applications, 2nd ed.* Wiley, New York.

Zang, H. & Ling, C.X. (2001), Learnability of augmented naive Bayes in nominal domains, in: C.E. Brodley & A.P. Danyluk, eds. *Proc. of the Eighteenth International Conference on Machine Learning*, (Morgan Kaufmann). pp. 617-623.