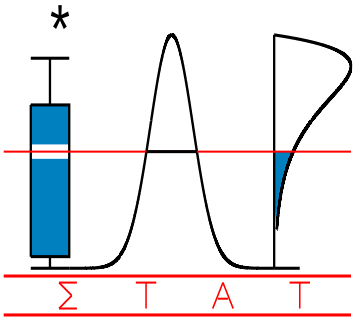
T E C H N I C A L
R E P O R T

0563

LOCAL LIKELYHOOD REGRESSION IN GENERALIZED LINEAR SINGLE-INDEX MODELS WITH APPLICATIONS IN MICROARRAY DATA

LAMBERT-LACROIX S.and J. PEYRE



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Local likelyhood regression in generalized linear single-index models with applications to microarray data

Sophie Lambert-Lacroix and Julie Peyre
LMC-IMAG, BP 53, 38041 Grenoble cedex 9, France

December 19, 2005

Abstract

Searching for an effective dimension reduction space is an important problem in regression, especially for high dimensional data such as microarray data. A major characteristic of microarray data consists in the small number of observations n and a very large number of genes p. This "large p, small n" paradigm makes the discriminant analysis for classification difficult. To answer this curse of dimensionality, a solution consists in reducing the dimension. In this paper, supervised classification is understood as a regression problem with a small number of observations and a large number of covariates. We propose a new approach for dimension reduction. Based on a semi-parametric approach, we use local likelihood estimates for single index generalized linear models. We consider asymptotic properties of our procedure and illustrate its asymptotic performances by simulations. Finally, we consider applications of our method when applied to binary and multi-class classification of three real data sets: Colon, Leukemia and SRBCT.

Keywords: Dimension reduction; Generalized linear models; Generalized linear single-index models; Local likelihood estimates; Nonparametric regression; Microarray data.

Availability: The software that implements the procedures and the data sets on which this paper focuses are freely available. Procedures for the binary case (`R` codes) can be downloaded from `http://www-lmc.imag.fr/lmc-sms/Julie.Peyre.html` and procedures for the multi-class problem (`MATLAB` codes) with data sets can be downloaded from `http://www-lmc.imag.fr/lmc-sms/Sophie.Lambert.html`

Contact: Sophie.Lambert@imag.fr

1 Introduction

Microarray technology generates a vast amount of data by measuring, through the hybridization process, the levels of virtually all the genes expressed in a biological sample. One can expect that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine.

One important goal of analyzing microarray data is to classify the samples. To cite a few, Golub *et al.* [12] have considered classification of acute leukemia, Alon *et al.* [2] have addressed the cluster analysis of tumor and normal colon tissues. The approaches developed in these papers consist in discrimination methods and machine learning methods (see [7] for a comparative study).

In microarray studies, the number of samples, n, is relatively small compared to the number of genes, p, usually in thousands. Unless a preliminary variable selection step is performed, standard statistical methods in classification perform poorly because there are far more variables than observations. One problem is multicolinearity: estimating equations become singular and have no unique and stable solution. Furthermore even if all genes can be used as in support vector machines, it does not seem to be sensible to use all the genes. This use allows presence of the noise associated with genes of little or no discrimination power, that could inhibit and degrade the performances of the classification rules in its application to unclassified tumor. In this situation, dimension reduction is needed to reduce the high p-dimensional gene space. In most previously mentioned works, the authors have used univariate methods for reducing the number of genes.

Alternative approaches to handle the dimension reduction problem can also be used. In particular, there exist parametric methods based on "Partial Least Squares" (PLS). In the microarray context, PLS yields orthogonal linear combinations of genes so reducing the dimension with few "super-genes". Nguyen et Rocke [21, 20] proposed using PLS for dimension reduction as a preliminary step to classification, based either on linear logistic discrimination, or linear or quadratic discriminant analysis. However, this seems to be intuitively unappealing because PLS is really designed to handle continuous responses and models that do not suffer from heteroscedasticity as it is the case for Bernoulli or multinomial data. More recently, Ding and Gentleman [6] proposed an approach based on the procedure of Marx [18]. They phrased the problem in a generalized linear models setting and applied Firth's procedure to avoid (quasi)separation. Indeed, for logistic regression, it is well known that convergence poses a long standing problem. Infinite parameter estimates can occur depending on the configuration of the sample points in the observation space ([1]). Fort and Lambert-Lacroix [11] proposed a new method combining Partial Least Squares (PLS) and ridge penalized logistic regression and applied this procedure to the microarray data classification.

There exist alternative semi-parametric approaches. Antoniadis *et al.* [3] proposed to use the Minimum Average Variance Estimation (MAVE, [25]) to reduce the dimension before applying (non) parametric logistic regression. As PLS, this procedure provides linear combinations of genes. It is based on a local least square criterion combined with a nonparametric estimation by local polynomial of the regression function. Even if this procedure handles any response variables, it does not take into account the particular generalized linear model structure. In particular it does not use the relationship between mean and variance and the fact that in generalized linear models, we usually consider criteria based on likelihood (which coincides with the least square criterion only for gaussian models).

In this paper, we view the classification problem as a regression one with few observations and many predictor variables. We propose a new approach for dimension reduction which we call GSIM. Based on a semi-parametric approach, GSIM use local likelihood estimates for single-index generalized linear models. This method is similar to

the procedure OPG (Outer Product of Gradients) of [25] which is a simplified version of MAVE but with comparable performances. The difference between GSIM and OPG stands in replacing the least square criterion by the likelihood one. So we use the particular structure of generalized linear models.

In all the dimension reduction methods mentioned above, the reduction is obtained selecting some linear combinations of covariables. The goal is to search some informative direction and to delete directions which contain only noise. There exist another family of methods: the variable selection. These approaches consist in selecting the most informative genes. For example in diagnostic context, it could be interesting to select some genes instead of linear combinations (using a priori all the genes). Nevertheless, the compression approaches are not incompatible with variable selection approaches. The procedure proposed here can be used to select variables by adapting for example the Recursive Feature Elimination (RFE) of Guyon *et al.* [14].

This paper is organized as follows. In Section 2, we consider generalized linear model. In particular, we recall classical parametric and nonparametric methods with some limitations in high dimension. Next we give the definition of the generalized linear single-index models that allows to overcome the dimensionality problem. Section 3 is devoted to our estimation method in these models. First, we propose a procedure in the asymptotic context and give its asymptotic properties. Next, we propose to modify it in the "large p and small n" case by introducing a ridge penalty. In Section 4, we illustrate the asymptotic performances of our procedure by simulations. We also consider applications of our methods when applied to binary and multi-class classification on three real data sets: Colon, Leukemia and SRBCT (that is in the "large p and small n" case). Section 5 contains the proofs of asymptotic properties.

2 Model and Notations

After recalling the definition of the generalized linear models, we present the maximum likelihood method. In particular, we point out the problem of existence of the maximum likelihood estimator for the logistic regression. A nice overview for generalized linear models can be found in [8]. We also recall the local likelihood method, viewed as a nonparametric approach. In both situations, we underline the limitations of these approaches in the case of "small n large p" and we propose to consider instead the generalized linear single-index models.

2.1 Notations

For two integers $l < m$, $l : m$ is the vector $(l, l+1, \ldots, m)$. The p-dimensional vector of components equal to one is denoted by $\mathbb{1}_n$; $(e_1, \cdots, e_n)$ is the canonical base of $\mathbb{R}^n$ and Id_n is $n \times n$ identity matrix. For a matrix A, we denote by $A_{i,j}$ the element (i, j) (when A is one vector, we use the notation A_i); $A_{i:j,:}$ is the matrix composed of A rows from i to j; A^T denotes its transpose and $|A|$ its determinant.

For $\underline{l} = (l_1, \ldots, l_p)$ a vector of $\mathbb{N}^p$, we introduce the following notations,

$$|\underline{l}| = \sum_{j=1}^{p} l_j, \quad \underline{l}! = \prod_{j=1}^{p} l_j!, \quad \forall x \in \mathbb{R}^p,\ x^{\underline{l}} = \prod_{j=1}^{p} x_j^{l_j}.$$

Let f be a function which has continuous partial derivatives in $\mathbb{R}^p$ up to order q. For $\underline{l} \in \mathcal{A}_q = \{\underline{l}; |\underline{l}| \leq q\}$, the partial derivate $\partial^{|\underline{l}|} f(\underline{x})/\partial \underline{x}^{\underline{l}}$ is denoted by $D^{\underline{l}} f(\underline{x})$. We also denote by ∇ the gradient operator.

2.2 Generalized linear models

2.2.1 Definition

Let $(X_1^T, Y_1), \ldots, (X_n^T, Y_n)$ be an independent random sample of the random pair (X^T, Y), where Y is a response variable of $\mathbb{R}^G$, where G is positive integer, and X is the associated covariate vector of $\mathbb{R}^p$. We assume that the conditional density of Y given $X = x$, belongs to a canonical exponential family [19]

$$f_{Y/X}(y) = \exp\left\{ \frac{y^T \theta(x) - b(\theta(x))}{a(\phi)} + c(y, \phi) \right\} \tag{1}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. The parameter $\theta(\cdot)$ is called the canonical parameter and ϕ is called the dispersion parameter. Recall that the law support of this distribution must be independent of $\theta(\cdot)$. Under the model (1), it can be shown that

$$\mathbb{E}(Y|X = x) = \nabla b(\theta(x)) = \mu(x), \quad \mathrm{Var}(Y|X = x) = a(\phi)\nabla^2 b(\theta(x)).$$

In parametric generalized linear models, some transformation of the regression function $\mu(x) = \mathbb{E}(Y|X = x)$ is supposed to be linear in the covariates:

$$g_k(\mu(x)) = \eta_k(x) = x^T \gamma^{(k)}, \quad k = 1, \ldots, G, \tag{2}$$

where the function $g : \mathbb{R}^G \to \mathbb{R}^G$ is called the link function. The choice of $g = (\nabla b)^{-1}$ allows to identify the linear predictor $\eta(x)$ with the canonical parameter $\theta(x)$, and this special link is called *canonical link*.

In some cases, the linear relation of the parametric approach is not guaranteed. To enhance the flexibility of the model, another approach consists in supposing that $\eta(x)$ is a nonparametric function.

2.2.2 Example: logistic regression

This model is used for the polychotomous discrimination problem which will be considered for applications to microarrays. The categorical outcome have $G + 1$ classes labeled $0, 1, \ldots, G$, and the response variable Y is coded in the following way. The k-th component of Y is equal to 1 and the others are zeros if and only if the label is k. The label 0 is coded by G components equal to 0. This model is a multinomial one with parameters $(\mu_0, \cdots, \mu_G)$ where μ_k is the probability of the k-th class. The case $G = 1$ correspond to the Bernoulli model.

That is a generalized linear models with $Y \in \{0,1\}^G$ and link function $g_k(\mu) = \eta_k = \ln \mu_k - \ln \mu_0$, $1 \leq k \leq G$. The canonical exponential family is defined by $a(\phi) = 1$, $b(\eta) = \ln(1 + \sum_{l=1}^{G} \exp(\eta_l))$ and $c = 0$. It follows that

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_{l=1}^{G} \exp(\eta_l)}, \qquad 1 \leq k \leq G. \tag{3}$$

2.3 The parametric approach: maximum likelihood

2.3.1 Log-likelihood function

In the model (2) the parameter vector of size $G(p+1)$ defined by $\gamma = ({\gamma^{(1)}}^T, \ldots, {\gamma^{(G)}}^T)^T$ is estimated by maximum likelihood. We suppose the model to be sufficiently regular as in the logistic regression case.

Here we need some additional notations. The observations (y_i, x_i) of (Y_i, X_i), $1 \leq i \leq n$, are collected in the vector of response variables $\mathbf{Y} \in \mathbb{R}^{Gn}$ and in the design matrix $\mathbf{X}^{(G)} \in \mathbb{R}^{Gn \times Gp}$. The k-th bloc of $\mathbf{Y}$ is given by

$$\mathbf{Y}_{\iota_k+1:\iota_k+G} = y_k, \qquad \text{with} \qquad \iota_k = (k-1)G, \quad k = 1, \ldots, n. \tag{4}$$

In the same way, the rows from $\iota_k + 1$ to $\iota_k + G$ of $\mathbf{X}^{(G)}$ are build from the realization x_k of X_k. Precisely, we have

$$\mathbf{X}^{(G)}_{\iota_k+1:\iota_k+G,:} = \begin{bmatrix} x_k^T & 0 & \cdots & 0 \\ 0 & x_k^T & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & x_k^T \end{bmatrix}. \tag{5}$$

The bloc matrix $\mathbf{Z}^{(G)} \in \mathbb{R}^{Gn \times G(p+1)}$ is defined from the vector $z_k = [1 \; x_k^T]^T$ in an analogous way. Notice that for $G = 1$, $\mathbf{X}^{(G)}$ is the usual design matrix $\mathbf{X}^{(1)}$ of size $n \times p$, and that $\mathbf{Z}^{(1)} = [\mathbb{1}_n \; \mathbf{X}^{(1)}]$.

The log-likelihood of the observations for the value γ, simply denoted by $\ell(\gamma)$, is given by (up to a multiplicative term independent of γ)

$$\ell(\gamma) = \mathbf{Y}^T \mathbf{Z}^{(G)} \gamma - \sum_{k=1}^{n} b\left((\mathbf{Z}^{(G)} \gamma)_{\iota_k+1:\iota_k+G} \right).$$

Such an estimate is a solution to the normal equation ${\mathbf{Z}^{(G)}}^T (\mathbf{Y} - \boldsymbol{\mu}(\gamma)) = 0$ and the means vector $\boldsymbol{\mu}$ is given by

$$\forall 1 \leq k \leq n, \qquad \boldsymbol{\mu}_{\iota_k+1:\iota_k+G}(\gamma) = \nabla b(\varepsilon) \Big|_{\varepsilon = (\mathbf{Z}^{(G)}\gamma)_{\iota_k+1:\iota_k+G}}. \tag{6}$$

Notice that $\ell(\gamma)$ depends on γ through the linear predictor $\eta = \mathbf{Z}^{(G)}\gamma$. To make identifiable the parameters relating to the constant in the model (γ_{kp+1}, $k = 0, \ldots, G-1$), $\mathbf{X}^{(1)}$ must be centered: $\mathbf{X}^{(1)} - \mathbb{1}_n \mathbb{1}_n^T \mathbf{X}^{(1)}$. It is also recommended to standardize the design matrix for numerical stability in the computations.

2.3.2 Iteratively Reweighted Least Squares

If $\mathbf{Z}^{(G)}$ is of full column-rank, the parameter γ is identifiable. In general, the normal equations are not linear in γ and are solved in an iterative way. When the estimate exists and is unique, it can be computed as a limit of a converging

Newton-Rapson sequence. This algorithm is known as the Iteratively Reweighted Least Squares (IRLS) algorithm ([13]). Let $\mathbf{W}(\gamma) \in \mathbb{R}^{Gn \times Gn}$ be a bloc diagonal matrix, with its k-th bloc equal to $\nabla^2 b\left((\mathbf{Z}^{(G)}\gamma)_{\iota_k+1:\iota_k+G}\right) \in \mathbb{R}^{G\times G}$, for $1 \leq k \leq n$. Since $\nabla^2\ell(\gamma) = -\mathbf{Z}^{(G)^T}\mathbf{W}(\gamma)\mathbf{Z}^{(G)}$, IRLS iterations lead to a sequence $(\gamma^t)_t$ given by

$$\begin{aligned}\gamma^{t+1} &= \gamma^t + \left[\mathbf{Z}^{(G)^T}\,\mathbf{W}(\gamma^t)\,\mathbf{Z}^{(G)}\right]^{-1}\mathbf{Z}^{(G)^T}\left(\mathbf{Y} - \boldsymbol{\mu}(\gamma^t)\right)\\ &= \left[\mathbf{Z}^{(G)^T}\,\mathbf{W}(\gamma^t)\,\mathbf{Z}^{(G)}\right]^{-1}\mathbf{Z}^{(G)^T}\,\mathbf{W}(\gamma^t)\,\left\{\mathbf{Z}^{(G)}\gamma^t + \left[\mathbf{W}(\gamma^t)\right]^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}(\gamma^t)\right)\right\}.\end{aligned}$$

That is γ^{t+1} are obtained by a weighted least square regression of the pseudo-variable

$$z^t = \mathbf{Z}^{(G)}\gamma^t + \left[\mathbf{W}(\gamma^t)\right]^{-1}\left(\mathbf{Y} - \boldsymbol{\mu}(\gamma^t)\right) \tag{7}$$

onto the columns of $\mathbf{Z}^{(G)}$. For a convergent sequence, we have $\hat{\gamma}^{\mathrm{MV}} = \lim_t \gamma^t$.

If $\mathbf{Z}^{(G)}$ is not full column-rank, the parameter γ is no longer identifiable. Nevertheless, we can always consider the full column-rank matrix $\mathbf{Z}^{(\mathrm{red},G)}$ obtained with common computation based on singular value decomposition. When it exists, we have then a ML estimator $\hat{\gamma}^{\mathrm{MV}} \in \mathbb{R}^{\mathrm{rank}(\mathbf{Z}^{(\mathrm{G})})}$ and $\hat{\gamma}^{\mathrm{MV}}$ is defined as the minimal norm vector among all vectors satisfying $\mathbf{Z}^{(G)}\gamma = \mathbf{Z}^{(\mathrm{red},G)}\hat{\gamma}^{\mathrm{MV}}$.

2.3.3 Particular case of the parametric logistic regression

When $\mathbf{Z}^{(G)}$ is of full column-rank, the ML estimator does not necessary exist. It depends on the configuration of the n sample points in the covariables space (cf. [1, 23, 17]). There are three exclusive cases: separate, quasi-separate and overlap situations. In the first case, there exist γ such that for $1 \leq k \leq n$, $1 \leq i \leq G$,

$$\left[(\mathbf{Z}^{(1)}\gamma^{(i)})_k > (\mathbf{Z}^{(1)}\gamma^{(l)})_k, \quad \forall l \in \{0, \cdots, G\} \setminus \{i\}\right] \qquad \text{if and only if} \qquad \mathbf{Y}_{\iota_k+i} = 1, \tag{8}$$

with $\gamma^{(0)} = 0$. The quasi-separation stands when the strict inequality in (8) is replaced by a large one. In both these cases, the function ℓ is maximal when $\|\gamma\| \to +\infty$ and the ML estimator does not exist. In the overlap case, normal equations possess an unique solution, which in practice, is computed by the iterative Newton-Raphson method. Notice that the choice $\gamma^0 = (G+1)^{-1}(3\mathbf{Y} + (\mathbb{1}_{Gn} - \mathbf{Y}))$ is a good initialization to obtain a convergent sequence in the IRLS algorithm ([8]).

In the applications considered in Subsection 4.2, we have $n << p$, such that in practice $\mathrm{rank}(\mathbf{Z}^{(\mathrm{G})}) = Gn$. The matrix $\mathbf{Z}^{(G)}$ is not of full column-rank. By reparametrization, the likelihood equations leads to $\mathbf{Y} = \mu$ that involves (cf. (6))

$$(\mathbf{Z}^{(\mathrm{red},G)}\hat{\gamma})_{\iota_k+i} = \ln\left(\frac{\mathbf{Y}_{\iota_k+i}}{1 - \sum_{l=1}^{G}\mathbf{Y}_{\iota_k+l}}\right), \qquad \forall 1 \leq k \leq n, 1 \leq i \leq G.$$

So we have $\|\hat{\gamma}\| = +\infty$ and it follows that the likelihood estimator may never exist. So we must consider a dimension reduction method to address the regression problem in a subspace of smaller dimension.

2.4 The nonparametric approach

When $\eta(\cdot)$ is modeled in a nonparametric way, we can consider an estimation method based on local likelihood [9]. This method is commonly presented for the case of $p = 1$ and $G = 1$. Due to the applications considered here, we present it for any p and G. The function $\eta(\cdot)$ is locally approximated by a polynomial of order q

$$\eta_k(u) \sim \sum_{\underline{l} \in \mathcal{A}_q} D^{\underline{l}} \eta_k(x)(u-x)^{\underline{l}}/\underline{l}! = \sum_{\underline{l} \in \mathcal{A}_q} a_{\underline{l}}^k (u-x)^{\underline{l}}, \quad k = 1, \ldots, G,$$

for u in a neighborhood of x. We denote by $a_{\underline{l}}$ the vector $(a_{\underline{l}}^1, \ldots, a_{\underline{l}}^G)^T$. Let K^p be a p-dimensional kernel, H a bandwidth matrix, and $K_H^p(\cdot) = |H|^{-1} K^p(H^{-1}\cdot)$ be the rescaling of K^p. The local likelihood is a weighted likelihood, with weights $K_H^p(X_i - x)$:

$$\sum_{i=1}^{n} \mathcal{L}\left[\sum_{\underline{l} \in \mathcal{A}_q} a_{\underline{l}} \, (X_i - x)^{\underline{l}}, Y_i\right] K_H^p(X_i - x). \tag{9}$$

Here $\mathcal{L}(u, Y)$ is the log-likelihood function in which $\eta(x)$ is replaced by its polynomial approximation u. The local likelihood leads to $\widehat{D^{\underline{l}}\eta}_k(x) = \underline{l}! \, \hat{a}_{\underline{l}}^k(x)$, where $\{\hat{a}_{\underline{l}}(x), \underline{l} \in \mathcal{A}_q\}$ maximizes (9) with respect $\{a_{\underline{l}}, \underline{l} \in \mathcal{A}_q\}$. In particular, we have

$$\hat{\eta}_k(x) = \hat{a}_{(0,\ldots,0)}^k(x), \quad \widehat{\nabla\eta}_k(x) = (\hat{a}_{e_1}^k(x) \ldots, \hat{a}_{e_p}^k(x))^T, \quad k = 1, \ldots, G.$$

The estimators $\{\hat{a}_{\underline{l}}(x), \underline{l} \in \mathcal{A}_q\}$ are determined by an iterative algorithm as IRLS with adequate design and weight matrices. We can find in [9] several methods in order to estimate the bandwidth matrix in the case $p = 1$ and $G = 1$.

Due to the curse of dimensionality, surface smoothing techniques are not very useful in practice when there are more than two or three predictors variables. Indeed this problem refers to the fact that a local neighborhood in higher dimensions is no longer local. To deal with the curse of dimensionality problem, we propose to consider the generalized linear single-index models.

2.5 Generalized linear single-index models

In order to overcome the dimensionality problem, a popular way consists in first projecting all the predictors X onto a linear space spanned by the predictors and in fitting a nonparametric curve to their linear combinations. As in the previous subsection, we introduce these models for any G although they are generally presented for $G = 1$. That leads to the linear single-index model

$$Y = \mu(X) + \epsilon, \quad \mu_k(X) = \tilde{\mu}_k({\beta^{(k)}}^T X), \quad k = 1, \ldots, G, \tag{10}$$

with $\mathbb{E}(\epsilon|X) = 0$ almost surely and where $\tilde{\mu}_k$ (resp. μ_k) are functions defined over $\mathbb{R}$ (resp. $\mathbb{R}^p$).

Clearly the scale of ${\beta^{(k)}}^T X$ in $\tilde{\mu}_k({\beta^{(k)}}^T X)$ can be chosen arbitrarily: for any $c_k > 0$ $(\beta^{(k)}, \tilde{\mu}_k(\cdot))$ and $(c_k\beta^{(k)}, \tilde{\mu}_k(\cdot/c_k))$ leads to the same regression function. On the other hand, we have

$$\mathbb{E}[\nabla\mu_k(X)] = \mathbb{E}\left[\nabla\{\tilde{\mu}_k({\beta^{(k)}}^T X)\}\right] = \mathbb{E}\left[\{\nabla\{\tilde{\mu}_k\}({\beta^{(k)}}^T X)\}\right] \beta^{(k)}.$$

For identifiability purposes we propose to set $\beta^{(k)} = \mathbb{E}(\nabla\mu_k(X))$.

3 Estimation method

Our goal is to estimate $\beta^{(k)}$ and η_k for $k = 1, \ldots, G$. First we consider our approach in the asymptotical context and give the properties of the resulting estimate. In the applications to microarray data, we are far from the asymptotical case, and we propose to modify our approach by introducing a quadratic regularization term as ridge penalization. After discussing the computational aspects and choice of the hyperparameters, we compare our procedure with (r)OPG [25]. Notice that even if we consider applications with small n, we think that it is important to study asymptotic properties of our estimator, to justify their use.

3.1 Asymptotical view

3.1.1 binary case

We first consider the case $G = 1$ and consequently we temporally omit the index k which takes only one value equal to 1. We have $\beta = \mathbb{E}((g^{-1})'(\eta(X))\nabla\eta(X))$. The idea developed here is to estimate η and $\nabla\eta$ by their maximum local likelihood estimator $\hat{\eta}$ and $\widehat{\nabla\eta}$. Then β is estimated by the empirical mean of the variables $(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i)$. To end, $\widehat{\widehat{\eta}}$ is given by maximum local likelihood estimator computed from the sample $(\widehat{\beta}^T X_1, Y_1), \ldots, (\widehat{\beta}^T X_n, Y_n)$ and $\hat{\mu}(x) = g^{-1}(\widehat{\widehat{\eta}}(\widehat{\beta}^T x))$.

Our procedure for estimating β and η is described in the following algorithm.

Algorithm 1

Step A: *For $j = 1, \ldots, n$, find $\hat{\eta}(X_j) = \hat{a}_{(0,\ldots,0)}(X_j)$ and $\widehat{\nabla\eta}(X_j) = (\hat{a}_{e_1}(X_j) \ldots, \hat{a}_{e_d}(X_j))^T$, by maximizing*

$$\sum_{i=1}^{n} \mathcal{L}\left[\sum_{\underline{l}\in\mathcal{A}_q} a_{\underline{l}}(X_i - X_j)^{\underline{l}}, Y_i\right] K_H^p(X_i - X_j), \tag{11}$$

with respect $a_{\underline{l}}$, $\underline{l} \in \mathcal{A}_q$. We put

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i).$$

Step B: *Find $\hat{\eta}(x) = \hat{a}_0$ by maximizing*

$$\sum_{i=1}^{n} \mathcal{L}\left[a_0 + a_1(\hat{\beta}^T(X_i - x)), Y_i\right] K_{h_B}^1(\hat{\beta}^T(X_i - x)), \tag{12}$$

with respect a_0 and a_1.

This procedure involves the choice of a smoothing parameter at two different levels and also the one of the order q of the polynomial approximation in step A. This point will be developed in Subsection 3.3.

We are now considering asymptotic properties of the direction estimator in Step A. We first introduce some notations. Let $l_i(u, v) = (\partial^i/\partial u^i)\mathcal{L}(g^{-1}(u), v)$; l_i is linear in v for fixed u, and

$$l_1(\eta(x), \mu(x)) = 0, \quad l_2(\eta(x), \mu(x)) = -\rho(x),$$

where $\rho(x) = \{g'(\mu(x))^2 V(\mu(x))\}^{-1}$ and $V(\mu(x)) = \mathrm{Var}(Y|X = x)$. We suppose that the following conditions are satisfied.

Conditions

1. The vector of covariates X have a density f with compact support S_X.
2. The functions $f(\cdot)$, $D^{\underline{k}}\eta(\cdot)$, $|\underline{k}| = q + 2$, $D^{\underline{k}}f(\cdot)$ and $D^{\underline{k}}\rho(\cdot)$ for $|\underline{k}| = 1$, are continuous for $x \in S_X$.
3. The function $l_2(u, v) < 0$ for $u \in \mathbb{R}$ and v in the range of the response variable.
4. The functions $l_1(u, v)$ and $(g^{-1})''$ are bounded and $(g^{-1})''$ is continuous.
5. $\mathrm{Var}(Y|X = x) \neq 0$, and $g'(\mu(x)) \neq 0$, for $x \in S_X$. We also assume that
$$\inf_{x \in S_X} (\rho(x) f(x)) > 0.$$
6. The Kernel K^p is a probability density having compact support S_{K^p}. The bandwidth is a diagonal matrix $H = h\mathrm{Id}_p$. We assume that $h = cn^{-\alpha}$, c being a constant, and we also assume that nh^{p+2} goes to infinity. This last condition leads to the constraint $\alpha < 1/(p + 2)$.
7. We assume the existence of $\delta > 0$ such that
$$\frac{1 - 2\delta}{2q + p + 2} < \alpha < \frac{1 - 2\delta}{p + 2}. \tag{13}$$

These conditions are quite classical for this kind of models. Conditions 1 and 2 correspond to classical assumptions on the covariates. The conditions 3 and 4 insure the concavity of the objective function $\mathcal{L}$ and regularity assumptions on $\mathcal{L}$ and the link function. If the canonical link is used and if the variance is correctly specified (see condition 5) then the condition 3 holds. In this case, condition 3 is equivalent to $b'' > 0$, which holds because of $\mathrm{Var}(Y|X = x) = b''(\theta(x))$. The condition 5 is not restrictive, it is a condition on the GLM, on the distribution of X and the second derivative of the objective function. The assumption on Kernel and bandwidth (condition 6) are quite usual. Note that conditions 3 and 5 imply that ρ is strictly positive over S_X, the support of covariables. The constraints given on alpha in the last two conditions are quite close to the choice that would be made for estimation without penalization, and the choice to achieve a good balance between bias and variance is

$$\alpha = \frac{1}{2q + p + 2}.$$

Theorem 1 (*Consistency of the estimator $\hat{\beta}$*) *Under the foregoing conditions, $\hat{\beta}$ is a consistent estimator of β which means that*

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} P(|\hat{\beta} - \beta| > \epsilon) = 0.$$

The proof of Theorem 1 is given in Appendix 5.2. In this theorem, we obtain the consistency of the direction estimator in stepA, but we do not have any convergence rate. We already tried to obtain a speed in $\sqrt{n}$ but our method of proof requires then constraints such as $q > p/2$. Considering that in practice we have $q = 1$ or 2, such a constraint is unacceptable, and the task of obtaining better rates is left for future research.

3.1.2 Multicategorial outcomes

When $G > 1$, we can estimate the directions as in the step A. Indeed, in the logistic regression model for example, we have

$$\beta^{(k)} = \mathbb{E}\left(\sum_{m=1}^{G} \psi_{k,m}(X)\nabla\eta_m(X)\right),$$

where

$$\psi_{k,m}(X) = \frac{\exp(\eta_k(X))}{1+\sum_{l=1}^{G}\exp(\eta_l(X))}\left[\delta_{k,m} - \frac{\exp(\eta_m(X))}{1+\sum_{l=1}^{G}\exp(\eta_l(X))}\right],$$

and $\delta_{k,m} = 1$ if $k = m$, 0 otherwise. So $\beta^{(k)}$ can be estimated by the corresponding empirical mean. On the other hand, the application of the step B is not straightforward. Recall that $a_0 + a_1(\hat{\beta}^T(X_i - x))$ in (12) corresponds to the local linear expansion of $\tilde{\eta}$ at $\hat{\beta}^T x$. The role of $K^1_{h_B}(\hat{\beta}^T(X_i - x))$ is to weight strongly the point for which this expansion is valid. For $G > 1$, at each point (X_i, Y_i), we have G local linear expansions: $\tilde{\eta}_k$ at $\hat{\beta}^{(k)T}x$ and the difficulty stands in the determination of the weight. A natural choice is a weight equal to $K^G_{H_B}(\{\hat{\beta}^{(k)T}(X_i - x)\}_{k=1,\ldots,G})$ but the determination of the matrix H_B is a difficult problem still open. So we propose to replace the nonparametric fit by a parametric one. This fit is slightly different from the one presented in the Subsection 2.3.1: the bloc $\mathbf{X}^{(G)}_{\iota_k+1:\iota_k+G,:}$ is replaced by

$$\mathbf{X}^{(G)}_{\iota_k+1:\iota_k+G,:} = \begin{bmatrix} \hat{\beta}^{(1)T}x_k & 0 & \cdots & 0 \\ 0 & \hat{\beta}^{(2)T}x_k & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \hat{\beta}^{(G)T}x_k \end{bmatrix}.$$

Notice that, under similar assumptions and using exactly the same arguments as for the binary case, we can prove the consistency of the resulting estimators $\hat{\beta}^{(k)}$, $k = 1, \ldots, G$.

3.2 Small n and large p case

The maximization of (11) corresponds to the research of a weighted log-likelihood maximum, with weights $(K^p_H(X_i - X_j))_i$ and with a particular design matrix with columns growing with q. In practice, it is determined by an algorithm as IRLS (Section 2.3.2). In Subsection 2.3.3, we have recalled that the maximum does not always exist. In particular for the logistic regression, the likelihood maximum may never exist when the rank of the design matrix is equal to n (Section 2.3.3). In the applications considered here, this rank is given by n. Two approaches can be considered to avoid this problem. Firstly, we can select some covariables. Secondly, we can introduce a ridge type penalty in the weighted log-likelihood.

One could have chosen to select some genes instead of introducing a penalty term. Nevertheless, we tried our algorithm without penalty for microarray data, by selecting genes, and then the best results were obtained for a very small number of genes (two or three). These results were less good than those obtained when introducing a penalty term. Moreover, the results were worsening quite fast when increasing the number of genes considered. In fact, the noise triggered by the increasing number of genes is not compensated by the weighting which is not sufficient to correct the variability introduced by the number of genes. To obtain a better stability, one needs a large value of h parameter.

So in this paper we chose to introduce a ridge penalty, solution which allows to consider all the covariables. Moreover, a ridge penalty for the gradient estimate has been introduced by [24] in the gaussian context when $p = 1$. This penalty allows to solve problems such as sparse or clustered design. We expect to encounter this situation in the classification problems. So we believe that this penalty is needed even for large n.

3.3 Computational view and hyperparameters selection

In the first step of the algorithm, we must chose the order q of the polynomial approximation. In practice we retain a linear fit as for the rOPG method introduced in [25]. Moreover, we use in the step A a product kernel in order to reduce the curse of dimensionality; that leads to take a diagonal bandwidth matrix H. The kernels are gaussian.

Recall that the covariables are not standardized to have unit sample variance. Let denote by Σ^2 the diagonal matrix such that $\Sigma^2_{i,i}$ is given by the empirical variance of the covariable X_i. When the covariables are measured in different units, it is typically recommended that the variables be first standardized to make the penalty meaningful. That yields to use the norm $\|u\|^2_{\Sigma^2} = u^T\Sigma^2 u$ for the coefficients associated with the gradient. Precisely, for example in the case of $G = 1$, that consists in replacing (11) by

$$\sum_{i=1}^{n} \mathcal{L}\left[a_{\vec{0}} + b^T(X_i - X_j), Y_i\right] K_H^p(X_i - X_j) - \frac{\lambda}{2} b^T\Sigma^2 b,$$

where $b = (a_{e_1}, \ldots, a_{e_p})^T$. Such a penalty strongly penalizes the gradient in the directions that are the most variable. Notice that we make this standardization even in the case of expression arrays where for example the variables are all measured in the same unit and where this standardization is optional. Let $\mathbf{X}_s^{(1)} = (\mathbf{X}^{(1)} - \mathbb{1}_n \mathbb{1}_n^T \mathbf{X}^{(1)}/n)\Sigma^{-1}$ be the standardized design matrix. We denote by $\hat{\beta}_s^\lambda$ (resp. $\hat{\beta}^\lambda$) the estimator corresponding to $\mathbf{X}_s^{(1)}$ (resp. $\mathbf{X}^{(1)}$) with bandwidth matrix H_s (resp. H) and the usual euclidian (resp. $\|\cdot\|^2_{\Sigma^2}$) norm in the penalty term. We can show that $\hat{\beta}^\lambda = \Sigma^{-1}\hat{\beta}_s^\lambda$ for $H_s = H\Sigma^{-1}$. Then it is natural to compute the estimator by using the standardized design matrix. So we consider $H_s = h_A \mathrm{Id}_p$, and then we reduce the number of hyperparameters to chose.

Our procedure involves the choice of a smoothing parameter at two different levels. At the first one, the aim is the estimation of η and its gradient, and the bandwidth h_A should be optimal in this respect. When we need to penalize the log-likelihood, we must also determinate the regularization parameter λ. We opt for cross-validation. At the second step, we want to estimate $\tilde{\eta}$ and h_B should be optimal for this task. For this most common choice, we opt for the plug-in method proposed by [9]. In the applications to microarrays, the projections of the covariables on

the estimated direction sometimes split into two groups with very close numerical values within each group. In this case, it is not possible to compute h_B by the plug-in method and it is natural to use a parametric fit.

For example when $G = 1$, the procedure, presently derived in $\mathbb{R}^p$ can be equivalently derived in $\mathbb{R}^r$ where $r = \text{rank}(\mathbf{X}^{(1)}) \leq n$. For this purpose, we compute the Singular Value Decomposition UDV' (SVD) of $\mathbf{X}_s^{(1)}$, the standardized design matrix and collect the first r columns of UD in $\mathbf{X}_s^{(\text{red},1)} = (UD)_{:,1:r}$. It is readily seen that GSIM, run by replacing $\mathbf{X}_s^{(1)}$ by $\mathbf{X}_s^{(\text{red},1)}$, yields an estimate $\hat{\beta}^{\text{red},\lambda}$ uniquely related to $\hat{\beta}_s^{\lambda}$ by $\hat{\beta}_s^{\lambda} = V\hat{\beta}^{\text{red},\lambda}$. Hence, up to a SVD, the procedure is independent of p which is of computational importance. When $G > 1$, we have to use the matrix $\mathbf{X}_s^{(\text{red},1)}$ to construct $\mathbf{X}_s^{(\text{red},G)}$ and $\hat{\beta}_s^{(k),\lambda} = V\hat{\beta}^{\text{red},(k),\lambda}$ for $k = 1, \ldots, G$.

3.4 Comparison with rOPG

The procedure (r)OPG, introduced by [25], estimates the estimated effective dimension reduction space spanned by κ ($\kappa \ll p$) orthogonal directions. When $\kappa = 1$, this method is equivalent to step A of GSIM but with a least square criterion used instead of log-likelihood one. The direction is estimated by the vector associated to the largest eigenvalues of the empirical estimation of $\mathbb{E}(\nabla\mu(X)\nabla\mu(X)^T)$. The refined version, denoted by rOPG, consists in iterating until convergence a step A' (after the step A) defined by the following instructions. Put $\{\hat{\beta}\}^0 = \hat{\beta}$ and at the k-th iteration, $\{\hat{\beta}\}^k$ is obtained as in step A with weight $K_h^1(\{\hat{\beta}\}^{(k-1)T}(X_i - X_j))$ instead of $K_H^p(X_i - X_j)$. Indeed, in the linear single-index model, the function η is constant in the direction orthogonal to β, so we can stretch the window in this direction. That makes the kernel $K_h^1(\beta^T(\cdot - X_j))$.

Even if the (r)OPG method can handle any type of response variable, the least square criterion seems to be well adapted to gaussian situations but not to a categorical response. For example, when applied to categorical responses, we have observed that the results depends on the labels. Precisely, if we switch the labels, the direction estimation changes. That could lead to classification results very different according to the labels order (see Subsection 4.2). That does not occur for GSIM.

In [25], the procedure (r)OPG is not applied in high dimension data problems. In such cases, for the gradient estimation, the method amounts in projecting (in some geometry inducted by the weights computed from the kernel) one vector of length n on a space of dimension n. So the projection is the vector itself and does not depend on the bandwidth. In fact, it is equivalent to consider the parametric model $\mu(X) = \alpha + \beta X$ and to estimate β by likelihood maximum. It seems to be not very interesting to do that in our situation. So we propose to introduce, as for GSIM, a ridge penalty. Notice that in the code proposed by the authors, there exists one term of numerical stabilization corresponding to one ridge penalty fixed to 0.0001. Our approach only makes this constant a parameter λ.

In a classification goal, we have applied the same step B of GSIM after estimating β to make comparable the methods. So we show that taking account the relationship between expectation and variance in GLM, we improve the results. As for GSIM, we use cross-validation technique to determine the hyperparameters and plug-in method for compute the bandwidth of the second step.

Notice that this classification rule is slightly different from the method used in [3] where the dimension reduction

is obtained with MAVE another method developed by [25]. In fact, rOPG is a simplified version of MAVE and we have observed that both these methods lead to comparable performances. We have considered rOPG since it is most directly related to the step A of GSIM.

Concerning the numerical implementation, the procedure (r)OPG is stable by reparameterization using SVD as is GSIM. Only OPG is invariant up to the columns standardization. However we have choosen to standardize the design matrix to compare fairly the methods.

4 Numerical examples

4.1 Asymptotic study

In this subsection, we use one binary regression example to demonstrate the relation between estimation errors and the bandwidth for step A of GSIM and to check the asymptotic performance of our estimation.

As in the rOPG method, we fix $\lambda = 10^{-4}$ to stabilize numerical computations. Then the parameter h_A is determined by cross-validation on the mean-squared error. Indeed ([16]) the choice of the error criterion depends mainly on the way that the model is used to predict future observations. Notice that when n is very large (500 for example) cross-validation requires very long computational time. In this case, since n is very large we can randomly split the sample into a learning set and a test set for which we compute the mean-squared error as above. We recall that the bandwidth h_B is determined by plug-in method.

We consider the following binary regression model:

$$\eta(X_i) = \tilde{\eta}(\beta^T X_i), \quad \beta = (-1, 1, 1, 1)^T/2, \quad \tilde{\eta}(u) = 3\sin(2u).$$

The covariables $X_i = (X_i^{(1)}, \ldots, X_i^{(4)})^T$ are such that $X_i^{(1)} \sim \mathcal{N}(0.5; 1)$ and $X_i^{(k)} \sim \mathcal{N}(0; 1)$, $k = 2, 3, 4$. The components of X_i are independent. We run 100 replicates of the observation sequence of sizes $n = 50, 100, 250$ and 500. Concerning step A, we define the estimation error as $\|\hat{\beta} - \beta\|^2$ (where $\|\cdot\|^2$ is the usual euclidian norm). With different sample sizes and bandwidths, the average errors are shown in Figure 1. The vertical lines are the corresponding average of cross-validation bandwidths. This figure shows that the estimation procedure works well and cross-validation bandwidth is applicable to parameters estimation. Simulation results are listed in Table 1.

In order to evaluate the step B, the performance of the estimator $\widehat{\tilde{\eta}}$ is assessed via the square root of average squared errors (RASE)

$$RASE = \left(\frac{1}{n_{grid}} \sum_{j=1}^{n_{grid}} [\widehat{\tilde{\eta}}(u_j) - \tilde{\eta}(u_j)]^2 \right)^{1/2},$$

where $\{u_j, j = 1, \ldots, n_{grid}\}$ are the grid points at which the function $\tilde{\eta}$ is estimated. Here we have taken 50 points uniformly spaced in the range $[-1.5; 1.5]$. The results are given in Table 1. Figure 2 summarizes typical performance of the estimators of the function $\tilde{\eta}$ for $n = 250$ and 500.

4.2 Applications to microarray data

4.2.1 Supervised learning

The goal of the supervised learning is to predict the labels of some sample (like tumor class) from its gene expression profile. The classes are predefined and the task is to understand the basis for the classification from a set of labeled objects (training or learning set). This information is then used to classify future observation. This classification problem can be viewed as a regression one. As seen, categorical outcomes belong to generalized linear models family. For a new gene expression profile x, we can compute from the learning set, $\hat{\eta}(x) \in \mathbb{R}^G$ where $\hat{\eta}$ is any estimator of the predictor η. Therefore the classification rule consists in predicting the class by that which gives the largest likelihood. This is equivalent to predict class i if and only if

$$[\hat{\eta}_i(x) \geq \hat{\eta}_l(x), \quad \forall l \in \{0, \cdots, G\}],$$

where $\hat{\eta}_0 = 0$ by convention. Hence methods such as rOPG or GSIM may be used in applications to classification of microarrays data.

4.2.2 Comparison methods

We compare the classification results from our procedure to those of other classifiers including rOPG, diagonal linear discriminant analysis (DLDA), diagonal quadratic discriminant analysis (DQDA) and k-nearest neighbors (KNN) based on the Euclidean distance (see [5] for an overview of these last three methods).

DLDA, DQDA and KNN are thus introduced in the present paper as "classical statistical method". Comparing our method with rOPG, we show how to improve the results obtained by rOPG by taking into account the GLM structure.

4.2.3 Data and pre-processing

We will consider in turn two data sets.

Colon[1]**:** The Colon data set contains 62 tissue samples with 2000 genes: 40 tumors tissues, coded 1, and 22 normal tissues, coded 0 (see [2] for more details).

Leukemia[2]**:** The Leukemia data set, contains 72 tissue samples with 7129 genes: 47 cases of acute lymphoblastic leukemia (ALL), coded 0, and 25 cases of acute myeloid leukemia (AML), coded 1 (see [12] for more details). Furthermore, we can also treat these data as a multi-class problem by considering both type B (38 samples) and T (9 samples) of the ALL case.

SRBCT[3]**:** This data set consists of microarray experiments of small round blue cell tumors (SRBCT) of childhood cancer (see [15]). It contains 88 samples with 2308 genes: 29 cases of Ewing sarcoma (EWS), coded 1, 11 cases of

[1] `http://microarray.princeton.edu/oncology/affydata/index.html`

[2] `http://www.broad.mit.edu/cancer/software/genepattern/datasets/`

[3] `http://www.thep.lu.se/pub/Preprints/01/lu_tp_01_06_supp.html`

Burkitt lymphoma (BL), coded 2, 18 cases of neuroblastoma (NB), coded 3, 25 cases of rhabdomyosarcoma (RMS), coded 4. A total of 63 training samples and 25 test samples are provided. Five of the test set are non-SRBCT and are not considered here.

Pre-processing: For Leukemia and Colon data, the pre-processing steps recommended in [7] are applied: thresholding (floor of 100 and ceiling of 16000)/ filtering (exclusion of genes with max/min$\leq$ 5 and (max-min)$\leq$ 500 / $\log_{10}$-transformation / standardization in row (each sample is centered and normalized). Notice that this last step is essential to have microarrays at the same scale. The goal of this standardization differs from the one of the standardization "in column" in order to avoid identifiability problem and to have good regression behavior. These pre-processing steps yield a resulting number of covariates depending on the subdivision Learning and Testing set, lower than the initial number of genes but still far larger than the number of observations. Notice that the SRBCT data do not need pre-processing.

4.2.4 Assessing prediction methods

Resampling study: It is common to assess the performance of the classification rules for a selected subset of genes by their errors on the test set and also by their leave-one-out cross-validated errors. Due to the instability of leave-one-out error rates, we perform a re-randomization study *i.e.* an out-of-sample analysis on 100 random subdivisions of the data set into a learning set and a test set. For the Colon data no learning and test sets are available and we have chosen a test set size equal to one third of the data (2:1 scheme of [7]). In each learning test, each subclass is represented with the same proportion as in the total population. For the Leukemia and SRBCT data, a test set is available and we randomly split the original data set into a training set and a test set of the same size as the original ones. Here each subclass is represented with the same proportion as in the original learning set (for example 19 ALL-Bcell, 8 ALL-Tcell and 11 AML for Leukemia data). We use the same subdivisions for Leukemia data, when we treat these data as two or three classes problem. Each subdivision yields a test set error rate for each predictor; Boxplots are used to summarize these error rates over the runs.

Hyperparameters choice:

The optimal number of neighbors k for KNN method is determined by a cross-validation technique based on the misclassified rate. The k range for is given by $\mathcal{K}_l = \{1, \ldots, 20\}$. Moreover, bandwidth and regularization parameter (for rOPG or GSIM step A) are simultaneously determined by cross-validation techniques based on misclassified rates. For rOPG (resp. GSIM), we use 5 log-linearly spaced points in the range $[10^{-3}; 30]$ (resp. $[10^{-3}; 30]$) for λ and in the range $[0.5; 6]$ (resp. $[0.8; 100]$) for h_A. The h_A range for the both methods differs. Indeed, the bandwidth for rOPG method correspond to univariate kernel since in the refined version of OPG, the kernel is computed over covariables after projection. Note that, to fairly evaluate and compare the methods, pre-processing and (hyper)parameters estimations are performed on the training set (at each step of the cross-validation process).

4.2.5 Discussion

Misclassification rates results are reported in tables 2 to 6 and boxplots are plotted in figures 3 to 6. For Colon data, one can observe that DLDA, DQDA and GSIM methods leads to similar misclassification rate with slightly best results for DLDA. However, for Leukemia data, one can notice that GSIM seems to provide best results than all the other methods. If DLDA remains good for two class data, when we consider three classes, GSIM misclassication rate is significantly smaller. For SRBCT data, the results of GSIM become very good compared to other methods. This method seems to be the most relevant method for categorical data.

When proposing the GSIM method, the aim was to improve rOPG results by developing a method adapted to categorical data. This purpose seems to be fully reached: the results are a lot better. Furthermore, it is important to point out a serious drawback of the rOPG method: if we switch the data labels (that is, for example, replacing 0 by 1 and 1 by 0 for two classes data), we observe differences in the results. For example, in table 5, we give the results for rOPG method, considering all the possible different labels orders for Leukemia data with 3 classes. Misclassification rate can be almost doubled depending on the labels order. We also observe differences for two classes data when switching labels: a mean error rate of 0.061 instead of 0.052 for Leukemia data and 0.201 instead of 0.199 for Colon data. Moreover, even when the misclassification rates are quite the same globally, they often differ in detail, the misclassified samples are not the same. That confirms that such a method is not suitable for categorical data.

The results obtained with KNN method are slightly less good than those obtained with DLDA, DQDA and GSIM methods. Besides, in practice we observe many cases of indecisions. We really believe that the frequent occurrence of the indecision case shows that KNN is not a pertinent method (for this kind of data sets). The weakness of this classical statistical method is clearly illustrated by the numerical results. This problem probably refers to that in higher dimensions, nearest neighbors are not in a local neighborhood.

We also want to stress on the importance of the standardization in row of the data, that we see as a normalization between the microarrays. Note that this treatment is not needed for SRBCT data which is already conveniently preprocessed. DLDA and DQDA methods are very sensitive to this standardization in row: if the results obtained are quite good when the standardization is done, they deteriorate when this pre-processing step is suppressed. For example in Colon data case, we obtain a mean error rate of 0.286 (instead of 0.144) for DLDA and 0.314 (instead of 0.154) for DQDA. GSIM method has showed a better stability respect to this standardization in row step, the results are almost the same when the standardization is not done: 0.170 (instead of 0.155).

Thus, GSIM method provides good results for all the considered data sets, especially for Leukemia data with three classes and SRBCT data. We may expect to still have good results even with more classes which is not the case for the other methods tested.

5 Appendix

In this section, we give the proof of Theroem 1. In order to obtain asymptotical properties of $\hat{\beta}$, we need some asymptotical properties of the estimators of $D^{\underline{k}}\eta(x)$. These properties are given by two lemma presented and proved in Section 5.1. Theorem is proved in Section 5.2.

5.1 Proof of two lemma

The proofs of the two lemmas presented in this appendix are quite similar as in the univariate case, notations are just becoming a little more complicated (for the univariate case, see [10, 9]). Here we assume that the conditions of the Theorem 1 are verified.

We consider the normalized estimator $\hat{a}^*(x)$ which is a vector of length equal to the cardinal of $\mathcal{A}_q$ and with component $\underline{l} \in \mathcal{A}_q$ given by

$$c_n^{-1} h^{|\underline{l}|}[\hat{a}_{\underline{l}}(x) - D^{\underline{l}}\eta(x)/\underline{l}!],$$

where $c_n = (nh^p)^{-1/2}$. It can easily be seen that $\hat{a}^*(x)$ maximizes

$$\sum_{i=1}^{n} \mathcal{L}\left[g^{-1}\left(\bar{\eta}(x, X_i) + c_n a^{*T} R(x, X_i)\right), Y_i\right] K^d(H^{-1}(X_i - x)),$$

as a function of a^*, where

$$\bar{\eta}(x, X_i) = \eta(x) + \sum_{\underline{k} \in \mathcal{A}_q \setminus \mathcal{A}_0} D^{\underline{k}}\eta(x)(X_i - x)^{\underline{k}}/\underline{k}!$$

and

$$R(x, X_i) = \left\{(H^{-1}(X_i - x))^{\underline{k}}\right\}_{\underline{k} \in \mathcal{A}_q}.$$

Equivalently, $\hat{a}^*(x)$ maximizes

$$\mathcal{L}_n(a^*) = \sum_{i=1}^{n} \left(\mathcal{L}\left[g^{-1}\left(\bar{\eta}(x, X_i) + c_n a^{*T} R(x, X_i)\right), Y_i\right] - \mathcal{L}\left[g^{-1}\left(\bar{\eta}(x, X_i)\right), Y_i\right]\right) K^p(H^{-1}(X_i - x)).$$

Condition 3 implies that the function $\mathcal{L}_n$ is concave in a^*. A Taylor series expansion of $\mathcal{L}\left(\left[g^{-1}(\cdot), Y_i\right]\right)$ leads to

$$\mathcal{L}_n(a^*) = W_x^T a^* + \frac{1}{2} a^{*T} A_n a^* + \frac{c_n^3}{6} \sum_{i=1}^{n} l_3(\eta_i, Y_i)(a^{*T} R(x, X_i))^3 K^p(H^{-1}(X_i - x)), \tag{14}$$

where η_i is between $\bar{\eta}(x, X_i)$ and $\bar{\eta}(x, X_i) + c_n a^{*T} R(x, X_i)$,

$$W_x = c_n \sum_{i=1}^{n} l_1\left[\bar{\eta}(x, X_i), Y_i\right] R(x, X_i) K^p(H^{-1}(X_i - x)),$$

and

$$A_n = (c_n)^2 \sum_{i=1}^{n} l_2\left[\bar{\eta}(x, X_i), Y_i\right] R(x, X_i)^T R(x, X_i) K^p(H^{-1}(X_i - x)).$$

Let $\mathcal{D}$ be the set defined by $\{u; x + Hu \in S_X\} \cap S_{K^p}$.

Also, define

$$\Sigma_x = \left\{\rho(x) f(x) \nu_{\underline{l}+\underline{k}}\right\}_{\underline{l}\in\mathcal{A}_q, \underline{k}\in\mathcal{A}_q}.$$

where $\nu_{\underline{l}} = \int_{\mathcal{D}} u^{\underline{l}} K^p(u) du$.

Lemma 1 *Under the foregoing conditions,*

$$\sup_{x\in S_X} \left|\hat{a}^* - \Sigma_x^{-1} W_x\right| \quad \xrightarrow[n\to\infty]{P} \quad 0 \tag{15}$$

Proof. First, we prove that $A_n = -\Sigma_x + o_P(1)$. This can be shown using the fact that, for $\underline{l}$ and $\underline{k}$ in $\mathcal{A}_q$,

$$(A_n)_{\underline{l},\underline{k}} = (\mathbb{E}A_n)_{\underline{l},\underline{k}} + O_P\left[\left\{\mathrm{Var}\,(A_n)_{\underline{l},\underline{k}}\right\}^{\frac{1}{2}}\right].$$

The mean in the above expression equals to

$$(\mathbb{E}A_n)_{\underline{l},\underline{k}} = \int_{\mathcal{D}} l_2\left[\bar{\eta}(x, x+Hu), \mu(x+Hu)\right] f(x+Hu) K^d(u) u^{\underline{l}+\underline{k}} du.$$

Because the support of K^p is compact,

$$\bar{\eta}(x, x+Hu) = \eta(x+Hu) - \sum_{|\underline{k}|=q+1} D^{\underline{k}}\eta(x)(Hu)^{\underline{k}}/\underline{k}! + o(h^{q+1}),$$

uniformly in x. Using a Taylor expansion of l_2 about $(\eta(x+Hu), \mu(x+Hu))$, we obtain

$$(\mathbb{E}A_n)_{\underline{l},\underline{k}} = -\int_{\mathcal{D}} \rho(x+Hu) f(x+Hu) K^p(u) u^{\underline{l}+\underline{k}} du + o(h^q).$$

Next the use of a Taylor expansion of ρf about x leads to

$$(\mathbb{E}A_n)_{\underline{l},\underline{k}} = -\rho(x) f(x) \nu_{\underline{l}+\underline{k}} + O(h) = -\Sigma_x + O(h),$$

uniformly in x. Similar arguments show that $\mathrm{Var}\left(\{A_n\}_{\underline{l},\underline{k}}\right) = O((nh^p)^{-1})$ and the last term of (14) is $O_P(\{nh^p\}^{-1/2})$, and the condition $nh^{p+2} \to \infty$ involves that:

$$\mathcal{L}_n(\underline{a}^*) = W_x^T \underline{a}^* - \frac{1}{2} \underline{a}^{*T} \Sigma_x \underline{a}^* + o_P(1), \tag{16}$$

uniformly in $x \in S_{\underline{X}}$.

By the Convexity Lemma (see [22]), the equation (16) holds uniformly in $a^* \in C$ for any compact set C and we can apply the Lemma A.1 of [4] which yields

$$\sup_{x\in S_X} \left|\hat{a}^* - \Sigma_x^{-1} W_x\right| \xrightarrow[n\to\infty]{P} 0.$$

Now we compute the first two moments of W_x. Precisely we have the following lemma:

Lemma 2 *Under the foregoing conditions,*

$$\mathbb{E}\left(\{W_x\}_{\underline{l}}\right) = \rho(x)f(x)\left[\left(nh^{2q+2+p}\right)^{\frac{1}{2}}\sum_{|\underline{k}|=q+1}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\nu_{\underline{l}+\underline{k}}\right.$$
$$\left. + \left(nh^{2q+4+p}\right)^{\frac{1}{2}}\left(\sum_{|\underline{k}|=q+1}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\sum_{|\underline{m}|=1}\frac{D^{\underline{m}}(\rho f)(x)}{(\rho f)(x)}\nu_{\underline{l}+\underline{k}+\underline{m}} + \sum_{|\underline{k}|=q+2}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\nu_{\underline{l}+\underline{k}}\right)\right]$$
$$+ o(\left[nh^{2p+4+d}\right]^{\frac{1}{2}})$$

$$Cov\left(\{W_x\}_{\underline{l},\underline{k}}\right) = \frac{f(x)\,Var(Y|X=x)}{[V\{\mu(x)\}g'\{\mu(x)\}]^2}\int_{\mathcal{D}}K^p(u)^2u^{\underline{l}+\underline{k}}du + o(1).$$

Remark Note that for this lemma we have kept two terms to express the expectation of $\{W_x\}_{\underline{l}}$, the first term is $O(nh^{2q+2+p})$ and the second one is $O(nh^{2q+4+p})$. Thus the second term is negligible compared to the first one; in some cases, problems can occur depending on the order q of the development. If one uses a symmetrical kernel (which is often the case), one can obtain some nul moments $\nu_{\underline{l}+\underline{k}}$. Thus, if q is even, all the $\nu_{\underline{l}+\underline{k}}$ will be zero and the second term becomes necessary. If q is odd, the second term is equal to zero. In the sequel of the proof, we will consider that we have $O(nh^{2q+2+p})$.

Proof. By Taylor's expansion,

$$l_1\left[\bar{\eta}(x,x+Hu),\mu(x+Hu)\right] = \rho(x+Hu)\sum_{q+1\leq|\underline{k}|\leq q+2}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}[Hu]^{\underline{k}} + o(h^{q+2}),$$

and we obtain

$$\mathbb{E}\left(\{W_x\}_{\underline{l}}\right) = c_n^{-1}\left[\sum_{q+1\leq|\underline{k}|\leq q+2}h^{|\underline{k}|}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\int_{\mathcal{D}}f(x+Hu)\rho(x+Hu)u^{\underline{l}+\underline{k}}K^p(u)du + o(h^{q+2})\right].$$

By a Taylor expansion of ρf about x, we obtain

$$\begin{aligned}\mathbb{E}\left(\{W_x\}_{\underline{l}}\right) &= c_n^{-1}\left[(\rho f)(x)\sum_{q+1\leq|\underline{k}|\leq q+2}h^{|\underline{k}|}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\nu_{\underline{l}+\underline{k}}\right.\\ &+ \left.\sum_{q+1\leq|\underline{k}|\leq q+2}h^{|\underline{k}|+1}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\sum_{|\underline{m}|=1}D^{\underline{m}}(\rho f)(x)\nu_{\underline{l}+\underline{k}+\underline{m}}\right] + o(c_n^{-1}h^{q+2}).\end{aligned}$$

That leads to

$$\begin{aligned}\mathbb{E}\left(\{W_x\}_{\underline{l}}\right) &= c_n^{-1}(\rho f)(x)\left[h^{q+1}\sum_{|\underline{k}|=q+1}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\nu_{\underline{l}+\underline{k}} + h^{q+2}\sum_{|\underline{k}|=q+2}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\nu_{\underline{l}+\underline{k}}\right.\\ &+ \left.h^{q+2}\sum_{|\underline{k}|=q+1}\frac{D^{\underline{k}}\eta(x)}{\underline{k}!}\sum_{|\underline{m}|=1}\frac{D^{\underline{m}}(\rho f)(x)}{(\rho f)(x)}\nu_{\underline{l}+\underline{k}+\underline{m}}\right] + o(c_n^{-1}h^{q+2}).\end{aligned}$$

Since $c_n^{-1}h^{q+1} = \left(nh^{2q+2+p}\right)^{\frac{1}{2}}$ and then $c_n^{-1}h^{q+2} = \left(nh^{2q+4+p}\right)^{\frac{1}{2}}$, we obtain the first result.
The covariance between the $\underline{l}^{th}$ and $\underline{k}^{th}$ component of W_x is

$$\mathbb{E}(\{W_x\}_{\underline{l}}\{W_x\}_{\underline{k}}) + O(h^{2q+2+p}).$$

We have

$$\mathbb{E}(\{W_x\}_{\underline{l}}\{W_x\}_{\underline{k}}) = nc_n^2\mathbb{E}\left(l_1^2\left[\bar{\eta}(x,X_1),Y_1\right]R(x,X_1)_{\underline{l}}R(x,X_1)_{\underline{k}}\left[K^p\left(H^{-1}(X_1-x)\right)\right]^2\right).$$

With $\bar{\eta}(x,X_1) = \eta(x) + O(h^q)$ and using a Taylor expansion of l_1 in the first variable about $\eta(x)$, we obtain:

$$l_1^2\left[\bar{\eta}(x,X_1),Y_1\right] = l_1^2\left[\eta(x),Y_1\right] + O(h^q).$$

With $nc_n^2 = h^{-p}$, we have

$$\begin{aligned}\mathbb{E}(\{W_x\}_{\underline{l}}\{W_x\}_{\underline{k}}) &= h^{-p}\mathbb{E}\left[l_1^2\left(\eta(x),Y_1\right)R(x,X_1)_{\underline{l}}R(x,X_1)_{\underline{k}}\left[K^p\left(H^{-1}(X_1-x)\right)\right]^2\right]\\ &+ h^{-p}O(h^q)\mathbb{E}\left[R(x,X_1)_{\underline{l}}R(x,X_1)_{\underline{k}}\left[K^p\left(H^{-1}(X_1-x)\right)\right]^2\right].\end{aligned}$$

We are now considering the second term of the right member of this equation. We note it T_2.

$$T_2 = h^{-p}O(h^q)\int_{S_X} R(x,X_1)_{\underline{l}}R(x,X_1)_{\underline{k}}\left[K^p\left(H^{-1}(X_1-x)\right)\right]^2 f(X_1)dX_1$$

which leads to

$$\begin{aligned}T_2 &= O(h^q)\int_{\mathcal{D}} u^{\underline{k}+\underline{l}}\left[K^p(u)\right]^2 f(x+Hu)du\\ &= O(h^q) = o(1).\end{aligned}$$

and then

$$\mathbb{E}(\{W_x\}_{\underline{l}}\{W_x\}_{\underline{k}}) = h^{-p}\,\mathbb{E}\left(l_1^2\left(\eta(x),Y_1\right)R(x,X_1)_{\underline{l}}R(x,X_1)_{\underline{k}}\left[K^p\left(H^{-1}(X_1-x)\right)\right]^2\right) + o(1).$$

Using the definition of l_1, we do a Taylor expansion of the function ρf about x, and with the definition of $\rho(x)$, we obtain the second result of the lemma.

5.2 Proof of Theorem

In this section, we give the proof of Theroem 1 using the results of the two lemmas in Section 5.1. Notations are those introduced in 5.1.

We have

$$\beta = \mathbb{E}\left[(g^{-1})'(\eta(X))\nabla\eta(X)\right]$$

$$\widehat{\beta} = \frac{1}{n}\sum_{i=1}^{n}(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i)$$

and therefore

$$\hat{\beta}-\beta = \frac{1}{n}\sum_{i=1}^{n}\left[(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i) - \mathbb{E}\left[(g^{-1})'(\eta(X))\nabla\eta(X)\right]\right].$$

By introducing the term $(g^{-1})'(\eta(X_i))\nabla\eta(X_i)$, we can write

$$\begin{aligned}\hat{\beta}-\beta &= \frac{1}{n}\sum_{i=1}^{n}\left[(g^{-1})'(\eta(X_i))\nabla\eta(X_i)-\mathbb{E}\left[(g^{-1})'(\eta(X))\nabla\eta(X)\right]\right]\\ &+ \frac{1}{n}\sum_{i=1}^{n}\left[(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i)-(g^{-1})'(\eta(X_i))\nabla\eta(X_i)\right].\end{aligned}$$

Thanks to the properties of empirical mean, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left[(g^{-1})'(\eta(X_i))\nabla\eta(X_i)-\mathbb{E}\left[(g^{-1})'(\eta(X))\nabla\eta(X)\right]\right]=o_P(1).$$

Now we consider the second term $(g^{-1})'(\hat{\eta}(X_i))\widehat{\nabla\eta}(X_i)-(g^{-1})'(\eta(X_i))\nabla\eta(X_i)$. By a Taylor expansion of $(g^{-1})'(\hat{\eta}(x))$ about $\eta(x)$ we obtain

$$(g^{-1})'(\hat{\eta}(x))=(g^{-1})'(\eta(x))+(g^{-1})''(\eta(x))(\hat{\eta}(x)-\eta(x))+O([\hat{\eta}(x)-\eta(x)]^2).$$

We introduce

$$\epsilon_n(x)=(g^{-1})''(\eta(x))\nabla\eta(x)(\hat{\eta}(x)-\eta(x))+(g^{-1})'(\eta(x))(\widehat{\nabla\eta}(x)-\nabla\eta(x))$$

and

$$r_n(x)=(g^{-1})'(\hat{\eta}(x))\widehat{\nabla\eta}(x)-(g^{-1})'(\eta(x))\nabla\eta(x)-\epsilon_n(x).$$

Then, we have

$$r_n(x) = (g^{-1})''(\eta(x))(\hat{\eta}(x)-\eta(x))(\widehat{\nabla\eta}(x)-\nabla\eta(x))+\widehat{\nabla\eta}(x)O([\hat{\eta}(x)-\eta(x)]^2).$$

The estimation error becomes

$$\hat{\beta}-\beta=\frac{1}{n}\sum_{i=1}^{n}[\epsilon_n(X_i)+r_n(X_i)]+o_P(1).$$

To obtain the result, we just have to prove that

$$\frac{1}{n}\sum_{i=1}^{n}r_n(X_i)=o_P(1)\ ,\qquad \frac{1}{n}\sum_{i=1}^{n}\epsilon_n(X_i)=o_P(1). \tag{17}$$

The first condition of (17) will be true if we have

$$\sup_{x\in S_X}|r_n(x)|=o_P(1).$$

Since $\widehat{\eta}(x)=\widehat{D^{\underline{k}}\eta}(x)$ with $|\underline{k}|=0$ and $\widehat{\nabla\eta}(x)=\widehat{D^{\underline{k}}\eta}(x)$ with $|\underline{k}|=1$ and since, by Condition 4 of the Theorem 1 $(g^{-1})''$ is bounded, the following condition is sufficient

$$\sup_{x\in S_X}|(\hat{\eta}(x)-\eta(x))(\widehat{D^{\underline{k}}\eta}(x)-D^{\underline{k}}\eta(x))|=o_P(1),\ \text{for } |\underline{k}|=0 \text{ and } |\underline{k}|=1 \tag{18}$$

provided that $\widehat{\nabla\eta}(x)$ is bounded when n goes to infinity. This will be ensured later since we will show in the following that $\sup_{x\in S_X} |\nabla\eta(x) - \widehat{\nabla\eta}(x)| = o_P(1)$. In order to prove the sufficient Condition (18), we first show that, for $|\underline{k}| = 0$ or 1, we have

$$\forall \epsilon > 0, \quad \sup_{x\in S_X} |\widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x)| = o_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon}).$$

We consider Δ_n a discretization of the hypercube S_X fine enough to ensure:

$$\sup_{x\in S_X} \inf_{x'\in\Delta_n} \left|\Sigma_x^{-1}W_x - \Sigma_{x'}^{-1}W_{x'}\right| = o_P(1)$$

Using Lemma 1, we have:

$$\sup_{x\in S_X} \left|\hat{a}^* - \Sigma_x^{-1}W_x\right| \xrightarrow[n\to\infty]{P} 0,$$

with

$$\hat{a}^*(x)_{\underline{k}} = \frac{c_n^{-1}h^{|\underline{k}|}}{\underline{k}!}\left[\widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x)\right].$$

Therefore, since $c_n = (nh^p)^{-1/2}$, we obtain, for $|\underline{k}| = 0$ or 1:

$$\begin{aligned}
\sup_{x\in S_X}\left|\widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x)\right| &= \frac{1}{h^{|\underline{k}|}\sqrt{nh^p}} \sup_{x\in S_X} \hat{a}^*(x)_{\underline{k}} \\
&= \frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}\left\{\sup_{x\in S_X} |(\Sigma_x^{-1}W_x)_{\underline{k}}| + o_P(1)\right\} \\
&= \frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}\left\{\sup_{x\in \Delta_n} |(\Sigma_x^{-1}W_x)_{\underline{k}}| + o_P(1)\right\}.
\end{aligned}$$

Let $D_{\underline{k},\underline{l}} = \sup_{x\in S_X} (\Sigma_x^{-1})_{\underline{k},\underline{l}}$. We know, using the constraint $\inf_{x\in S_X}(\rho(x)f(x)) > 0$ (see Condition 5), that all the $D_{\underline{k},\underline{l}}$ are finite. Thus, with the results of Lemma 2, we have:

$$\begin{aligned}
\sup_{x\in S_X}\left|\widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x)\right| &\leq \sum_{\underline{l}\in A_q} D_{\underline{k},\underline{l}} \sup_{x\in\Delta_n}\left|\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}(W_x)_{\underline{l}} - \mathbb{E}\left[\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}(W_x)_{\underline{l}}\right]\right| \\
&+ O(h^{q+1-|\underline{k}|}) + o_P(\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}). \qquad (19)
\end{aligned}$$

In the previous expression, one can notice that the term $O(h^{q+1-|\underline{k}|})$ gives us the order of magnitude of the expectation of W_x. As pointed out in a remark after Lemma 2, one may have a more refined term $O(h^{q+2-|\underline{k}|})$ depending on the parity of q. As we just need to obtain an upper bound, the term $O(h^{q+1-|\underline{k}|})$ is sufficient. We note

$$d_x = \left|\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}(W_x)_{\underline{l}} - \mathbb{E}\left[\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}(W_x)_{\underline{l}}\right]\right|.$$

W_x was defined by

$$W_x = c_n \sum_{i=1}^{n} l_1\left[\bar{\eta}(x, X_i), Y_i\right] R(x, X_i) K^p(H^{-1}(X_i - x)).$$

We note

$$\Psi(x, X_i, Y_i) = l_1\left[\bar{\eta}(x, X_i), Y_i\right] R(x, X_i) K^p(H^{-1}(X_i - x)).$$

We then obtain

$$W_x = c_n \sum_{i=1}^{n} \Psi(x, X_i, Y_i)$$

and

$$d_x = \left| \sum_{i=1}^{n} \left\{ \frac{1}{nh^{|\underline{k}|+p}} \left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} - \mathbb{E}\left[\frac{1}{nh^{|\underline{k}|+p}} \left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} \right] \right\} \right|$$

We now use Bernstein's inequality in order to obtain, for a given $\tau > 0$ and for $\underline{l} \in \mathcal{A}_q$:

$$\mathrm{P}\left(d_x > \tau\right) \leq 2\, exp\left(\frac{-\tau^2}{2\sum_{i=1}^{n} Var\left(\frac{1}{nh^{|\underline{k}|+p}} \left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} \right) + \frac{2}{3nh^{|\underline{k}|+p}} M\tau} \right).$$

with M a constant such as

$$\mathrm{P}\left(\left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} - \mathbb{E}\left[\left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} \right] \leq M \right) = 1.$$

For $v > 0$ and $\epsilon > 0$, let τ be

$$\tau = n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon} v.$$

Bernstein's inequality is then written:

$$\mathrm{P}\left(d_x > n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon} v \right) \quad \leq$$

$$2\, exp\left(\frac{-n^{1+\alpha p+2\alpha|\underline{k}|+2\epsilon}}{2\sum_{i=1}^{n} Var\left(\frac{1}{h^{|\underline{k}|+p}} \left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} \right) + \frac{2n}{3h^{|\underline{k}|+p}} M n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon} v} \right).$$

Lemma 2 gives $Var(\{W_x\}_{\underline{l}} = O(1)$ which implies

$$Var\left(\frac{1}{h^{|\underline{k}|+p}} \left\{ \Psi(x, X_i, Y_i) \right\}_{\underline{l}} \right) = O(n^{\alpha p+2\alpha|k|}).$$

Thus, we obtain, for the right member of the inequality, an expression in $exp(-C * n^{2\epsilon})$ where C is a positive constant. Consequently, in the inequality (19), the first term of the right member is upper-bounded by

$$2 \sum_{\underline{l} \in A_q} D_{\underline{k},\underline{l}} exp(-Cn^{2\epsilon})$$

which goes to zero when n goes to infinity. For all $\tau > 0$, we can write

$$|d_x| \quad \leq \quad \tau \mathbb{1}_{|d_x| \leq} \tau + |d_x| \mathbb{1}_{|d_x| > \tau}.$$

With $\tau = n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon} v$, we have seen that $\mathrm{P}\left(|d_x| > \tau\right)$ is upper-bounded by a term in $2exp(-Cn^{2\epsilon})$. We obtain:

$$\sup_{x \in S_X} |d_x| \leq O_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon}).$$

So, for $\epsilon > 0$,

$$\begin{aligned} \sup_{x \in S_X} \left| \widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x) \right| &\leq O_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon}) + O(h^{q+1-|\underline{k}|}) + o_P(\frac{1}{h^{|\underline{k}|}\sqrt{nh^p}}) \\ &\leq O_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon}) + O(h^{q+1-|\underline{k}|}) + o_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|}) \\ &\leq O_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\epsilon}) + O(h^{q+1-|\underline{k}|}). \end{aligned}$$

The previous inequality is true for every $\epsilon > 0$. Let $\epsilon = \delta$ with δ defined in the Condition 7 of the Theorem 1. We use this value δ for the term with $|\underline{k}| = 0$ and for the term with $|\underline{k}| = 1$. Now let's show that the term in $O_P(n^{-(1-\alpha p)/2+\alpha|\underline{k}|+\delta})$ is dominating. We want to show $h^{q+1} = o(n^{-(1-\alpha p)/2+\delta})$. That will be true only if

$$\alpha > \frac{1-2\delta}{2q+p+2}$$

for $|\underline{k}| = 0$ or 1. δ has been choosen in such a manner that it verifies this condition. Thus, for $|\underline{k}| = 0$ or 1:

$$\sup_{x \in S_X} |(\hat{\eta}(x) - \eta(x))(\widehat{D^{\underline{k}}\eta}(x) - D^{\underline{k}}\eta(x))| \quad \leq \quad O_P(n^{-1+\alpha(p+|\underline{k}|)+2\delta}). \tag{20}$$

Here, we want to have the right member of the inequality (20) going to zero when n goes to infinity for $|\underline{k}| = 0$ or 1. So we will need to have:

$$\alpha < \frac{1-2\delta}{p+|\underline{k}|}$$

For $|\underline{k}| = 0$ or 1, this condition is verified because it is less strict than Condition 7 of the Theorem 1. We have proved that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} r_n(X_i) = o_P(1).$$

It remains to study the term $\frac{1}{n}\sum_{i=1}^{n} \epsilon_n(X_i)$ (second condition of (17)). It is sufficient to show that

$$\sup_{x \in S_X} |\epsilon_n(x)| = o_P(1).$$

By definition, we have

$$\epsilon_n(x) = (g^{-1})''(\eta(x))\nabla\eta(x)(\hat{\eta}(x) - \eta(x)) + (g^{-1})'(\eta(x))(\widehat{\nabla\eta}(x) - \nabla\eta(x)).$$

With the previous calculus and with the assumptions of the Theorem 1, we have

$$\begin{aligned} \sup_{x \in S_X} |\epsilon_n(x)| &\leq O_P(n^{-(1-\alpha p)/2+\delta}) + O_P(n^{-(1-\alpha p)/2+\alpha+\delta}) \\ &\leq O_P(n^{-(1-\alpha p)/2+\alpha+\delta}). \end{aligned}$$

And, thanks to the condition $\alpha < (1-2\delta)(p+2)$, we obtain the result. That finishes the proof.

Acknowledgements

We are really grateful to A. Antoniadis for constructive and fruitful discussions that substantially improved this article.

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24.

References

[1] A. Albert and J. Anderson. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1):1–10, 1984.

[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96(12):6745–6750, 1999.

[3] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective Dimension Reduction Methods for Tumor Classification using gene Expression Data. *Bioinformatics*, 19(5):563–570, 2003.

[4] R. Carroll, J. Fan, I. Gijbels, and M. Wand. Generalized partially linear single index models. *J. Am. Statist. Ass.*, 92:477–489, 1997.

[5] L. Devroye, L. Gyorfi, and G. Lugosi. *theory of pattern recognition.* Springer-Verlag, New York, 1996.

[6] B. Ding and R. Gentleman. Classification Using Generalized Partial Least Squares. *J. Comp. Graph. Stat.*, 2005. A paratre.

[7] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, 97:77–87, 2002.

[8] L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models. 2nd ed.* Springer Series in Statistics. New York, 2001.

[9] J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications.* Chapman and Hall, London, 1996.

[10] J. Fan, N. Heckman, and M. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Stat. Assoc.*, 90:141–150, 1995.

[11] G. Fort and S. Lambert-Lacroix. Classification using Partial Least Squares with Penalized Logistic Regression. *Bioinformatics*, 21(7):1104–1111, 2005.

[12] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[13] P. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *J.R. Statist.Soc. B*, 46(2):149–192, 1984.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.*, 46(1-3):389–422, 2002. Erratum : http://clopinet.com/isabelle/Papers/index.html.

[15] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medecine*, 7(6):673–679, 2001.

[16] S. Le Cessie and J. C. Van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[17] E. Lesaffre and A. Albert. Partial separation in logistic discrimination. *J. R. Stat. Soc., Ser. B*, 51(1):109–116, 1989.

[18] B. D. Marx. Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression. *Technometrics*, 38(4):374–381, 1996.

[19] P. McCullagh and J. Nelder. *Generalized Linear Models. 2nd ed.* New-York : Chapman & Hall, 1989.

[20] D. Nguyen and D. Rocke. Multi-class cancer classification via Partial Least Squares with gene expression profiles. *Bioinformatics*, 18(9):1116–1226, 2002.

[21] D. Nguyen and D. Rocke. Tumor classification by Partial Least Squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[22] D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7:186–199, 1991.

[23] T. Santner and D. Duffy. A note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 73(3):755–758, 1986.

[24] B. Seifert and T. Gasser. Finite-Sample Variance of Local Polynomials : Analysis" and Solutions. *J. Amer. Stat. Assoc.*, 91(433):267–275, 1996.

[25] Y. Xia, H. Tong, W. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc., Ser. B*, 64(3):363–410, 2002.

Tables and Figures

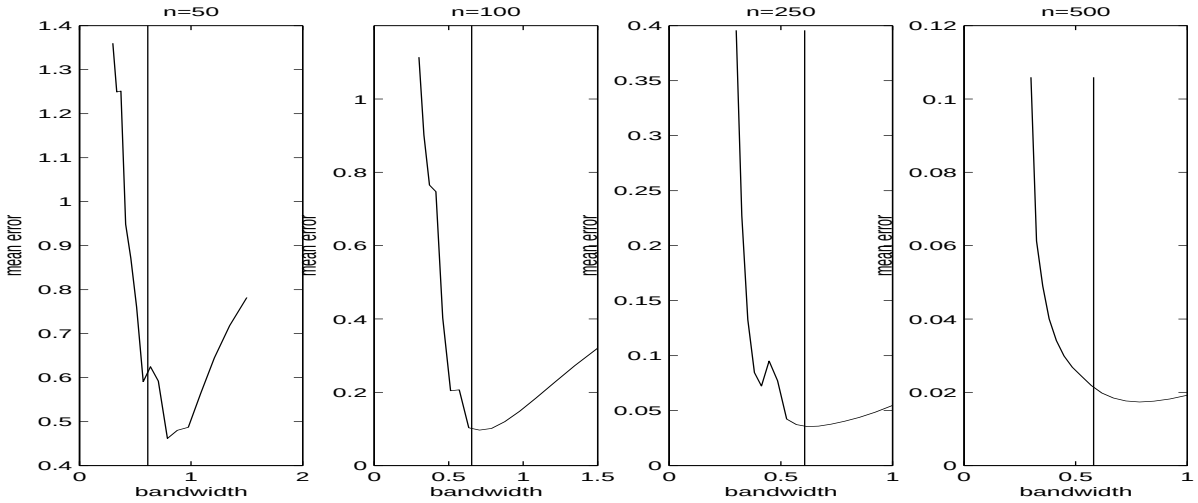


Figure 1: `Simulation results.` The solid line are means of the estimation errors from 100 replications. The vertical lines are means of corresponding cross-validation bandwidths.

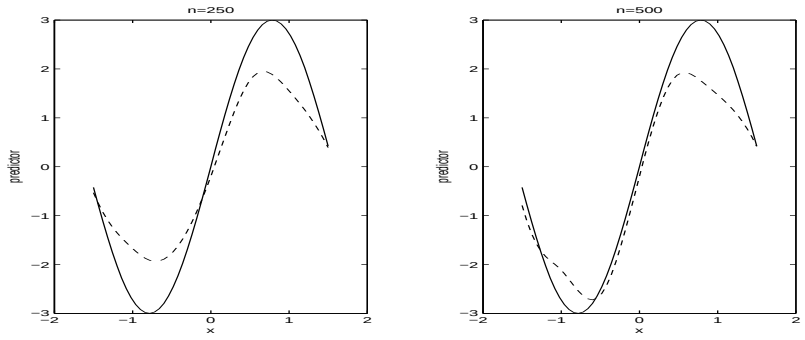


Figure 2: `Curve estimate.` Solid curve corresponds to $\tilde{\eta}$ and the dotted curve is its estimate ($n = 250$, 500).

n	β_1	β_2	β_3	β_4	RASE
50	-0.5338 (0.2024)	0.2558 (0.3620)	0.3306 (0.3598)	0.38116 (0.3124)	1.5491 (0.3016)
100	-0.4961 (0.1097)	0.4718 (0.1310)	0.4774 (0.1363)	0.4828 (0.1522)	1.1577 (0.2490)
250	-0.4757 (0.1285)	0.4861 (0.1465)	0.4716 (0.1402)	0.4833 (0.1568)	1.1847 (0.2487)
500	-0.4967 (0.0757)	0.5266 (0.0494)	0.4802 (0.0829)	0.4747 (0.0717)	0.8972 (0.0736)

Table 1: `Simulation results.` Mean and mean squared error (in parentheses) of estimated parameters and RASE.

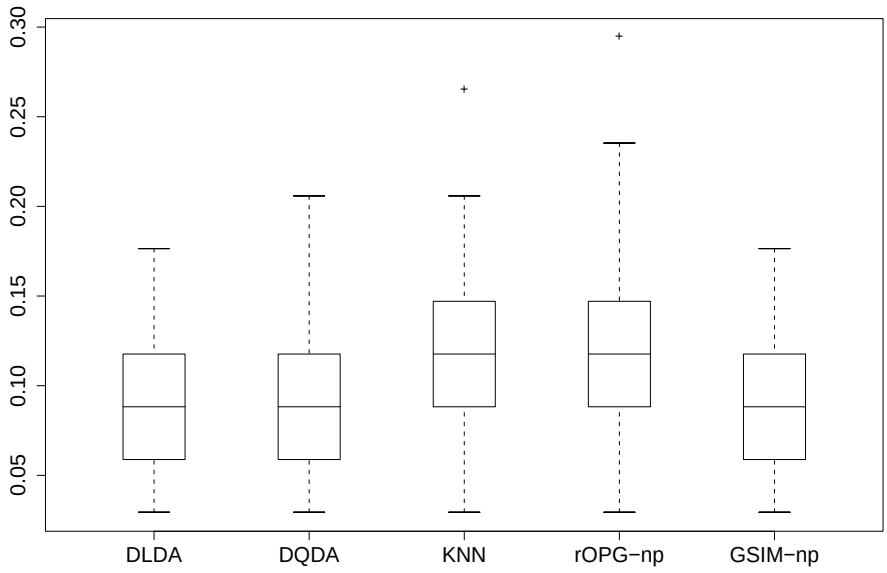


Figure 3: `Colon`. Resampling analysis: boxplot.

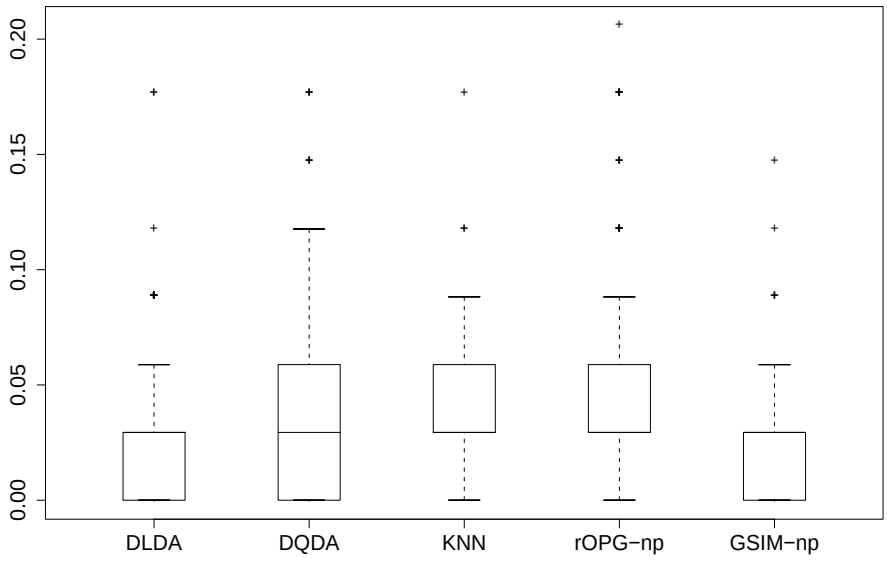


Figure 4: `Leukemia with 2 classes`. Resampling analysis: boxplot.

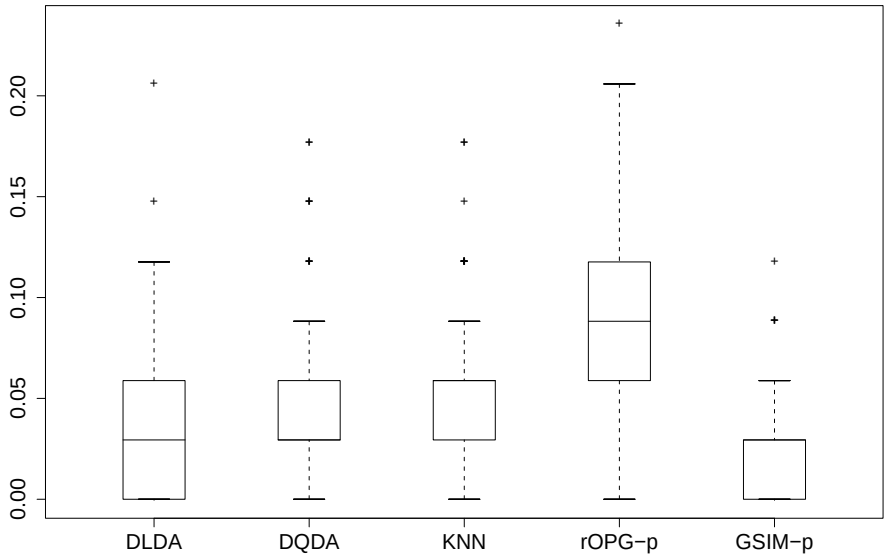


Figure 5: `Leukemia with 3 classes.` Resampling analysis: boxplot.

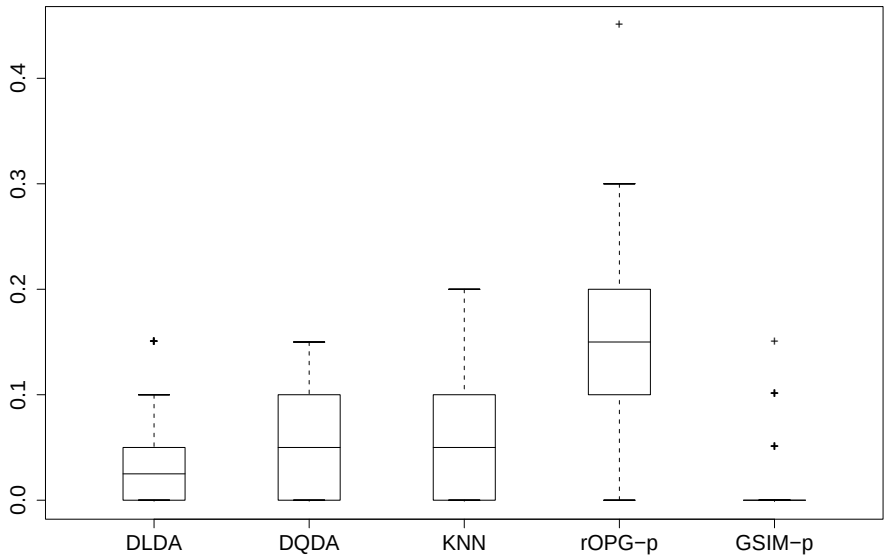


Figure 6: `SRBCT (with 4 classes).` Resampling analysis: boxplot.

	DLDA	DQDA	KNN	rOPG-np	GSIM-np
moy	0.144	0.154	0.205	0.201	0.155
std	0.057	0.064	0.072	0.077	0.056

Table 2: `Colon.` Resampling analysis: mean and standard-deviation.

	DLDA	DQDA	KNN	rOPG-np	GSIM-np
moy	0.032	0.046	0.046	0.061	0.027
std	0.034	0.044	0.032	0.072	0.026

Table 3: `Leukemia with 2 classes.` Resampling analysis: mean and standard-deviation.

	DLDA	DQDA	KNN	rOPG-p	GSIM-p
moy	0.039	0.046	0.055	0.088	0.025
std	0.037	0.039	0.036	0.048	0.023

Table 4: `Leukemia with 3 classes.` Resampling analysis: mean and standard-deviation.

moy	0.088	0.155	0.117	0.124	0.086	0.153
std	0.048	0.067	0.074	0.083	0.052	0.063

Table 5: `Leukemia with 3 classes.` Resampling analysis: mean and standard-deviation for `rOPG-p` according to different labels order.

	DLDA	DQDA	KNN	rOPG-p	GSIM-p
moy	0.040	0.046	0.065	0.149	0.008
std	0.047	0.045	0.053	0.069	0.026

Table 6: `SRBCT (with 4 classes).` Resampling analysis: mean and standard-deviation.