University of Salerno
Department of Economics and Statistics

Doctoral thesis in
"Economics and Policy Analysis of Markets and Firms "

Cycle: XXXII
Curriculum: Statistical Methods

A screening selection procedure for nonparametric regression and survival analysis

Candidate:
Sara Milito

Supervisor:
Prof. Francesco Giordano

PhD Coordinator:
Prof. Alessandra Amendola

ACADEMIC YEAR 2018-2019

Contents

Introduction

This thesis aims at proposing a new method of solving the nonparametric and non-additive regression problem in presence of ultra-high dimensional data. In this context, there are two relevant aspects: variable selection and structure discovery, such as identification of the variables that affect the response variable and the type of effects (linear or non linear), respectively.

In this thesis we propose a nonparametric method of variable selection that works in two stages. At the first stage, a screening procedure is performed: selecting a subset of variables which contains the true covariates with probability 1. In the second, we transform the screening step into variable selection using a non-penalized approach. In this way we take advantage of the simplicity of screening and we overcome the problem of estimating penalty parameters. Furthermore, our screening approach has the potential to distinguish linear and non-linear covariates, therefore it also succeeds in structure discovery.

Chang et al. (2016a), without requiring a specific parametric form of the underlying data model, proposed a screening method using empirical likelihood and local polynomials. Once the estimate of the marginal function between a particular variable and the response is obtained, they used empirical likelihood to test whether this function is significantly different from zero. Despite the excellent results in terms of dimensionality achieved, the authors did not perform any variable selection and structure discovery. To solve these problems, we propose to complicate their approach by estimating the first marginal derivative rather than the marginal function. In this way, we obtain a new fully nonparametric screening method, called *Derivative Empirical Likelihood Sure Independence Screening*(D-ELSIS). In order to transform our screening selection procedure into variable selection procedure, we use the subsample technique. In particular, we propose to apply the subsample idea not on the results of a variable selection procedure, as in Meinshausen and Bühlmann (2010), but after a screening procedure. With this tool, the variables selected through the D-ELSIS are then further evaluated to investigate their probability, in terms of relative frequency, to be chosen when the data are randomly sampled. Furthermore, although thresholds are used in this approach, these do not need to be estimated.

In summary, the potential of the proposed approach is threefold. First, we obtain a screening method that is totally non-parametric and that works in the context of nonparametric and

non-additive regression. Second, we transform the screening into variable selection without estimation of penalty parameters. Third, by estimating the first derivatives, we are able to distinguish the effects of the selected covariates on the response variable.

In this thesis the theoretical properties of our new D-ELSIS approach as a screening method will be analyzed. Moreover, simulation study and empirical application on real dataset will be described to evaluate the performance of the proposal. In particular, the consistency property is achieved with an exponential rate. Moreover, we pay something in order to estimate the first marginal derivative. The theoretical results will also be presented to support the transformation from a screening method to a variable selection method. We will aim at analyzing the structure discovery property which opens up future research perspectives.

Furthermore, we extend our proposal to time-to-event analysis. High-dimensional data are available due to the rapid growth of technology. In recent years, technology has also experienced strong growth in the medical and genetic fields. Variable selection is fundamental in survival analysis, where we find time-to-event outcome variable, which is a different type of outcome variables because the outcome of interest is not only whether event occurred, but also when that event occurred. Most of the methods in the literature consider a conditional estimate of the survival function, using the Kaplan and Meier estimator (Kaplan and Meier, 1958). Since this does not take into account the direct effect of covariate on survival probability, it has some disadvantages. We have managed to highlight and justify the possibility of applying the D-ELSIS method also in this context. With our D-ELSIS procedure, we obtain a fully nonparametric screening procedure without the use of the Kaplan and Meier estimate of survival function. This is the fundamental difference among our method and the other screening techniques. Furthermore, based on our knowledge, in survival context, a screening method that combines empirical likelihood and local polynomial regression has never been used.

The thesis is divided into two parts. In the first part, the regression problem will be addressed with high-dimensional data, our proposal will be examined from a theoretical point of view and the results of some simulations and an empirical study will be presented. In the second part our proposal will be applied in the context of survival analysis. Also in this case the results of the application of our new approach on simulated data will be reported.

Part I

Nonparametric regression

Introduction

The remarkable development of computing power and other technologies in recent decades has allowed scientists to collect data of unprecedented size and complexity. High dimensional data analysis has become increasingly frequent and important in various fields of sciences, such as genomic, health sciences, economics, finance, engineering and machine learning. Such a demand from applications presents many new challenges as well as opportunities for statistics.

Statistical accuracy, model interpretability and computational complexity are three important pillars of any statistical procedures. In conventional studies, the number of observations n is much larger than the number of variables p. In such cases, none of the three aspects has to be sacrificed for the efficiency of others. The traditional methods, however, face significant challenges when the dimensionality p is comparable to or larger than the sample size n. These challenges include: (i) how to implement statistical procedures that are more efficient for inference; (ii) how to derive the asymptotic or non-asymptotic theory; (iii) how to make the estimated models interpretable; (iv) how to make the statistical procedures computationally efficient and robust.

A mainstream statistical problem is to model the relationship between one or more output variables Y and their associated covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ based on a sample of dimensions n in the regression analysis. The textit variable selection problem occurs when there is uncertainty about which subset of the covariates p we should use. This situation is particularly interesting when p is large (or greater than n) and $X_1, \ldots, X_p$ are believed to contain many redundant or irrelevant variables. So, variable selection is the process of selecting a subset of relevant variables to use in model construction: we detect the relevant variables all together and, contextually, we also estimate the coefficients for the parametric model and the function in the nonparametric setting. Sometimes, it is difficult to find the true set of relevant variables because the number of candidate variables is very large. One possible solution is to first run textit variable screening to reduce the number of predictors and then use a known variable selection method. In fact, variable screening is the process of filtering out irrelevant variables, with the aim to reduce the dimensionality of the problem, while all relevant variables survive with probability tending to 1. In variable screening the problem is to consider each variable one by one and to detect which variable is relevant in the explanation of the response. In this

case we detect the set of important variables without estimating of each component. Since we do not know which variables are relevant, the idea of screening procedure is to order all the variables based on some measures of their effect on the response and to keep only the top ones. The main difference between variable and screening procedures is in their results. The first one aims to estimate the exact set of relevant variables, whereas the screening only reaches a suboptimal result, because it finds a rough set of candidates which includes the relevant covariates with high probability.

In order to analyse the variable selection problem it is necessary to consider different aspects. First, we need to define the relation between the dimensionality p and the sample size n, especially when $p >> n$. Second, we need to understand the relationships among the explanatory variables, that is to pay particular attention to the design matrix. Third, we need to check the conditions under which the procedure is able to have good estimation and selection properties.

The purpose of the first part of this thesis is to find a completely nonparametric new procedure capable of selecting the relevant covariates in the presence of ultra-high dimensional data, without imposing conditions on the underlying model. To this end, we have considered a very general model, that is, a nonparametric and non-additive one. Our new proposal, called *Derivative Empirical Likelihood Sure Independence Screening* (D-ELSIS), works as follows. First, it uses local polynomial regression to estimate the first marginal derivatives of the regression function, with respect to all variables in $\mathbf{X}$ (so, p derivatives in total). Then, it checks by the empirical likelihood (a nonparametric inference method, see Owen (2001)) if these derivatives are uniformly zero (or not) on the support of each variable. Those variables for which the test is passed are chosen as relevant covariates. Based on our knowledge, no other screening method uses the marginal estimate of the first derivative and empirical likelihood for screening purposes.

From a theoretical point of view, we will demonstrate that, under some regularity conditions, D-ELSIS has screening and variable selection properties with a exponential growth rate. This will also be demonstrated with simulations comparing our approach with those existing in literature. Furthermore, we will show theoretically how it is possible to transform our screening selection procedure into variable selection procedure, using the subsample technique.

In Chapter 1 we introduce the variable selection problem both in linear regression and in nonparametric regression models and we describe some of the structure discovery's techniques. We explain the idea of screening and the different methodologies that can be found in the literature. In Chapter 2 we introduce our new estimator, called D-ELSIS , to carry out screening in the context of nonparametric and non-additive models and the possibility of using it to achieve variable selection. Chapter 3 studies its theoretical properties under some regularity conditions. In Chapter 4, we carry out extensive numerical simulations to asses the

performance of the proposed D-ELSIS screener and compare it with other existing approaches. Finally we present an empirical study.

Chapter 1

Variable selection problems

When $p < n$, there is a rich literature on variable selection that can be used to identify non-zero components of the coefficient vector. The common approach is to use the Ordinary Least Square (OLS) in linear setting. When $p > n$, we find two different definitions for the dimensionality of the problem. *High* dimensional data is usually classified as the situation where p tends to infinity as n tends to infinity at polynomial rate, $p = O(n^{\alpha})$, with some $\alpha > 1$. *Ultra-high* dimensional data is the situation where p grows at a exponential rate in n, namely $\log(p) = O(n^{\alpha})$ for $\alpha \in (0, 1)$.

In the high dimensional setting, the design matrix, which contains the observations for a set on candidate explanatory variables, often denoted by $\boldsymbol{X}$, has an important role. Each row of this matrix represents an individual object, with the successive columns corresponding to the variables and their specific values for that object. When p is greater than n, the design matrix is rectangular, having more columns than rows. A notorious difficulty of high dimensional model selection comes from the collinearity among the predictors, as shown for example in Fan and Lv (2008). In their paper they showed how the collinearity can easily be spurious in a high dimensional geometry, which can lead to selecting a wrong model. Any variable can be well-approximated even by a couple of spurious variables, and can even be replaced by them when the dimensionality is much higher than the sample size. If that variable is a signature predictor and is replaced by spurious variables, we choose wrong variables to associate the covariates with the response and, even worse, the spurious variables can be independent of the response at population level, leading to completely wrong scientific conclusions. The maximum spurious correlation grows with dimensionality. Collinearity also gives rise to issues of over-fitting and model mis-identification. In variable selection we need to consider some conditions on the design matrix, in order to have the *oracle property*. A method possess the oracle property if it selects the correct subset of predictors with probability tending to one and estimates the non-zero parameters as efficiently as could be possible if we knew in advance which variables were uninformative (Fan and Li, 2001).

A consistent estimation procedure in terms of parameter estimation will not necessarily also be consistent in terms of selecting the correct model, where the opposite is also true, as shown in Zhao and Yu (2006). The consistency in terms of parameter estimation requires that the estimations tend in probability to the true parameters as the sample size n tends to infinity. On the other hand, the consistency in model selection requires that the set of selected variables tends to the true set of relevant variables, with probability 1, as n tends to infinity. In general, we desire an estimator to have both kind of consistencies.

In variable selection problems it is possible to distinguish between two contexts: parametric regression and nonparametric regression. When we know the functional relationship between the covariates and the response, we are in the parametric case. On the other hand, when we have no information about such relationships, we are in the nonparametric setting. In practice, there is often a little prior information that the effects of the covariates take a linear form or belong to any other finite-dimensional parametric family, so it is possible to use a nonparametric model to avoid incorrect specifications of the model. In the context of nonparametric models, particular attention is given to additive models. These increase substantially the flexibility of the ordinary parametric model and allow a data-analytic transform of the covariates to enter into the linear model. Also in the context of nonparametric regression, it is possible to define the oracle property: a nonparametric regression estimator has the nonparametric oracle property if it selects the correct subset of predictors with probability tending to one and estimates the regression functional form at the optimal nonparametric rate (Storlie et al., 2011).

What makes high dimensional statistical inference feaseble is the assumption that the regression function lies in a low dimensional manifold, as is shown in Fan and Lv (2010). In such cases, the p-dimensional regression parameters are assumed to be sparse with many components being zero, where non-zero components indicate the relevant variables. They assessed that, with sparsity, variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance the model interpretability with parsimonious representation. It can also help to reduce the computational cost when sparsity is very high. Following this sparsity principle, numerous variable selection approaches have been developed in the literature for high and ultra-high dimensional feature space.

1.1 Variable selection in linear regression models

In regression analysis, the linear model has been commonly used to link a response variable to explanatory variables for data analysis. One major reason for this is that the resulting OLS estimators have a closed form that is easy to compute. However, in the high-dimensional setting this closed form breaks down. In this situation there are many methods that outperform

OLS. The traditional linear regression model has the following formulation

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots, n \tag{1.1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are $i.i.d$, independent of $\boldsymbol{X}_i, i = 1, \ldots, n$ and such that $E(\epsilon_i) = 0$. The vector $\beta = (\beta_o, \beta_1, \ldots, \beta_p)^T$ represents the vector of coefficients. We are going to focus our attention in case when the number of predictors, p, is large relative to the number of observations, n. Classical variable selection procedures perform model selection and parameter estimation simultaneously. The majority of these procedures select variables by minimizing a penalized objective function with the following form:

$$Loss function + penalisation.$$

The penalty part is used to reduce the complexity of the model and to encourage sparsity in the final model. The most well known of these procedures is the LASSO (Least Absolute Shrinkage and Selection Operator) of Tibshirani (1996), which imposes an L_1 penalty ($||\cdot||_1$) on the coefficients under the assumption that the vector β is sparse. In fact, the LASSO estimator $\widehat{\beta}$ minimizes the following penalized sum of squares

$$\sum_i (Y_i - \boldsymbol{X}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1.2}$$

with $\boldsymbol{X}_i^T = (X_{i1}, \ldots, X_{ip})$.

The parameter $\lambda \geq 0$ controls the amount of regularization applied to the estimate and the penalty function is $p_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$. Setting $\lambda = 0$ reverses the LASSO problem to OLS which minimizes the unregularized empirical loss. On the other hand, a very large λ will completely shrink the parameters to 0 thus leading to the empty or null model. In general, moderate values of λ will cause shrinkage of the solutions towards 0, and some coefficients may end up being exactly 0.

In order to have an accurate variable selection we have to require some restrictions: the *beta-min condition*, which demands that the non-zero regression coefficients are sufficiently large, and the *irrepresentable condition* for the design matrix. The irrepresentable condition depends mainly on the covariance of the predictor variables and states that LASSO selects the true model consistently if the variables present in the true model can neither be too strongly correlated with each other nor with the noise variables. In fact, if there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to select only one variable from the group and does not care which one is selected. Those two conditions are restrictive but non-checkable, however are essentially necessary (Bühlmann and

Van De Geer, 2011). Zhao and Yu (2006) showed that, under the irrepresentable condition, the LASSO is consistent for variable selection provided p is not too large compared with n and the penalty parameter λ grows faster than $n^{1/2}$. Specifically, p is allowed to be as large as $\exp(n^a)$ for some $0 < a < 1$ when the errors have Gaussian tails. However, the value of λ required for variable selection consistency over-shrinks the non-zero coefficients, which leads to asymptotically biased estimates. Thus the LASSO is variable selection consistent if

$$\lim_{n\to\infty} P(M_* = \widehat{M}_n) = 1,$$

where $M_* = \{j : \beta_j \neq 0\}$ is the set of indices of all variables present in the true model, and $\widehat{M}_n = \{j : \widehat{\beta}_n \neq 0\}$ is the set of indices with every parameter estimates unequal to zero. Moreover, if the LASSO is variable selection consistent, it is not efficient for estimating the non-zero parameters, so these considerations confirm that the LASSO does not possess the oracle property (Fan and Li, 2001). However, retrieving all variables from the true model, whether or not accompanied by some noise variables, is a desirable property in itself. We will refer to this as the *variable screening property*:

$$\lim_{n\to\infty} P(M_* \subseteq \widehat{M}_n) = 1.$$

For this property to hold we again need the sparsity assumption and the beta-min condition, but the irrepresentable condition can be relaxed, leaving us with a less strong assumption on the design matrix, namely the *restricted eigenvalue condition* (Bühlmann and Van De Geer, 2011) which is technical but not overly restrictive in sparse problems. So, under the sparsity assumption, where the true variables have corresponding coefficients above some detection limit (the beta-min assumption), the LASSO has the ability to select them all. Even if there are some variables in the true model with coefficients that are too small to detect, one could still argue that the LASSO is able to find the influential and for that reason most relevant variables.

Efron et al. (2004) proposed a fast and efficient Least Angle Regression (LARS) algorithm for variable selection, a simple modification of which produces the entire LASSO solution path $\{\widehat{\beta}(\lambda) : \lambda > 0\}$ that optimizes (1.2). The computation is based on the fact that the LASSO solution path is piecewise linear in λ. The LARS algorithm starts from a large value of λ which selects only one covariate that has the greatest correlation with the response variable and decreases the λ value until the second variable is selected, at which the selected variables have the same correlation (in magnitude) with the current working residual as the first one.

Numerous alternatives and extensions have been suggested to deal with the problem of variable selection in linear regression model, with different forms of penalisation function. A few examples include SCAD (Fan and Li, 2001), the Elastic Net (Zou and Hastie, 2005) and the Dantzig selector (Candes and Tao, 2007). Fan and Li (2001) proposed to use the Smoothly

Clipped Absolute Deviation (SCAD) penalty function p_λ, which is a non-decreasing quadratic spline on $[0, \infty)$, linear on $(0, \lambda)$ and constant on $[a\lambda, \infty)$ for some $a > 2$:

$$p_\lambda(\beta) = \lambda \left\{ I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda) \right\}.$$

For the L_1 penalty, $a = \lambda$. The authors showed that the root-n consistency for any penalized approach requires that $\lambda = O(n^{-1/2})$. On the other hand, the oracle property requires that $\sqrt{n}\lambda \to \infty$. These two conditions for LASSO cannot be satisfied simultaneously, so the oracle property does not hold. The authors showed that this property holds for the SCAD. Kim et al. (2008) proved the oracle property in the case where the dimension p grows at a certain polynomial rate that depends on the moment condition of the noise, provided that the true model is sparse. Moreover, with a Gaussian noise, they showed that the dimension p can grow exponentially fast.

Zou and Hastie (2005) considered mixed norms in their approach called Elastic Net. The penalty function in this case has the following formula:

$$p_\lambda(\ \beta) = \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2.$$

This method overcomes the issue of collinearity because it favours selection of correlated regressors simultaneously while LASSO tends to select only one out of them. In fact, the Elastic Net can be solved as a LASSO using slight modification of the LARS algorithm. *Elastic net irrepresentable condition* (EIC) is crucial for the Elastic net's model selection consistency. In the standard case when the dimension p and the number of the relevant variables, s, are fixed, EIC is necessary and sufficient for the Elastic net to consistently select the true model (Yuan and Lin, 2007). When p and s both grow as n grows, EIC is not sufficient any more. Some conditions on the relationship among p, s and n are required. For consistency results, it is required that n grows at a rate faster than $s \log(p - s)$ (Jia and Yu, 2010).

The L_1 regularization has also been used in the Dantzig selector proposed by Candes and Tao (2007), which is defined as the solution to

$$\min \| \beta \|_1 \ \text{ subject to } \ \| n^{-1} \boldsymbol{X}^T (Y - \boldsymbol{X}\beta) \|_\infty \leq \lambda,$$

where $\lambda \geq 0$ is the regularization parameter. The Dantzig selector uses the L_∞ norm of the covariance vector $n^{-1}\boldsymbol{X}^T(Y - \boldsymbol{X}\beta)$, i.e., the maximum absolute covariance between a covariate and the residual vector $Y - \boldsymbol{X}\beta$, for controlling the model fitting. This L_∞ constraint can be viewed as a relaxation of the normal equation $\boldsymbol{X}^T Y = \boldsymbol{X}^T \boldsymbol{X}\beta$, finding the estimator that has the smallest L_1-norm in the neighbourhood of the least squares estimate. Under the uniform uncertainty principle on the design matrix $\boldsymbol{X}$, an assumption on the conditioning number requiring that all sub matrices of $\boldsymbol{X}$ are uniformly close to orthonormal matrices,

which can be stringent in high dimensions, Candes and Tao (2007) showed that, with high probability, the Dantzig selector mimics the risk of the oracle estimator up to a logarithmic factor $\log p$. Although the Dantzig selector is in a certain sense asymptotically equivalent to the LASSO (Bickel et al., 2009), their estimation consistency requires different conditions on the correlations between predictors because the Dantzig selector is related to an estimating equation, whereas the LASSO requires a specific likelihood or an objective function. The two methods depend on different correlation structures of predictors for sign consistency. Dicker and Lin (2013) considered random design of predictors and suggested Irrepresentable Conditions for the Dantzig selector in the fixed p case. Their method, however, cannot be extended to handle the case of p growing with n or the $p > n$ paradigm. Gai et al. (2013) considered fixed design with both fixed p and diverging p, even $p = \exp(n^a)$ for some constant $a > 0$. Irrepresentable Conditions are provided for the sign consistency of model selection. They established that these conditions are sufficient for a strong sign consistency and necessary for a weak sign consistency. Moreover, after shrinking the ultra-high dimension to a value that is smaller than the sample size, they also provided the conventional consistency of estimation when the dimension s of significant predictors is of a rate of $o(n)$.

1.2 Variable selection in nonparametric regression models

In practice, there is often little prior information that the effects of the covariates take a linear form or belong to any finite-dimensional parametric family. The nonparametric regression model

$$Y_i = m(\boldsymbol{X}_i) + \epsilon_i \tag{1.3}$$

where $m(\cdot)$ is a general smooth function, relaxes the strong assumptions that are made by a linear model but is much more challenging in high dimensions. In order to consider a nonparametric setting, it is possible to use a flexible class of nonparametric models, such as the additive model

$$Y_i = \sum_{j=1}^{p} m_j(X_{ij}) + \epsilon_i \tag{1.4}$$

introduced by Stone (1985). This additive combination of univariate functions - one for each covariate X_j - is less general than joint multivariate nonparametric models but can be more interpretable and easier to fit. In fact, Stone (1985) showed that the estimates based on the spline approximation achieve the optimal rate of convergence under a general fixed number of components p as for $p = 1$, that is smaller than n, because each component can be expressed as a linear combination of a set of basis functions whose coefficients must be either killed or selected simultaneously.

Lin and Zhang (2006) proposed the Component Selection and Smoothing Operator (COSSO) method for model selection and estimation in multivariate nonparametric regression models in the framework of smoothing spline ANOVA. For fixed p, they showed that the COSSO estimator in the additive model converges at the rate $n^{-l/(2l+1)}$, where l is the order of smoothness of the components. They also showed that, in the special case of a tensor product design, the COSSO correctly selects the non-zero additive components with high probability. Considering $\boldsymbol{\mathcal{M}}$ the function space

$$\boldsymbol{\mathcal{M}} = 1 \oplus \mathcal{M}_1 \text{ with } \mathcal{M}_1 = \bigoplus_{\alpha=1}^{p} \mathcal{M}^{\alpha}$$

where $\mathcal{M}^1, \ldots, \mathcal{M}^p$ are p orthogonal subspaces of $\boldsymbol{\mathcal{M}}$, the COSSO procedure finds $m \in \mathcal{M}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(\boldsymbol{X}_i)\}^2 + \tau_n^2 J(m) \text{ with } J(m) = \sum_{\alpha=1}^{n} \parallel P^{\alpha} m \parallel$$

where τ_n is a smoothing parameter and $P^{\alpha}m$ is the orthogonal projection of $m(\cdot)$ into $\mathcal{M}^{\alpha}$. The penalty term $J(m)$ is a sum of Reproducing Kernel Hilbert Spaces (RKHS) norms, instead of the squared norm employed in the traditional smoothing spline method.

The Sparse Additive Model (SpAM) approach of Ravikumar et al. (2009) imposed a sparsity constraint $\lambda \sum_{j=1}^{p} ||m_j||_2$ on the index set of functions $m_j(\cdot)$ that are not identically zero. The SpAM has the selection property using a particular form of smoothing, a truncated orthogonal basis and some constraints on the design matrix. In their theoretical results, they required that the eigenvalues of a design matrix must be bounded away from zero and infinity, where the design matrix is formed from the basis functions for the non-zero components. Another required condition is similar to the irrepresentable condition of Zhao and Yu (2006). It is not clear for what type of basis functions this condition is satisfied (Huang et al., 2010).

A particular nonparametric and non-additive regression model is the Gaussian regression model

$$Y_i = m(\boldsymbol{X}_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1.5}$$

where the input variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are n *i.i.d.* random variables with values in $\mathbb{R}^p$, the error terms $\epsilon_1, \ldots, \epsilon_n$ are n *i.i.d.* Gaussian random variables with mean zero, variance σ^2 independent of the $\boldsymbol{X}_i$'s and $m(\cdot)$ is the unknown regression function. Some procedures follow similar approaches focusing on the point wise estimation of the regression function. The basic idea is to start from a locally linear (or polynomial) point-wise estimator $m_n(x)$ at a point x obtained from the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - w(\boldsymbol{X}_i - x))^2 K_H(\boldsymbol{X}_i - x), \tag{1.6}$$

where $K_H(\cdot)$ is a localizing window function depending on a matrix (or a vector) H of smooth-

ing parameters. Different techniques are used to (locally) select variables.

Lafferty and Wasserman (2008) assumed the unknown regression function to be four times continuously differentiable with bounded derivatives and the density $f(\cdot)$ of the covariates to be uniform on the unit cube. The algorithm they proposed, called RODEO, is a greedy procedure performing simultaneously local bandwidth choice, function estimation and variable selection. RODEO is shown to converge when the ambient dimension p is $O\left(\frac{\log n}{\log(\log n)}\right)$ while the intrinsic dimension s, the number of relevant variables, does not increase with n. In the RODEO algorithm, the localizing window function depends on one smoothing parameter per variable and the partial derivative of the local estimator with respect to the smoothing parameter is used to select variables. It is based on the idea that bandwidth and variable selection can be simultaneously performed by computing the infinitesimal change in a nonparametric estimator as a function of the smoothing parameters, and then thresholding these derivatives to get a sparse estimate. The statistic used is

$$Z_j(h) = \frac{\partial \widehat{m}_h(x)}{\partial h_j}$$

where $\widehat{m}_h(x)$ denote an estimator of $m(x)$, with x fixed point, based on a vector of smoothing parameters $h = (h_1, \ldots, h_p)$. $Z_j(h)$ should discriminate between relevant and irrelevant covariates: if X_j is irrelevant, then we expect that changing the bandwidth h_j for that variable should cause only a small change in the estimator $\widehat{m}_h(\boldsymbol{X})$; while, if X_j is relevant, then we expect that changing the bandwidth h_j for that variable should cause a large change in the estimator.

Bertin and Lecué (2008) proposed a procedure based on the L_1-penalisation of local polynomial estimators and proved its consistency when $s = O(1)$, but p is allowed to be as large as $\log n$, up to a constant. They used two steps in their approach. In the first one, they determined the set of indices of relevant variables, and in the second they constructed an estimator of the value $m(x)$. To determine the set of indices, based on the principle of local linear regression, under the assumption m to be α-Holderian around x with $\alpha > 0$, denoted by $m \in \Sigma(\alpha, x)$, they consider the following set of vectors:

$$\bar{\Theta}(\lambda) = \arg\min_{\theta \in \mathbb{R}^{p+1}} \left[\frac{1}{nh^p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} m_j(\boldsymbol{X}_i)\theta_j \right)^2 K\left(\frac{X_i - x}{h} \right) + 2\lambda \parallel \theta \parallel_1 \right],$$

where h is a bandwidth, and $K(\cdot)$ is a symmetric kernel function. The convergence in the second steps has rate $n^{-2\alpha/(2\alpha+s)}$, that is the fastest convergence rate. Indeed, in all the above works the emphasis is in the theoretical analysis quantifying the estimation error of the proposed methods. It is shown in Lafferty and Wasserman (2008) that the RODEO algorithm is a nearly optimal point wise estimator of the regression function, these results are further

improved in Bertin and Lecué (2008) where optimal rates are derived.

Comminges and Dalalyan (2012) made a summary of the dimensionality that can be achieved to have the consistency in the selection of the model in linear regression and in nonparametric case. A careful statistical analysis is proposed considering different regimes for n, p and s in nonparametric regression when both n and s tend to infinity, even if s grows extremely slowly. The main results in this sense are the following:

- When the number of relevant variables s is fixed and the sample size n tends to infinity, there exist positive real number c^* such that no estimator of the sparsity pattern may be consistent if $(\log p)/n \geq c^*$.
- When the number of relevant variables s tends to infinity with $n \to \infty$, then there exist real number $\bar{c}_i, i = 1, 2$ such that $\bar{c}_1 > 0$, no estimator of the sparsity pattern may be consistent if $\bar{c}_1 s + \log\log(p/s) - \log n > \bar{c}_2$.
- In particular, if p grows not faster than a polynomial in n, an estimator of the sparsity pattern may be consistent if $s = o(\log n)$.

1.3 Structure discovery

As we said in the previous sections, there are two important classes of regression models for the analysis of statistical data, the linear and the nonparametric model. It is possible to consider some advantages for each class. In the context of linear models, which is the simplest, the interpretation of the parameters is very easy and the estimates are more efficient if the linear assumption is valid. In nonparametric models the assumption on the functional form of the model is less stringent. Indeed, not only can we find covariates with a linear effect on the response variable, as in linear models, but variables can have nonlinear and intersection effects. Between linear and nonparametric models there is another particular class of models, called partially linear models. This type of models have wide applications thanks to their flexibility and the advantages that derive from both linear and non-parametric model, allowing some covariates to be linear and others to not be linear. One natural question about this model is, given a set of covariates, how one decides which covariates have linear effects and which covariates have nonlinear effects.

The structure selection problem is fundamentally important, as the validity of the fitted model, and its inference heavily depends on whether the model structure is specified correctly. The model selection problem is divided in identifying the kind of effect of each variable (linear or non-linear) and the presence of interaction effect.

Compared to the linear model selection, the structure selection for partially linear models is much more challenging because the models involve multiple linear and non-linear functions and a model search needs to be conducted within some infinite-dimensional function space.

Furthermore, the difficulty level of model search increases dramatically as the data dimension grows due to the *curse of dimensionality*. Most works assume the partially linear model. The formula of this model is

$$Y_i = b + \boldsymbol{X}_i^T \beta + f(\boldsymbol{X}_i) + \epsilon_i$$

where b is the intercept, β is a vector of unknown parameters for linear terms, $f(\cdot)$ is an unknown function from $\mathbb{R}_s$ to $\mathbb{R}$, and ϵ_i's are $i.i.d.$ random errors with mean zero and variance σ^2, is given or known. In practice, data analysts often-times have to rely on their experience, historical data, or some screening tools to make an educated guess on the function forms for individual covariates. Two methods in common use are the screening and hypothesis testing procedures. The screening method first conducts univariate nonparametric regression for each covariate or unstructured additive models and then determines linearity or non-linearity for each term by visualizing the fitted function. This method is useful in practice but lacks theoretical justifications. The second method is to test linear null hypotheses against non-linear alternatives, sequentially or simultaneously, for each covariate. However, proper test statistics are often hard to construct and the tests may have low power when the number of covariates is large. In addition, these methods handle the structure selection problem and the model estimation separately, making it difficult to study inferential properties of the final estimator.

The main purpose of Zhang et al. (2011) was to distinguish linear and non-linear terms for partially linear models automatically and consistently, proposing an approach, called the LAND (Linear And Non-linear Discoverer), to identify model structure and estimate the regression function simultaneously. By solving a regularization problem in the frame of smoothing spline ANOVA, the LAND is able to distinguish linear and non-linear terms, remove uninformative covariates from the model, and provide a consistent function estimate. Under some assumptions, the LAND estimator has a rate of convergence $n^{-2/5}$ if the tuning parameters are chosen appropriately. Finally, adding the requirement that the density for $\boldsymbol{X}$ is bounded away from zero, assuming a tensor product design for the observations, the LAND achieve the property of selection consistency in the space of periodic component function, with fixed number of covariates and non-high dimensional setting.

Huang et al. (2012) proposed a semi-parametric regression pursuit method for identifying linear and non-linear effects. They embed partially linear models into a nonparametric additive model. By approximating the nonparametric components using spline series expansions, they transformed the problem of model specification into a group variable selection problem. They then determined the linear and non-linear components with a penalized approach, using a minimax concave penalty imposed on the norm of the coefficients in the spline expansion. They referred to this penalized approach as the group MCP method. In fact, they considered a truncated series expansion for approximating the function in the additive model. They showed

that the proposed approach is model pursuit consistent, meaning that it can correctly determine which covariates have a linear effect and which do not, with high probability under some conditions. The proposed approach has the same asymptotic property as the nonparametric estimator in the nonparametric additive model. They also showed that the estimated coefficients of linear effects are asymptotically normal, with the same distribution as the estimator assuming the true model were known in advance.

Promising as these methods are, they share a couple of drawbacks. Firstly, they only considered predictors of a fixed dimension which may not be an appropriate assumption in applications with high dimensional data. Secondly, their model only dealt with continuous response variables and thus excluded many interesting cases with binary or count responses. Lian et al. (2014) considered the double penalty structure recovery in a much more general regression setting. Besides the exponential family distribution generalization for the response, they also allowed the dimension p of the covariates growing at an exponential order of the sample size n. The estimation procedure is carried out in the framework of optimizing a doubly penalized quasi-likelihood. Similar to Huang et al. (2012) they started with a nonparametric additive model and used a spline series approximation to the component functions. The SCAD penalties (Fan and Li, 2001) on L_2-norms of the component functions and their second derivatives are used to identify respectively the zero and linear components. The spline series approximation transforms these L_2-norms to the norms of the coefficient vectors. An iterated procedure, combining the local quadratic approximation to the SCAD penalties and the reiterated weighted least squares, is used to obtain the final estimate. Considering some assumptions they showed that the method achieves selection consistency as well as the optimal convergence rates for the estimates of the non-zero components and the asymptotic normality for the estimates of the linear components, allowing a dimensionality $\log p = o(n^{d/(2d+1)})$, where $d > 1/2$ is the smoothness of the component functions.

All of the previous works face with the problem of the type of covariates that we have to include in the model. In many contests one wishes to allow for the possibility of interactions among the predictors. This poses serious statistical and computational difficulties when p is large, as the number of candidate interaction terms is of order p^2. The approach named Variable selection using Adaptive Non-linear Interaction Structure in High dimension (VANISH), of Radchenko and James (2010), combined the interaction terms and the additive non-linear model. This criterion enforced the heredity constraint, so if an interaction term is added to the model, then the corresponding main effect is automatically included. It is possible to express the additive non-linear model as

$$Y = \sum_{j=1}^{p} \boldsymbol{m}_j + \sum_{j=1}^{p} \sum_{k=j+1}^{p} \boldsymbol{m}_{jk} + \varepsilon \tag{1.7}$$

where $\boldsymbol{m}_j = (m_j(X_{1j}), \ldots, m_j(X_{nj}))^T$ are the main effect terms, for $j = \ldots, p$, $\boldsymbol{m}_{jk} = (m_{jk}(X_{1j}, X_{1k}), \ldots, m_{jk}(X_{nj}, X_{nk}))^T$ are the two-way interaction terms, and Y and ϵ are n-dimensional vectors corresponding to the response and the error terms, respectively. The penalised function, added to the loss function, is

$$p_\lambda(\boldsymbol{m}) = \lambda_1 \sum_{j=1}^{p} \left(||\boldsymbol{m}_j||^2 + \sum_{k:k\neq j}^{p} ||\boldsymbol{m}_{jk}||^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} ||\boldsymbol{m}_{jk}||, \tag{1.8}$$

where $||\cdot||$ is the Euclidean norm. In this case, λ_1 can be interpreted as the weight of the penalty for each additional predictor included in the model, while λ_2 corresponds to an additional penalty on the interaction terms for the reduction in interpretability of a non-additive model. In the linear setting VANISH has the selection property for p as large as $\exp(n^{1-\epsilon})$, with arbitrarily small positive ϵ, while in non linear setting for p as large as $\exp(n^{3/5-\epsilon})$. VANISH could be extended to higher order interaction term, for example including the third order interactions. The main practical limitation is that one would need to fit of order p^3 terms which may not be possible for large p.

1.4 Screening

Although the previous methods have been successfully applied in many high-dimensional analyses, it is difficult to apply them directly to those ultra-high dimensional statistical problems, due to their computational complexity. For example, the irrepresentable condition for the LASSO can be rather stringent in ultra-high dimension.

Fan and Lv (2008) proposed a Sure Independent Screening (SIS) procedure in linear regression models with Gaussian covariates and responses which screens variables by ranking their marginal correlations with the response variable. Differently from existing methods that minimize a penalized objective function, SIS uses the statistic $w_j = \frac{1}{n} X_j^T Y$ which represents a simple marginal correlation between response and standardized covariate. Considering the covariate one by one, the procedure calculates the statistic and provides a ranking of the X_j. The set $\widehat{M}$ of relevant features is determinated by a simple threshold:

$$\widehat{M} = \{1 \leq j \leq p : |w_j| \text{ is among the top } d \text{ largest ones}\}.$$

The SIS is method in which screening is first applied to reduce the dimensionality from p to a moderate one d, say, below sample size n, and inference is then conducted on the much reduced feature space. Fan and Lv (2008) established the desirable *sure screening property*, that is, most of the important features are retained with probability approaching one as the sample size diverges to ∞, even if the dimensionality of the features is allowed to grow exponentially fast

with the sample size. Assuming that $2\kappa + \tau < 1$, with $\kappa \in [0, 1/2)$ and $\tau > 0$, they proved the screening property achieving a dimensionality $\log p = O(n^{\xi})$ with $\xi \in (0, 1-2\kappa)$ and Gaussian noise $\epsilon \sim N(0, \sigma^2)$ for some $\sigma > 0$. They assume that $var(Y) = O(1)$, $\lambda_{max}(\boldsymbol{\Sigma}) = O(n^{\tau})$, $\min_{j \in M_*} \mid \beta_j \mid \geq c_1 n^{-\kappa}$, and $\min_{j \in M_*} \mid cov(\beta_j^{-1} Y, X_j) \mid \geq c_2$, where $\boldsymbol{\Sigma} = cov(\boldsymbol{X})$, with $c_1, c_2 > 0$ and the p-dimensional covariate vector has an elliptical distribution with the random matrix $\boldsymbol{X}\boldsymbol{\Sigma}^{1/2}$ having a so called concentration property that holds for Gaussian distributions. The condition on λ_{max}, the maximum eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ of predictors $\boldsymbol{X}$, rules out the case of strong collinearity, τ controls the rate of probability error in recovering the true sparse model, κ controls the signals of the parameters and the last condition, on the covariance, rules out the situation in which an important variable is marginally uncorrelated with Y, but jointly correlated with Y. Under the above conditions, Fan and Lv (2008) showed that

$$P(M_* \subset \widehat{M}) \to 1 \text{ as } n \to \infty$$

where $M_* = \{1 \leq j \leq p : \beta_j \neq 0\}$ is the true set of relevant covariates. Fan and Lv (2008) suggested to be conservative in the choice of d, for instance $n-1$ or $n/\log n$: a larger d means larger probability of including the true model M_* in the final model $\widehat{M}$.

Since the marginal utilities are employed to rank the importance of features, SIS can suffer from some potential issues associated with independence learning. First, some noise covariates strongly correlated with the important ones can have higher marginal utilities than other important ones. Second, some important covariates that are jointly correlated but marginally uncorrelated with the response can be missed after the screening step. To address these issues, Fan and Lv (2008) further introduced an extension of the SIS method, called the Iterative SIS (ISIS). The main idea is to iteratively update the estimated set of important variables using SIS conditional on the estimated set of variables from the previous step.

Independence feature screening is a class of rapidly developing approaches that is particularly useful in preliminary analysis for pre-processing data to reduce the scale of high-dimensional statistical problems. Since the work of Fan and Lv (2008), feature screening for ultra-high-dimensional data received a lot of attentions in the literature. Many authors have developed various sure independence screening procedures, many of which can be classified into two categories: model-based and model-free.

1.4.1 Model-based Screening

Fan and Song (2010) extended the SIS procedure to generalized linear models and presented a more general version of the independent learning by ranking the maximum marginal likelihoods or the maximum marginal likelihood estimates. Consider the generalized linear

model (GLM) with canonical link. That is, the conditional density is given by

$$f(y|x) = \exp\{y\theta(x) - b(\theta(x)) + c(y)\}$$

for some known functions $b(\cdot), c(\cdot)$, and $\theta(x) = x^T\beta$. The penalized likelihood is

$$-n^{-1}\sum_{i=1}^{n} l(\boldsymbol{X}_i^T\beta, Y_i) - \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

where $l(\theta, y) = b(\theta) - y\theta$. The maximum marginal likelihood estimator (MMLE) $\widehat{\beta}_j^{M_*}$ is defined as the minimizer of the component wise regression

$$\widehat{\beta}_j^{M_*} = arg\min_{\beta_0,\beta_j} \sum_{i=1}^{n} l(\beta_0 + \beta_j X_{ij}, Y_i)$$

where X_{ij} is the ith observation of the jth variable. Fan and Song (2010) select a set of variables whose marginal magnitude exceeds a predefined threshold value γ_n:

$$\widehat{M}_{\gamma_n} = \{1 \le j \le p : |\widehat{\beta}_j^{M_*}| \ge \gamma_n\}.$$

This is equivalent to ranking features according to the magnitude of MMLEs $|\widehat{\beta}_j^{M_*}|$. Taking the population version of the minimizer of the component wise regression

$$\beta_j^{M_*} = arg\min_{\beta_0,\beta_j} E\{l(\beta_0 + \beta_j X_j, Y)\}$$

they show that $\beta_j^{M_*} = 0$ if and only if $\text{cov}(X_j, Y) = 0$, and under some additional conditions if $|cov(X_j, Y)| \ge c_1 n^{-\kappa}$ for $j \in M_*$, for given positive constants c_1 and $\kappa \in [0, 1/2)$, then there exists a constant c_2 such that

$$\min_{j\in M_*} |\beta_j^{M_*}| \ge c_2 n^{-\kappa}.$$

So, as long as X_j and Y are somewhat marginally correlated, the marginal signal $\beta_j^{M_*}$ is detectable. They proved further the sure screening property:

$$P(M_* \subset \widehat{M}_{\gamma_n}) \to 1$$

if $\gamma_n = c_3 n^{-\kappa}$ with a sufficiently small c_3. For the Gaussian linear model with sub-Gaussian covariate tails, the dimensionality can be as high as $\log p = o(n^{(1-2\kappa)/4})$, a weaker result than that in Fan and Lv (2008) in terms of condition on p, but a stronger result in terms of the conditions on the covariates. For logistic regression with bounded covariates, the dimensionality can be as high as $\log p = o(n^{1-2\kappa})$. The authors also discussed the size of the selected

model $\widehat{M}_{\gamma_n}$ in the asymptotic sense. Under some regularity conditions, they showed that with probability approaching one, $|\widehat{M}_{\gamma_n}| = O\{n^{2\kappa}\lambda_{max}(\boldsymbol{\Sigma})\}$, where the constant κ determines how large the threshold λ_n is, and $\lambda_{max}(\boldsymbol{\Sigma})$ controls how correlated the predictors are. If $\lambda_{max}(\boldsymbol{\Sigma}) = O(n^{\tau})$, the size of $\widehat{M}_{\gamma_n}$ has the order $O(n^{2\kappa+\tau})$.

A general transformation regression model is defined to be

$$H(Y_i) = \boldsymbol{X}_i^T\beta + \epsilon_i.$$

Li et al. (2012a) proposed the rank correlation as a measure of the importance of each predictor by imposing an assumption on strict monotonicity on $H(\cdot)$. They proposed the marginal rank correlation

$$\omega_j = \frac{1}{n(n-1)}\sum_{i\neq l}^{n} I(X_{ij} < X_{lj})I(Y_i < Y_l) - \frac{1}{4}$$

to measure the importance of the jth predictor X_j. According to the magnitudes of all ω_j's, the feature screening procedure based on the rank correlation selects a sub model

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : |\omega_j| > \gamma_n\}$$

where γ_n is the predefined threshold value. Li et al. (2012a) referred this rank correlation based feature screening procedure to as a Robust Rank Correlation Screening (RRCS) procedure to deal with ultra-high-dimensional data. From the definition of the marginal rank correlation, it is robust against heavy-tailed distributions and invariant under monotonic transformation, which implies that there is no need to estimate the transformation $H(\cdot)$. When $H(\cdot)$ is an unspecified strictly increasing function, supposing that the minimum of the mean of the true covariates is a positive constant free of p, Li et al. (2012a) proved that the RRCS enjoys the sure screening property, provided that $\gamma_n = c_3 n^{-\kappa}$ for some constant c_3. The dimensionality achieved is $p = O(\exp(n^{\delta}))$ for some $\delta \in (0,1)$ satisfying $\delta + 2\kappa < 1$ for any $\kappa \in (0, 1/2)$.

Fan et al. (2011) developed a Nonparametric Independence Screening (NIS) method by ranking the importance of predictors via the magnitude of nonparametric components in sparse ultra-high dimensional additive models. They suggested estimating the nonparametric components marginally with spline approximation, and ranking the importance of predictors using the magnitude of nonparametric components. An intuitive population level marginal screening utility is $E(f_j^2(X_j))$, where $f_j(X_j) = E(Y|X_j)$ is the projection of Y onto X_j. With the sample $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, $f_j(x)$ can be estimated via a normalized B-spline basis $B_j(x) = \{B_{j1}(x), \ldots, B_{jd_n}(x)\}^T$, with

$$\widehat{f}_{nj}(x) = \widehat{\beta}_j^T B_j(x),\ 1 \leq j \leq p,$$

where $\widehat{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$ is obtained through the component wise least squares regression:

$$\widehat{\beta}_j = arg \min_{\beta_j \in \mathrm{R}^{d_n}} \sum_{i=1}^{n} (Y_i - \beta_j^T B_j(X_{ij})).$$

Thus the screened model index set is

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : \| \widehat{f}_{nj} \|_n^2 \geq \gamma_n\}$$

for some predefined threshold value γ_n, with $\| \widehat{f}_{nj} \|_n^2 = n^{-1} \sum_{i=1}^{n} \widehat{f}_{nj}(X_{ij})^2$. Fan et al. (2011) advocated the sure screening property of NIS based on a set of conditions: the r-th derivative of f_j is Lipschitz of order α for some $r > 0$, $\alpha \in (0, 1]$ and $q = r + \alpha > 1/2$, the marginal density function of X_j is bounded away from 0 and infinity, the signal of the active components do not vanish, i.e., $\min_{j \in M_*} E\{f_j^2(X_j)\} \geq c_1 d_n n^{-2\kappa}$ with $0 < \kappa < q/(2q+1)$ and $c_1 > 0$, the sup norm $\| m(\cdot) \|_\infty$ is bounded, the number of spline basis d_n satisfies $d_n = o(n^{1/3})$ for some $c_2 > 0$ and the *i.i.d.* random error ϵ_i satisfies the sub-exponential tail probability, so for any $B_1 > 0$, $E\{\exp(B_2|\epsilon_i|)|\boldsymbol{X}_i\} < B_2$ for some $B_2 > 0$. Under this conditions,

$$P(M_* \subset \widehat{M}_\gamma) \to 1$$

for $p = \exp\{n^{1-4\kappa} d_n^{-3} + n d_n^{-3}\}$. In addition, if $var(Y) = O(1)$, then the size of the selected model $|\widehat{M}_\gamma|$ is bounded by the polynomial order of n and depends on the largest eigenvalue of the covariance matrix. In the special case in which $\lambda_{max}(\boldsymbol{\Sigma}) = O(n^\tau)$, the size of the selected variables is of order $O(n^{2\kappa+\tau})$, that is the same result of Fan and Song (2010).

The empirical likelihood approach (Owen, 2001) is demonstrated effective in scenarios with less restrictive distributional assumptions for statistical inferences, but this approach encounters substantial difficulty when data dimensionality is high. More specifically, the data dimensionality p cannot exceed the sample size n in the conventional empirical likelihood construction. The properties of marginal empirical likelihood approach, where the available features are assessed one at a time individually, are systematically studied in Chang et al. (2013a) for linear regression models and generalized linear models, proposing a screening procedure based on the marginal Empirical Likelihood approach (EL-SIS). In this case the dimensionality problem of the empirical likelihood is solved, because they use one covariate after the other. The study of Chang et al. (2013a) contributes to the sure independence feature screening for high-dimensional data analysis from the following two aspects. First, a fundamental difference of this approach compared to all the previous approaches in literature is that the marginal empirical likelihood ratio statistic is a self-studentized quantity (Owen, 2001) while other screening methods usually rely on the ranking of covariates based on mag-

nitudes of some marginal estimators. The EL-SIS approach manages to further integrate the level of uncertainty resulting from the use of the conditions of finite sample. This special quality is of fundamental importance because in practice the levels of uncertainty corresponding to the different covariates may be different when contributing to the response variable of interest. Not considering standard errors could confuse the ranking for screening of characteristics based on the marginal estimators themselves, especially in high-dimensional statistical problems. Second, this screening procedure does not require restrictive assumptions about the distribution of errors or of the response variable. The authors show that EL-SIS represents a unified framework for feature screening in linear regression models and generalized linear models. Thanks to this second characteristic, errors can not be normally distributed in linear models, while in generalized linear models the response does not necessarily have to follow a distribution that belongs to the exponential family. To apply a marginal empirical likelihood approach for the linear regression model, Chang et al. (2013a) considered the marginal moment condition of the least squares estimator:

$$E\{X_j(Y - X_j\beta_j^{M_*})\} = 0 \ \ (j = 1, \ldots, p) \tag{1.9}$$

where $\beta_j^{M_*}$ is interpreted as the marginal contribution of covariate X_j to Y. From (1.9), considering that the explanatory variables are standardized, it is possible to see that $\beta_j^{M_*} = E(X_jY)$ is the covariance between X_j and Y so that $\beta_j^{M_*} = 0$ is equivalent to that Y and X_j are marginally uncorrelated. Therefore, because $E(X_j^2) = 1$, (1.9) is equivalent to

$$E(X_jY - \beta_j^{M_*}) = 0. \tag{1.10}$$

Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ be collected independent data, $g_{ij}(\beta) = X_{ij}Y_i - \beta$ $(j = 1, \ldots, p)$. Based on (1.10), Chang et al. (2013a) defined the following marginal empirical likelihood:

$$EL_j(\beta) = \sup\left\{\prod_{i=1}^n \omega_i : \omega_i \geq 0, \sum_{i=1}^n \omega_i = 1, \sum_{i=1}^n \omega_i g_{ij}(\beta) = 0\right\}$$

for $j = 1, \ldots, p$. The empirical likelihood estimates are calculated by maximizing the empirical likelihood function subject to constraints based on the estimating function and the trivial assumption that the probability weights w_i of the likelihood function sum to 1 (details about empirical likelihood are in section 2.3 of this thesis). For any given β in the convex hull of $\{X_{ij}Y_i\}_{i=1}^n$, the marginal empirical likelihood ratio was defined as

$$l_j(\beta) = -2\log\{EL_j(\beta)\} - 2n\log n = 2\sum_{i=1}^n \log\{1 + \lambda g_{ij}(\beta)\}$$

where λ is the Lagrange multiplier satisfying

$$0 = \sum_{i=1}^{n} \frac{g_{ij}(\beta)}{1 + \lambda g_{ij}(\beta)}.$$

The authors proved that the same function $g_{ij}(\cdot)$ can be applied for both linear models and generalised linear models with centred response variable Y. The value $l_j(0)$, the marginal empirical likelihood ratio evaluated ad $\beta = 0$, should not be large if $\beta_j^{M_*} = 0$, so Chang et al. (2013a) used $l_j(0)$ as a device for variable screening. In fact, $l_j(0)$ has a very clear practical interpretation by noting that it can be used to test the null hypothesis $H_0 : \beta_j^{M_*} = 0$. The authors evaluated $l_j(0)$ for all $j = 1, \ldots, p$, and, given a threshold γ_n, they selected a set of variables by

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j(0) \geq \gamma_n\}.$$

The authors also showed that this approach has the screening property under some conditions: the variable Y must to have bounded variance, there exists a positive constant c_1 such that

$$\min_{j \in M_*} |E(X_j Y)| = \min_{j \in M_*} |\text{cov}(Y, X_j)| \geq c_1 n^{-\kappa},$$

with $\kappa \in [0, 1/2)$, and there are positive constants K_1, K_2, γ_1 and γ_2 such that

$$P\{|X_j| \geq u\} \leq K_1 \exp(-K_2 u^{\gamma_1}) \text{ for each } j = 1, \ldots, p \text{ and any } u > 0,$$

$$P\{|Y| \geq u\} \leq K_1 \exp(-K_2 u^{\gamma_2}) \text{ for any } u > 0.$$

In case of linear model, the dimensionality may grows as $\log(p) = o(n^{1/2-\kappa})$, which is weaker than in Fan and Lv (2008) where $\log(p) = o(n^{1-2\kappa})$, because in this case the authors payed the price for allowing non-normal covariates and a more general error distribution; in generalised linear models the allowed dimensionality is $\log(p) = o(n^{(1-2\kappa)\gamma_1/(2\gamma_1+2)})$, which is a stronger result than that in Fan and Song (2010). The EL-SIS is also selection consistent, so

$$P\{\widehat{M}_{\gamma_n} = M_*\} \to 1 \text{ as } n \to \infty$$

if $\rho_j = E(X_j Y) = 0$ for any $j \notin M_*$, with $\log p = o(n^{\min((\gamma/6),((1-2\kappa)\gamma/(\gamma+2))})$, where $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$. In case in which $\max_{j \notin M_*} |\rho_j| = o(n^{-\eta})$, where $\eta > \kappa$, $\min_{j \notin M_*} E(X_j^2 Y^2) \geq c_2$ for some $c_2 > 0$, for any $\tau \in (1/2 - \eta, 1/2 - \kappa)$, the size of the selected model $|\widehat{M}_{\gamma_n}|$ is under control, considering $\gamma_n = c_2 n^{2\tau}$.

If there is some additional knowledge about the importance of a certain set of covariates, it is helpful to use this prior information and rank the importance of features by replacing simple marginal correlations with the marginal correlations conditional on such a set of variables. In many applications, researchers often know this set of certain predictors $\boldsymbol{X}_C$ related to the

response Y in advance. As shown in Barut et al. (2016), conditional information can help reducing the correlation among the variables. They proposed a Conditional Sure independence Screening (CSIS) from known active predictors which allows to recover the hidden importance variables and reduce the number of false negatives. But the CSIS has a strongly restrictive assumption for distributional model and needs to estimate β_C repeatedly when individually measuring the strength of the conditional contribution of the remaining variables given $\boldsymbol{X}_C$.

Hu and Lin (2017) proposed a Conditional Sure Screening feature procedure by Conditional Marginal Empirical Likelihood Ratio (CMELR - CSIS), which can be equally applied in both linear models and generalized linear models. This screening procedure gives better results than both EL-SIS and CSIS when the heteroscedastic models have hidden important variables or unimportant variables that are highly marginal correlated with the response. As a result, the procedure not only inherits the advantages of EL-SIS and CSIS, but also has flexibility in practice. In fact, it is able to identify the remaining features that contribute to the response when their number grows exponentially with the sample size. Hu and Lin (2017) define two index sets as

$$A = \{k : \beta_k \neq 0\},\ \bar{A} = \{k : \beta_k = 0\}$$

where A is the active index set that corresponds to the active predictors, and $\bar{A}$ is the complement set of A. Without loss of generality, they suppose that these known active predictors are the first s_C components $X_1, \ldots, X_{s_C}$ of $\boldsymbol{X}$. Denoting $\boldsymbol{X}_C = (X_1, \ldots, X_{s_C})^T$, $\boldsymbol{X}_D = (X_{s_{C+1}}, \ldots, X_p)^T$, and partitioning the parameters β as $\beta = (\beta_C^T, \beta_D^T)^T$, correspondingly, CMELR - CSIS tries to identify the set $D \cap A = \{j \in D : \beta_j \neq 0\}$. They show that it is possible to use a unified conditional marginal moment condition for linear model and generalised linear model as

$$E\{[X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)]Y\} - \alpha_j = 0$$

where α_j is denoted as the correlation coefficient between the centralized variable $X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)$ and the response Y. The authors show that α_j can be used as a tool for recruiting the corresponding index j. In fact, if the centralized variables, $X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)$ and $X_k - E(X_k|\boldsymbol{X}_C^T\beta_C)$, are uncorrelated, where $j \neq k$, $j \in D$ and $k \in D \cap A$, then

$$E\{[X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)][X_k - E(X_k|\boldsymbol{X}_C^T\beta_C)]\} = 0, \quad j \neq k, \quad j \in D, \quad k \in D \cap A.$$

Since $E(X_j|\boldsymbol{X}_C^T\beta_C)$ is unknown, Hu and Lin (2017) constructed an estimator of this quantity and then they used the function $\widehat{g_{ij}^{(c)}}(\alpha) = [X_{ij} - \widehat{E}(X_j|\boldsymbol{X}_{iC}^T\beta_C)]Y_i - \alpha_j$ to obtain the estimated

conditional marginal empirical likelihood ratio at zero as

$$\widehat{l_j}(0) = 2\sum_{i=1}^{n} \log\{1 + \widehat{\lambda}\widehat{g_{ij}^{(c)}}(0)\}$$

where $\widehat{\lambda}$ is the Lagrange multiplier satisfying

$$0 = \sum_{i=1}^{n} \frac{\widehat{g_{ij}^{(c)}}(0)}{1 + \widehat{\lambda}\widehat{g_{ij}^{(c)}}(0)}.$$

So, they selected the index set of active variables as

$$\widehat{D \cap A}_{\gamma_n} = \{j \in D : \widehat{\lambda}(0) \geq \gamma_n\}.$$

Finally, Hu and Lin (2017) established the sure screening property under some conditions. If there are positive constants K_1, K_2, γ_1 and γ_2 such that

$$P\{|X_j - E(\ X_j|\boldsymbol{X}_C^T\boldsymbol{\beta}_C)| > u\} \leq K_1 \exp\{-K_2 u^{\gamma_1}\}$$

for any $j \in D$ and any $u > 0$ and

$$P\{|Y| > u\} \leq K_1 \exp\{-K_2 u^{\gamma_2}\}$$

for any $u > 0$, $\max_i |X_{ik}Y_i| = O_p(n^{\omega})$ where $\omega < 1/2 - \kappa$, for $\tau \in (0, 1/2 - \kappa)$ and $\gamma_n = c_1^2 n^{2\tau}$, it is possible to show that

$$P\{D \cap A \subset \widehat{D \cap A}_{\gamma_n}\} \to 1 \quad \text{as} \quad n \to \infty$$

achieving the same dimensionality of Chang et al. (2013a). However, adding some conditions, $\max_{j \notin D \cap A} |E\{[X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)]Y\}| = O(n^{-\eta})$ where $\eta > \kappa$ and $\min_{j \notin D \cap A} E\{[X_j - E(X_j|\boldsymbol{X}_C^T\beta_C)]^2 Y^2\} \geq c_3$ for some $c_3 > 0$, for any $j \notin D \cap A$ and any $\tau \in (\max((1/2 - \eta), \omega), 1/2 - \kappa)$, they also controlled the size of the selected set of variables.

A popular model selection method is Best Subset Regression. However, one downfall of this method is that the computational cost increases exponentially in p. Thus, one can imagine the difficulty this would propose in an ultra-high dimensional setting. An alternative to best subset selection is Forward Regression. Wang (2009) proposed using Forward Regression (FR) in ultra-high dimensional settings as a way to shrink the dimension down to a manageable size in a classical linear regression model. All of the preceding methods employ an independent screening process where the utility for a given predictor does not depend on any other predictor.

FR deviates from this trend. Using FR, the result is a sequence of nested models each with one more predictor than the last. These variables are added according to which one will decrease the regression sum of squares the most. Defined $C_0 = \emptyset$, for each $k = 1, \ldots, n$ the author repeats the following steps. With each $j \in F \backslash C_{k-1}$, with $F = \{1, \ldots, p\}$, Wang (2009) fitted a model using $C_{k-1} \cup \{j\}$ as the set of predictors. From this model the author computed the residual sum of squares

$$RSS_j^{(k-1)} = Y^T(I_n - \boldsymbol{H}_{j(k-1)})Y$$

where $\boldsymbol{H}_{j(k-1)} = \boldsymbol{X}_{C_{k-1}\cup\{j\}}(\boldsymbol{X}^T_{C_{k-1}\cup\{j\}}\boldsymbol{X}_{C_{k-1}\cup\{j\}})^{-1}\boldsymbol{X}^T_{C_{k-1}\cup\{j\}}$ is the projection matrix. Now, a smaller RSS is more desirable. The author selected

$$a_k = arg \min_{j \in F\backslash C_{k-1}} RSS_j^{(k-1)}$$

and created the next set in the solution path $C_k = C_{k-1} \cup \{a_k\}$. The final result is $\boldsymbol{C} = \{C_k : 1 \leq k \leq n\}$ where $C_k = \{a_1, \ldots, a_k\}$. Wang (2009) imposed some technical conditions to prove the screening property with a dimensionality that diverges to infinity at an exponential rate: the normality assumption for $\boldsymbol{X}$ and ϵ, a constraint on the smallest and the largest eigenvalues of the covariance matrix, i.e. with $0 < \tau_{min} < \tau_{max} < \infty$, $2\tau_{min} < \lambda_{min}(\boldsymbol{\Sigma}) \leq \lambda_{max}(\boldsymbol{\Sigma}) < 2^{-1}\tau_{max}$ and the minimal size of the non-zero coefficients $\min_{j\in M_*} |\beta| \geq \nu_\beta n^{-\xi_{min}}$, with some constant ν and ξ_{min}. The author proved that the FR algorithm can detect all relevant predictors within a finite number of steps, much smaller than the sample size n. Another option is to select a set based on some criterion. A possible option would be the BIC criterion proposed by Chen and Chen (2008)

$$BIC(M_*) = \log \widehat{\sigma}^2_{(M_*)} + n^{-1} \mid M_* \mid (\log n + 2 \log d)$$

where $\widehat{\sigma}^2_{(M_*)} = n^{-1}RSS(M_*)$. Let $\widehat{m} = arg \min_{1\leq m\leq n} BIC(C^{(m)})$ and $\widehat{C} = C^{(\widehat{m})}$, Wang (2009) pointed out that this criterion has only been proven to have the consistency in selection when $p = O(n^\alpha)$ for some $\alpha > 0$, but has the screening property for $\log(p) = O(n^\xi), \xi > 0$, when $q \leq \nu n^{\xi_0}$, so

$$P(M_* \subset \widehat{C}) \to 1.$$

The partially linear model,

$$Y = \boldsymbol{X}^T\beta + g(U) + \epsilon \tag{1.11}$$

where U is an univariate explanatory variable in $[0, 1]$ (for simplicity), and $g(U)$ is an unknown smooth function of U, with $(\boldsymbol{X}^T, U)^T$ and ϵ independent, is important in the context of semi-parametric regression. In regression analysis, the profile least squares approach is useful to convert the semi-parametric model to the least squares setting. Following this approach, it is

possible to verify that

$$Y_i - E(Y_i|U_i) = \sum_{j=1}^{d} \beta_j \{X_{ij} - E(X_{ij}|U_i)\} + \epsilon_i.$$

Then, defining the profiled response and the profiled predictor as $Y_i^* = Y_i - E(Y_i|U_i)$ and $\boldsymbol{X}_i^* = \boldsymbol{X}_i - E(\boldsymbol{X}_i|U_i)$ respectively, the partial linear model (1.11) reduces to the classical linear regression model

$$Y_i^* = \boldsymbol{X}_i^{*T}\beta + \epsilon_i. \tag{1.12}$$

To implement the linear model (1.12) in practice, however, the unknown functions $E(Y_i|U_i)$ and $E(\boldsymbol{X}_i|U_i)$ need to be estimated nonparametrically. Liang et al. (2012) used the local linear regression technique of Fan and Gijbels (1996) to estimate $E(Y_i|U_i)$, and using the same steps of Wang (2009), they proposed a Profiled Forward Regression (PFR) to variable screening. Adding some conditions on nonparametric regression to that assumed in Wang (2009), Liang et al. (2012) showed that the performance of PFR can be asymptotically as good as FR and the screening property can hold at a rate sharper than the rate given in Wang (2009).

Hao and Zhang (2014) considered Interaction-selection procedure featured with FORward selection, which is referred as iFOR, in a regression model with linear and second order terms

$$Y_i = \beta_0 + \boldsymbol{X}_i^t\beta^{(1)} + Z_i^T\beta^{(2)} + \epsilon_i$$

where the vector Z contains quadratic and two-way interaction terms. They proposed an algorithm, called iFORT. This is a two stage procedure, which at first stage selects only main effects by FS, obtaining a set $\widehat{M}$, while in the second stage, the interaction terms generated under the heredity condition are considered, so they expanded the set $\widehat{M}$, adding all the two-way interactions within $\widehat{M}$ and then implemented FS on the extended set. Also in this case, to select the optimal model from the path, they used the BIC proposed by Chen and Chen (2008). Under the same condition of Wang (2009), adding a strong heredity condition $\beta_{kl} \neq 0 \Rightarrow \beta_k\beta_l \neq 0$, Hao and Zhang (2014) stated the sure screening property for interaction selection for ultra-high dimensional setting, with $\log p = \nu n^{\xi}, \xi < 1/2$.

1.4.2 Model-free Screening

The aforementioned screening methods only work well when the models are correctly specified, but, in the presence of incorrect model specifications, they are unable to select all the relevant variables. In particular, these models focus on identifying covariates that have a particular effect on the response variable, not taking into account the possibility that different covariates give different effects. In practice, we typically have data with a huge number of candidate variables, but we have little information that the actual model is linear

or follows any other specific parametric, nonparametric or semi parametric form. Thus, it is of great interest to develop model-free feature screening procedures for ultra-high-dimensional data. By model-free, it means that one does not need to impose a specific model structure on regression functions to carry out a screening procedure. One way to achieve the model-free goal is to develop feature screening procedures for a general class of models which include most commonly-used parametric, nonparametric and semi parametric models as special cases.

Zhu et al. (2011) proposed a Sure Independent Ranking and Screening (SIRS) procedure to screen the significant explanatory variables under a unified model framework, which includes a lot of parametric and nonparametric models. This flexibility is achieved by using a marginal utility measure that is concerned with the entire conditional distribution of the response given the predictors. Another strategy to achieve model-free is to employ the measure of independence to efficiently detect linearity and non-linearity between predictors and the response variable and construct feature screening procedures for ultra-high-dimensional data. Li et al. (2012b) proposed a SIS procedure based on the Distance Correlation (DC-SIS) and showed the sure screening property without assuming any particular regression function. Nonparametric quantile regression is useful to analyse the heterogeneous data, by separately studying different conditional quantiles of the response given the predictors. The Quantile-Adaptive screening (QA) (He et al., 2013) improved the robustness of NIS by allowing heteroscedasticity in the model.

Model-free using the conditional distribution function

The Kolmogorov filter (KF) of Mai and Zou (2013) is a fully nonparametric robust screening method. It deals with binary classification problems and uses the Kolmogorov–Smirnov test statistic to screen covariates. Considering $F_{+j}(x)$ and $F_{-j}(x)$ the conditional cumulative probability functions of X_j given $Y = c(1, -1)$, respectively, and defining

$$K_j = \sup_{-\infty<x<\infty} |F_{+j}(x) - F_{-j}(x)|$$

for which the sample version is defined as $K_{nj} = \sup_{-\infty<x<\infty} |\widehat{F}_{+j}(x) - \widehat{F}_{-j}(x)|$, Mai and Zou (2013) ranked all variables by the K_{nj} statistics and selected the subset

$$\widehat{M}_{d_n} = \{j : K_{nj} \text{ is among the first } d_n \text{ largest of all } K_{nj}\}$$

where the default value for $d_n = n/\log(n)$ or, with a more conservative choice, $d_n = n$. The Kolmogorov filter significantly outperforms other existing screening methods for binary classification problems, it works with all types of covariates and is invariant under univariate monotone transformations of the covariates. It have the sure screening property even when the covariates are strongly dependent on each other. In fact, the screening property holds

with probability going to one if

$$\delta_{M_*} = \min_{j \in M_*} K_j - \max_{j \in M_*^c} K_j \gg \{\log(p)/n\}^{1/2}.$$

This result is very promising because it was commonly believed before Mai and Zou (2013) that marginal screening methods tend to work well if and only if the noise variables are weakly correlated with the relevant variables. The limitation of KF is that this procedure is designed for binary classification problems and is inapplicable when the response variable can take more than two values.

Mai and Zou (2015) developed the Fused Kolmogorov Filter (FKF), a fully nonparametric model-free variable screening method that could provide a unified solution to variable screening problems emerging from a wide variety of applications such as binary classification, multi class classification, regression and Poisson regression, among others. This method should also work with discrete, categorical or continuous covariates and it is invariant under univariate monotone transformations of response variable or covariates or both. As the name suggests, the fused Kolmogorov filter is built upon two main ideas, the Kolmogorov–Smirnov test statistic, as used in Mai and Zou (2013), and fusion. When the response variable is binary, the fused Kolmogorov filter is exactly the KF proposed in Mai and Zou (2013), and fusion is not needed. The fusion part becomes critically important when the response variable is continuous. Following the KF, Mai and Zou (2015) considered

$$K_j^* = \sup_{y_1, y_2} \sup_x |F_j(x|Y = y_1) - F_j(x|Y = y_2)|$$

where K_j^* is a natural generalization of K_j. In fact, $K_j^* = 0$ if and only if X_j is independent of Y. In order to use K_j^*, they found an empirical version of K_j^*. This step is trivial for the binary response case, but it is much more difficult when Y takes infinite values because it requires the knowledge of $F_j(x|y)$ for all possible values y. Mai and Zou (2015) found an approximation of K_j^* by slicing the response into multiple slices, considering a partition

$$\boldsymbol{G} = \left\{ [a_l, a_{l+1}) : a_l < a_{l+1}, l = 0, \ldots, G-1 \text{ and } \bigcup_{l=1}^{G-1} [a_l, a_{l+1}) \setminus \{a_0\} = \mathbb{R} \right\}$$

where $a_0 = -\infty$, $a_G = \infty$ and $\mathbb{R}$ is the support of Y. They then defined a random variable $H \in \{1, \ldots, G\}$ such that $H = l + 1$ if and only if Y is in the lth slice. Mai and Zou (2015) computed a Kolmogorov–Smirnov test statistic for each pair of slices and then took the supreme of all pairwise Kolmogorov–Smirnov test statistics:

$$K_j^{\boldsymbol{G}} = \max_{l,m} \sup_x |F_j(x|H = l) - F_j(x|H = m)|$$

where $F_j(x|H=l) = P(X_j \leq x|H=l)$. In fact, X_j is independent of Y if and only if $K_j^{\boldsymbol{G}} = 0$ when Y takes finite values and each possible value forms a slice. The authors showed that it is possible to use $K_j^{\boldsymbol{G}}$ to evaluate the dependence between Y and X_j even if Y is continuous and they stated that $K_j^{\boldsymbol{G}}$ is a better measure for variable screening than K_j^*. Mai and Zou (2015) estimated $K_j^{\boldsymbol{G}}$ for all p variables using a partition $\boldsymbol{G}$ by

$$\widehat{K}_j^{\boldsymbol{G}} = \max_{l,m} \sup_x |\widehat{F}_j(x|H=l) - \widehat{F}_j(x|H=m)|$$

where

$$\widehat{F}_j(x|H=l) = \frac{1}{n_l} \sum_{H^i=l} \mathbb{I}(X_j^i \leq x),$$

n_l is the sample size within the lth slice and $H^i = l$ if Y_i is in the lth slice. To make the method insensitive to the slicing scheme, Mai and Zou (2015) repeated the procedure for different ways of slicing and then took the sum of their outcomes as the final screening statistic. They considered N different partitions $\boldsymbol{G}_i$ for $i = 1, \ldots, N$, where each partition $\boldsymbol{G}_i$ contains G_i intervals. They suggested an intuitive uniform slicing to partition data into G slices. If Y is categorical with levels $1, \ldots, G$, or Y is discrete with finite possible values $1, \ldots, G$, they set $H = Y$. If Y is discrete and can take infinite values, they set $H = Y + 1$ if $Y < G - 1$ and $H = G$ if $Y \geq G - 1$. When Y is continuous, they consider the intervals bounded by the $\frac{l}{G}$th sample quantiles of Y for $l = 0, \ldots, G$. They considered multiple uniform slicing $\boldsymbol{G}_i, 1 \leq i \leq N$ where $\boldsymbol{G}_i$ has G_i many slices. At the end, Mai and Zou (2015) combined the information of all $\boldsymbol{G}_i$ and computed the final Kolmogorov Filter statistic as

$$\widehat{K}_j = \sum_{i=1}^{N} \widehat{K}_j^{\boldsymbol{G}_j}.$$

Finally, they ranked each covariate by its fused Kolmogorov statistic and screened out those covariates at the bottom of the rank list:

$$\widehat{M} = \{j : \widehat{K}_j \text{ is among the } d_n\text{'th largest }\}.$$

Under the following two conditions:

- there exists a set S such that $M_* \subset S$

$$\Delta_S = \min_i \left(\min_{j \in S} K_j^{(o)}(G_i) - \max_{j \notin S} K_j^{(o)}(G_i) \right) > 0$$

  where the slicing is built on the theoretical quantiles of Y, when the distribution of Y is known, so the jointly important predictors should also be marginally important;

- considering $G_{min} = \min_i G_i$, then for any b_1, b_2 such that $P(Y \in [b_1, b_2)) \leq 2/G_{min}$, it is
$$|F_j(x|y_1) - F_j(x|y_2)| \leq \frac{\Delta_S}{8}$$
for all x, j and $y_1, y_2 \in [b_1, b_2)$, so the sample quantiles of Y are close enough to the population quantiles of Y;

the sure screening property holds with probability tending to one if

$$\Delta_S \gg \sqrt{\frac{\log n \log(pN \log n)}{n}}.$$

In addition, if there exist $0 < \kappa < 1$ such that $\Delta_s \gg n^{-\kappa}$, the FKF can handle the same order of dimension as SIS

$$\log p \ll n^{\xi} \text{ with } \xi \in (0, 1 - 2\kappa)$$

without imposing any parametric assumptions. But the FKF screening procedure is computationally heavy since the calculation of the Kolmogorov–Smirnov statistic involves the numerical optimization problem and it is sensitive to the selection of the number of slices.

Cui et al. (2015) proposed a sure independence screening using Mean Variance index (MV-SIS) for ultra-high-dimensional discriminant analysis based on the empirical conditional distribution function. The procedure is robust to model misspecification, heavy-tailed distributions of explanatory variables and outliers, but they only studied the scenario where response variable is categorical and explanatory variables are continuous. It not only retains the advantages of the Kolmogorov filter, but also allows the categorical response having a diverging number of classes in the order of $O(n^{\kappa})$ with some $\kappa \geq 0$. The MV-SIS is applicable for the setting in which the response is continuous, but the feature variables are categorical, in a nonparametric additive model. The authors considered Y a categorical response with R classes $\{y_1, y_2, \ldots, y_R\}$, and X_j a continuous covariate with a support R_{X_j}. To investigate the dependence relationship between X_j and Y, they considered the conditional distribution function of X_j given Y, denoted by $F(x|Y) = P(X_j \leq x|Y)$. Denoting by $F(x) = P(X_j \leq x)$ the unconditional distribution function of X_j and $F_r(x) = P(X_j \leq x|Y = y_r)$ the conditional distribution function of X_j given $Y = y_r$, if $F_r(x) = F(x)$ for any $x \in R_{X_j}$ and $r = 1, 2, \ldots, R$, then X_j and Y are independent. This motivated Cui et al. (2015) to consider the index

$$MV(X_j|Y) = E_{X_j}[\mathrm{Var}_{\,Y}(F(X_j|Y))]$$

to measure the dependence between X_j and Y. They showed that $MV(X_j|Y) = 0$ if and only if X_j and Y are statistically independent, so they used the $MV(X_j|Y)$ as a marginal utility

for feature screening. They also found an estimator of this quantity

$$\widehat{MV}(X_j|Y) = \frac{1}{n}\sum_{r=1}^{R}\sum_{j=1}^{n}\widehat{p}_r[\widehat{F}_r(X_j) - \widehat{F}(X_j)]^2$$

where $\widehat{p}_r = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{Y_i = y_r\}$, $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\boldsymbol{X}_i \leq x\}$, and $\widehat{F}_r(x) = \frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{X}_i \leq x, Y_i = y_r\}/\widehat{p}_r$. Without specifying a regression model, they defined the active predictor subset by

$$M_* = \{k : F(y|x) \text{ functionally depends on } X_k \text{ for some } y = y_r\}$$

They applied the MV index for each pair (X_k, Y)

$$\omega_k = MV(X_k|Y)$$

and used

$$\widehat{\omega}_k = \widehat{MV}(X_k|Y)$$

to choose the index set

$$\widehat{M} = \{k : \widehat{\omega}_k \geq cn^{-\kappa}, \text{ for } 1 \leq k \leq p\}.$$

Therefore, Cui et al. (2015) showed the sure screening property :

$$P\{M_* \subset \widehat{M}\} \to 1$$

imposing that there exist two positive constants c_1 and c_2 such that $c_1/R_n \leq \min_{1\leq r\leq R_n} p_r \leq \max_{1\leq r\leq R_n} p_r \leq c_2/R_n$, assuming that $R_n = O(n^\kappa)$ for $\kappa \geq 0$ is the diverging number of classes for the response, and that there exist positive constants $c > 0$ and $0 \leq \tau < 1/2$ such that $\min_{k\in D}\omega_k \geq 2cn^{-\tau}$, the minimum true signal. MV-SIS can handle a dimensionality equal to $\log p = O(n^\alpha)$, where $\alpha < 1 \leq 2\tau - \kappa$ with $0 \leq \kappa < 1 - 2\tau$. If R_n is fixed, that is, $\kappa = 0$, then MV-SIS can handle the even larger dimensionality $\log p = O(n^\alpha)$, where $\alpha < 1 - 2\tau$.

In Yan et al. (2018) the purpose is to develop an effective and computationally feasible feature screening procedure for ultra-high dimensional data analysis. The proposed screening procedure can be available for various types of covariates and response variable including discrete, categorical and continuous variables, and is robust to model misspecification, outliers and heavy-tailed distributions of explanatory variables. It is also model-free, without specifying a regression model of explanatory variables and response variable, and is easily implemented without involving the numerical optimization problem. To this end, the authors proposed a marginal slicing feature screening procedure, which is referred to as the slicing Fused Mean-Variance Filter (FMV) screening, based on the empirical conditional distribution function of explanatory variable given response variable. They tried to combine two ideas:

the measure of the dependence between each covariate and the response variable of Cui et al. (2015), and the slicing technique of Mai and Zou (2015). In fact, they extended the MV-SIS method for a categorical response to a continuous response variable and then they used an empirical version of the distribution functions (for the explanatory variables and for the response) to estimate their index. Yan et al. (2018) defined the following index:

$$MV_j = E_{X_j}[\mathrm{Var}_Y F(X_j|Y)] = \int\int \{F_j(x|Y=y) - F_j(x)\}^2 dF_j(x) dF_Y(y).$$

It is possible to use this index for identify the significant explanatory variables in ultra-high dimensional data analysis because $MV_j = 0$ if and only if X_j is independent of Y. In order to solve the integral problem of continuous variable, they transformed it into a tractable sum problem using the slicing method of FKF. In fact, it is quite difficult to estimate MV_j when Y is a continuous random variable or a discrete random variable having countable values because it involves evaluating $F_j(x|y)$ for all possible values y. To address the issue, the authors approximated MV_j by slicing the response Y on its support R_Y. To this end, they defined the following partition of the support R_Y for a given positive integer S:

$$\boldsymbol{S} = \{[a_g, a_{g+1}) : a_g < a_{g+1}, g = 1, \ldots, S\}$$

where $a_1 = \inf\{y : F_Y(y) < 1\}$ and $a_{S+1} = \sup\{y : F_Y(y) < 1\}$. They also defined a random variable $G = \{1, \ldots, S\}$ such that $G = g$ if and only if Y is in the gth slice $[a_g, a_{g+1})$ for $g = 1, \ldots, S$. In particular, when Y is a discrete variable, they took $G = Y$. Although they cannot evaluate $F_j(x|Y = y)$ for all possible values y under our considered case, they approximate $F_j(x|Y = y)$ on a slice $G = g$ (i.e., $a_g \leq Y < a_{g+1}$) by using $F_j^S(x|G = g)$, where $F_j^S(x|G = g) = P(X_j \leq x|G = g)$. Thus, the sliced MV_j can be approximated by

$$MV_j^S = \sum_{g=1}^{S} p_g^S \int \{F_j^S(x|G=g) - F_j(x)\}^2 dF_j(x)$$

where $p_g^S = P(G = g)$ and $F_j^S(x|G = g) = P(X_j \leq x, G = g)/p_g^S = P(X_j \leq x, a_g \leq Y < a_{g+1})/P(a_g \leq Y < a_{g+1})$. MV_j^S enjoys the same property as MV_j since $MV_j^S = 0$ if and only if X_j is independent of Y when Y takes countable values and each possible value of Y forms a slice. However, when Y is a continuous random variable, MV_j^S is a consistent estimator of MV_j for feature screening, assuming $F_j(x|y)$ continuous in y, $\max_{g=1,\ldots,S} P(G = g) \to 0$ and $\lim_{S\to 0} SP(G = g) \to 1$. The sample version of MV_j^S can be estimated by

$$\widehat{MV}_j^S = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{S} \widehat{p}_g^S \{\widehat{F}_j^S(X_{ij}|G=g) - \widehat{F}_j(X_{ij})\}^2.$$

Following the fusion in Mai and Zou (2015), Yan et al. (2018) considered K different slicing schemes, computed the MV_j^S for each K and then they took the sum. The fused mean variance filter $FMV_j = \sum_{k=1}^{K} MV_j^S$ is approximated by

$$\widehat{FMV}_j = \sum_{k=1}^{K} \widehat{MV}_j^S.$$

The authors showed the sure screening property under some regularity conditions:

- there exists a set E such that $M_* \subset E$ and

$$\Delta_E = \min_k \{\min_{j \in E} MV_j^{S_k} - \max_{j \notin E} MV_j^{S_k}\} > 0;$$

- given $S_{min} = \min_k S_k$, for any b_1 and b_2 such that $P(Y \in [b_1, b_2)) \leq (1 + \Delta_E)/S_{min}$, the

$$\sup_{x \in R_{X_j}} |F_j(x|y_1) - F_j(x|y_2)| \leq \Delta_E/8$$

  for any $j \in 1, \ldots, p$ and $y_1, y_2 \in [b_1, b_2)$; and moreover $S_k = O(n^\kappa)$ for $\kappa \geq 0$.

If $\Delta_E \geq Cn^\tau$ with some positive constant C, the FMV procedure can be used to deal with the dimensionality $\log p = O(n^\xi)$, where $\xi < 1 - 2\tau - \kappa$, $0 \leq \tau \leq 1/2$ and $0 \leq \kappa < 1 - 2\tau$, which depends on the minimum true signal strengthen and the number of slices. If the number of slices is not growing with n, namely $\kappa = 0$, the dimensionality achieved is $\log p = O(n^\xi)$, where $\xi < 1 - 2\tau$ with $0 \leq \tau < 1/2$. If $S_k = O(\log(n))$ (Mai and Zou, 2015), the FMV filter enjoys the sure screening property with the probability tending to one only if $\Delta_E \gg \sqrt{\log(n)\log(Kp)/n}$.

Model-free using empirical likelihood

Chang et al. (2016a) considered an independence feature screening method for a general class of regression problems covering the nonparametric and semi-parametric families. This approach directly targets at quantifying the strength of data evidence against the null hypothesis that explanatory variables are not locally contributing to the response variable. Moreover, the statistic in this approach is self-studentized, automatically incorporating variance of the marginal statistical approach. The authors considered the set $M_* = \{1 \leq j \leq p : E(Y|\boldsymbol{X})$ varies with the value of $X_j\}$ as the set of contributing explanatory variables, and, without loss of generality, they assumed $E(Y) = 0$ that implies $E\{m(\boldsymbol{X})\} = 0$, since $\boldsymbol{X}$ is high-dimensional. Without any prior information on which of the covariates are contributing in explaining Y, Chang et al. (2016a) investigated the marginal contribution from each explanatory variable in explaining Y to justify whether it is relevant. For such a purpose, they

considered marginal nonparametric regression problems:

$$\min_{f_j \in L_2} E[\{Y - f_j(X_j)\}^2] \quad \text{with} \quad j = 1, \ldots, p$$

where L_2 denotes the class of square integrable functions. Noting that $E(Y|X_j)$ is the minimizer of the nonparametric regression, they used $f_j(x) = E(Y|X_j = x)$ to evaluate the marginal contribution of X_j locally at $X_j = x$. If an explanatory variable X_j is not contributing to Y marginally, then $f_j(x) = 0$ for all x in the support of X_j, $\mathcal{X}$. The authors considered the Nadaraya–Watson (NW) estimator for $f_j(x)$

$$\widehat{f}_j(x) = \frac{n^{-1}\sum_{i=1}^n K_h(X_{ij} - x)Y_i}{n^{-1}\sum_{i=1}^n K_h(X_{ij} - x)}$$

although this choice does not compromise the general applicability of the marginal empirical likelihood with other nonparametric approaches, for example, the local linear estimator (Fan and Gijbels, 1996), etc. For assessing $f_j(x) = 0$ at a given x without distributional assumptions, they constructed the following empirical likelihood:

$$EL_j(x, 0) = \sup\left\{\sum_{i=1}^n \omega_i : \omega_i \geq 0, \sum_{i=1}^n \omega_i = 1, \sum_{i=1}^n \omega_i K_h(X_{ij} - x)Y_i = 0\right\}.$$

By applying the Lagrange multiplier method, the authors obtained the empirical likelihood ratio:

$$l_j(x, 0) = -2\log\{EL_j(x, 0)\} - 2n\log n = 2\sum_{i=1}^n \log\{1 + \lambda K_h(X_{ij} - x)Y_i\}.$$

Since the denominator of NW converges to the density of X_j evaluated at x, a large value of $l_j(x, 0)$ is taken as evidence against $f_j(x) = 0$ provided that the density of X_j is bounded away from 0 at x. Hence, $l_j(x, 0)$ is indeed a statistic for testing whether or not the numerator of NW has zero mean locally at x. For assessing $E(Y|X_j) \equiv 0$, Chang et al. (2016a) proposed to use

$$l_j(0) = \sup_{x \in X_n} l_j(x, 0)$$

for each $j = 1, \ldots, p$, where $\mathcal{X}_n$ is a partition of the support $\mathcal{X}$ into several intervals. For feature screening purposes, they proposed selecting the set of explanatory variables by

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j(0) \geq \gamma_n\}.$$

The screening property holds in this case under some conditions:

- the continuity of each $f_j(x)$, in fact each f_j has to belong in $C^r(\mathcal{X})$ with bounded derivative, and, if $r = 0$, f_j's satisfy the Lipschitz condition with an order $\alpha \in [0, 1)$;

- the density of each X_j does not vanish on its support and it implies bounded support of the explanatory variables;
- the minimal signal strength of M_*, measured by $\| f_j \|_\infty$, cannot be too weak and depends on the continuity of f_j via r, so $\min_{j \in M} \| f_j \|_\infty \geq c_1 n^{-\kappa}$, with $\kappa \in [0, \frac{g_1}{2g_1+2})$;
- the partition of the support $\mathcal{X}$ has to be of size at least $O(n^{-\xi})$, with $\xi > 0$;
- a condition on the tail distribution of Y, such that $P(|Y| \geq u) \leq K_1 \exp -K_2 u^\gamma$, with K_1, K_2 and $\gamma > 0$;
- the requirement for the kernel function so that the bias due to kernel smoothing is not dominating and $h \asymp n^{-\omega}$ with $\omega \in [\frac{\kappa}{g_1}, 1)$;

with a dimensionality $\log p = o(n^\epsilon)$ for $\epsilon = \min\left\{1 - 2\kappa - \frac{\kappa}{g_1}, \left(\frac{1}{2} - \kappa - \frac{\kappa}{g_1}\right)\gamma\right\}$, with $g_1 = \max\{r, \alpha\}$. If Y follows a normal or sub-normal distribution, then $\gamma = 2$, and the highest dimensionality achieved is $\log p = o(n^{1-2\kappa-\frac{2\kappa}{g_1}})$, while if f_j has derivatives of all the orders, such that $g_1 = r = \infty$, then the highest dimensionality become $\log p = o(n^{1-2\kappa})$. Chang et al. (2016a) show that this procedure can control the size of $\widehat{M}_{\gamma_n}$ and has the selection consistency in the ideal case where

$$\max_{j \notin M_*} \| f_j \|_\infty = o(n^{-\kappa})$$

and imposing $\omega = \frac{\kappa}{g_1}$. In particular, when Y has a compact support, the selection consistency holds if $\log p = o(n^{1-2\kappa-\frac{2\kappa}{g_1}})$; when all $f_j \in C^\infty(\mathcal{X})$ the selection consistency holds if $\log p = o(n^{1-2\kappa})$. When Y has a normal o sub-normal distribution and in presence of orthogonal condition (Huang et al., 2010), the selection consistency holds if $\log p = o(n^{\min\{\frac{1}{2}-\kappa-\frac{\kappa}{g_1}, \frac{1}{3}-\frac{\kappa}{3g_1}\}})$.

Chu and Lin (2018) proposed a method by combining EL, conditional technique and SIRS. They took the correlation among the variable into account and applied the marginal empirical likelihood method on the conditional SIRS: this method is called CSIRS.

Chapter 2

Independence screening by marginal empirical likelihood and local polynomial derivatives

2.1 Introduction of the method

Suppose that we have a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ from the data model

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i \tag{2.1}$$

where Y is the response variable, $\boldsymbol{X}$ is the vector of p candidate variables and ε is the error, with $E(\varepsilon|\boldsymbol{X}) = 0$. Finally, $m(\cdot)$ is the unknown regression function. In our method, we don't impose any particular form to the function $m(\cdot)$, so we consider a general class of regression problems, including parametric and nonparametric, additive and non-additive models. As regard the dimensionality p of the variable vector $\boldsymbol{X}$, this can grow exponentially with the sample size n, and, without loss of generality, we assume that $E(Y) = 0$ implying that $E\{m(\boldsymbol{X})\} = 0$. Let us denote with

$$M_* = \{1 \leq j \leq p : \text{the } j\text{-th variable in } \boldsymbol{X} \text{ is relevant for explanation of } Y\}$$

the set of s true relevant covariates in model (2.1). Moreover, we consider a very sparse model: only a small fraction of the explanatory variables contribute to the response ($s \ll p$).

In order to identify the s relevant covariates in M_* that contribute to the response variable in high-dimensional nonparametric regression analysis, we propose an independence model-free feature screening technique that combines two different elements: the local polynomial regression and the empirical likelihood. We apply the local polynomial regression to estimate

a marginal derivative with respect to the covariate X_j, for $j = 1, \ldots, p$ (so, p derivatives in total) in the regression model (2.1). Once we have this estimation, we use the empirical likelihood to verify if this derivative is zero uniformly in the covariate's support. Until now, based on our knowledge, no other screening method uses marginal derivatives to evaluate the effective incidence of covariates on the dependent variable.

With the use of derivatives, we investigate the marginal contribution from each explanatory variable in explaining Y to justify whether it is relevant or not. In fact, the partial derivative $\frac{\partial m(\boldsymbol{X})}{\partial X_j}$ says in what way the value of $m(\boldsymbol{X})$ changes if you increase X_j by a small amount, while holding the rest of the arguments fixed. We can evaluate partial derivatives using the tools of single-variable calculus: to calculate $\frac{\partial m(\boldsymbol{X})}{\partial X_j}$ simply compute the (single variable) derivative with respect to X_j, treating the rest of the arguments as constants. If an explanatory variable X_j is not contributing to Y marginally, then the derivative $\frac{\partial m(\boldsymbol{X})}{\partial Xj}(x) = 0$ for all $x \in \mathcal{X}_j$, where $\mathcal{X}_j$ is the support of X_j. With this idea in mind, we attempt a feature screening procedure that it is capable to determine whether $\frac{\partial m(\boldsymbol{X})}{\partial X_j} \equiv 0$ or not for each $j = 1, \ldots, p$. The details will be given in section 2.4.

Because we impose no restriction on the structure of the model, we need a nonparametric statistical tool to estimate this partial derivatives. Our choice has fallen on the use of local polynomial regression: with this method we are able to evaluate the contribution of the explanatory variable X_j locally at $X_j = x$.

2.2 Derivative estimation by local polynomials

Local polynomial fitting method has many notable features both from theoretical and practical point of view. Local polynomial fitting adapts to various types of designs (random and fixed, highly clustered and nearly uniform), and there is an absence of boundary effects. In fact, compared to other kernel estimators, with local polynomial fitting no boundary modifications are required. Furthermore, the local polynomial approximation method is appealing on general scientific grounds because it uses the least squares principle.

We now start the exploration of the method of local polynomial fitting. We introduce the framework for this particular smoothing technique in the case of one-dimensional explanatory variables $X_1, \ldots, X_n$, following the notation of Fan and Gijbels (1996). Consider the bivariate data $(X_i, Yi), \ldots, (X_n, Y_n)$, which form an independent and identically distributed sample from a population (X, Y). Of interest is to estimate the regression function $m(x_0) = E(Y|X = x_0)$ and its derivatives $m'(x_0), \ldots, m^{(d)}(x_0)$. To help us understand the estimation methodology, we can regard the data as being generated from the model

$$Y = m(X) + \sigma(X)\varepsilon,$$

where $E(\varepsilon) = 0$, $Var(\varepsilon) = 1$, and X and ε are independent. We always denote the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X, i.e. the design density, by $f(\cdot)$. Suppose that the $(d+1)^{th}$ derivative of $m(x)$ at the point x_0 exists and is continuous. We approximate locally the marginal function $m(x)$ using Taylor's Expansion by a polynomial of order d (Fan and Gijbels, 1996):

$$m(x) \approx m(x_0) + m^{(1)}(x_0)(x - x_0) + \cdots + \frac{m^{(d)}(x_0)}{d!}(x - x_0)^d.$$

Then, we can estimate the expansion terms using weighted least squares by minimizing the following equation for $\beta_v := \beta_v(x_0) = m^{(v)}(x_0)/v!$:

$$\sum_{i=1}^{n} \left[Y_i - \sum_{v=0}^{d} \beta_v(x_0)(X_i - x_0)^v \right]^2 K_h(X_i - x_0) \tag{2.2}$$

where h, called the bandwidth, controls the size of the neighbourhood around x_0, $K_h(\cdot)$ controls the weights, with $K_h(x) \equiv K(x/h)/h$ and K a kernel function satisfying $\int K(x)dx = 1$. Let $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)$ be the solution of the minimization problem in (2.2), then $\widehat{m}^{(v)}(x_0) = v!\widehat{\beta}_v$ is an estimator for $m^{(v)}(x_0)$, with $v = 0, \ldots, d$. In matrix notation, let $\mathbf{X}$ be the design matrix centred at x_0:

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \ldots & (X_1 - x_0)^d \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \ldots & (X_n - x_0)^d \end{pmatrix}, \tag{2.3}$$

$Y = (Y_1, \ldots, Y_n)^T$ and $\mathbf{W}$ a diagonal matrix of weights with diagonal elements $K_h(X_i - x_0)$, for $i = 1, \ldots, n$. Then, the local estimate of $m^{(v)}(x)$ with a dth degree polynomial is

$$\widehat{m}^{(v)}(x; d, h) = v!\boldsymbol{e}_{v+1}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}Y = v!\sum_{i=1}^{n} W_{id}(x)Y_i \tag{2.4}$$

for $v = 0, \ldots, d$, where $W_{i,d}(x) = \boldsymbol{e}_{v+1}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{e}_i$. Here e_r is the $(d+1) \times 1$ vector having 1 in the rth entry and zeros elsewhere.

To estimate the function $m^{(v)}(\cdot)$ we need to solve the weighted least square problem for all points x_0 in the domain of interest. We remark that we do not need to know whether $Var(Y|X = x)$ remains constant or not, because we fit (2.2) locally and the variance is approximately the same in a local neighbourhood. This is a great advantage of the local polynomial fitting. The matrix $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is positive definite as long as there are at least $d+1$ local effective design points. This assumption is granted with probability tending to one since we always assume that $nh \to \infty$ (Masry and Fan, 1997).

The estimate for $m(x)$ (when we consider $v = 0$) is therefore computed as

$$\widehat{m}(x; d, h) := \widehat{\beta} = \boldsymbol{e}_1^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} Y = \sum_{i=1}^{n} W_{i,d}(x) Y_i$$

where

$$W_{i,d}(x) := \boldsymbol{e}_1^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{e}_i$$

and $\boldsymbol{e}_i$ is the i-th canonical vector. So, the local polynomial estimator for the function $m(x)$ is a weighted combination of the responses. When $d = 0$, the local polynomial estimator is the Nadaraya-Watson estimator with this explicit weights formulation

$$W_{i,0}(x) = \frac{K_h(X_i - x)}{\sum_{l=1}^{n} K_h(X_l - x)},$$

when $d = 1$, it is the local linear estimator, which has weights equal to

$$W_{i,1}(x) = \frac{1}{n} \frac{\widehat{s}_2 - \widehat{s}_1 (X_i - x)}{\widehat{s}_2 \widehat{s}_0 - \widehat{s}_1^2} K_h(X_i - x)$$

where $\widehat{s}_r = \widehat{s}_r(x; h) := \frac{1}{n} \sum_{i=1}^{n} (X_i - x)^r K_h(X_i - x)$.

The extension of local polynomial fitting ideas to estimation of the vth derivative is straightforward. One can estimate $m^{(v)}(x)$ via the intercept coefficient of the vth derivative of the local polynomial being fitted at x, assuming that $v \leq d$. For example, the local polynomial estimate of $m'(x)$ is simply the slope of the local polynomial fit. In general, the local estimate of $m^{(v)}(x)$ with a dth degree polynomial is

$$\widehat{m}^{(v)}(x; d, h) = v! \boldsymbol{e}_{v+1}^T (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} Y \tag{2.5}$$

for all $v = 0, \ldots, d$. As before, e_{v+1} is the $(d + 1) \times 1$ vector having 1 in the $(v + 1)$th entry and zeros elsewhere.

There are three critical parameters whose choice can have an effect on the quality of the fit. These are the bandwidth, h, the order of the local polynomial being fit, d, and the kernel or weight function, K (often denoted K_h to emphasize its dependence on the bandwidth):

- a too large bandwidth under-parametrizes the regression function, causing a large modelling bias, while a too small bandwidth over-parametrizes the unknown function and results in noisy estimates. Ideal theoretical choices of the bandwidth are easy to obtain, but is not directly practically usable since it depends on unknown quantities;

- for a given bandwidth h, a large value of d would expectedly reduce the modelling bias, but would cause a large variance and a considerable computational cost. Fan and Gijbels (1996) showed that there is a general pattern of increasing variability: for estimating

$m^{(v)}(x_o)$, there is no increase in variability when passing from an even (i.e. $d-r$ even) $d = v + 2q$ order fit, with $q \in \mathbb{N}$, to an odd $d = v + 2q + 1$ order fit, but when passing from a odd $d = v+2q+1$ order fit to the consecutive even $d = v+2q+2$ order fit there is a price to be paid in terms of increased variability. Therefore, even order fits $d = v + 2q$ are not recommended. Since the bandwidth is used to control the modelling complexity, it is recommend the use of the lowest odd order, i.e. $d = v+1$, or occasionally $d = v+3$;

- since the estimate is based on the local regression, no negative weight K should be used. In fact, for all choices of d and v the optimal weight function is $K(z) = \frac{3}{4}(1-z^2)_+$, the *Epanechnikov* kernel, which minimizes the asymptotic MSE of the resulting local polynomial estimators (Fan and Gijbels, 1996).

It is possible to use the notation in (2.2) to express the conditional mean and variance of $\widehat{\beta}$:

$$E(\widehat{\beta}|\boldsymbol{X}) = \beta + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\mathbf{r}$$

$$Var(\widehat{\beta}|\boldsymbol{X}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$$

where $\boldsymbol{r} = (m(X_1), \ldots, m(X_n)) - \boldsymbol{X}\beta_j$ is the vector of residuals of the local polynomial approximation, and $\boldsymbol{\Sigma} = diag\{K_h^2(X_i - x_0)\sigma^2(X_i)\}$. These exact bias and variance expressions are not directly usable, since they depend on unknown quantities: the residuals $\boldsymbol{r}$ and the diagonal matrix $\boldsymbol{\Sigma}$.

When $(X_i, Y_i), \ldots, (X_n, Y_n)$ is an *i.i.d.* sample from the population (X, Y), Theorem 3.1 of Fan and Gijbels (1996) shows the following result on the approximation of bias and variance, using this notation. The moments of K and K^2 are denoted respectively by $\mu_j = \int u^j K(u)du$ and $\nu_j = \int u^j K^2(u)du$. Considering some matrices and vectors of moments

$$S = (\mu_{j+l})_{0\le j,l\le d} \qquad c_d = (\mu_{d+1}, \ldots, \mu_{2d+1})^T$$
$$\bar{S} = (\mu_{j+l+1})_{0\le j,l\le d} \qquad \bar{c}_d = (\mu_{d+2}, \ldots, \mu_{2d+2})^T$$
$$S^* = (\nu_{j+l})_{0\le j,l\le d}$$

and the unit vector $e_{v+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ with 1 on the $(v+1)^{th}$ position. Assume $f(x_0) > 0$, $f(\cdot)$, $m^{(d+1)}(\cdot)$ and a $\sigma^2(\cdot)$ are continuous in a neighbourhood of x_0. Further, assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\widehat{m}^v(x_0)$ is given by

$$Var\{\widehat{m}^{(v)}(x_0)|X\} = e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{v!\sigma^2(x_0)}{f(x_0)nh^{1+2v}} + o_P\left(\frac{1}{nh^{1+2v}}\right).$$

The asymptotic conditional bias for $d-v$ odd is given by

$$Bias\{\widehat{m}^{(v)}(x_0)|X\} = e_{v+1}^T S^{-1} c_d \frac{v!}{(d+1)!} m^{(d+1)}(x_0) h^{d+1-v} + o_P(h^{d+1-v}).$$

Further, for $d-v$ even the asymptotic conditional bias is

$$Bias\{\widehat{m}^{v}(x_0)|X_j\} = e_{v+1}^T S^{-1} \bar{c}_d \frac{v!}{(d+2)!} \left\{ m^{(d+2)}(x_0) + (d+2) m^{(d+1)}(x_0) \frac{f'(x_0)}{f(x_0)} \right\} h^{d+2-v} + o_p(h^{d+2-v})$$

provided that $f'(\cdot)$ and $m^{(d+2)}(\cdot)$ are continuous in a neighbourhood of x_0 and $nh^3 \to \infty$. From the above result it is possible to see that there is a theoretical difference between the cases $d-v$ odd and $d-v$ even. For $d-v$ odd the asymptotic bias has a simpler structure and does not involve $f'(x_0)$, a factor appearing in the asymptotic bias when $d-v$ is even. For this reason, the Nadaraya-Watson estimator, that is a special case of polynomial regression that uses the local constant fit ($d=0$) for estimating the regression function ($v=0$), has an additional term in the asymptotic bias.

In order to give an exhaustive summary on the theoretical results in the literature on the use of local polynomials, we consider the case of dependent data. When the data are dependent and $d-v$ is odd, Masry and Fan (1997) obtain the expression of bias and variance of this estimator using some additional conditions. First of all, they introduce the mixing coefficients. Let F_i^k be the σ-algebra of events generated by the random variables $\{(X_j, Y_j), i \leq j \leq k\}$ and denote by $L_2(F_i^k)$ the collection of all random variables which are F_i^k-measurable and have finite second moment. The stationary process $\{(X_j, Y_j)\}$ is called strongly mixing or α-mixing if

$$\sup_{A \in F_{-\infty}^0, B \in F_k^\infty} |P(AB) - P(A)P(B)| = \alpha(k) \to 0 \quad \text{as} \quad k \to \infty.$$

The mixing conditions indicate basically the maximum dependence between two time events at least k steps apart. Local polynomial fitting techniques continue to apply under the weak dependence in medium or long term, namely, when k is large. The short term dependence does not have much effect on the local smoothing method. The reason is that for any two given random variables X_i and X_j and a point x, the random variables $K_h(X_i - x)$ and $K_h(X_j - x)$ are nearly uncorrelated as $h \to 0$. This property is, however, not shared by parametric estimators.

Masry and Fan (1997) in their Theorem 5 state that under certain mixing conditions, local polynomial estimators for dependent data have the same asymptotic behaviour as for independent data. Note that the bias arguments are unaffected, whereas the variance calculations are affected under dependence. Let $f(x)$ be the density of X_l and $\sigma^2(x) = Var(Y|X_l = x)$. Let S, S^* and c_p denote the same moment matrices and vector as those introduced before. Under conditions:

1. the kernel K is bounded with bounded support;

2. $f_{X_o,X_l|Y_0,Y_l}(x_0,x_l|y_0,y_l) \leq A-1<\infty \quad \forall l \geq 1$;

3. with α-mixing processes for some $\delta > 2$ and $\alpha < 1-2/\delta$,

$$\sum_l l^{\alpha}[\alpha(l)]^{1-2\delta} < \infty, \quad E|Y_0|^{\delta} < \infty, \quad f_{X_0|Y_0}(x|y) \leq A_2 < \infty$$

4. for α-mixing processes there exists a sequence of positive integers satisfying $s_n \to \infty$ and $s_n = o\{(nh_n)^{1/2}\}$ such that

$$(n/h_n)^{1/2}\alpha(s_n) \to 0, \text{as} \quad n \to \infty$$

if $h_n = O(n^{1/(2d+3)})$, then as $n \to \infty$,

$$\sqrt{nh_n^{2v+1}}\left\{\widehat{m}^{(v)}(x) - m^{(v)}(x) - \frac{e_{v+1}^T S^{-1} c_d v! m^{(d+1)}(x)}{(d+1)!}h_n^{d+1-v}\right\} \xrightarrow{\text{d}} N\left(0, \frac{e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1}(v!)^2\sigma^2(x)}{f(x)}\right)$$

at continuity points of σ^2, f with $f(x) > 0$.

With this consideration, we need a polynomial of order two to estimate the marginal first derivative for our nonparametric model. In fact, for $v = 1$ (the order of the derivative that we consider), we need a polynomial of order $d = v + 1$. When $d = 2$ the local polynomial estimator is called *local quadratic estimator* and it has this formulation for its weights

$$W_{i,2}(x) = \frac{1}{n}S(x;h)K_h(X_i - x) \tag{2.6}$$

where

$$S(x;h) = \frac{\widehat{s}_2\widehat{s}_3 - \widehat{s}_1\widehat{s}_4 + [\widehat{s}_0\widehat{s}_4 - \widehat{s}_2^2](X_i - x) + [\widehat{s}_1\widehat{s}_2 - \widehat{s}_0\widehat{s}_3](X_i - x)^2}{\widehat{s}_0\widehat{s}_2\widehat{s}_4 + 2\widehat{s}_1\widehat{s}_2\widehat{s}_3 - \widehat{s}_2^3 - \widehat{s}_0\widehat{s}_3^2 - \widehat{s}_1^2\widehat{s}_4}$$

Following the previous formulas, the bias and the variance are equal both in the case of dependent data and in that of independent data and have those expressions:

$$Bias\{\widehat{m}'(x_0)|X\} = \frac{1}{3!}e_2^T S^{-1} c_2 m^{(3)}(x_0)h^2 + o_P(h^2)$$

$$Var\{\widehat{m}'(x_0)|X\} = e_2^T S^{-1} S^* S^{-1} e_2 \frac{\sigma^2(x_0)}{f(x_0)nh^3} + o_P\left(\frac{1}{nh^3}\right)$$

2.2.1 The choice of the bandwidth

A theoretical optimal local bandwidth for estimating $m^{(v)}(x_0)$ is obtained by minimizing the conditional Mean Squared Error (MSE) given by

$$[Bias\{\widehat{m}^{(v)}(x_0)|X\}]^2 + Var\{\widehat{m}^{(v)}(x_0)|X\}$$

This ideal choice of a local bandwidth can be approximated by the asymptotically optimal local bandwidth, i.e. the bandwidth which minimizes the asymptotic MSE. Using the expressions for the bias and variance introduced before, it is possible to obtain the asymptotic MSE, whose minimization leads to

$$h_{opt}(x_0) = C_{v,d}(K)\left[\frac{\sigma^2(x_0)}{\{m^{(d+1)}(x_0)\}^2 f(x_0)}\right]^{1/(2d+3)} n^{-1/2d+3}$$

where

$$C_{v,d}(K) = \left[\frac{(d+1)!^2(2v+1)\int K_v^{*2}(t)dt}{2(d+1-v)\{\int t^{d+1}K_v^*(t)dtP\}^2}\right]^{1/(2d+3)}$$

and K^* is the equivalent Kernel (Fan and Gijbels, 1996).

A commonly used, simple measure of global loss is the weighted Mean Integrated Squared Error (MISE). Minimization of the conditional weighted MISE

$$\int \left(\left[Bias\{\widehat{m}^{(v)}(x)|X\}\right]^2 + Var\{\widehat{m}^{(v)}(x)|X\}\right) w(x)dx$$

with $w(\cdot) \geq 0$ some weight function, leads to a theoretical optimal constant bandwidth. Using again the asymptotic expressions for bias and variance we find an asymptotically optimal constant bandwidth given by

$$h_{opt} = C_{v,d}(K)\left[\frac{\int \sigma^2(x)w(x)/f(x)dx}{\int\{m^{(d+1)}(x)\}^2 w(x)dx}\right]^{1/(2d+3)} n^{-1/(2d+3)}.$$

It is understood that the integrals are finite and that the denominator does not vanish. These asymptotically optimal bandwidths depend on unknown quantities such as the design density $f(\cdot)$, the conditional variance $\sigma^2(\cdot)$ and the derivative function $m^{(d+1)}(\cdot)$, and hence further work is needed for achieving practical bandwidth selection procedures.

In Wand and Jones (1994) there is a rich review of the different type of bandwidth selector based on the minimization of the MISE. Since the purpose of this thesis is not to find the optimal bandwidth, we will use the Leave-one-out method that is a conceptually simple and appealing bandwidth selector. Leave-one-out cross validation uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation

data. Although this method often leads to an overestimation of the bandwidth, the results of our simulations encourage us to consider that this problem does not affect the results in terms of relevant variables screening.

2.3 Empirical likelihood

Empirical likelihood is a nonparametric method of statistical inference that uses likelihood methods, without having to assume that the data come from a known family of distributions. An excellent review on this topic can be found in Owen (2001).

Likelihood methods can be used to find efficient estimators, to construct tests with good power properties and when the data are incompletely observed, or distorted, or sampled with a bias, they can be used to offset or correct for these problems. Likelihood can be used to pool information from different data sources. In fact, it is possible to incorporate knowledge arising from outside of the data. This knowledge may take the form of constraints that restrict the domain of the likelihood function, or it may be in the form of a prior distribution to be multiplied by the likelihood function. In parametric likelihood methods, the joint distribution of all available data is assumed to have a known form, apart from one or more unknown quantities. The problem behind parametric approaches lies in the choice of the parametric family to use. Indeed, there is no reason to suppose that a newly encountered dataset belongs to one of the well-known parametric families. Such incorrect specification can render likelihood-based estimates inefficient and, consequently, the confidence intervals and the corresponding tests may completely fail. To solve this problem, many statisticians have turned to nonparametric inferences to avoid having to specify a parametric family for the data.

The advantages of empirical likelihood arise because it combines the reliability of the nonparametric methods with the flexibility and effectiveness of the likelihood approach. The name "empirical likelihood" was adopted because the empirical distribution of the data plays a central role. It was not called nonparametric likelihood, so as not to assume that it would be the only way to extend nonparametric maximum likelihoods to likelihood ratio functions.

The empirical cumulative distribution function is a nonparametric maximum likelihood estimate (NPMLE). In fact, for a random variable $X \in \mathbb{R}$, the cumulative distribution function (CDF) is the function $F(x) = P(X \leq x)$, for $-\infty < x < \infty$. We use $F(x-)$ to denote $P(X < x)$ and so $P(X = x) = F(x) - F(x-)$. Let $X_1, \ldots, X_n \in \mathbb{R}$, the empirical cumulative distribution function (ECDF) of $X_1, \ldots, X_n$ is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$

for $-\infty < x < \infty$. Moreover, assumed $X_1, \ldots, X_n \in \mathbb{R}$ independent with common CDF F_0,

the nonparametric likelihood of the CDF F is

$$L(F) = \sum_{i=1}^{n} (F(X_i) - F(X_i-)).$$

This is the probability of getting exactly the observed sample values $X_1, \ldots, X_n$ from the CDF F. One consequence is that $L(F) = 0$ if F is a continuous distribution. To have a positive nonparametric likelihood, a distribution F must place positive probability on every one of the observed data values. Owen (2001) in Theorem 2.1 proves that the nonparametric likelihood is maximized by the ECDF. Thus the ECDF is the NPMLE of F.

There are some analogies between parametric and nonparametric likelihood methods. In parametric models, given $\widehat{\eta}$ the MLE of η, the MLE of $\theta(\eta)$, where θ is a particular function, will be $\widehat{\theta} = \theta(\widehat{\eta})$. In the nonparametric setting, it is possible to consider $\theta = T(F)$, where F is a continuous distribution and T is a real-valued function of distributions. The true unknown parameter is $\theta_0 = T(F_0)$. Proceeding by analogy, the NPMLE of θ will be $\widehat{\theta} = T(F_n)$. Thus, if the function T is the mean function of X when X has the distribution F_0, *i.e.* $\theta_0 = \int x dF_0(x)$, then by analogy the NPMLE of θ_0 will be the mean of F_n. This mean is of course $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$. For a subset $A \subset \mathbb{R}$, the NPMLE of $P(X \in A)$ will be the sample fraction of X_i in A.

In parametric inference we may base hypothesis tests and confidence regions on the likelihood ratio. If $L(\eta)$ is much smaller than $L(\widehat{\eta})$, then we reject the hypothesis that $\eta_0 = \eta$, and exclude η from our confidence region for η_0. Wilks's theorem provides that $-2\log(L(\eta_0)/L(\widehat{\eta}))$ tends to a chi-squared distribution as $n \to \infty$, under mild regularity conditions, allowing us to decide just how small $L(\eta)$ must be in order for η to get rejected. The degrees of freedom in the chi-squared distribution are usually equal to the dimension of the set of η values. When we want a confidence region for θ we take the image of a confidence region for η. That is

$$\{\theta(\eta) | L(\eta) \geq cL(\widehat{\eta})\}$$

where the threshold c is chosen using Wilks's theorem, with degrees of freedom equal to the dimension of the set of θ values. We may also use ratios of the nonparametric likelihood as a basis for hypothesis tests and confidence intervals. For a distribution F, define

$$R(F) = L(F)/L(F_n)$$

through the nonparametric likelihood $L(F)$. We proceed by analogy with parametric likelihood. Suppose that we are interested in a parameter $\theta = T(F)$ for some function T of distributions. This F is a member of a set $\boldsymbol{F}$ of distributions. Define the profile likelihood

ratio function:

$$R(\theta) = \sup\{R(F)|T(F) = \theta, F \in \boldsymbol{F}\}.$$

Empirical likelihood hypothesis tests reject $H_0 : T(F_0) = \theta_0$, when $R(\theta_0) < r_0$ for some threshold value r_0. Empirical likelihood confidence regions are of the form

$$\{\theta|R(\theta) \geq r_0\}.$$

In many settings, the threshold r_0 may be chosen using an empirical likelihood theorem (ELT), a nonparametric analogue of Wilks's theorem. In particular, the Theorem 2.2 of Owen (2001) for the empirical likelihood used to the univariate mean shows this result: considering $X_1, \ldots, X_n$ independent random variables with common distribution F_0, $\mu_0 = E(X_i)$, and supposing that $0 < Var(X_i) < \infty$, then $-2\log(R(\mu_0))$ converges in distribution to $\chi^2_{(1)}$ as $n \to \infty$. It is possible to consider two aspects of this theorem. First, the chi-squared limit is the same as we typically find for parametric likelihood models with one parameter. Second, there is no assumption that X_i are bounded random variables. They only need to have a bounded variance, which constrains how fast the sample maximum and minimum can grow as n increases.

Supposing that there are no ties in the data, let $w_i = F(X_i)$, $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$, the nonparametric likelihood has the form

$$L(F) = \prod_{i=1}^n w_i \quad \text{and} \quad L(\widehat{F}) = \prod_{i=1}^n \frac{1}{n},$$

the nonparametric likelihood ratio is

$$R(F) = \prod_{i=1}^n nw_i$$

and the profiled likelihood is

$$R(\theta) = \sup\left\{\prod_{i=1}^n nw_i|T(F) = \theta\right\}.$$

2.3.1 Empirical likelihood for the mean

To test whether $\mu = \mu_0$, we need to compute $R(\mu_0)$. To set confidence limits for μ, we need to find the two values of μ that solve the equation $R(\mu) = r_0$, given a threshold value r_0. To compute the curve $R(\mu)$, let the ordered sample values be $X_{(1)} \leq, \ldots, \leq X_{(n)}$. First we eliminate the trivial cases: if $\mu < X_{(1)}$ or $\mu > X_{(n)}$ then there are no weights $w_i \geq 0$ summing to 1 for which $\sum_{i=1}^n w_i X_i = \mu$. In such cases we take $\log R(\mu) = -\infty$, and $R(\mu) = 0$

by convention. Similarly if $\mu = X_{(1)} < X_{(n)}$ or $\mu = X_{(n)} > X_{(1)}$ we take $R(\mu) = 0$, but if $X_{(1)} = X_{(n)} = \mu$, we take $R(\mu) = 1$.

Considering the non trivial case, with $X_{(1)} < \mu < X_{(n)}$, we seek to maximize $\prod_{i=1}^{n} nw_i$, or equivalently $\sum_{i=1}^{n} \log(nw_i)$ over $w_i \geq 0$ subject to the constraints that $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i X_i = \mu$. So, the profiled likelihood has the form

$$R(\mu) = \sup_{w} \left\{ \prod_{i=1}^{n} nw_i | w_i > 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i X_i = \mu \right\}.$$

The objective function $\sum_{i=1}^{n} \log(nw_i)$ is a strictly concave function on a convex set of weight vectors. Accordingly, a unique global maximum exists. We also know that the maximum does not have any $w_i = 0$, so it is an interior point of the domain.

Proceeding with the method of Lagrange multipliers, it is possible to find that

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda(X_i - \mu)}$$

and the value λ depends on the value of μ, and solves

$$\frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0$$

2.4 The proposed procedure

As mentioned before, we estimate the first marginal derivative of our nonparametric model (2.1) using the local quadratic estimator (2.5) with weights (2.6). In order to use the univariate local quadratic estimator, we consider $f_j(x) = E(Y|X_j = x)$, that is the marginal contribution of X_j locally at $X_j = x$. On one hand, if $m(\cdot)$ is an additive function as in (1.4), then

$$\frac{\partial m(\boldsymbol{X})}{\partial X_j}|_{X_j = x} = \frac{\partial m_j(\boldsymbol{X})}{\partial X_j}|_{X_j = x} = f_j'(x), \tag{2.7}$$

where $m_j(\cdot)$ is the part of $m(\cdot)$ relative to the X_j variable alone and $f_j'(\cdot)$ is the first derivative of $f_j(\cdot)$. On the other hand, if $m(\cdot)$ is not additive, the equality (2.7) is not true, but we can evaluate again the marginal incidence of the explanatory variable X_j by $f_j(\cdot)$. In fact, for an explanatory variable X_j that is not contributing to Y marginally, $f_j'(x) = 0$ for all $x \in \mathcal{X}_j$, with $\mathcal{X}_j$ is the support of X_j. This suggests to investigate a feature screening procedure by assessing whether $f_j' \equiv 0$ or not for each $j = 1, \ldots, p$. We remark that we have chosen to work on the univariate marginal derivative to efficiently apply the univariate estimate of the derivative using local polynomials. This, although it significantly alters the regression function, does not entail significant consequences in terms of the choice of variables, since both the original

(non-additive) model and the one on which we work (its marginalized additive approximation) depend on the same set of relevant variables.

For assessing $f_j'(x) = 0$ at given x without distributional assumptions, we construct the following empirical likelihood, miming the same steps of Chang et al. (2016a):

$$EL_j(x,0) = \sup_w \left\{ \prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i W_{i,2}(x) Y_i = 0 \right\}. \tag{2.8}$$

By applying the Lagrange multiplier method for solving (2.8), we obtain the empirical likelihood ratio

$$l_j(x,0) = -2\log\{EL_j(x,0)\} - 2n\log n = 2\sum_{i=1}^n \log\{1 + \lambda W_{i,2}(x) Y_i\} \tag{2.9}$$

where λ is the univariate Lagrange multiplier solving $\sum_{i=1}^n \frac{W_{i,2}(x)Y_i}{1+\lambda W_{i,j,2}(x)Y_i} = 0$.
Since $\sum_{i=1}^n W_{i,2}(x)Y_i$ converges to the marginal derivative of X_j evaluated ad x, a large value of $l_j(x,0)$ is taken as evidence against $f_j'(x) = 0$. Then, $l_j(x,0)$ is a statistic for testing whether or not (2.5) with $W_{i,2}$ defined in (2.6) has zero mean locally at x. For assessing $f_j'(x) \equiv 0$ uniformly on $\mathcal{X}_j$, we use

$$l_j(0) = \sup_{x \in \mathcal{X}_j} l_j(x,0)$$

for each $j = 1, \ldots, p$.

For feature screening purpose, we sort l_j for all $j = 1, \ldots, p$ in decreasing order, and we take the first γ_n covariates. In this way, we create a set

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j \geq \gamma_n\}$$

We will specify later the value of γ_n for which the proposed approach is capable to identify the true relevant covariates, the so called sure screening property.

In order to implement the proposed method, we evaluate the statistic l_j using $l_j(0) = \max_{1 \leq i \leq n} l_j(X_{ij}, 0)$, where X_{ij} is the ith observation of the jth explanatory variable. With this expedient, we can use the univariate optimisation to solve (2.9) using the Lagrange multiplier method.

2.5 From screening to variable selection

The substantial difference between variable selection and screening selection lies in the specification of a threshold γ_n. In fact, using variable selection procedure, a set of covariates that is exactly the true one is selected. Thus the exact value of γ_n is known. The so called screening property ensures that the result of the screening procedure is a set of covariates

which contains the true relevant ones. This set relies on the threshold γ_n, but its value is not known exactly because depends on some unknown quantities. Since independence feature screening was introduced by Fan and Lv (2008), in every method proposed in literature, it has always been difficult to choose the γ_n value in practice.

In variable selection literature, nonparametric and non-additive methods present some drawback. For example, in the Rodeo procedure (Lafferty and Wasserman, 2008) the variables have uniform distribution and the dimensionality of p is $O(\frac{\log n}{\log(\log n)})$ while the intrinsic dimension s, the number of relevant variables, does not increase with n. We propose to transform our screening method in a variable selection method that also works with a larger dimensionality and, moreover, allows the relevant covariates to grow with n. Furthermore, it can be used with any type of covariate's distribution.

As suggested by Hall and Miller (2009), it is possible to use the variable screening results for variable selection. In each variable screening method, model-based or model-free, the important covariates are likely to be ranked ahead of the irrelevant ones. Usually, to obtain exactly the relevant covariates, two steps are performed. In the first one a variable screening is performed, in the second one a variable selection is made on the top ranked covariates resulting from the screening. Instead of using this two step procedure, we propose a method that uses the subsample idea to transform a screening selection technique in a variable selection technique.

Meinshausen and Bühlmann (2010) proposed the Stability selection in order to find an estimation of the λ value in the LASSO of Tibshirani (1996), using the subsample technique. Stability selection is based on subsampling in combination with (high dimensional) selection algorithms. The method is extremely general, in fact, in the first stage, a chosen variable selection technique is applied to randomly picked subsamples of the data of size $\lfloor n/2 \rfloor$. In the second one, the variables which are most likely to be selected at the first stage, using a prespecified threshold, are taken as the final estimate of the set of important variables. They prove for the randomized lasso that stability selection will be variable selection consistent, even if the necessary conditions for consistency of the original method are violated. Moreover, stability selection will asymptotically select the right model in scenarios where the lasso fails. In short, stability selection is the marriage of subsampling and high dimensional selection algorithms.

We propose to use the same sumbsample idea not on the result of a variable selection procedure, as in Meinshausen and Bühlmann (2010), but after a screening procedure in order to evaluate the stability of the top ranked screened variables. With this method, the variables selected through the D-ELSIS are then further evaluated to investigate their probability to be chosen when the data are randomly sampled. Some considerations motivate this proposal. As the screening property suggests, in each screening result the probability that the true

covariates fill the first positions in the ranking tends to 1:

$$P(M_* \subseteq \widehat{M}_{\gamma_n}) \to 1.$$

For every subset of $I \subseteq M_*$, the probability still tends to 1:

$$P(I \subseteq \widehat{M}_{\gamma_n}) \to 1.$$

While if we consider a subset $I \not\subset M_*$, so a subset of the variables containing also irrelevant covariates, this probability tends to zero

$$P(I \subseteq \widehat{M}_{\gamma_n}) \to 0.$$

In fact, when we consider a subsample of the variables containing also irrelevant covariates (as a result after screening), the probability that they will consistently exhibit an high influence over the dependent variable Y over many subsamples of the data is small. Following this considerations, we can choose a threshold π greater than zero and less that 1 to discriminate between relevant and not relevant covariates.

In order to identify the true set of relevant covariates, we implement the following procedure. At the first step, we perform D-ELSIS screening using all the observations in the dataset, obtaining a ranking of covariates. We choose only the first $p*$ of these, where $p*$ is a previously chosen threshold, obtaining the set $\widehat{M}_{p*}$. At the second step, we randomly create G subsample of the dataset of size $m = \lfloor n/2 \rfloor$. On these G sets, we carry out the screening procedure again, obtaining a G different ranking of the p covariate. As done for the ranking on the whole dataset, we consider only the first $p*$ ranked covariates from each subsample, obtaining the sets $\widehat{M}^{(i)}_{p*}$ for $i = 1, \ldots, G$. At the third step, we consider a set K with only k covariates included in $\widehat{M}_{p*}$ for $k = 1, \ldots, p*$. For every set K (that includes at most the $p*$ screened variables) it is computed the probability :

$$\pi_n(K) = P(K \subseteq \widehat{M}_{p*}). \tag{2.10}$$

where $\pi_n(K)$ is estimated as the relative frequency that $K \subset \hat{M}_{p*}$ over all G subsets of size m. To obtain an estimator of (2.10), we compute the relative frequency that the same variables that are in K, they are also in $\widehat{M}^{(i)}_{p*}$ for $i = 1, \ldots, G$:

$$\widehat{\pi}_{n,m,G}(K) = \frac{\sum_{i=1}^{G} \mathbb{I}(K \subseteq \widehat{M}^{(i)}_{p*})}{G}. \tag{2.11}$$

Defining a further threshold, π, between 0 and 1, the set of variables K whose relative frequency $\pi_n(K)$ will exceed this threshold, will be that one to be considered as relevant, thus the stable

covariates.

In practice, at the third step we proceed in the following way. Consider all subsets K of size 1, thus we have p^* subsets formed by a single covariate of $\widehat{M}_{p^*}$. We call this subsets K_i^*, for $i = 1, \ldots, p^*$. For each of these, their relative frequency is calculated by (2.11). We choose the set K_i^* whose relative frequency is higher than the threshold π. In the event that more than one set exceeds this threshold, we choose the one with the highest relative frequency. We call this set K_1 and we denote with X_{j_1} the stable covariate, thus $K_1 = \{X_{j_1}\}$. We update the subset K_1 adding to X_{j_1} all the other $p^* - 1$ variables, obtaining the sets K_i^*, for $i = 1, \ldots, p^* - 1$, of two covariates. We repeat the procedure for calculating the relative frequency for each of these. If the relative frequency of at least one of these subsets, for example the subset $\{X_{j_1}, X_{j_2}\}$, exceeds the threshold π, then we will have the set $K_2 = \{X_{j_1}, X_{j_2}\}$ as a stable set. This set will still be updated, forming $p^* - 2$ sets with three covariates. And it will go on until at a certain step $s \leq p^*$ at which the threshold π is not exceeded by any subset K_s formed by s variables. The stable variables will be the variables chosen in the previous step, therefore those contained in K_{s-1}.

We want to underline that the choice of the thresholds p^* and π does not condition the result of our procedure. The result of the screening, as we have already discussed extensively, is a subset which contains not only the true relevant covariates, but also not relevant ones. In the literature, a threshold $\gamma_n < n$ is chosen conservatively. To be uniform, it is possible to choose $p^* = n$ and implement the proposed procedure. Or choose a lower value. In fact, given the principle of sparsity, the number of true covariates will be much lower than the number of available variables. For the value of the threshold π, since for any value of π between 0 and 1, we are able to separate the real relevant covariates from those mistakenly considered as such given the nature of the screening procedure, we do not need to estimate this quantity. A higher value will only speed up the iterative procedure.

Chapter 3

Theoretical results

Screening property of D-ELSIS

In this section we show theoretically that our D-ELSIS procedure has the sure screening property and its dimensionality. We recall the fundamental quantities for our procedure.

We have a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ from the data model

$$Y_i = m(\boldsymbol{X}_i) + \varepsilon_i \tag{3.1}$$

where Y is the response variable, $\boldsymbol{X}$ is the vector of p candidate variables, ε is the error, with $E(\varepsilon|\boldsymbol{X}) = 0$ and $m(\cdot)$ is the unknown regression function. We consider a general class of regression problems without imposing any particular form to the function $m(\cdot)$, and, we assume that $E(Y) = 0$ implying that $E\{m(\boldsymbol{X})\} = 0$. Let us denote with

$$\boldsymbol{M}_* = \{1 \leq j \leq p : \text{the } j\text{-th variable in } \boldsymbol{X} \text{ is relevant for explanation of } Y\}$$

the set of s true relevant covariates in model (2.1). Moreover, we consider a very sparse model, so $s \ll p$.

Considering $f_j(x) = E(Y|X_j = x)$, that is the marginal contribution of X_j locally at $X_j = x$, with our D-ELSIS we estimate the first marginal derivative of the nonparametric model using the local quadratic estimator with a polynomial of order $d = 2$:

$$\widehat{f}_j^{\,\prime}(x; 2, h) = \sum_{i=1}^{n} W_{i,2}(x) Y_i \tag{3.2}$$

where $W_{i,2}(x)$ is defined in equation (2.6).

Throughout this thesis, we use $||\cdot||_\infty$ to denote the sup-norm and $C^r(\mathcal{I})$ denotes the class of all continuous functions defined over $\mathcal{I}$ that are r time differentiable. We assume the following conditions.

(A1) $\{f_j\}_{j=1}^p$ belong to $C^r(\mathcal{X}_j)$, where $r \geq 3$ and $\mathcal{X}_j$ is the support of X_j. In addition, there exists a constant K_1 such that $|f_j^{(r)}(x)| \leq K_1$ for any $x \in \mathcal{X}_j$ and $j = 1, \ldots, p$.

(A2) The marginal density function g_j of X_j satisfies $0 < K_2 \leq g_j(x) \leq K_3 < \infty$ on $\mathcal{X}_j$ for $j = 1, \ldots, p$. In addiction, there exists $g_j^{(2)}(x)$ for any $x \in \mathcal{X}_j$ and $j = 1, \ldots, p$.

(A3) There exist nonnegative constants $c_1 > 0$ and $\kappa \in [0, \frac{r-1}{2r+1})$ such that $\min_{j \in M_*} ||f_j'||_\infty \geq c_1 n^{-\kappa}$, where r is given in assumption (A1).

(A4) There exist positive constants K_5, K_6, γ_1 and γ_2 such that $P(|Y| \geq u) \leq K_5 \exp(-K_6 u^{\gamma_1})$ for any $u > 0$ and $P(|X_j| \geq u) \leq K_5 \exp(-K_6 u^{\gamma_2})$ for each $j = 1, \ldots, p$ and any $u > 0$.

(A5) The kernel function $\mathcal{K}(\cdot)$ is continuous, bounded and symmetric with bounded support. In addition, $\mathcal{K}(\cdot)$ is of order 2, that is, $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$ and $\int u^2\mathcal{K}(u)du = \mu_2 < \infty$.

Here, (A1) is a condition describing the continuity of each $f_j(x) = E(Y|X_j = x)$. This condition is necessary in order to apply the univariate polynomial regression for the estimation of first marginal derivative with a polynomial of order $p = 2$. Assumption (A2) is standard for kernel regression implying that the density of X_j does not vanishing on its support and implies bounded support of the explanatory variables. The condition in (A3) is for identifying M_*, which require that the minimal signal strength measured by $||f_j'||_\infty$ cannot vanish at a rate faster than $n^{-1/2}$ (Stone, 1982). Assumption (A4) on the tail distribution of the response and the explanatory variables is a conventional technical requirement for Cramér-type large deviation. For example, $\gamma_1 = 2$ if the response variable Y is a normal or sub-Gaussian distribution and $\gamma_1 = \infty$ if Y has a compact support. Assumption (A5) specifies the requirement for the kernel function so that the bias due to the kernel smoothing is not dominating.

Meanwhile, we assume that the bandwidth h satisfies $h \asymp n^{-\omega}$ for some positive ω whose specification is discussed later.

Instead of using the exact local quadratic estimator, we propose to consider a new estimator,

$$\hat{\beta}(x) = \frac{1}{nh^2} \boldsymbol{e}_2^T \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

to discriminate between relevant and non relevant covariates. As it is possible to see from its formulation, this estimator comes from the estimator for the first derivatives (2.5) with weights (2.6), for which we delete the inverted matrix $(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ and we divide for nh^2. The following Lemma states that the mean of $\hat{\beta}$ is different from zero when the covariate is relevant, while its mean is zero with non relevant covariates.

Lemma 1. *Under assumptions (A1)-(A3) and (A5), assuming that the bandwidth h satisfies $h \asymp n^{-\omega}$ for some positive $2\omega \geq \kappa$ where κ is given in (A3), if $f_j(x) \neq \frac{1}{g_j(x)} \forall x \in \mathcal{X}_j$, then*

1. $E(\hat{\beta}(x)) = 0$ *for any* $j \notin M_*$,

2. $\exists\ x \in \mathcal{X}_j : E(\hat{\beta}(x)) \neq 0$ *for any* $j \in M_*$,

where $\hat{\beta}(x) = \frac{1}{nh^2} \boldsymbol{e}_2^T \mathbf{X}^T \mathbf{W} \mathbf{Y} = \frac{1}{nh^2} \sum_{i=1}^n \mathcal{K}_h(X_{ij} - x)(X_{ij} - x) Y_i$.

Proof. We can simplify the matrix formulation of our $\hat{\beta}$ estimator, noting that

$$\boldsymbol{e}_2^T \mathbf{X}^T \mathbf{W} \mathbf{Y} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n \mathcal{K}_h(X_i - x) Y_i \\ \sum_{i=1}^n \mathcal{K}_h(X_i - x)(X_i - x) Y_i \\ \sum_{i=1}^n \mathcal{K}_h(X_i - x)(X_i - x)^2 Y_i \end{pmatrix} = \sum_{i=1}^n \mathcal{K}_h(X_i - x)(X_i - x) Y_i$$

This means that

$$\hat{\beta}(x) = \frac{1}{nh^2} \sum_{i=1}^n \mathcal{K}_h(X_i - x)(X_i - x) Y_i$$

Now we consider the mean of $\hat{\beta}(x)$. Note that

$$\begin{aligned} E(\hat{\beta}(x)) =& E\left\{ \frac{1}{nh^2} \sum_{i=1}^n \mathcal{K}_h(X_{ij} - x)(X_{ij} - x) Y_i \right\} \\ =& \frac{1}{h^2} E\left\{ \mathcal{K}_h(X_{ij} - x)(X_{ij} - x) Y_i \right\} \\ =& \frac{1}{h} \int \mathcal{K}(v) f_j(x + vh) g_j(x + vh) v dv \end{aligned}$$

by Taylor expansion we have

$$f_j(x + vh) = f_j(x) + f_j'(x) vh + \frac{f_j^{(2)}(x) v^2 h^2}{2!} + O(h^3)$$

$$g_j(x + vh) = g_j(x) + g_j'(x) vh + O(h^2)$$

and by (A.5), we achieve

$$\begin{aligned} & \frac{1}{h} \int \mathcal{K}(v) f_j(x + vh) g_j(x + vh) v dv \\ & = f_j'(x) g_j(x) \mu_2 + f_j(x) g_j'(x) \mu_2 + O(h^2) \end{aligned}$$

On the one hand, if $j \notin M_*$, this means $f_j'(x) = 0$ and $f_j(x) = 0$, so we achieve $E(\hat{\beta}) = 0$. On the other hand, if $j \in M_*$, since from (A2) we know the density $g_j(x)$ is bounded away from 0, $g_j'(x)$ is bounded and from (A5) $\mu_2 < \infty$, then $E(\hat{\beta}) \neq 0$ provided that $f(x) \neq \frac{1}{g(x)}\ \forall x \in \mathcal{X}_j$ and $2\omega \geq \kappa$ because h^2 must go to zero at a faster rate than $f_j'(\cdot)$. □

Remark. If the number of relevant variable is finite, $\kappa = 0$, then the required condition $\omega > \kappa$

is always satisfied.

Using Lemma 1, we can simplify the empirical likelihood, noting that

$$\begin{aligned} EL_j(x,0) =& \sup_w \left\{ \prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i W_{i,2}(x) Y_i = 0 \right\} \\ =& \sup_w \left\{ \prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i U_{ij} = 0 \right\}. \end{aligned}$$

where $U_{ij} = \frac{1}{nh^2} K_h(X_{ij} - x)(X_{ij} - x)Y_i$. Using the Lagrange multiplier method for solving the above equation, we obtain the empirical likelihood ratio:

$$l_j(x,0) = -2\log\{EL_j(x,0)\} - 2n\log n = 2\sum_{i=1}^n \log\{1 + \lambda U_{ij}\}$$

where λ is the univariate Lagrange multiplier. This $l_j(x,0)$ is a statistic for testing whether or not $f_j'(\cdot)$ has zero mean locally at x. For assessing $f_j'(x) \equiv 0$ uniformly on the support $\mathcal{X}_j$ of X_j, we use

$$l_j(0) = \sup_{x \in \mathcal{X}_j} l_j(x,0)$$

for each $j = 1, \ldots, p$, where $\mathcal{X}_j$ is the support of X_j.

For feature screening purpose, we sort l_j for all $j = 1, \ldots, p$ in decreasing order and we take the first γ_n covariates. In this way, we create a set

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j \geq \gamma_n\}.$$

In order to obtain the screening property of D-ELSIS, we need the following lemmas.

Lemma 2. *Under assumptions (A1)-(A3) and (A5), assuming that the bandwidth h satisfies $h \asymp n^{-\omega}$ for some positive $2\omega \geq \kappa$ and $f_j(x) \neq \frac{1}{g_j(x)} \forall x \in \mathcal{X}_j$, there exist two positive constants C_1 and C_2 such that*

$$K_2\mu_2 \left| |f_j'(x)| - C_2|f_j(x)| \right| \leq \left| E\left\{ \frac{1}{h^2} \mathcal{K}_h(X_{ij} - x)(X_{ij} - x)Y_i \right\} \right| \leq C_1\mu_2 \left| |f_j'(x)| + |f_j(x)| \right|$$

for any $j \in M_*$.

Proof. Without loss of generality, we assume $f_j'(x) > 0$. From Lemma 1, we know that

$$\frac{1}{h} \int \mathcal{K}(v) f_j(x + vh) g_j(x + vh) v dv = f_j'(x) g_j(x) \mu_2 + f_j(x) g_j'(x) \mu_2 + O(h^2)$$

By assumption (A2), it follows that $0 \leq |g_j'(x)| \leq K_3' < \infty$. Let $C_1 = \max\{K_3, K_3'\}$. There-

fore,

$$\mu_2 \left| |f_j'(x)|K_2 - |f_j(x)|K_3'| \right) \leq \left| E\left\{ \frac{1}{h^2} \mathcal{K}_h(X_{ij} - x)(X_{ij} - x)Y_i \right\} \right| \leq C_1 \mu_2 \left| |f_j'(x)| + |f_j(x)| \right|$$

provided that $2\omega \geq \kappa$. Now, if we set $C_2 = K_3'/K_2$, the result follows. □

Lemma 3. *For given j, define $Z_{ij} = K_h(X_{ij} - x)(X_{ij} - x)Y_i$. Under assumptions (A4), (A5) and assuming that the bandwidth h satisfies $h \asymp u^{-\omega}$ for some positive $\omega > \kappa$, then*

$$P\{|Z_{ij}| > u\} \leq K_5 \exp\{-K_6 u^{\gamma(1-\omega)}\} \quad \text{for any} \quad j = 1, \ldots, p$$

where $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$.

Proof. Pick $\epsilon > 0$,

$$\begin{aligned} P\{|Z_{ij}| > u\} =& P\left\{|K_h(X_{ij} - x)(X_{ij} - x)| > u^{\epsilon}, |K_h(X_{ij} - x)(X_{ij} - x)Y_i| > u\right\} + \\ &+ P\left\{|K_h(X_{ij} - x)(X_{ij} - x)| \leq u^{\epsilon}, |K_h(X_{ij} - x)(X_{ij} - x)Y_i| > u\right\} \\ &\leq P\left\{|K_h(X_{ij} - x)(X_{ij} - x)| > u^{\epsilon}\right\} + P\{|Y_i| > u^{1-\epsilon}\} \end{aligned}$$

The distribution of the product $K_h(X_{ij} - x)(X_{ij} - x)$ is a sub-exponential distribution with parameter γ equal to infinity. In fact, this product gives a distribution that has non-zero values only on a bounded support, because the kernel used is bounded. If $0 < h < c$ with $c > 0$, this product is zero uniformly. But, when $h \to 0$ with rate ω,

$$\begin{aligned} P\left\{|K_h(X_{ij} - x)(X_{ij} - x)| > u^{\epsilon}\right\} =& P\left\{ \left| \frac{1}{h} K\left(\frac{X_{ij} - x}{h} \right)(X_{ij} - x) \right| > u^{\epsilon} \right\} \\ =& P\left\{ \left| K\left(\frac{X_{ij} - x}{h} \right)(X_{ij} - x) \right| > u^{\epsilon} h \right\} \\ &\leq K_5 \exp(-K_6 u^{\gamma_1(\epsilon - \omega)}) \end{aligned}$$

So, we have

$$P\{|Z_{ij}| > u\} \leq K_5 \exp(-K_6 u^{\gamma_1(\epsilon-\omega)}) + K_5 \exp(-K_6 u^{\gamma_2(1-\epsilon)})$$

In order to get the best rate for the right-hand side of above inequality, we need $\gamma_1(\epsilon - \omega) = \gamma_2(1 - \epsilon))$. It means that $\epsilon = \frac{\gamma_2 + \gamma_1 \omega}{\gamma 1 + \gamma 2}$. Hence,

$$P\{|Z_{ij}| > u\} \leq K_5 \exp(-K_6 u^{\gamma(1-\omega)})$$

where $\gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}$.

□

Remark. The part of condition (A4) for X_j variable, is only sufficient thus it should be relaxed. In fact, without the sub-exponential condition for the distribution of all the explanatory variables, the result in Lemma 3 can still achieved, with a γ that depends only on γ_1.

Lemma 4. *For given j, define $U_{ij} = K_h(X_{ij} - x)(X_{ij} - x)Y_i$ and $\mu_{0j} = E[(U_{ij})]$. Under assumption (A1), (A2) and (A5), then*

$$E[(U_{ij} - \mu_{0j})^2] \leq Ch$$

with $C > 0$.

Proof. Note that

$$E(U_{ij}^2|X_{ij}) = \frac{1}{h^2}\mathcal{K}^2\left(\frac{X_{ij}-x}{h}\right)(X_{ij}-x)^2E(Y^2|X_{ij}) \leq c\frac{1}{h^2}\mathcal{K}^2\left(\frac{X_{ij}-x}{h}\right)(X_{ij}-x)^2$$

since $E(Y^2|X_{ij}) \leq c, \forall X_{ij}$, with a constant $c > 0$. So,

$$\begin{aligned} E[E(U_{ij}^2|X_{ij})] \leq & \frac{c}{h^2}\int \mathcal{K}^2\left(\frac{X_{ij}-x}{h}\right)(X_{ij}-x)^2 dX_{ij} \\ = & hc\int \mathcal{K}^2(v)v^2dv = Ch \end{aligned}$$

with $C = c\int \mathcal{K}^2(v)v^2dv$. Since $E[(U_{ij} - \mu_{0j})^2] \leq E[U_{ij}^2] = E[E(U_{ij}^2|X_{ij})]$, we achieve the result.

□

Proposition 1. *Under assumptions (A1)-(A5), pick $\omega \in (\kappa, 1)$, then there exists a uniform constant C_1 depending only on K_5, K_6, γ_1 and γ_2 appeared in assumption (A4), such that for any $j \in M_*$ and $L \to \infty$,*

$$P\left\{l_j(0) < \frac{c_1^2K_2^2n^{1-2\kappa-4\omega}\mu_2^2}{2L^2}\right\} \leq \begin{cases} \exp(-Cn^{1-2\kappa-3\omega}) + \exp(-CL^{\gamma(1-\omega)}), \\ \quad \textit{if} \quad (1-2\kappa-3\omega)\delta < \kappa+\omega \\ \exp(-Cn^{\frac{1-\kappa-2\omega}{1+\delta}}) + \exp(-CL^{\gamma(1-\omega)}), \\ \quad \textit{if} \quad (1-2\kappa-3\omega)\delta \geq \kappa+\omega \end{cases}$$

where $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$ and $\delta = \max\left(\frac{2}{\gamma(1-\omega)} - 1, 0\right)$.

Proof. Given $j \in M_*$, for any $t > 0$,

$$P\{l_j(0) < 2t\} \leq P\{l_j(x, 0) < 2t\}.$$

We will give an upper bound for the one on the right-hand side of above inequality. Define $U_{ij} = \mathcal{K}_h(X_{ij} - x)(X_{ij} - x)Y_i$ and $\mu_{0j} = E(U_{ij})$. If $2\omega \geq \kappa$, by Lemma 2, we can always find the constant C_2 such that $|\mu_{0j}| \geq h^2 K_2 \mu_2 \left| |f_j'(x)| - C_2 |f_j(x)| \right| > 0$. Without loss of generality, we assume $\mu_{0j} > 0$. If $\mu_{0j} < 0$, let $\tilde{U}_{ij} = -U_{ij}$. Note that

$$\begin{aligned} EL_j(x,0) &= \sup\left\{ \prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i U_{ij} = 0 \right\} \\ &= \sup\left\{ \prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \tilde{U}_{ij} = 0 \right\}. \end{aligned}$$

This means $l_j(x,0) = -2\log\{EL_j(x,0)\} - 2n\log n$ does not depend on the sign of μ_{0j}. Since

$$l_j(x,0) = 2 \max_{\lambda \in \Lambda_{n,j}} \sum_{i=1}^n \log(1 + \lambda U_{ij}),$$

where $\Lambda_{n,j} = \{\lambda : 1 + \lambda U_{ij} \geq n^{-1} \text{ for all } i = 1, \ldots, n\}$ (Owen, 2001), we can choose $\lambda = n^{-\epsilon} / \max_l |U_{ij}|$ for some $\epsilon > 0$, such that for n sufficiently large $\lambda \in \Lambda_{n,j}$. We obtain that

$$P\{l_j(x,0) < 2t\} \leq P\left\{ \sum_{i=1}^n \log\left(1 + \frac{U_{ij}}{n^\epsilon \max_l |U_{ij}|}\right) < t \right\}.$$

Hence, using the Taylor expansion, as in the proof of Proposition 1 of Chang et al. (2016b), by (A4) and Lemma 3, we have

$$\begin{aligned} P\{l_j(x,0) < 2t\} &\leq P\left\{ \sum_{i=1}^n (U_{ij} - \mu_{0j}) < (tn^\epsilon + n^{1-\epsilon}) \max_l |U_{lj}| - n\mu_{0j} \right\} \\ &\leq P\left[\frac{1}{n^{\frac{1}{2}}\sigma} \sum_{i=1}^n (U_{ij} - \mu_{0j}) < \frac{(tn^{\epsilon - \frac{1}{2}} + n^{\frac{1}{2} - \epsilon})M - n^{\frac{1}{2}}\mu_{0j}}{\sigma} \right] \\ &\quad + C\exp(-CM^{\gamma(1-\omega)} + \log n), \end{aligned}$$

where $\sigma^2 = E\{(U_{ij} - \mu_{0j})^2\}$. For $L \to \infty$, pick ϵ satisfies $n^\epsilon = L/\mu_{0j}$. Choose $\eta \in \left(0, \frac{4}{5}\right)$ and let $M = \eta L$ and $2t = \frac{n\mu_{0j}^2}{2L^2}$, then

$$\frac{tn^\epsilon M}{n\mu_{0j}} = \frac{\eta}{4} \quad \text{and} \quad \frac{n^{1-\epsilon}M}{n\mu_{0j}} = \eta.$$

By Lemma 4 we have that $\sigma^2 \leq Ch$. Since $f_j'(x)$ goes to zero at a slower rate than $f_j(x)$ (Stone, 1982), so the rate of μ_{0j} has the quantity $|f_j'(x)|$ as dominant part. Therefore, with a

n sufficiently large, by Lemma 1 of Chang et al. (2013b),

$$
\begin{aligned}
&P\left\{l_j(x,0) < \frac{c_1^2 K_2^2 n^{1-2\kappa} h^4 \mu_2^2}{2L^2}\right\} \le P\left\{l_j(x,0) < \frac{n\mu_{0j}^2}{2L^2}\right\} \\
&\le P\left\{\frac{1}{n^{\frac{1}{2}}\sigma}\sum_{i=1}^{n}(U_{ij}-\mu_{0j}) < \frac{(\frac{5}{4}\eta-1)n^{\frac{1}{2}}\mu_{0j}}{\sigma}\right\} + C\exp(-CM^{\gamma(1-\omega)}+\log n) \\
&\le \begin{cases} \exp(-Cn^{1-2\kappa}h^3)+\exp(-CL^{\gamma(1-\omega)}), & \text{if} \quad (1-2\kappa-3\omega)\delta < \kappa+\omega \\ \exp(-Cn^{\frac{1-\kappa}{1+\delta}}h^{\frac{2}{1-\delta}})+\exp(-CL^{\gamma(1-\omega)}), & \text{if} \quad (1-2\kappa-3\omega)\delta \ge \kappa+\omega \end{cases}
\end{aligned}
$$

where $\delta = \max(\frac{2}{\gamma(1-\omega)}-1, 0)$. □

Proposition 1 gives a uniform result for all explanatory variables contributing in the true model. With large probability and uniformly for all $j \in M_*$, the diverging rate of $l_j(0)$ is not slower than $n^{1-2\kappa-4\omega}L^{-2}$. If $j \notin M_*$, that is, the explanatory variable X_j does not have the marginal contribution to Y (i.e., $f_j' = 0$), following the argument of Owen (2001) and Chang et al. (2013a), it can be shown that the corresponding $l_j(0)$ is $O_p(1)$. Hence, $n^{1/2-\kappa-2\omega}L^{-1}$ is required to diverge as $n \to \infty$ for sure independent screening. Furthermore, we note that the requirement for the bandwidth used in Proposition 1 is mild, which can be naturally satisfied by the conventional optimal bandwidth $h = O(n^{-1/7})$ selected by cross-validation method.

Let $L = n^{1/2-\kappa-2\omega-\tau}$ for some $\tau \in (0, \frac{1}{2}-\kappa-2\omega)$, we obtain the following corollary, based on Proposition 1, more specifically summarising that the set M_* can be distinguished by the statistic $l_j(0)$.

Corollary 1. *Under assumptions (A1)-(A5), pick $\omega \in [\frac{\kappa}{2}, \frac{1}{4}-\frac{\kappa}{2})$, $\tau \in (0, \frac{1}{2}-\kappa-2\omega)$ with $\kappa < \frac{1}{4}$, then there exists a uniform constant C_1 depending only on K_5, K_6, γ_1 and γ_2 appeared in assumption (A4) such that*

$$
\begin{aligned}
\max_{j\in M_*} P\{l_j(0) < c_1^2 K_2^2 n^{2\tau}\mu_2^2\} \le &\exp(-C_1 n^{(1/2-\kappa-2\omega-\tau)(1-\omega)\gamma)}) \\
&+ \exp(-C_1 n^{\min\{1-2\kappa-3\omega,(1-\kappa-2\omega)/(1+\delta)\}})
\end{aligned}
$$

where $\delta = \max\{\frac{2}{\gamma(1-\omega)}-1, 0\}$, C_1 is given in Proposition 1 and γ is given in Lemma 3.

Remark. In order to have $\frac{\kappa}{2} < \frac{1}{4}-\frac{\kappa}{2}$, we need to impose $\kappa < \frac{1}{4}$.

We establish the sure property of our approach in the following theorem based on Corollary 1.

Theorem 1. *Under assumptions (A1)-(A5), pick $\omega \in [\frac{\kappa}{2}, \frac{1}{4}-\frac{\kappa}{2})$ and $\gamma_n = c_1^2 K_2^2 n^{2\tau}\mu_2^2$ for some $\tau \in (0, \frac{1}{2}-\kappa-2\omega)$ with $\kappa < \frac{1}{4}$, then there exists a uniform constant C_1 depending only*

on K_5, K_6, γ_1 and γ_2 appeared in assumption (A4) such that

$$P\{M_* \subset \widehat{M}_{\gamma_n}\} \geq 1 - s\exp(-C_1 n^{(1/2-\kappa-2\omega-\tau)(1-\omega)\gamma}) - s\exp(-C_1 n^{\min\{1-2\kappa-3\omega,(1-\kappa-2\omega)/(1+\delta)\}})$$

where $\delta = \max\left\{\frac{2}{\gamma(1-\omega)} - 1, 0\right\}$, C_1 is given in Proposition 1 and γ is given in Lemma 3.

Proof. Consider a threshold level $\gamma_n = c_1^2 K_2^2 n^{2\tau}\mu_2^2$ for the estimated set $\widehat{M}_{\gamma_n}$ and note that

$$\begin{aligned} P\{M_* \subsetneq \widehat{M}_{\gamma_n}\} &= P\{\text{There exists } j \in M_* \text{ such that } l_j(0) < c_1^2 K_2^2 n^{2\tau}\mu_2^2\} \\ &\leq s \max_{j \in M_*} P\{l_j(0) < c_1^2 K_2^2 n^{2\tau}\mu_2^2\} \end{aligned}$$

where $s = |M_*|$. Using Corollary 1, we achieve the result. □

Theorem 1 implies that D-ELSIS method can handle the following non-polynomial dimensionality: $\log p = o(n^\epsilon)$ for $\epsilon = \min\{(1/2 - \kappa - 2\omega - \tau)(1-\omega)\gamma, 1 - 2\kappa - 3\omega\}$. By noting that $2\omega \geq \kappa$ and that the rate of ϵ is increasing as ω increases, then we can choose as the best rate for the bandwidth the value $\omega = \frac{\kappa}{2}$. In this case, considering τ close enough to zero, the highest dimensionality is achieved with the optimal $\epsilon = \min\{(\frac{1}{2} - 2\kappa)(1 - \frac{\kappa}{2})\gamma, 1 - \frac{7}{2}\kappa\}$. If Y and X_j for $j = 1, \ldots, p$ follow the normal or sub-Gaussian distribution such that $\gamma = 1$, the corresponding highest dimensionality satisfies $\log p = o(n^{(\frac{1}{2}-2\kappa)(1-\frac{\kappa}{2})})$. If Y and X_j for $j = 1, \ldots, p$ have a compact support which means $\gamma = \infty$, the corresponding highest dimensionality satisfies $\log p = o(n^{1-\frac{7}{2}\kappa})$.

In what follows, we consider the size of the selected $\widehat{M}_{\gamma_n}$ under the ideal case that

$$\max_{j \notin M_*} ||f_j'||_\infty = o(n^{-\kappa}). \tag{3.3}$$

Now we need to investigate how large is the set $\widehat{M}_{\gamma_n}$. This question is closely related to the probabilistic behaviour of $P\{l_j(0) \geq c_1^2 K_2^2 n^{2\tau}\mu_2^2\}$ for each $j \notin M_*$. We need the following lemmas.

Lemma 5. *Under assumption (A2) and (A4), suppose that $h \asymp n^{-\omega}$ for some $\omega < 1$ and there exists a positive constant ρ such that $\inf_{u \in [a,b]} E(Y^2 | X_j = u) \geq \rho$ for any $j \notin M_*$, then*

$$P\{S_j^2 \leq \frac{1}{2}E(U_{ij}^2)\} \leq \begin{cases} \exp(-Cn^{1-\omega}), & \text{if } \gamma(1-\omega) \geq 4 \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{4}}), & \text{if } 0 < \gamma(1-\omega) < 4 \end{cases}$$

where $S_j^2 = \frac{1}{n}\sum_{i=1}^n U_{ij}^2$, C is uniform for any $j = 1, \ldots, p$ and γ is given in Lemma 3.

Proof. Note that

$$P\{S_j^2 \le \frac{1}{2}E(U_{ij}^2)\} = P\left[\frac{1}{n}\sum_{i=1}^{n}\{U_{ij}^2 - E[U_{ij}^2]\} \le -\frac{1}{2}E[U_{ij}^2]\right]$$
$$= P\left[\frac{1}{n^{-\frac{1}{2}}\tilde{\sigma}_j}\sum_{i=1}^{n}\{U_{ij}^2 - E[U_{ij}^2]\} \le -\frac{n^{\frac{1}{2}}E[U_{ij}^2]}{2\tilde{\sigma}_j}\right]$$

where $\tilde{\sigma}_j^2 = E\left[\{U_{ij}^2 - E(U_{ij}^2)\}^2\right]$. By the same way in the proof of Lemma 4, we can get that $\tilde{\sigma}^2 \le Ch$ for all $j = 1, \ldots, p$. On the other hand, by (A2),

$$E(U_{ij}^2) \ge E\left\{\sum_{i=1}^{n} K_h^2(X_{ij} - x)(X_{ij} - x)^2 E(Y^2|X_{ij}))\right\} \ge K_2^2 h\rho \int K^2(v)v^2 dv \ge Ch$$

Then, by Lemma 1 of Chang et al. (2013b),

$$P\{S_j^2 \le \frac{1}{2}E(U_{ij}^2)\} \le \exp(-C(nh)^{\frac{1}{1+\delta}})$$

where $\delta = \max(\frac{4}{\gamma(1-\omega)} - 1, 0)$. Note that the positive constant C is uniform for any $j \notin M_*$. □

Lemma 6. *Under assumption (A2) and (A5), suppose that $h \asymp n^{-\omega}$ for some $\omega < 1$ and there exists a positive constant ρ such that $\inf_{u\in[a,b]} E(Y^2|X_j = u) \ge \rho$ for any $j \notin M_*$. If $\max_{j\notin M_*} |\mu_{0j}| = O(n^{-\psi})$ for some $\psi > 2\omega$, then there exists a uniform positive constant C for any $j \notin M_*$*

$$P\left(|\lambda_j| > \frac{4|n^{-1}\sum_{i=1}^{n} U_{ij}|}{3S_j^2}\right) \le \begin{cases} \exp\left(-Cn^{(\psi-2\omega)(1-\omega)\gamma}\right), & \text{if } \psi - 2\omega < \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \\ \exp\left(-Cn^{\frac{(1-\omega)^2\gamma}{\max((1-\omega)\gamma,2)+2}}\right), & \text{if } \psi - 2\omega \ge \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \end{cases} \tag{3.4}$$

where λ_j is defined by $n^{-1}\sum_{i=1}^{n} \frac{U_{ij}}{1+\lambda_j U_{ij}} = 0$ and γ is given in Lemma 3.

Proof. Using Lemma 3 of Chang et al. (2016b) and Lemma 5, we obtain the result. □

Lemma 7. *Under assumption (A2) and (A5), suppose that $h \asymp n^{-\omega}$ for some $\omega < 1$ and there exists a positive constant ρ such that $\inf_{u\in[a,b]} E(Y^2|X_j = u) \ge \rho$ for any $j \notin M_*$. Choose $\tau \in (0, \frac{1}{2} - k - 2\omega)$, for a given $j \notin M_*$. If $\max_{j\notin M_*} |\mu_{0j}| = O(n^{-\psi})$ for some $\psi > 2\omega$ and $\psi + \tau - \frac{\omega}{2} - \frac{1}{2} > 0$, then*

$$P\{l_j(x,0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\} \leq \left\{ \begin{array}{c} \exp\left(-Cn^{(\psi-2\omega)(1-\omega)\gamma}\right), \text{ if } \psi - 2\omega < \frac{1-\omega}{\max((1-\omega)\gamma,2)+2} \\ \exp\left(-Cn^{\frac{(1-\omega)^2\gamma}{\max((1-\omega)\gamma,2)+2}}\right), \text{ if } \psi - 2\omega \geq \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \end{array} \right.$$
$$+ \left\{ \begin{array}{c} \exp(-Cn^{2\tau}), \text{ if } 2\tau < \frac{\gamma(1-\omega)^2}{6} \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{6}}), \text{ if } 2\tau \geq \frac{\gamma(1-\omega)^2}{6} \end{array} \right.$$

where C is a uniform positive constant not depending on j and x and γ is given in Lemma 3.

Proof. We follow the same line of proof of Lemma 4 of Chang et al. (2016b). Consider $|c_{i1}| < 1$ and $|c_{i2}| < 1$ for all $i = 1, \ldots, n$. Then by Taylor expansion, we have

$$\begin{aligned} l_j(x,0) =& n\left(\frac{1}{n}\sum_{i=1}^n U_{ij}^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n U_{ij}\right)^2 - n\left(\frac{1}{n}\sum_{i=1}^n U_{ij}^2\right)^{-1}\left\{\frac{1}{n}\sum_{i=1}^n \frac{\lambda_j^2 U_{ij}^3}{(1+c_{i2}\lambda_j U_{ij})^3}\right\} \\ &+\frac{2}{3}\sum_{i=1}^n \frac{\lambda_j^3 U_{ij}^3}{(1+c_{i1}\lambda_j U_{ij})^3} \\ =:& I_1 + I_2 + I_3, \end{aligned}$$

Now, we can define

$$\mathcal{A} = \left\{ |\lambda_j| < \frac{4|n^{-1}\sum_{i=1}^n U_{ij}|}{3n^{-1}\sum_{i=1}^n U_{ij}^2} \text{ and } \left|\frac{1}{n}\sum_{i=1}^n U_{ij}\right| \cdot \max_j |U_{ij}| < \frac{1}{4n}\sum_{i=1}^n U_{ij}^2 \right\},$$

Using Lemma 6 we have

$$P\left(\mathcal{A}^c\right) \leq \left\{ \begin{array}{c} \exp\left(-Cn^{(\psi-2\omega)(1-\omega)\gamma}\right), \text{ if } \psi - 2\omega < \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \\ \exp\left(-Cn^{\frac{(1-\omega)^2\gamma}{\max(\gamma(1-\omega),2)+2}}\right), \text{ if } \psi - 2\omega \geq \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \end{array} \right. . \tag{3.5}$$

We can note that, if $\mathcal{A}$ holds, we obtain

$$|I_3| \leq C\left(\sum_{i=1}^n |U_{ij}|^3\right)\left|\frac{1}{n}\sum_{i=1}^n U_{ij}\right|^3\left(\frac{1}{n}\sum_{i=1}^n U_{ij}^2\right)^{-3}.$$

Noting that

$$\begin{aligned} P\{l_j(x,0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\} \leq& P(I_1 + I_3 \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}) \\ \leq& P\left(I_1 \geq \frac{1}{2}c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right) + P\left(I_3 \geq \frac{1}{2}c_1^2 K_2^2 \mu_2^2 n^{2\tau}, \mathcal{A} \text{ holds }\right) + P(\mathcal{A}^c), \end{aligned}$$

we only need to give an upper bounds for the quantities appeared on the right-hand side

respectively. Using Lemma 1 of Chang et al. (2013b) and Lemma 5, considering $\tau+\psi-\frac{\omega}{2}-\frac{1}{2}>0$, we have

$$P(I_1 \geq \frac{1}{2}c_1^2K_2^2\mu_2^2n^{2\tau}) \leq \left\{ \begin{array}{c} \exp(-Cn^{2\tau}),\ \text{if } 2\tau < \frac{1-\omega}{1+2\delta} \\ \exp(-Cn^{\frac{2\tau-\omega+1}{2+2\delta}}),\ \text{if } 2\tau \geq \frac{1-\omega}{1+2\delta} \end{array} \right. + \left\{ \begin{array}{c} \exp(-Cn^{1-\omega}),\ \text{if } \gamma(1-\omega)\geq 4 \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{4}}),\ \text{if } 0<\gamma(1-\omega)<4 \end{array} \right.$$

and

$$\begin{aligned} &P(I_3 \geq \frac{1}{2}c_1^2K_2^2\mu_2^2n^{2\tau}, \mathcal{A} \text{ holds }) \\ &\leq \left\{ \begin{array}{c} \exp(-Cn^{\frac{4\tau-\omega+1}{3}}),\ \text{if } 2\tau < \frac{(1-\delta)(1-\omega)}{1+2\delta} \\ \exp(-Cn^{\frac{2\tau-2\omega+2}{3+3\delta}}),\ \text{if } 2\tau \geq \frac{(1-\delta)(1-\omega)}{1+2\delta} \end{array} \right. + \left\{ \begin{array}{c} \exp(-Cn^{1-\omega}),\ \text{if } \gamma(1-\omega)\geq 6 \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{6}}),\ \text{if } 0<\gamma(1-\omega)<6 \end{array} \right. \end{aligned}$$

where $\delta = \max(\frac{2}{(\gamma(1-\omega))}-1, 0)$. Hence, noting that $\tau < \frac{1}{2}-\kappa-2\omega$,

$$\begin{aligned} &P(I_1 \geq \frac{1}{2}c_1^2K_2^2\mu_2^2n^{2\tau}) + P(I_3 \geq \frac{1}{2}c_1^2K_2^2\mu_2^2n^{2\tau}, \mathcal{A} \text{ holds }) \\ &\leq \left\{ \begin{array}{c} \exp(-Cn^{2\tau}),\ \text{if } \gamma(1-\omega)\geq 6 \\ \exp(-Cn^{2\tau}),\ \text{if } 0<\gamma(1-\omega)<6 \text{ and } 2\tau < \frac{\gamma(1-\omega)^2}{6} \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{6}}),\ \text{if } 0<\gamma(1-\omega)<6 \text{ and } 2\tau \geq \frac{\gamma(1-\omega)^2}{6} \end{array} \right. . \end{aligned}$$

We complete the proof of this lemma. □

Proposition 2. *Under assumptions (A1)-(A2) and (A4)-(A5), suppose* $\max_{j\notin M_*}||f_j'||_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{5}{4}\kappa$. *Pick* $\omega \in [\frac{\kappa}{2}, \min(\frac{1}{4}-\frac{\kappa}{2}, 2(\eta-\kappa))$, $\tau \in (\max(\frac{1}{2}-\eta-\frac{3\omega}{2}, 0), \frac{1}{2}-\kappa-2\omega)$. *If* $\inf_{u\in[a,b]} E(Y^2|X_j=u) \geq \rho$ *for some positive* ρ *for any* $j \notin M_*$, *then there exists a uniform positive constant* C_2 *such that for any* $j \notin M_*$,

$$P\left\{l_j(0) \geq c_1^2K_2^2\mu_2^2n^{2\tau}\right\} \leq \exp(-C_2n^{\min\{\eta(1-\omega)\gamma, (1-\omega)^2\gamma/[\max(\gamma(1-\omega),2)+2], 2\tau, \gamma(1-\omega)^2/6\}})$$

where γ *is given in Lemma 3.*

Proof. Given $j \notin M_*$, for any $t>0$,

$$P(l_j(0)\geq t) \leq \sum_{x\in\mathcal{X}_j} P\{l_j(x,0)\geq t\}.$$

We need to bounded $P\{l_j(x,0)\geq t\}$ for each $x\in\mathcal{X}_j$. Note that $\max_{j\notin M_*}||f_j'||_\infty = O(n^{-\eta})$, then $|\mu_{0j}| \leq Ch^2n^{-\eta}$ for any $j\notin M_*$. Note that $|\mu_{0j}| = O(n^{-\psi})$, then $\psi = \eta+2\omega$. As

$$P\left\{l_j(0)\geq c_1^2K_2^2\mu_2^2n^{2\tau}\right\} \leq \sum_{x\in\mathcal{X}_j} P\left\{l_j(x,0)\geq c_1^2K_2^2\mu_2^2n^{2\tau}\right\}$$

noting that the number of x in $\mathcal{X}_j$ is $O(n)$, by Lemma 7 we can obtain

$$P\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\}$$
$$\leq \begin{cases} \exp\left(-Cn^{\eta(1-\omega)\gamma} + \log n\right), \text{ if } \eta < \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \\ \exp\left(-Cn^{\frac{(1-\omega)^2\gamma}{\max(\gamma(1-\omega),2)+2}} + \log n\right), \text{ if } \eta \geq \frac{1-\omega}{\max(\gamma(1-\omega),2)+2} \end{cases} + \begin{cases} \exp(-Cn^{2\tau} + \log n), \text{ if } 2\tau < \frac{\gamma(1-\omega)^2}{6} \\ \exp(-Cn^{\frac{\gamma(1-\omega)^2}{6}} + \log n), \text{ if } 2\tau \geq \frac{\gamma(1-\omega)^2}{6} \end{cases}$$

Since all the exponents of n are positive, we can delete $\log n$ in all the quantities. In this way, we complete the proof. □

Using Proposition 2, we can find the corresponding upper bound for $P\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\}$. The result is given in the following Corollary.

Corollary 2. *Under assumptions (A1)-(A2) and (A4)-(A5), suppose* $\max_{j \notin M_*} ||f_j'||_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{5}{4}\kappa$. *Pick* $\omega = \frac{\kappa}{2}$, $\tau \in (\max(\frac{1}{2} - \eta - \frac{3\kappa}{4}, 0), \frac{1}{2} - 2\kappa)$. *If* $\inf_{u \in [a,b]} E(Y^2 | X_j = u) \geq \rho$ *for some positive* ρ *for any* $j \notin M_*$, *then there exists a uniform positive constant* C_3 *such that for any* $j \notin M_*$,

$$P\left\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right\} \leq \exp(-C_3 n^{\min\{\eta(1-\frac{\kappa}{2})\gamma, (1-\frac{\kappa}{2})^2\gamma/[\max(\gamma(1-\frac{\kappa}{2}),2)+2], 2\tau, \gamma(1-\frac{\kappa}{2})^2/6\}})$$

where γ *is given in Lemma 3.*

From Corollary 2, we obtain the following theorem for the size of $\widehat{M}_{\gamma_n}$.

Theorem 2. *Under assumptions (A1)-(A2) and (A4)-(A5), suppose* $\max_{j \notin M_*} ||f_j'||_\infty = O(n^{-\eta})$ *for some* $\eta > \frac{5}{4}\kappa$. *Pick* $\omega = \frac{\kappa}{2}$ *and* $\gamma_n = c_1^2 K_2^2 \mu_2^2 n^{2\tau}$ *for some* $\tau \in (\max(\frac{1}{2} - \eta - \frac{3\kappa}{4}, 0), \frac{1}{2} - 2\kappa)$. *If* $\inf_{u \in [a,b]} E(Y^2 | X_j = u) \geq \rho$ *for some positive* ρ *holds for any* $j \notin M_*$, *then*

$$P(|\widehat{M}_{\gamma_n}| > s) \leq p \exp(-C_3 n^{\min\{\eta(1-\frac{\kappa}{2})\gamma, (1-\frac{\kappa}{2})^2\gamma/[\max(\gamma(1-\frac{\kappa}{2}),2)+2], 2\tau, \gamma(1-\frac{\kappa}{2})^2/6\}})$$

where γ *is given in Lemma 3 and* C_3 *is given in Corollary 2.*

Proof. By noting that

$$|\widehat{M}_{\gamma_n}| = \sum_{j \in M_*} \mathbb{I}\left\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right\} + \sum_{j \notin M_*} \mathbb{I}\left\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right\} \leq s + \sum_{j \notin M_*} \mathbb{I}\left\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right\}$$

we have $P(|\widehat{M}_{\gamma_n}| > s) \leq \sum_{j \notin M_*} \mathbb{I}\left\{l_j(0) \geq c_1^2 K_2^2 \mu_2^2 n^{2\tau}\right\}$. Using Corollary 2 we achieve the result. □

This theorem shows that our screening procedure can really control the set size of the selected variables. With large probability, the number of the selected variables is not larger

than the true size s. From Theorem 1 and Theorem 2, we have that

$$P(\widehat{M}_{\gamma_n} = M_*) \to 1 \text{ as } n \to \infty$$

provided that $\log p = o(n^{\min\{\eta\gamma(1-\frac{\kappa}{2}),(1-\frac{\kappa}{2})^2\gamma/[\max(\gamma(1-\frac{\kappa}{2}),2)+2],2\tau,\frac{\gamma(1-\frac{\kappa}{2})^2}{6},1-\frac{7}{2}\kappa,(\frac{1}{2}-2\kappa-\tau)(1-\frac{\kappa}{2})\gamma\}})$. This selection consistency property demonstrates that, under condition (3.3), our approach performs very well by distinguishing the true relevant variables from false ones. In order to obtain the optimal diverging rate for p, we can select

$$\tau = \begin{cases} \frac{\gamma(1-\frac{\kappa}{2})}{\gamma(1-\frac{\kappa}{2})+2}\left(\frac{1}{2}-2\kappa\right) \text{ if } \eta > \frac{1}{\gamma(1-\frac{\kappa}{2})+2} + \frac{11\kappa\gamma(1-\frac{\kappa}{2})+6\kappa}{4[\gamma(1-\frac{\kappa}{2})+2]} \\ \frac{1}{2}-\eta-\frac{\kappa}{4}+\varsigma \text{ if } \eta \leq \frac{1}{\gamma(1-\frac{\kappa}{2})+2} + \frac{11\kappa\gamma(1-\frac{\kappa}{2})+6\kappa}{4[\gamma(1-\frac{\kappa}{2})+2]} \end{cases} \tag{3.6}$$

where ζ can be chosen to be positive and converging to 0 as $n \to \infty$. Hence, $P(\widehat{M}_{\gamma_n} = M_*) \to 1$ as $n \to \infty$ provided that

$$\log p = \begin{cases} o(n^{\min\{(1-\frac{\kappa}{2})^2\gamma/[\max((1-\frac{\kappa}{2})\gamma,2)+2],\frac{\gamma(1-\frac{\kappa}{2})^2}{6},(1-4\kappa)\frac{\gamma(1-\frac{\kappa}{2})}{\gamma(1-\frac{\kappa}{2})+2}\}}) \\ \text{if } \eta > \frac{1}{\gamma(1-\frac{\kappa}{2})+2} + \frac{11\kappa\gamma(1-\frac{\kappa}{2})+6\kappa}{4[\gamma(1-\frac{\kappa}{2})+2]} \\ o(n^{\min\{(1-\frac{\kappa}{2})^2\gamma/[\max((1-\frac{\kappa}{2})\gamma,2)+2],\frac{\gamma(1-\frac{\kappa}{2})^2}{6},(\eta-\frac{5}{4}\kappa)(1-\frac{\kappa}{2})\gamma\}}) \\ \text{if } \eta \leq \frac{1}{\gamma(1-\frac{\kappa}{2})+2} + \frac{11\kappa\gamma(1-\frac{\kappa}{2})+6\kappa}{4[\gamma(1-\frac{\kappa}{2})+2]} \end{cases} \tag{3.7}$$

More specifically, if Y and X_j for $j = 1, \ldots, p$ have a compact support which means $\gamma = \infty$, the above selection consistency holds if $\log p = o(n^{1-4\kappa})$. When $\gamma = 1$ and $\eta = \infty$, that is when Y and X_j for $j = 1, \ldots, p$, follow normal or sub-Gaussian distribution and in presence of partial orthogonal condition (Huang et al., 2010), the selection consistency holds if $\log p = o(n^{\min\{\frac{1}{24}(2-\kappa)^2,\frac{1-4\kappa}{6-\kappa}(2-\kappa)\}})$.

By comparing our optimal diverging rate of p with that achieved by the other competitor screening procedures, it is possible to note that our approach achieves a lower dimensionality. So, we need more observations to have the selection consistency property, compared, for example, to Chang et al. (2016a), which represents our direct competitor. In Chang et al. (2016a) when Y follows a normal or sub-Gaussian distribution and in presence of partial orthogonal condition, the selection consistency holds if $\log p = o(n^{\min\{\frac{1}{2}-\kappa-\frac{\kappa}{g_1},\frac{1}{3}-\frac{\kappa}{3g_1}\}})$, where ϱ_1 characterizes the continuity of the marginal projections $f_j(x)$. In fact, their rate depends on the continuity of the marginal $f_j(x)$. The greater the smoothness of the function, the better the dimensionality achieved. For example, if $f_j(x) \in C^\infty$ their dimensionality becomes $\log p = o(n^{1-2\kappa})$. Local polynomial theory suggests that a consistent estimate of the first derivative can be achieved if $f_j(x) \in C^r$, with $r \geq 3$. Increasing the smoothness of the function does not increase our dimensionality.

Since the estimation of first marginal derivative $f_j'(x)$ with local polynomials needs more

observations than the estimation of $f_j(x)$ with the NW, as in Chang et al. (2016a), this result was expected. Furthermore, the ultra-high dimensional rate of our competitor is achieved under some very stringent assumptions. Compared to Chang et al. (2016a), we do not regularize the maximum distance between two point in the support $\mathcal{X}_j$ and we use a Kernel of order 2 instead of a Kernel which the order depends on the parameter that characterizes the smoothness of the function. This is a substantial advantage of our method. In fact, it is very difficult to manage with a Kernel of order greater that 6.

Screening vs variable selection

As said in the previous chapters, the threshold γ_n is unknown and it is difficult to estimate it. For this reason we fix a tuning parameter, $p* > s = |M_*|$, such that we choose $p*$ covariates with the largest values of our statistic l_j. Denote with $\widehat{M}_{p*}$ such a set of selected covariates. In this way, we get the D-ELSIS Screening Property by using assumptions (A1)-(A5). It means that all the relevant covariates belong to the set $\widehat{M}_{p*}$.

Let $I \subset \{1, \dots, p\}$ and define $\pi_n(I) = P\left(I \subseteq \widehat{M}_{p*}\right)$. The transformation of our screening selection method in a variable selection method is based on the following ssumptions:

(a1) $\max_{j \notin M_*} ||f_j'||_\infty = o(n^{-\kappa})$;

(a2) If $j, j' \notin M_*$, then $\min_{j,j'} P\left(||\widehat{f'}_j||_\infty > ||\widehat{f'}_{j'}||_\infty\right) \to c_3$ and
$\max_{j,j'} P\left(||\widehat{f'}_j||_\infty > ||\widehat{f'}_{j'}||_\infty\right) \to c_4$ as $n \to \infty$ with $0 < c_3 < c_4 < 1$.

Assumption (a1) says that all the marginal derivatives of the irrelevant covariates are smaller than the minimum coefficient of the relevant covariates. Assumption (a2) states that all the irrelevant covariates can be exchanged. Then, there does not exist one irrelevant covariate which is dominant with respect to the other irrelevant ones.

Theorem 3. *Suppose that assumptions (A1)-(A5), (a1) and (a2) hold. If $p \equiv p_n \to \infty$ as $n \to \infty$ and $|M_*| < \infty$, then*

$$\pi_n(M_*) \to 1 \quad \textit{and} \quad \pi_n(I) \to 0$$

as $n \to \infty$ with any $I \not\subseteq M_*$.

Proof. Since $|M_*| < \infty$, we can set $p* > |M_*|$ such that $p* < \infty$. By assumptions (A1)-(A5) and by the D-ELSIS Screening Property, it follows that $\pi_n(M_*) \to 1$ as $n \to \infty$.
Now, consider a set $I \not\subseteq M_*$. Without loss of generality, suppose that $I = M_*$ except for a covariate, say $j_* \notin M_*$. For absurd, suppose that $P(I \subseteq \widehat{M}_{p*}) \to c > 0$ as $n \to \infty$. It implies that also $P(j_* \in \widehat{M}_{p*})$ converges to a positive quantity. By using the Borel-Cantelli Lemma

and $p = p_n \to \infty$ as $n \to \infty$, it follows that

$$\sum_{j \notin M_*, j \neq j_*} P\left(||\widehat{f}'_j||_\infty > ||\widehat{f}'_{j_*}||_\infty\right) < \infty$$

as $n \to \infty$. So $P\left(||\widehat{f}'_j||_\infty > ||\widehat{f}'_{j_*}||_\infty\right) \to 0$ for any $j \neq j_*$ and $j \notin M_*$. Since both j_* and j are irrelevant covariates, then the last result is an absurd by assumption (a2). Thus, we can conclude that $P(I \subseteq \widehat{M}_{p*}) \to 0$ as $n \to \infty$. Since $p*$ is finite, then the result follows for any $I \nsubseteq M_*$ and $\widehat{M}_{p*}$. □

Only for simplicity we assume in Theorem 3 that the number of relevant covariates is finite, i.e. $|M_*| < \infty$.

Now, we need to estimate $\pi_n(I)$. For this purpose, we can use the subsampling technique with $m < n$, the size of each subsample. In this way, we apply the D-ELSIS Screening procedure to each subsample of size m. The total number of subsample is $\binom{n}{m}$. This number can be very large. So, we can randomly draw without replacement G subsamples. Therefore, for each subsample, we have the set of covariates $\widehat{M}^{(i)}_{p*}$, $i = 1, \ldots, G$, which is built by using the statistic l_j, with m observations and G is the number of subsample. Then we can estimate $\pi_n(I)$ by $\widehat{\pi}_{n,m,G}(I) = \frac{1}{G}\sum_{i=1}^{G} \mathbb{I}\left(I \subset \widehat{M}^{(i)}_{p*}\right)$.

Now, the next theorem states the consistency of $\widehat{\pi}_{n,m,G}(I)$ for $\pi_n(I)$.

Theorem 4. *Suppose that the assumptions of Theorem 2 hold. Then*

$$|\widehat{\pi}_{n,m,G}(I) - \pi_n(I)| \xrightarrow{p} 0,$$

when $m \to \infty$ as $n \to \infty$ and $G \to \infty$.

Proof. By Theorem 3, we have that $\pi_n(I) \to \pi(I)$ as $n \to \infty$, where $\pi(I) = 1$ if $I \subset M_*$ and $\pi(I) = 0$ if $I \nsubseteq M_*$.

Let $E^*(\cdot), Var^*(\cdot)$ and $P^*(\cdot)$ be $E(\cdot|\mathcal{X}_n), Var(\cdot|\mathcal{X}_n)$ and $P(\cdot|\mathcal{X}_n)$, respectively, with $\mathcal{X}_n = \{(Y_1, X_1'), \ldots, (Y_n, X_n')\}$. Moreover, let $Z_i = \mathbb{I}\left(I \subseteq \widehat{M}^{(i)}_{p*}\right)$. So, we can write

$$E^*(\widehat{\pi}_{n,m,G}(I)) = \frac{1}{G}\sum_{i=1}^{G} E^*(Z_i^*) = \frac{1}{\binom{n}{m}}\sum_{i=1}^{\binom{n}{m}} Z_i \equiv \widehat{\pi}_{n,m}(I) \xrightarrow{\text{p}} \pi(I) \tag{3.8}$$

as $m \to \infty$ and $n \to \infty$. Z_i^* is the same as Z_i except that it is randomly chosen over all the different subsamples. Remember that the conditional mean of Z_i, given the set $\widehat{M}^{(i)}_{p*}$, is $\pi_m(I)$.

Since each subsample is drawn without replacement, the conditional variance $Var^*(\cdot)$ has to be considered with respect to the hypergeometric distribution. So, it follows that

$$Var^*(\widehat{\pi}_{n,m,G}(I)) \leq \frac{1}{G}\widehat{\pi}_{n,m}(I)(1 - \widehat{\pi}_{n,m}) \xrightarrow{\text{p}} 0,$$

as $m \to \infty$, $n \to \infty$ and $G \to \infty$. By using the conditional Chebyshev inequality, we have that

$$P^*(|\widehat{\pi}_{n,m,G}(I) - E^*(\widehat{\pi}_{n,m,G}(I))| > \epsilon) \leq \frac{Var * (\widehat{\pi}_{n,m,G}(I))}{\epsilon^2} \xrightarrow{\text{p}} 0$$

for any $\epsilon > 0$ as $m \to \infty$, $n \to \infty$ and $G \to \infty$. Finally, $\widehat{\pi}_{n,m,G}(I) \xrightarrow{\text{p}} \pi(I)$ by (3.8). The proof is complete. □

Generally, in order to have a consistent estimator using the subsample technique, we need to impose the condition $\frac{m}{n} \to 0$. In our case, choosing $m = \lfloor n/2 \rfloor$, this condition is not satisfied. Theorem 4 shows that, even in this case, $\widehat{\pi}_{n,m,G}(I) = \frac{1}{G}\sum_{i=1}^{G} \mathbb{I}\left(I \subset \widehat{M}_{p*}^{(i)}\right)$ is a consistent estimator of $\pi_n(I)$. So, the condition $\frac{m}{n} \to 0$ is not necessary in our case.

Chapter 4

Simulations and empirical study

Several simulation studies are conducted to investigate the performance of the proposed D-ELSIS method in terms of the following three criteria: (i) the median of the minimum model size (MMSs, i.e., the smallest number of the selected covariates including all the active explanatory variables) for 100 repetitions; (ii) the IQR divided by 1.34 (SD), that is the robust measure of the standard error of MMS; (iii) the true positive rate in percentage (TPR) that control the precision measuring the proportion of actual relevant variables that are correctly identified as such. To calculate the TPR we consider that the predicted relevant variables are the first 20. In order to have a very good method, the MMS should be equal to the number of the true active variable, with small SD and high TPR. We set $n = (500, 750, 1000)$ and $p = (100, n/2, 2n)$.

For comparison, we also considered other three previous screening methods for nonparametric models: the Fused Kolmogorov Filter (FKF) of Mai and Zou (2015), the Fused Mean-Variance (FMV) of Yan et al. (2018) and the local Empirical Likelihood SIS (ELSIS) of Chang et al. (2016a), described in Chapter 1 of this thesis. For the implementation of the best bandwidth in the kernel regression estimation for ELSIS and D-ELSIS, we used the R package **NonpModelCheck** of Zambom et al. (2017). Among the various options of the package, we have chosen for both models the cross-validation leave-one-out, which performs satisfactorily. Instead, as regard the likelihood estimation for empirical likelihood, we used the R package **emplik** of Zhou (2018). Furthermore, for the FMV method, we used the code that the authors provided in their paper (Yan et al., 2018).

We consider the following experiments in the simulation study.

Example 1 : Additive model with uniform covariates

This example is taken from Example 3 of Fan et al. (2011) and from Example 2 of Chang et al. (2016a). In this case we are interested in evaluate the performance of our approach in detecting relevant variables in nonparametric model with additive component when the signal to noise ratio increases. Data are generated from model

$$Y = 5X_1 + 3(2X_2 - 1)^2 + 4\frac{\sin(2\pi X_3)}{2 - \sin(2\pi X_3)} +$$

$$6\left[0.1\sin(2\pi X_4) + 0.2\cos(2\pi X_4) + 0.3(\sin(2\pi X_4))^2 + 0.4(\cos(2\pi X_4))^3 + 0.5(\sin(2\pi X_4))^3\right] + \sigma\epsilon$$

Here predictors X_j's are $i.i.d$ random variables of $U(0,1)$ distribution, and $\varepsilon \sim N(0,1)$ is independent of X_j's. In this case we have $s = 4$ relevant covariates. We consider four different signal to noise ratios by varying σ^2, as in Chang et al. (2016a). The results are in Table 4.1.

Example 2 : Non-additive model with uniform covariates

This example is taken from Example 4.1 of Lafferty and Wasserman (2008). We consider a nonparametric non-additive model and covariates with bounded support. Data are generated from model

$$Y = 5X_1^2 X_2^2 + \epsilon.$$

In this case, predictors X_j's are $i.i.d$ random variables of $U(0,1)$ distribution, and $\epsilon \sim N(0, \sigma^2)$, with $\sigma = 0.5$, is independent of X_j's. The results are in Table 4.2

Example 3 : Linear model with correlated normal covariates

We are interested in assessing whether the presence of correlation between predictors influences the performance of our procedure. Data are generated from model

$$Y = X_1 + X_2 + X_3 + X_4 + \epsilon$$

where $X \sim N(0, \Sigma)$ and $\varepsilon \sim N(0,1)$ is independent of each X_j with $j = 1, \ldots, p$. In this model we have $s = 4$ relevant linear covariates, all of which with the same parameter 1. We set the variance-covariance matrix $\Sigma = (\sigma_{kj})$ with $\sigma_{kk} = 1$ and $\rho = \sigma_{kj} = c(0, 0.5)$: so we consider both independent and correlation cases between active and non-active covariates. The results are in Table 4.3.

Example 4 : Single index model with independent normal covariates

This example is taken from Example 4 of Chang et al. (2016a). Data are simulated from model

$$Y = m(\boldsymbol{X}) + \sigma\epsilon$$

where $m(\boldsymbol{X})$ is generated from $\exp\left\{-\frac{1}{2}\left(\frac{X_1^2}{0.8^2} + \frac{X_2^2}{0.9^2} + \frac{X_3^2}{1} + \frac{X_4^2}{1.1^2}\right)\right\}$ by appropriately scaling it to have zero mean and unit variance, predictors are independently gener-

ated from standard normal distribution and $\epsilon \sim N(0,1)$ is independent from X_i's. We set the noise levels as 0.5. The results are in Table 4.4.

4.1 Simulation results

Overall, in all the settings are considering here, D-ELSIS typically offers similar and sometimes better performance than its competitors.

In Example 1, where we have a nonparametric additive model, D-ELSIS and ELSIS are equivalent in each considered combination of number of observations and covariates. Both approaches are able to correctly identify the set of true variables in the first 20 top ranked covariates for all the signal to noise ratios considered. Furthermore, the percentage of relevant covariates included from the top ranked is always 100, so both the methods do not make the error of excluding the relevant among the first 20. On the other hand, the other two competitors considered, FMV and FKF, fail to achieve the same performance. This happens in each of the cases evaluated, even when the number of observations is greater than the number of covariates, as in the case $n = 500$ and $p = 100$. Furthermore, the TPR rate is always lower compared to TPRs of the first two approaches. So, with this example, both approaches based on the fused technique fail to be competitive.

In Example 2 and Example 4 the four approaches are practically equivalent, both in terms of MMS and precision. In Example 2 we considered uniform covariates in a nonparametric and non-additive model, while in Example 4 a Single index model with independent standard normals. The two models have in common the fact of being composed of a single non-additive function, depending on two or more independent covariates. The equivalence in the four different approaches is substantially due to the independence between the covariates. This is also confirmed by the results of Example 3.

In Example 3, we have a linear model with normal correlated covariates. When the covariates are independent, the four methods are equivalent. In case of correlation among active and inactive covariates, so when $\rho = 0.5$, D-ELSIS is equivalent to methods based on the fused technique, FKF and FMV, while its performance is higher than the ELSIS model, even when the number of covariates is lower than that of observations.

In their simulation Chang et al. (2016a) use the combination $n = 100$, $p = 1000$, while in this thesis we set the following values: $n = (500, 750, 1000)$ and $p = (100, n/2, 2n)$, so a different proportion contemplating also a not-so-high dimensional case. We have not chosen the same proportion to give uniformity to the various simulations, since this proportion has brought satisfactory results compared to the other procedures, especially respect to those based on the fused technique and in the presence of correlation. Given this good performance results with low proportion between the number of covariates and the number of observations,

Table 4.1: Simulation results from Example 1

$s=4$	$n=500$						$n=750$						$n=1000$					
	$p=100$		$p=250$		$p=1000$		$p=100$		$p=375$		$p=1500$		$p=100$		$p=500$		$p=2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method	$\sigma^2=1$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	9 (8.96)	95.75	13 (16.42)	92.00	37 (70.52)	83.25	5 (2.99)	98.50	8 (7.65)	95.25	25 (37.50)	86.50	4 (0.76)	100.00	5 (5.04)	97.00	15 (20.34)	91.25
FKF	10 (10.45)	94.00	12 (21.27)	90.50	54 (104.29)	81.75	6 (3.73)	97.25	10 (16.41)	92.00	30 (59.51)	86.25	4 (0.75)	99.75	6 (4.66)	96.75	11 (34.33)	91.25
	$\sigma^2=1.74$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	7 (4.66)	97.25	12 (14.55)	91.25	46 (82.84)	82.25	5 (2.24)	99.00	8 (8.21)	94.75	17 (25.37)	88.75	4 (0.75)	100.00	5 (3.17)	98.25	13 (12.69)	93.25
FKF	9 (8.58)	95.25	20 (29.29)	88.00	60 (111.75)	82.25	5 (2.99)	99.25	8 (10.63)	94.00	22 (44.78)	87.25	4 (0.75)	99.75	6 (5.97)	97.50	8 (13.99)	93.25
	$\sigma^2=2$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.75)	98.75	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	8 (8.96)	95.50	16 (21.64)	89.75	32 (61.19)	81.75	5 (3.73)	98.75	8 (6.90)	96.00	18 (25.00)	88.25	4 (0.75)	99.75	5 (2.99)	98.75	10 (13.62)	92.75
FKF	10 (9.14)	95.25	18 (20.90)	89.00	45 (83.21)	83.00	6 (4.67)	98.25	9 (13.43)	93.25	17 (30.41)	88.25	4 (0.75)	99.75	5 (4.66)	98.50	12 (18.66)	82.00
	$\sigma^2=3$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	99.75	4 (0.19)	98.75	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	99.75	4 (0.19)	98.75	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	8 (9.14)	95.50	17 (18.47)	89.25	37 (73.13)	83.50	5 (2.24)	99.50	8 (5.97)	95.50	19 (28.73)	87.75	4 (0.75)	99.75	6 (3.73)	98.50	9 (16.79)	92.25
FKF	10 (11.19)	94.00	15 (19.03)	89.75	39 (60.82)	83.25	5 (3.17)	98.50	13 (18.84)	91.50	19 (47.76)	88.00	4 (2.24)	99.75	5 (4.48)	96.75	11 (14.55)	91.25

Table 4.2: Simulation results from Example 2

$s = 2$	$n = 500$						$n = 750$						$n = 1000$					
	$p = 100$		$p = 250$		$p = 1000$		$p = 100$		$p = 375$		$p = 1500$		$p = 100$		$p = 500$		$p = 2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method																		
D-ELSIS	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00
ELSIS	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00
FMV	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00
FKF	2 (0.00)	100.00	2 (0.00)	100 .00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00	2 (0.00)	100.00

Table 4.3: Simulation results from Example 3

$s=4$	$n=500$						$n=750$						$n=1000$					
	$p=100$		$p=250$		$p=1000$		$p=100$		$p=375$		$p=1500$		$p=100$		$p=500$		$p=2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method	$\rho=0$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FKF	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
	$\rho=0.5$																	
D-ELSIS	4 (0.00)	100.00	4 (0.75)	96.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.75)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.75
ELSIS	5 (2.99)	99.00	8 (9.89)	94.75	18 (25.56)	87.50	4 (0.75)	100.00	5 (2.24)	98.00	5 (2.24)	98.00	4 (0.00)	100.00	4 (0.75)	99.75	5 (5.22)	98.75
FMV	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FKF	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00

Table 4.4: Simulation results from Example 4

$s = 4$	$n = 500$						$n = 750$						$n = 1000$					
	$p = 100$		$p = 250$		$p = 1000$		$p = 100$		$p = 375$		$p = 1500$		$p = 100$		$p = 500$		$p = 2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method																		
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.75	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FMV	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
FKF	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00

for a higher proportion we expect to improve the performance of our procedure compared to competitors. Furthermore, we could not consider a lower dimensionality for the number of observations in the dataset because to obtain consistency using local polynomials to estimate the first derivative requires more observations than the estimate of the simple function. Some tests, which are not in this thesis, were done with a higher proportion and the performances were satisfactory. Finally, we have also chosen cases where the number of observations is higher with respect to the number of the covariates, since an approach that fails to correctly select the relevant covariates in this convenient case, will surely achieve worse results for high-dimensional data.

4.2 Empirical study

In this section we use a real dataset to illustrate our new variable selection method, that combine the D-ELSIS and the subsample technique. In the dataset used Italian firms that operated in the building sector in the period 2006-2017 are reported. The financial information are collected from the Orbis database, provided by Bureau van Djik.

One problem of interest is to discriminate companies based on their (high / low) probability of failure and predict bankruptcy before it occurs. In these cases, the dependent variable, which indicates whether a firm is distressed, is binary, and its knowledge is crucial for estimating the model parameters. We evaluate the probability of business failure in terms of financial stability by using profitability ratios (e.g., return on assets [ROA] and return on equity [ROE]). These ratios are suitable measures of a firm's performance and business stability, as is shown in Amendola et al. (2017). Since the period includes the economic instability, we opt for analyzing the years 2007-2009 in order to evaluate if our proposal is robust to choose the most relevant covariates in presence of financial shock. The sample consists of two subsets:

- the disease group is composed of those industrial firms that had undertaken the juridical procedure of bankruptcy in Italy between 2007 and 2009;
- the reference group consists of firms that were still active as of 2009 and has provided full information for the years between 2007 and 2009.

For each company, all financial statements and information available in the database are collected. The predictor database for the years of interest is elaborated starting from the financial statements of each firm included in the sample, for a total of 41,203 balance sheets. In particular, we compute 124 indicators selected as potential predictors according to three criteria (see Table 4.5):

- they have a relevant financial meaning in the failure context;
- they have been widely used in the failure prediction literature;

- the information needed to calculate these ratios is available.

Table 4.5: Variables used in the study divided according to the area.

Area	#
Turnover	24
Liquidity	19
Efficiency	19
Profitability	36
Solvency	26

To take into account the possible status changes of firms over the time considered, the binary variable indicating if failure has occurred is set equal to one only in the year in which a bankruptcy filing occurred. Thus, it is possible to use all available information to produce bankruptcy probability estimates for all firms at each time point.

A preliminary analysis is performed to prepare the data for a detailed statistical analysis. First, the missing data are identified, and those variables for which the percentage of missing values is high are deleted. Second, the financial ratios with a high number of values equal to zero are analyzed to determine if their values are affected by input errors. If the financial statements of a firm have excessive null information (i.e., zero or not available values), they are deleted. Third, the correlation between the surrogate dependent variable and other covariates was computed to throw out those ratios with a correlation greater than or equal to 0.98. Finally, the data are standardized in order to have zero mean and unit variance, and it was verified that the new design matrix is well defined.

For each year under analysis, we proceed as follows. First, we perform on a certain sample the D-ELSIS procedure selecting the first 10 top ranked variables. Second, we randomly sample 500 observations for 100 times (which means that we have 100 subsamples with 500 observations), and then for each subset we perform the D-ELSIS procedure again selecting only the first 10 top ranked variables, as in the first step. For the variables identified in the first step, we calculate their relative frequency of being selected from the top 10 positions in the various sabsample. From Theorem 3 in Chapter 3, we can choose a threshold π between zero and one to distinguish the relevant covariates from the noise ones. We set this threshold $\pi = 0.8$. We identify as relevant only those variables for which the probability of being selected from the top 10 positions in the various sabsample, exceeds the threshold. Table 4.6 reported the variables selected by our procedure.

As a result of our variable selection, we obtain 9 of the covariates of the dataset. Some of these are aligned with what was established in the literature by the paper of Amendola et al. (2017). It can be noted that the number of variables extracted changes across years. In fact, we observe that there are 6 variables that are common to all years. These are the variables that over time will affect business failure, that is measured by our dependent variable ROA. These

Table 4.6: Selected variable for each year

Variable	2007	2008	2009
X_2	√	√	√
X_{27}	√	√	√
X_{29}	√		√
X_{30}	√		
X_{31}	√	√	√
X_{57}	√	√	√
X_{76}	√	√	√
X_{90}	√	√	√
X_{91}		√	

X_2 is the age of the company, X_{27} is the EBIT to fixed assets, X_{29} is the net quick assets to Inventory, X_{30} is the Ebit to total sales, X_{31} is the Ebitda to Sales, X_{57} is the other shareholders fund to total assets, X_{76} is the operating cash flow, X_{90} is the inventory to Operating income and X_{91} is the gross profit to sales.

variables take into account the age of the company (X_2), its profitability (X_{27}, X_{31}, X_{90}), its solvency (X_{57}) and its liquidity (X_{76}). The remaining 3 variables are X_{29}, X_{30} and X_{91}. The first regards liquidity while the second and the third concern profitability. The first is selected for the years 2007 and 2009. We can note that the opposite situation occurs instead for the variable X_{91}, which represents an index of profitability. Furthermore, the profitability index X_{30} is selected only among the relevant variables only in the 2007. These different sets of identified covariates are mainly due to the financial crisis that has taken place in the years considered, which certainly had a strong impact on the bankruptcy of Italian companies.

Chapter 5

Conclusions

After a review of the literature for variable and screening selection methods in ultra-high dimensional setting, we have proposed and investigated a new procedure D-ELSIS in a nonparametric and non-additive context, which combines the local polynomials with the empirical likelihood. The innovative element of this approach consists in the estimation of the first marginal derivative with the use of local polynomials. By derivatives, we investigate the marginal contribution of each variable X_j in explaining Y, to justify whether it is relevant or not. Furthermore, we use the empirical likelihood to detect if this partial derivative is zero uniformly in the covariate's support.

Our theoretical results suggest that D-ELSIS has the screening and variable selection properties with a nonpolynomial dimensionality. Unfortunately, we achieve a dimensionality that, despite being ultra-high, does not exceed the dimensionality of our direct competitor. We expected this result because the use of local polynomials to estimate the first marginal derivative requires a lot number of observations. Furthermore, we have not imposed maximum distance between two adjacent observations. For any order of smoothness of the function, our procedure uses a Kernel of order 2 instead of a Kernel whose order depends on the parameter that characterizes the smoothness. This is a substantial advantage of our method compared to the competitor, as it is very difficult to manage with a kernel of order greater than 6.

The simulations results, show that D-ELSIS has really good screening performance compared to other model-free screener method in literature. Unlike the other approaches, D-ELSIS is able to select the true relevant covariates both when the underlying model is nonparametric and when it is linear. Furthermore, it manages to have significant results even in the presence of correlation between relevant and non-relevant covariates. As in all screening method in literature, in D-ELSIS the problem is the identification of the threshold γ_n, that detects how many covariates one needs to select in order to obtain the true set of active variables. We have shown theoretically that it is possible to transform our screening method in a variable selection method, using the subsample technique. In fact, with the subsample procedure, we

select the variables through the D-ELSIS and, after, we investigate their probability to be chosen when the data are randomly sampled. In this way, we can achieve a variable selection procedure without penalisation in the ultra-high setting, even if the model is nonparametric and non-additive.

Another important aspect of regression is the type of impact of independent covariates on the dependent covariate, the so called structure discovery. The literature focuses on partially linear models, or on nonparametric but additive models, often using methods with penalty. Also in this case, as for variable selection, the dimensionality is not very high, especially when we consider various order of interactions. For example, VANISH (Radchenko and James, 2010) could be extended to high-order interaction term, including the third order interactions, but it may not be possible for large p. We can use D-ELSIS also for structure discovery. In fact, a variable is linear in the regression model when the difference between the marginal derivative of the covariate and its mean is zero for each point:

$$f_j'(x_0) - E(f_j'(X_j)) = 0 \quad \forall x_0 \in \mathcal{X}_j$$

where $\mathcal{X}_j$ is the support of X_j. If this difference is null, for each point, then it can be said that the covariate taken into consideration is linear. With D-ELSIS we have already the estimation of marginal derivatives. Then, we need only to test if the difference is identically null. Our choice of which method to use to test this difference is, again, the empirical likelihood. We hope to work along this direction with our D-ELSIS procedure, finding its theoretical results for structure discovery.

Finally, D-ELSIS not only manages to transform itself into a method of variable selection with the subsample technique, but can also distinguish linear covariates from non-linear ones. The other approaches in screening literature, including our direct competitor Chang et al. (2016a), do not have this advantage. All this, however, has a price: the lowest achievable dimensionality.

Part II

Regression problem in survival analysis

Introduction

Modern biomedical studies generate a large amount of survival data or "time to event outcome variable" with high dimensional biological indicator for various scientific purposes. In practice, many covariates are often available as potential risk factors. To enhance model predictability and interpretation, a parsimonious model is always desirable. For this reason, variable selection is also vital in survival analysis. Thus, selecting significant variables plays crucial role in model building and is particularly challenging in the presence of a large number of predictors.

In literature there are statistical techniques to analyse a time-to-event outcome variable, which is a different type of outcome variable than those considered in the previous chapters. A time-to-event variable reflects the time until a participant experiences an event of interest (e.g., heart attack, goes into cancer remission, death). Statistical analysis of time-to-event variables requires different and more specific assumptions than those described thus far for other types of outcomes because of the unique features of these variables. The statistical analysis, in this case, is called *time-to-event analysis* or *survival analysis*, even though the outcome is not always death. Some questions of interest in survival analysis are: What is the probability that a participant survives 5 years? Are there differences in probability between groups (e.g., between those assigned to a new versus a standard drug in a clinical trial)? How do certain personal, behavioral or clinical characteristics affect participants' chances of survival? For instance, identifying genomic profiles that are associated with a particular disease may help with understanding its progression processes and designing more effective therapies.

With the advent of new biotechnologies, the emergence of gene expressions, methylation and next-generation RNA sequencing, have increased the dimensionality of data leading to larger and larger scale (Hong and Li, 2017). In these cases, the dimensionality of covariates may grow exponentially with the sample size and such data has been commonly referred as *ultra-high dimensional data*. Thus, the context is very similar to that introduced in the first part of the thesis. If the event occurred in all individuals, many methods of analysis seen in the previously part of this thesis would be applicable. However, it is usual that, at the end of observational time, some individuals have not experiences the event of interest and, thus, their true time-to-event is unknown. When this happens, the survival time is said to be *censored*.

In addition, survival data are rarely distributed normally, are usually distorted, and events of interest typically take place at the beginning of study time, while relatively few events occur at the end of that period. These features make the construction of special methods for survival analysis necessary.

When the number of covariates p is less than the sample size n, many *ad-hoc* statistical tools have been developed for survival data. The parametric and semiparametric regressions, such as the *Accelerated Failure Time model* (AFT) and the *Cox Proportional Hazards model*, respectively, have been routinely used for modelling censored outcome data in many practical settings. When $p > n$, penalized likelihood methods for variable selection have been proposed by various authors since the paper of Tibshirani (1997), and the oracle properties and statistical error bounds of estimation have been established (Huang et al., 2013). However, when $p \gg n$, computational issues inherent in these methods make them not applicable to ultra-high dimensional survival data because of serious challenges from computational cost, statistical accuracy and stability. In order to overcome these problems, also in survival analysis, many authors (such as Zhao and Li (2012); Gorst-Rasmussen and Scheike (2013); Song et al. (2014)) suggest to use the screening technique: obtaining a set of covariates that with probability 1 contains the set of true relevant ones. Even in this particular context it is possible to consider different statistical approaches, related to parametric and non-parametric models.

The aim of the part of this thesis is to find a model-free screening method in ultra-high dimensional setting different from the existent ones. Most of the model-free methods in the literature consider a conditional estimate of the survival function, using the Kaplan and Meier estimator (Kaplan and Meier, 1958). This estimator has some disadvantages, especially with continuous covariates. Our intent is to manage conditioning appropriately, taking into account the direct effect of covariate on survival. We have managed to highlight and justify the possibility of applying the D-ELSIS method, proposed in the first part of this thesis, also in this context. The element of innovation lies in the use of the marriage between local polynomials and empirical likelihood. In fact, no other time-to-event variable screening method uses these methods for screening purposes. The results demonstrate that the proposed nonparametric screener selects the true relevant covariates, especially in the presence of correlation.

In Chapter 6 we introduce the variable selection problem in linear and nonparametric regression models for survival analysis and the different methodologies for screening and variable selection purposes available in the literature. In Chapter 7 we explain how it is possible to use our D-ELSIS method to carry out screening in the context of nonparametric models with time-to-event data. In Chapter 8, we carry out extensive numerical simulations to asses the performance of the proposed D-ELSIS screener, also comparing it with other existing approaches.

Chapter 6

Variable selection in survival analysis

Survival analysis is a branch of statistics for analysing data which times of interest from a well-defined time origin until the occurrence of some particular events or end-points are investigated. In medical research, the time origin often corresponds to the recruitment of an individual into an experimental study, such as a clinical trial to compare two or more treatments. If the end-point is the death of the patient, the resulting data are literally *survival times* (Collett, 2015). This type of analysis, in the last years, is applied not only in medicine, but also in a number of others research areas, such as public health, socio-economic science, and engineering. In socio-economic research, it is used to investigate complex phenomena such as unemployment, employment, inflation, supply and demand for bank loans, life expectancy of the products, etc. The engineering sciences have also contributed to the development of survival analysis called *reliability analysis* where the main focus is in modelling the lifetimes of machines or electronic components (Kleinbaum and Klein, 1996). The standard statistical procedures cannot be applied in survival data analysis for several reasons:

- survival data are not symmetrically distributed (typically positively skewed);
- there is a presence of censored data, which means that for an individual the end-point of survival time has not been observed for different reasons, i.e. individual may drop out of a study, he/she may have a different event (an accident, a first childbirth, leaving school), he/she may decease or its lost to follow-up.

The censoring should be taken into account in the estimation of survival model. Three main type of censoring exist (Figure 6.1):

- *left censoring*: the start date of the event is not observed and the exact length of the survival time is not known (the subject was at risk for the event being studied before the start of the study);

- *right censoring*: at the time of observation the relevant event has not yet occurred (or the study ends before the event has occurred), and the total length of time between entry and exit from the state is unknown;
- *interval censoring*: the event time is only known to fall into an interval.

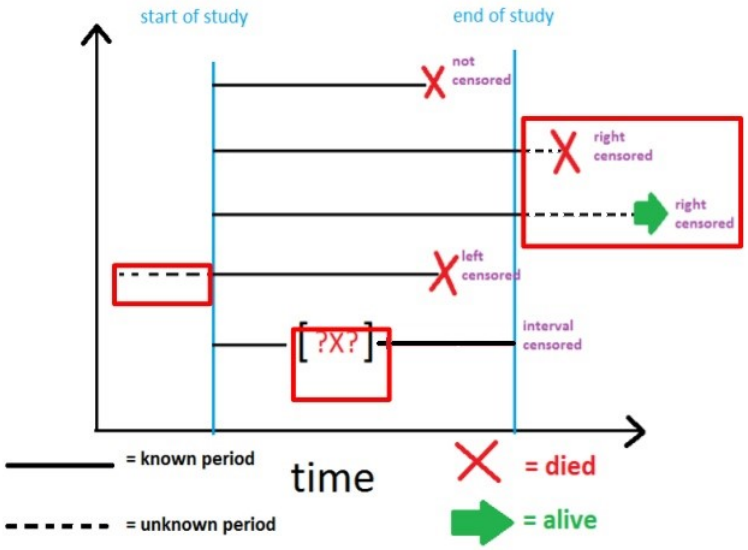

Figure 6.1: Type of censoring data

An important assumption in the analysis of censored survival data is that the actual survival time of an individual is independent from any mechanism that may cause the individual's survival time to be censored. *Independent censoring* essentially means that within any subgroup of interest, the subjects who are censored at time should be representative of all the subjects in that subgroup who remained at risk at that time with respect to their survival experience. In other words, censoring is independent provided that it is random within any subgroup of interest. Another assumption concerns the censoring mechanism that is assumed to be *non-informative*. A non-informative censoring occurs if the distribution of survival times does not provide any information about the distribution of censorship times, and vice versa.

6.1 Estimation of survival function

6.1.1 Basic concepts

Let T be a non-negative random variable representing the time until the occurrence of the event of interest. Its distribution function represents the probability that the survival time is

less than same value t and is given by

$$F(t) = P(T < t) = \int_0^t f(u)du$$

where $f(u)$ is the probability density function.

The *Survival Function* $S(t)$ is defined as the probability that the survival time is greater than or equal to t

$$S(t) = P(T \geq t) = 1 - F(t)$$

i.e., the probability that an individual survives from the time origin to some times beyond t.

The *Hazard Function* $h(t)$ is the risk or hazard of death at some times t and is given by the limiting value of the probability that an individual dies at time t, conditional on the individual having survived to that time:

$$h(t) = \lim_{\delta \to 0} \left\{ \frac{P(t \leq T < t + \delta | T \geq t)}{\delta} \right\}, \tag{6.1}$$

where the numerator is the probability that T lies between t and $t + \delta$, given that T is grater than or equal to t, while the denominator is the time interval δ. In literature, the hazard function is also called as *hazard rate*, *instantaneous death rate*, *intensity rate* or *force of mortality*. From equation (6.1), $h(t)\delta$ is the approximate probability that an individual dies in the interval $(t, t+\delta)$, conditional on that individual having survived to time t (Collett, 2015).

There are some useful relationships between the survival and hazard functions. The conditional probability in (6.1) can be expressed as

$$P(t \leq T < t + \delta | T \geq t) = \frac{P(t \leq T < t + \delta)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)},$$

then,

$$h(t) = \lim_{\delta \to 0} \left\{ \frac{F(t + \delta) - F(t)}{\delta} \right\} \frac{1}{S(t)}.$$

Noting that

$$f(t) = \lim_{\delta \to 0} \left\{ \frac{F(t + \delta) - F(t)}{\delta} \right\}$$

where $f(t)$ is exactly the derivative of $F(t)$ with respect to t, we have

$$h(t) = \frac{f(t)}{S(t)}.$$

Then

$$h(t) = -\frac{d}{dt} \log S(t)$$

and so

$$S(t) = e^{-H(t)}$$

where

$$H(t) = \int_0^t h(u)du.$$

The function $H(t)$ is called *integrated* or *cumulative hazard*. From equation (6.1.1), the cumulative hazard can be obtained from the survival function, since

$$H(t) = -\log S(t).$$

These three functions give mathematically equivalent specification of the distributions of the survival time T. In fact, if one of them is known, the other two are determined. One of these functions can be chosen as the basis of statistical analysis according to the particular situations: the survival function is the most useful for comparing the survival progress of two or more groups, while the hazard function gives a more convenient description of the risk of failure at any time.

Below we introduce notation that will be used throughout this thesis. Suppose we have n observations with p covariates. Denote by X_{ij} the jth covariate for subject i, and let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be a p-dimensional vector of covariates for the i-th individual. We add the assumption that the p variables are time-independent. Let T_i and C_i be the underlying survival and censoring times, respectively. We only observe $Y_i = \min(T_i, C_i)$, and the event indicator $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is the indicator function. In general, we consider right censoring time and we assume independent and non-informative censoring. We assume $(Y_i, \delta_i, \boldsymbol{X}_i)$ are *i.i.d.* In particular, we assume $(T_i, X_{ij}),\ i = 1, \ldots, n$, are *i.i.d* copies of (T, X_j), the random variables that underlie the survival time and covariates.

6.1.2 Kaplan-Meier estimator

The Kaplan-Meier (KM) estimator, developed by Kaplan and Meier (1958), is a nonparametric method used to estimate the survival function from lifetime data. The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. Its plot is a step function, where the estimated survival probabilities are constant between adjacent death times and only decreases at each death. For this estimator only the times at which the event happens are considered, while the censored times are ignored, taking into account however that the number of subjects at risk decreases in the presence of the censored. In order to estimate the KM, a series of time intervals is constructed such that only one death occurs per interval with the time of death indicating the start of an interval.

Suppose that there are n individuals with observed survival times $t_1, t_2, \ldots, t_n$. Some of these observations may be right-censored. We therefore suppose that there are r death times

amongst the individuals, with $r \leq n$. Let $t_{(1)} < t_{(2)} < ... < t_{(r)}$ be the ordered death times in non-decreasing order, where the jth is denoted $t_{(j)}$, for $j = 1, 2, ..., r$. The number of individuals who are alive just before time $t_{(j)}$, including those who are about to die at this time, will be denoted n_j, for $j = 1, 2, ..., r$, and d_j will denote the number who die at this time. The time interval from $t_{(j)} - \delta$ to $t_{(j)}$, where δ is an infinitesimal time interval, then includes one death time. Since there are n_j individuals who are alive just before $t_{(j)}$ and d_j deaths at $t_{(j)}$, the probability that an individual dies during the interval from $t_{(j)} - \delta$ to $t_{(j)}$ is estimated by d_j/n_j. The corresponding estimated probability of survival through that interval is then $(n_j - d_j)/n_j$. Assuming that the deaths of the individuals in the sample occur independently of another, the estimated survivor function at any time t in the kth constructed time interval from $t_{(k)}$ to $t_{(k+1)}$, $k = 1, 2, ..., r$, where $t_{(k+1)}$ is defined to be ∞, will be the estimated probability of surviving beyond $t_{(k)}$. This is actually the probability of surviving through the interval from $t_{(k)}$ to $t_{(k+1)}$ and all preceding intervals, and it leads to KM estimate of the survivor function, given by

$$\widehat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{6.2}$$

for $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r$, with $\widehat{S}(t) = 1$ for $t < t_{(1)}$ and where $t_{(r+1)}$ is taken to be ∞. If the largest observation is censored, for example t^*, $\widehat{S}(t)$ is undefined for $t > t^*$. On the other hand, if the largest observed survival time, $t_{(r)}$, is uncensored, $n_{(r)} = t_{(r)}$, and $\widehat{S}(t)$ is zero for $t \geq t_{(r)}$.

6.1.3 Cox Proportional Hazard model

Regression can be used to determine whether a characteristic of subjects affects the survival and, if so, how much and in what direction (to increase or decrease). Survival prediction could be difficult if some relevant risk factors are neglected. It is therefore necessary to identify those variables that affect the survival. A model that links the survival and covariates is called *Cox Proportional Hazard regression.*

Assuming that we have n individuals under observation, the Cox Proportional Hazards model (Cox, 1972) is given by

$$h(t|\boldsymbol{X}) = h_0(t) \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) = h_0(t) \exp\left(\sum_{i=1}^{n} \beta^T \boldsymbol{X}_i \right), \tag{6.3}$$

where $h_0(t)$ is called *baseline hazard function*, which is the hazard function for an individual for whom all the variables included in the model are zero, $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is the p-dimensional vector of explanatory variables for a particular individual, and $\beta^T =$

$(\beta_1, \beta_2, \ldots, \beta_p)$ is the p-dimensional vector of unknown coefficients.
The corresponding survival functions are given by

$$S(t|\boldsymbol{X}) = S_0(t)^{\exp(\sum_{j=1}^{p} \beta_j X_j)}.$$

This model does not make any assumptions about the form of $h_0(t)$ and assumes parametric form for the effect of the predictors on the hazard. For this reason it is a semi-parametric model: $h_0(t)$ is the nonparametric part of model while the exponential is the parametric part. The first part depends only on t, but not on $\boldsymbol{X}$, and it summarizes the pattern of duration dependence, assumed to be common to all individuals. The second part is an individual specific non-negative function of covariates $\boldsymbol{X}$, which does not depend on t (under the assumption that the covariates are time-independent), which scales the baseline hazard function common to all units. The beauty of the Cox approach is that this indeterminateness does not create any problems in the estimation. Even though the baseline hazard is not specified, we can still get a good estimate for regression coefficients and hazard ratio.

The measure of the effect is called *hazard ratio*. The hazard ratio of two individuals with different set of covariates $\boldsymbol{X}$ and $\boldsymbol{X}^*$ is

$$\widehat{HR} = \frac{h_0(t)\exp(\widehat{\beta}^T \boldsymbol{X})}{h_0(t)\exp(\widehat{\beta}^T \boldsymbol{X^*})} = \exp\left(\widehat{\beta}^T(\boldsymbol{X} - \boldsymbol{X^*})\right).$$

Under the assumption that the covariates are time-independent, this hazard ratio is also time-independent. For this reason the Cox model is called the *proportional hazards* model. A value of β_i greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases. An hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival (Walters, 1999). Essentially, this model is a multiple linear regression of the logarithm on the hazard on the variables, with the baseline hazard as the "intercept" term that varies over time (Bradburn et al., 2003).

Since the baseline hazard is unspecified, the Cox model can be still estimated by the method of *partial likelihood*, developed by Cox in 1972. Despite the resulting estimates are not as efficient as maximum likelihood estimates for a correctly specified parametric hazard regression model, there is a compensative virtue of this specification. In fact, with the advantage that the partial likelihood doesn't depend on the baseline, we overcome the problem of baseline misspecification. Having fit the model, it is possible to extract an estimate of the baseline hazard (Fox, 2002). Under regular conditions, the maximum partial likelihood estimator behaves the same as the ordinary maximum likelihood estimator of *i.i.d.* random samples in terms of asymptotic consistency, asymptotic normality and asymptotic efficiency (Cox, 1975).

Suppose that we have data for n individuals, among whom there are r different death times (so $n - r$ are the right censored times). The r ordered death times will be denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ (so $t_{(j)}$ is the jth ordered death time). The set of subjects that are at risk at time $t_{(j)}$ will be denoted by $R(t_{(j)})$, called *risk set*. This set is taken over individuals who have died and for whom the times of death have been recorded.

Cox (1975) showed that the relevant likelihood function for the proportional hazard model in (6.3) is given by

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta^T \boldsymbol{X}_{(j)})}{\sum_{l \in R_{(t_{(j)})}} \exp(\beta^T \boldsymbol{X}_{(j)})}, \tag{6.4}$$

where $\boldsymbol{X}_{(j)}$ is the vector of covariates for the individual who dies at the jth ordered death time, $t_{(j)}$. The denominator is a sum of the values of $\exp(\beta^T \boldsymbol{X})$ over all individuals who are at risk at time $t_{(j)}$. The product is taken over the individuals for whom death time have been recorded. The individuals for whom the survival times are censored do not contribute to the numerator of the log-likelihood function, but they do enter into the summation over the risk sets at death times that occur before a censored time. Moreover, the likelihood function depends only on the ranking of the death times, since this determinates the risk set at each death time. Consequently, also the inferences about the effect of explanatory variables on the hazard function depends only on the rank order of the survival times.

It is possible to give another form of this likelihood function with the same results. In fact, considering again n observed survival times, $t_1, t_2, \ldots, t_n$, and considering δ_i, the event indicators (which is zero if the ith survival time $t_i, i = 1, 2, \ldots, n$ is right-censored), the likelihood function in equation (6.4) can then be expressed in the form

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta^T \boldsymbol{X}_i)}{\sum_{l \in R(t_i)} \exp(\beta^T \boldsymbol{X}_l)} \right\}^{\delta_i}, \tag{6.5}$$

where $R(t_i)$ is the risk set at time t_i. In this last expression (6.5), the likelihood function is calculated using information about all the individuals in the dataset, and not, like the previous one (6.4), those referred only to the uncensored individuals.

6.2 Variable selection in Cox model

An important and challenging task is to efficiently select a subset of significant variables upon which the hazard function depends. There are many variable selection techniques in linear regression models. Some of them, such as Akaike Information Criterion (AIC) of Akaike (1974) and Bayesian Information Criterion (BIC) of Schwarz et al. (1978), can be easily extended to survival analysis. Volinsky and Raftery (2000) extended the BIC to the Cox model. They proposed a modification of the penalty term in the BIC, defining it in terms of the number of

uncensored events instead of the number of observations. Moreover, it is possible to extend to the context of survival data analysis the best subset variable selection and stepwise procedure (see Collett (2015)). The latter is a combination of forward and backward selection. In forward selection, variables are added to the model one at time. At each stage, the variable added is the one that gives the largest decreases in the value of $-2\log\widehat{L}$ on its inclusion, where $\widehat{L}$ is the estimation of likelihood function. The stopping rule happens when the next candidate for inclusion in the model does not reduce the $-2\log\widehat{L}$ by more than a particular quantity, chosen at the beginning of the procedure. In backward selection, a model that contains the largest number of variables under consideration is first fitted. Variables are then excluded one at time: at each stage the variable omitted is the one that increases the value of $-2\log\widehat{L}$ by the smallest quantity on its exclusion. The procedure ends when the next candidate for deletion increases the value $-2\log\widehat{L}$ by more than a pre-specified quantity. The most general procedure, the stepwise, works as follows. Variables are added to the model one at time and a variable that has been included in the model can be considered for exclusion at a larger stage. So, after adding a variable in the model, the procedure checks whether any previously included variable can be deleted. These procedures have two main disadvantages: they lead to the identification of one particular subset, rather than a set of a equally good ones and they also depend on the stopping rule (Collett, 2015). This procedure works well only in low-dimensional predictor scenario, while in high-dimension it is preferable to adopt different approaches.

When $p > n$, it is possible to extend the penalised partial likelihood methods for variable selection in the Cox model, with the following formulation

$$\log L(\beta) - n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

using the same idea of Section (1.1) of Chapter 1, where is presented the penalised objective function. When $p_\lambda(.) \equiv 0$, this is reduced to the partial likelihood function of Cox (1975). The penalized likelihood estimate of β is derived by maximizing the penalized partial likelihood with respect to β. With a proper choice of p_λ, many of the estimated coefficients will be zero and hence their corresponding variables do not appear in the model. Tibshirani (1997) extends the LASSO method imposing $p_\lambda(|\beta_j|) = \lambda(|\beta_j|)$, while Fan et al. (2002) propose a SCAD penalty. However, both algorithms were theoretically tested only when $p \ll n$. It is rather unlikely that the two assumptions of LASSO, the beta-min and irrepresentable conditions, hold (i.e. in genomics data). Since the first condition is sufficient and (essentially) necessary condition for model selection consistency, in general we cannot expect that the selected set, as retrieved by the LASSO, will be the true set of variables.

Bradic et al. (2011) addressed the problem of existence of an oracle estimator and regu-

larization estimator under an ultra-high dimensionality setting, where the full dimensionality might grow exponentially or non-polynomially fast with the sample size, in the order of $\log p = O(n^{\delta})$ for some $\delta > 0$ and the order of true regressors goes to infinity, in the order of $s = O(n^{\alpha})$ for $\alpha \in (0, 1)$. With bounded covariates, it is possible to find a strong oracle inequality for the LASSO and SCAD in Cox model, imposing some conditions on covariance matrix and on the score vectors of the log partial likelihood. For LASSO, it is needed to impose a version of irrepresentable condition for censored data very stringent. The restriction has the following formulation: $\beta_n^* \gg \sqrt{s}n^{-0.5+(0.5a+a_1-1)_+ + a_2}$, where $\beta_n^* = \min\{|\beta_j|, j \in M\}$ is the minimum signal strength and $a_2, a_2 > 0$ are constants, and $(x)_+$ is the positive part of x. With this condition the oracle property holds with $s = O(n^{1/3})$. For the SCAD, there is the oracle property, with a less stringent version of irrepresentable condition and with $\beta_n^* \gg \sqrt{s}n^{-0.5}$ and $\lambda_n \gg \sqrt{s}n^{-0.5+(0.5a+a_1-1)_+ + a_2}$.

Wit et al. (2014) extend the dgLARS method of Augugliaro et al. (2013) to Cox model. The basic idea underlying the dgLARS method is to use the differential geometrical structure of a generalized linear model (GLM) to generalize the LARS method originally proposed in Efron et al. (2004).

6.3 Screening

Zhao and Li (2012) generalized the sure independence screening of Fan and Lv (2008) for the Cox proportional hazards model with p covariates. This screening procedure is called the Principled Cox Sure Independence Screening (PSIS). Assuming a marginal Cox model, possibly misspecified, on each X_j, namely,

$$h_{0,j}^*(t) \exp(X_{ij}\beta_j^*),$$

they obtained the maximum partial likelihood estimate of β_j^* , denoted by $\widehat{\beta}_j$. Then, the importance of X_j is measured by a Wald type statistic for testing $\beta_j^* = 0$. As a result, the estimated M_*, where M_* is the true set of explanatory variables, is given by

$$\widehat{M_*} = \{j : I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \geq \lambda_n\}$$

, where $j = 1, \ldots, p$, λ_n is a pre-specified cut-off that depends on n and $\widehat{\beta}_j$ solves the partial likelihood score equation

$$U_j(\beta) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\nu}\left\{X_{ij} - \frac{\sum_{i=1}^{n} X_{ij}\exp(X_{ij}\beta)\tilde{Y}_i(t)}{\sum_{i=1}^{n}\exp(X_{ij}\beta)\tilde{Y}_i(t)}\right\} dN_i(t) = 0,$$

where $I_j(\widehat{\beta}_j) = -\frac{\partial U_j(\beta)}{\partial \beta} \mid_{\beta=\widehat{\beta}_j}$ is the observed information at $\widehat{\beta}_j$, $N_i(t) = I\{\min\{T_i, C_i\} \leq$

$t, \delta_i = 1\}$ is the observed failure process and $\tilde{Y}_i(t) = I(\min\{T_i, C_i\} \geq t)$ is the at-risk process. Here, $\nu > 0$ is the study duration, which is assumed to be long enough to ensure that sample events are observed during the interval $[0, \nu]$. When the true model size $|M_*| = s$, the expected false positive rate can be written as

$$E\left(\frac{|M_*^c \cap \widehat{M}_*|}{|M_*^c|}\right) = \frac{1}{p-s}\sum_{j \in M_*^c} P\{I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j| \geq \lambda_n\}.$$

Moreover, when $\beta_j = 0$ or $j \in M_*^c$, $I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j|$ converges in distribution to a standard normal variable and λ_n controls the expected false positive rate at $2\{1 - \Phi(\lambda_n)\}$, where Φ is the standard normal cumulative distribution function. In order to decrease the false positive rate to 0 as p increases with n, Zhao and Li (2012) fixed the number of false positives (FP) that they are willing to tolerate, which would correspond to a false positive rate of $FP/(p-s)$. Because s is usually unknown, it is possible to be conservative choosing $\lambda_n = \Phi^{-1}\{1 - q/2\}$ where $q = FP/p$, so the expected false positive rate is $2\{1 - \Phi(\lambda_n)\} = q \leq FP/(p-s)$, which is close to the desirable false positive rate, $FP/(p-s)$. To study the sure screening property, Zhao and Li (2012) first established the following β-min condition (that is, the true signals have enough marginal strengths): there exist constants $c_1 > 0$ and $0 < \kappa < 1/2$ such that $\min_{j \in M_*} |cov[X_{ij}, E\{F_T(C_i|\boldsymbol{X}_i)|\boldsymbol{X}_i\}]| \geq c_1 n^{-\kappa}$, where $F_T(\cdot|\boldsymbol{X}_i)$ is the cumulative distribution function of T_i given $\boldsymbol{X}_i$. Then Zhao and Li (2012) proved that

$$\min_{j \in M_*} |\beta_j| \geq c_2 n^{-\kappa}$$

where c_2 is a positive constant. This result leads to the sure screening property

$$P(M_* \subset \widehat{M}_*) \to 1$$

for the non-polynomial dimensionality problem $\log(p) = O(n^{1-2\kappa})$. However, given that PSIS stems from a Wald test based on a Cox model, its performance is unclear when the underlying assumption of a Cox model (i.e. the proportionality of hazard) fails.

With the goal of making the screening procedure less model-centric, Gorst-Rasmussen and Scheike (2013) proposed a Feature Aberration at Survival Times (FAST) statistic that measures the aberration of each covariate relative to its at-risk average. Specifically, for covariate X_j, the FAST statistic is defined as

$$d_j = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\nu}\left\{X_{ij} \; - \frac{\sum_{i=1}^n X_{ij}\tilde{Y}_i(t)}{\sum_{i=1}^n \tilde{Y}_i(t)}\right\} dN_i(t)$$

where $t \in [0, \nu]$. With standardized covariates, the population version of d_j is

$$\tilde{d}_j = E(d_j) = cov\{X_j, F_T(\nu|\boldsymbol{X}_i)\} + \int_0^\nu cov\{X_j, F_T(t|\boldsymbol{X}_i)\}K(t)dt,$$

where $F_T(t|\boldsymbol{X}_i) = P(T_i \leq t|\boldsymbol{X}_i)$ and $K(.)$ is a strictly positive function. Thus, $\tilde{d}_j$ is large if $cov\{X_j, F_T(t|\boldsymbol{X}_i)\}$ has a constant sign throughout $t \in [0, \nu]$. Thus, it is reasonable to consider the magnitude of d_j as a marginal utility to rank the importance of X_j. FAST can be also viewed as a score test statistic based on a Cox model. To study the sure screening property, Gorst-Rasmussen and Scheike (2013) assumed that the true hazard function is of the single-index form

$$h_i(t) = h(t, \boldsymbol{X}_i^T\beta), i = 1, \ldots, n,$$

requiring the resulting survival function $\exp\{\int_0^t h(s, .)ds\}$ to be strictly monotonic for each $t \geq 0$. They proposed to estimate the true set of variable by

$$\widehat{M_*} = \{j : |d_j| > \lambda_n\}$$

for a given λ_n. The authors showed that there exists a threshold $\zeta_n > 0$ such that $\min_{j \in M_*} |\tilde{d}_j| \geq \zeta_n$ and $\max_{j \notin M_*} |\tilde{d}_j| = 0$, assuming a linear regression property, which holds for Gaussian features and, more generally, for features following an elliptically contoured distribution, a restriction on the censoring mechanism that have to be partially random in the sense of depending only on irrelevant features and the partial orthogonality condition that was also used by Fan and Song (2010). Thus, the signals $\tilde{d}_j$ when $j \in M_*$ are stronger than those when $j \notin M_*$. They further assumed that $|cov(X_j, \boldsymbol{X}^T\beta)| \geq c_1 n^{-\kappa}$, $j \in M_*$, for some $c_1 > 0$ and $0 \leq \kappa < 1/2$. Then they showed that, by taking $\lambda_n = c_2 n^{-\kappa}$ for some constant $0 < c_2 \leq c_1/2$, the sure screening property holds even when p grows exponentially fast with n. Like the SIS, FAST assumes that the covariates present in the true model M_* are independent of the irrelevant covariates. This assumption is often violated in practice. To account for possible correlations between variables, Gorst-Rasmussen and Scheike (2013) proposed an iterated FAST procedure.

6.3.1 Model-free screening

It is desirable for screening tools to possess invariance properties under transformations of variables ($\boldsymbol{X}_i$ or T_i) and robustness against outliers. Researchers proposed Kendall's τ based screening methods (Li et al., 2012a), since Kendall's τ, a widely used measure of correlation, is robust against heavy tailed distributions and invariant under monotonic transformations. To accommodate survival data, Song et al. (2014) considered the concordance between failure time T and covariate X_j in the presence of censoring. This procedure is called the Censored

Rank Independence Screening (CRIS). Defining $\tau_j = P(X_{ji} > X_{ji'}, T_i > T_{i'}) - 1/4$, that measures the association between T and X_j. In fact τ_j is 0 if T and X_j are independent. Let $\phi_j = \delta_{ii'} I(X_{ij} > Xi'j, Y_i > Y_{i'})/S^2(Y_{i'})$, it can be easily shown that $E(\phi_j) = P(X_{ij} > X_{i'j}, T_i > T_{i'})$. Thus, a natural estimate of τ_j is:

$$\widehat{\tau}_j = \binom{n}{2}^{-1} \sum_{i<i'} \frac{\delta_{i'}}{\widehat{S}^2(Y_{i'})} I(X_{ij} > X_{i'j}, Y_i > Y_{i'}) - 1/4,$$

where $\widehat{S}$ is the Kaplan–Meier estimator of $S(t) = P(C_i \geq t)$. Define the true set as $M_* = \{j : p(T > t|\boldsymbol{X}) \text{ functionally depends on } X_j\}$. Then, it is estimated by a set of important predictors with large $\widehat{\tau}_j$:

$$\widehat{M_*} = \{j : |\widehat{\tau}_j| > \lambda_n\},$$

where λ_n is a predefined threshold value. Song et al. (2014) showed that $\widehat{\tau}_j$ is a consistent estimator for τ_j. Moreover, the sure screening property with a dimensionality $\log p = o(n^{1-2\kappa})$ is achieved taking $\lambda_n = c_7 n^{-\kappa}$ with $c_7 \leq c_0/2$ and when $\min_{j \in M_*} |P(X_{1j} > X_{2j}, T_1 > T_2) - 1/4| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$. The latter assumption states that the minimal marginal rank correlation between the active variables and the response variable should exceed a certain threshold. Model selection consistency can be achieved if there is a gap between signal variables and noise variables. In this case, a sufficient condition for model selection consistency is that $\boldsymbol{X}_{\widehat{M_*}}$ (the relevant variables) and $\boldsymbol{X}_{\widehat{M_*^c}}$ (the irrelevant variables) are independent. However, the computation of $\widehat{\tau}_j$ requires the comparison of all possible pairs of samples. This exceedingly heavy computational burden may hamper its applicability when the sample size is large.

The validity of model-based screening methods, such as PSIS, often hinges upon the assumptions of the underlying models. For example, when the proportional hazards assumption fails, the model-based approaches may incur a large number of false negatives and lead to invalid results. To develop a model-free framework that can be applicable to a more general class of survival models, He et al. (2013) proposed the Quantile Adaptive sure independence screening (QA). This approach performs screening based on the disparity between unconditional and conditional quantiles given each covariate. However, since not every quantile is estimable under censoring, its performance under heavy censoring is unclear.

QA and CRIS are model-free screeners, but they may not capture the full-range impact of covariates on the overall survival since QA focuses on a specific quantile level and CRIS relies on a summarized value of association. To more fully capture the overall influence of a covariate on the outcome distribution, Li et al. (2016) proposed a new metric called the Survival Impact Index (SII), which evaluates the absolute deviation of the covariate-stratified survival distribution from the unstratified survival distribution. Specifically, for

each $X_j, j = 1, \ldots, p$, SII is defined as

$$\xi_j = \int_{t \in T, x \in X} W_\xi(t, x) |S(t|X_j > x) - S(t)| dx dt,$$

where $W_\xi(t, x)$ is a pre-determined weight function introduced to capture the covariate impact on either early or late survival. The authors argued that if, for at least one t and one x, the survival function stratified on $X_j > x$ differs from the unstratified survival function at t, then ξ_j will be non-zero under mild conditions. On the other hand, if T and X_j are independent, then $\xi_j = 0$. To estimate ξ_j, Li et al. (2016) proposed to use

$$\widehat{\xi}_j = \int_{t \in T, x \in X} W_\xi(t, x) |\widehat{S}(t|X_j > x) - \widehat{S}(t)| dx dt,$$

where $\widehat{S}(t|X_j > x)$ is the Kaplan-Meier estimator based on sub-sample $X_j > x$ and $\widehat{S}(t)$ is the Kaplan-Meier estimator for the survival function of T. The set of important predictors is defined by

$$\widehat{M_*} = \{j : \widehat{\xi}_j > \lambda_n\}.$$

Under some regularity conditions, the estimated survival impact index $\widehat{\xi}_j$ is uniformly consistent to ξ_j. If $p = O(\exp(n^c))$ for some $0 < c < 1$ and $\min_{j \in M} \xi_j > c_0 n^{-\alpha}$ for some constants $c_0 > 0$, $0 \leq \alpha < (1 - c)/2$, and if the information collected from the region $T \times \boldsymbol{X}$ can produce a rather stable estimation of X_j's impact on the distribution of T, Li et al. (2016) proved that

$$p(M_* \subset \widehat{M_*}) \to 1$$

by taking $\lambda_n = bn^{-\alpha}$ with $b \leq c_0/2$.

In a survival setting, non-parametric variable screeners have focused on discerning how each candidate variable influences survival functions. One way of detecting such influence is by studying the variability of survival functions for the sub-populations or strata defined by each variable. The difference patterns, however, may vary across covariates. Specifically, the differences may occur either during the early or late period in the follow-up due to disease-related characteristics. Therefore, screening approaches that rely on a single screening criterion may not be able to capture the complex difference patterns and may lead to false non-discovery. In order to capture the differences during the periods, Hong et al. (2018) proposed to consider the following Integrated Powered Density (IPOD):

$$\int_0^t f^\gamma(s) ds,$$

where a power index $\gamma (> 0)$ inflates either early ($\gamma > 1$) or late differences ($\gamma < 1$) during the life span and thus it gives more flexibility in detecting distributional differences. IPOD

resembles the cumulative distribution function (CDF) and satisfies the basic properties of CDFs, except that it does not necessarily approach one when $t \to \infty$. To derive the screening criterion, first consider a discrete X_j with R_j categories. The unique property of IPOD motivates the following marginal utility to detect distributional differences:

$$I_j^{(\gamma)} = \max_{r_1, r_2 \in \{1,\ldots,R_j\}} \sup_{t \in [0,\nu]} \left| \int_0^t f_{T|X_j}^{\gamma}(s|X_j = r_1)ds - \int_0^t f_{T|X_j}^{\gamma}(s|X_j = r_2)ds \right|,$$

where $f_{T|X_j}(s|X_j = r)$ denotes the conditional density function of T given $X_j = r$. Since $I_j^{(\gamma)} = 0$ if and only if T and X_j are independent, it serves as a measure of marginal utility for each X_j. The framework of IPOD accommodates different γ's. For example, when $\gamma = 1$, $I(\gamma)_j$ is simply the classical Kolmogorov difference: $\max_{r_1,r_2\in\{1,\ldots,R_j\}} \sup_{t\in[0,\nu]} |F_T|X_j(t|X_j = r_1) - F_T|X_j(t|X_j = r_2)|$. So this framework is general, including the Kolmogorov filter (Mai and Zou, 2015) as a special case. Denote by $h_n > 0$ the bandwidth of a kernel function $K(\cdot)$, $I_j^{(\gamma)}$ can be estimated by

$$\widehat{I}_j^{(\gamma)} = \max_{r_1, r_2 \in \{1,\ldots,R_j\}} \sup_{t \in [0,\nu]} \left| \int_0^t \widehat{f}_{T|X_j}^{\lambda}(s|X_j = r_1)ds - \int_0^t \widehat{f}_{T|X_j}^{\lambda}(s|X_j = r_2)ds \right|,$$

with

$$\widehat{f}_T(t) = \sum_i K((t - t_i)/h_n)(\widehat{S}_T(t_{i-1}) - \widehat{S}_T(t_i)),$$

where $\widehat{S}(t)$ is the Kaplan–Meier estimator for the survival function of T and the conditional density estimator $\widehat{f}_{T|X_j}(t|X_j = r)$ can be obtained similarly as $\widehat{f}_T(t)$ by restricting samples to $X_j = r$. Hong et al. (2018) defined the true important feature set as

$$M_* = \{j : S(t|\boldsymbol{X}) \text{ functionally depends on } X_j \text{ for some } t \in (0, \infty)\}$$

estimated by $\widehat{M}_{1*} = \left\{j : \widehat{I}(\gamma)_j > \lambda_n\right\}$ where $\lambda_n > 0$. This procedure is referred to as the IPOD screening.

When a covariate X_j is continuous, it can be discretized into R_j slices. Suppose there are N different ways of slicing a continuous covariate X_j by using the percentiles of the empirical distribution of X_j, denoted by $\Lambda_{ju}, u = 1, \ldots, N$, with each slice Λ_{ju} contains R_{ju} intervals. Denoting as $\widehat{I}(\gamma)_{j,\Lambda_{ju}}$ the IPOD screening statistic corresponding to the slicing scheme of Λ_{ju}, they proposed the following fused IPOD screening statistic that collects all information from each slice:

$$\tilde{I}_j^{(\gamma)} = \sum_{u=1}^{N} \widehat{I}_{j,\Lambda_{ju}}^{(\gamma)}.$$

This statistic leads to the following screening criterion, $\widehat{M}_{2*} = \left\{j : \tilde{I}_j^{(\gamma)} > \lambda_n\right\}$, where $\lambda_n > 0$

is a pre-specified constant.

For large sample results, Hong et al. (2018) stipulated the following assumptions. Let Λ_{juo} be the partition based on the theoretical quantiles $q_{ju(r)}$ of X_j and $I_{jo}^{(\gamma)} = \sum_{u=1}^{N} I_{j,\Lambda_{juo}}^{(\gamma)}$ and assume that there exist $c > 0$ and $0 < v < 1/2$ such that $\min_{j \in M} I_{jo}^{(\gamma)} \geq 2cn^{-v}$ for a specific γ. When covariates include both continuous and discrete values, under the above conditions, for $0 < \alpha < 1 - 3\kappa - 2v - 2\mu - 2\rho$, if $N = O(\log n)$ and $\log p = O(n^{\alpha})$, the fused IPOD has the sure screening property. IPOD enjoys the invariance property like other non-parametric screeners such as SII and CRIS, but it is more computationally efficient with increasing n. The performance of the method depends on how well the distribution function can be estimated on each covariate-defined stratum. Hence, it may not work well for small sample sizes.

To accommodate censoring in ultra-high-dimensional survival data, Liu et al. (2018) replaced the conditional distribution of each covariate given a response variable in FKF of Mai and Zou (2015) with a conditional distribution of a response variable given each covariate, and then they used the Kaplan–Meier to estimate the unknown conditional distributions. This Kolmogorov–Smirnov statistic-based independence screening method can deal with discrete, categorical or continuous covariates and it is called the fused Kolmogorov–Smirnov statistic-based Sure Independence Screening (KS-SIS). Liu et al. (2018) proposed the following statistic:

$$K_j^{G_j} = \max_{l_1, l_2} \sup_{0 \leq t \leq \tau} |S_j(t|I_j = l_1) - S_j(t|I_j = l_2)|$$

where $G_j = \{[a_l^j, a_{l+1}^j) : a_l^j < a_{l+1}^j, l = 0, \ldots, G_j - 1 \text{ and } \cup_{l=0}^{G_j - 1} [a_l^j, a_{l+1}^j) \setminus \{a_0^j\} = \mathbb{R}\}$. When X_j takes finite values such that each possible value forms a slice, X_j is independent of T if and only if $K_j^{G_j} = 0$. When X_j is continuous or discrete, it is independent of T if and only if $K_j^{G_j} = 0$ for any partition G_j. The estimate of $K_j^{G_j}$ is defined as

$$\widehat{K}_j^{G_j} = \max_{l_1, l_2} \sup_{0 \leq t \leq \tau} |\widehat{S}_j(t|I_j = l_1) - \widehat{S}_j(t|I_j = l_2)|,$$

where $\widehat{S}_j(t|I_j = l_b)$, with $b = c(1, 2)$, is the Kaplan-Meier estimator of $S_j(t|I_j = l_b)$. Furthermore, they used the idea of fusion to improve the efficiency of the Kolmogorov-Smirnov measure. Then

$$\widehat{K}_j = \sum_{k=1}^{N_j} \widehat{K}_j^{G_{kj}}$$

is an estimate of $K_j = \sum_{k=1}^{N_j} K_j^{G_{kj}}$ considering N_j different partition $G_{kj}, k = 1, \ldots, N_j$. Then based on $\widehat{K}_j$, they defined the estimated active set as

$$\widehat{A}(d_n) = \{1 \leq j \leq p : \widehat{K}_j \text{ is amongst the first } d_n \text{ largest of all } \widehat{K}_j\}$$

where d_n is a prespecified positive integer. Under the following two conditions

- there exists a set B such that $A \subset B$

$$\Delta_B = \min_{j \in B} \min_{1 \le k \le N_j} K_j^{(o)}(G_{kj}) - \max_{j \notin B} \max_{1 \le k \le N_j} K_j^{(o)}(G_{kj}) > 0$$

  where the slicing is built on the theoretical quantiles of X_j, so the jointly important predictors should also be marginally important;

- if X_j is continuous, then for any d_2, d_2 such that $P(X_j \in [d_1, d_2)) \le 2/\min_k G_{kj}$, they have

$$|S_j(t|x_1) - S_j(t|x_2)| \le \frac{\Delta_B}{8}$$

  for all t, j and $x_1, x_2 \in [d_1, d_2)$,

the sure screening property holds with probability tending to one if

$$\Delta_B \gg \sqrt{\frac{\log n \log(pN \log n)}{n}}$$

with a dimensionality $p = O(n^c)$ with $c < 1$.

Chapter 7

D-ELSIS in survival analysis

Let T, C and $\boldsymbol{X}_{(n\times p)}$ be respectively the survival time, the censoring time and their associated covariates, with $n \ll p$. Correspondingly, let $Z = \min(T, C)$ be the observed time and let $\delta = I(T \leq C)$ be the censoring indicator. To analyse the association between T and $\boldsymbol{X}$ in a statistical setting, we consider the observed data $\{(\boldsymbol{X}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ as an *i.i.d.* random sample from a population $(\boldsymbol{X}, Z, \delta)$. Further, we consider two classical assumptions on the censoring times:

(B1) the independence censoring, so T and C are independent given $\boldsymbol{X}$;

(B2) the noninformative censoring, so the conditional distribution of C given $\boldsymbol{X}$ does not involve the parameters of interest.

The requirement (B1) essentially means that the uncensored subjects under follow-up have to be representative of the surviving population. This condition that is satisfied when censoring occurs independently of the survival time. When there are covariates, then the independent censoring assumption is made conditional on the covariate information. The assumption (B2) assumes that participants should drop out of the study only for reasons unrelated to the study. These two assumptions are classical in survival analysis.

Following the notation in Fan and Gijbels (1996), it is possible to study the contribution of the risk factors via the conditional hazard rate function. For a given covariate X_j, the hazard rate at a given time t is

$$h(t|X_j) = \lim_{\delta\to 0} \frac{P(t \leq T < t + \delta | T \geq t, X_j)}{\delta}$$

representing the risk that an individual fails immediately after time t given survived at the that time. There are two popular models based on the hazard rate: the Accelerated Failure Time model (AFT) and the Cox Proportional Hazards model (PH-COX). The AFT assumes

that the hazard rate has the form

$$h(t|x) = h_0(t\psi(x))\psi(x)$$

and the PH-COX describes the hazard rate via

$$h(t|x) = h_0(t)\exp(\psi(x)),$$

where $h_0(\cdot)$ is the baseline hazard function and $\psi(x)$ is the function depicts the contribution of covariates x.

Assume for simplicity that the covariates $\boldsymbol{X}$ are not time varying and that the random variable T is absolutely continuous. Note that

$$h(t|x) = f(t|x)/S(t|x), \quad \text{with} \quad S(t|x) = 1 - F(t|x)$$

being the conditional survivor function, where $f(t|x)$ and $F(t|x)$ are respectively the conditional density and distribution function of T given $X = x$. From the property of the distribution function, the survival function is continuous and non-increasing. The conditional survivor function can be represented as

$$S(t|x) = \exp\left\{-\int_0^t h(u|x)du\right\} \tag{7.1}$$

for a nonnegative random variable T. Hence, modelling the hazard rate function is equivalent to assume a specific form for the conditional distribution or for the conditional survival function.

Regarding the considerations state above, it is possible to introduce a more general approach to study how the variables affect the event of interest via the regression model:

$$Y = g(T) = m(\boldsymbol{X}) + \epsilon$$

for a given transformation $g(\cdot)$, often a logarithmic function, with $\sigma^2(\boldsymbol{X}) \equiv 1$. This approach attempts to assess the contributions of risk factors via a mean response function $m(\cdot)$, where $m(\cdot)$ is an unknown function.

Also in this context, as we did in the previous chapters, we are looking for a method that allows screening without imposing conditions on the functional form that links time, which in this case is the response variable, to the other covariates. To do this, we therefore need a nonparametric estimator.

7.1 Limitation of Kaplan Meier estimator

As we said in Chapter 6, the Kaplan Meier (KM) estimator is a nonparametric statistical method used to estimate the probability of survival over time, given by

$$\widehat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \tag{7.2}$$

for $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r$, with $\widehat{S}(t) = 1$ for $t < t_{(1)}$ and where $t_{(r+1)}$ is taken to be ∞. Kaplan Meier plots visualize the probability that a patient survives a certain time. For each time interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk. Subjects who have died, dropped out, or move out are not counted as at risk, and since they are considered as censored, they are not included in the denominator. Total probability of survival till that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time. The Kaplan Meier estimator is well defined for all time points less than the largest observed study time. If the largest study time corresponds to a death time, the estimated survival curve is zero beyond this point. If the largest time point is censored, the value of $S(t)$ beyond this point is undetermined.

As it is possible to see from the formula (7.2), this estimator does not involve covariates. However, we know that there are significant factors that contribute to different survival times. With additional information we can have more accurate survival estimates to individual patients. In fact, when we have a study for example on the effect of a drug for a particular disease, we want to evaluate whether this medicine is effective or not. To do this, we can consider the difference between survival functions, taking into account patients with and without the treatment. When the covariate is discrete, the KM can be used to compare survivals. Two survivals are obtained, one for the case-subjects and one for those under control. With the help of a graph, we can evaluate which of the two functions obtained through the KM estimate is higher. If the subjects who took the treatment have a higher survival, then it is possible to say that the medicine is able to treat the patients. We underline that Kaplan Meier estimator does not test the difference of survival functions. In fact, the Kaplan Meier estimator estimates survival probabilities and does not compare them. It does not compare them. To make inferences on these survival probabilities we need a test. In the literature there are various procedures that allow to test whether the difference between the two curves is significant or not. Among these, the well known test is the *log-rank test* (Peto and Peto, 1972).

When we estimate $S(t)$ function conditioned on a continuous variable, we need to discretize the variable and get as many survival functions as classes in which we have divided the variable. Evaluating the comparison of the classes with the plot, in this case, becomes much

more complicated. Furthermore, the KM estimator cannot be used to estimate the survival function conditioned by the effect of several variables at the same time.

Some model-free screening methods, such as KS-SIS (Liu et al., 2018) and IPOD (Hong et al., 2018), use the KM estimator in order to discern how each candidate variable influences survival function, even in the presence of continuous covariates. In both cases, the continuous covariates are discretized into slices and the authors compare the survival functions (estimated on each slice using KM) with the Kolmogorov-Smirnov statistic. KM is estimated considering the number of events that occurred and the number of subjects that survived up to that point. Substantially, when they calculate the survival function on the slice, they take into account a subset of the observations, so only a subset of the events and a subset of the subjects at risk. If instead of considering more slices for the continuous covariate we consider only one slice, we cannot actually estimate the effect of the covariate on survival, because the set that we select is the whole set of observations. Therefore, considering a single slice, we obtain an estimation of the survival function which is not conditioned. It is possible to estimate directly the covariate's effect on survival function using a different approach not involving the KM estimator, as we will see in the next section.

7.2 D-ELSIS in survival analysis

For the survival-based screening, we intend to recover a sparse subset

$$M_* = \{1 \leq j \leq p : \text{the } j\text{-th variable in } \boldsymbol{X} \text{ is relevant for explanation of } Y\}.$$

In biomedical studies, it is reasonable to stipulate a sparsity condition that only a small number of covariates are relevant. That is, the cardinality s of M_* is small relative to p.

In order to construct a nonparametric screening procedure, we need a nonparametric estimator different from KM that does not suffer from the problems considered in the previous section. Specifically, the KM estimator conditioned by continuous covariates depends on how the subsets are chosen. To solve the problem of the conditional estimation of the survival function, we consider the general regression model in survival analysis:

$$Y = g(T) = m(\boldsymbol{X}) + \epsilon \tag{7.3}$$

with a generic function $g(\cdot)$ strictly monotone. Depending on the form of $g(\cdot)$, we can obtain the COX and AFT models. If $g(t) = t$ we model the time directly.

This model has the same structure as the nonparametric model considered in the first part of this thesis, in Chapter 2, since we do not impose condition on the functional form of $m(\cdot)$. The substantial difference is that dependent variable is a function of time and not time directly. We need the following assumptions:

(C1) there exists a $\nu > 0$ such that $S(\nu|X_j) > \theta_1 > 0$ for $1 \leq j \leq p$, where θ_1 is a positive constant;

(C2) for any $t \in [0, \nu]$, $f_{T|X_j}(t|x)$ is greater than a positive constant $\hat{c}_0$ for $j \in M_*$.

Condition (C1) is imposed to avoid problems with estimating the tail of the conditional survival functions. Because (C1) is satisfied in many clinical settings, it is widely used in literature (Peng and Fine, 2009). In practise, ν is often chosen to be the study duration. Condition (C2) states that the conditional density is positive.

Using these two assumptions and the strict monotonicity of $g(\cdot)$, the relevant covariates in the model (7.3) will also be relevant for the survival function. In fact, the strictly monotonicity of $g(\cdot)$ ensures that a variation of t leads to a variation of $g(t)$. When a variable X_j influences $Y = g(t)$ in the regression model (7.3), this variable is relevant. Since the survival function $S(t|x)$ and the density $f_{T|X_j}(t|x)$ are positive, as ensured by (C1) and (C2), respectively, a variation of t, due to a variation of X_j in (7.3), leads to a variation of $S(t|X_j = x)$. In this way we can handle the problem of the conditioned survival function in a better way than using the KM estimator. Therefore, looking for the covariates relevant for the regression model is sufficient to find the covariates being also relevant for the survival probability.

Since we have the same regression model structure of (2.1), we can use the same procedure considered in Chapter 2. In fact, in order to identify explanatory variables that contribute to the response variable in high-dimensional non-parametric regression in survival analysis, we consider our independence model-free feature screening technique D-ELSIS, proposed in Chapter 2. With our D-ELSIS procedure, we obtain a model-free screening procedure without the use of the KM estimate of survival function. This is the fundamental difference among our method and the other model-free screeners in literature, such as IPOD of Hong et al. (2018) and KS-SIS of Liu et al. (2018). Furthermore, based on our knowledge, in survival context, a screening method that combines empirical likelihood and local polynomial regression has never been used.

As regards the assumptions in Chapter 3 for the screening property of D-ELSIS, we need to adequately adopt them in this context. Consider the marginal contribution $f_j(x) = E(Y|X_j = x) = E(g(t)|X_j = x)$ of explanatory variable X_j on $g(t)$. Since $g(\cdot)$ is strictly monotone, this is the same that considers a marginal contribution of an explanatory variable X_j on t. We apply the local polynomial regression to estimate the first partial derivative with respect to the covariate X_j in the regression model, for $j = 1, \ldots, p$. Once we get this estimation, we use the empirical likelihood to verify if this derivative is zero uniformly in the covariate's support. With the use of partial derivative, we investigate the marginal contribution from each explanatory variable in explaining Y, that is a function of time, to justify whether it is relevant or not, following the same idea of Chapter 2. In fact, if $f_j'(\cdot) \equiv 0$, the covariate X_j is not relevant, otherwise X_j is an active variable.

As we mentioned in Section 2.4 of this thesis, we estimate the first marginal derivative of our nonparametric model (7.3) using the local quadratic estimator

$$f_j'(x) = \sum_{i=1}^{n} W_{i,2}(x) Y_i = \sum_{i=1}^{n} \frac{1}{n} S(x;h) K_h(X_i - x) Y_i. \tag{7.4}$$

As we explained in Lemma 1 of Chapter 3, for assessing $f_j'(x) = 0$ at given x without distributional assumptions, we can use a simplified version of the local quadratic estimator in the following empirical likelihood:

$$EL_j(x,0) = \sup_{w} \left\{ \prod_{i=1}^{n} w_i : w_i \geq 0, \sum_{i=1}^{n} w_i = 1, \sum_{i=1}^{n} w_i U_{ij} = 0 \right\}, \tag{7.5}$$

where $U_{ij} = \frac{1}{h^2} K_h(X_{ij} - x)(X_{ij} - x) Y_i$. By applying the Lagrange multiplier method for solving (7.5), we obtain the empirical likelihood ratio

$$l_j(x,0) = -2\log\{EL_j(x,0)\} - 2n\log n = 2\sum_{i=1}^{n} \log\{1 + \lambda U_{ij}\}, \tag{7.6}$$

where λ is the univariate Lagrange multiplier solving $\sum_{i=1}^{n} \frac{U_{ij}}{1+\lambda U_{ij}} = 0$. The $l_j(x,0)$ is a statistic for testing whether or not (7.4) has zero mean locally at x. For assessing $f_j'(\cdot) \equiv 0$, we use

$$l_j(0) = \sup_{x \in \mathcal{X}_j} l_j(x,0)$$

for each $j = 1, \ldots, p$, where $\mathcal{X}_j$ is the support of variable X_j.

For feature screening purpose, we sort l_j for all $j = 1, \ldots, p$ in decreasing order, and we take the first γ_n covariates. In this way, we create a set

$$\widehat{M}_{\gamma_n} = \{1 \leq j \leq p : l_j \geq \gamma_n\}$$

In order to implement the proposed method, we evaluate the statistic l_j using $l_j(0) = \max_{1 \leq i \leq n} l_j(X_{ij}, 0)$.

As we did in the previous part of this thesis, we can transform the screening selection procedure in a variable selection procedure with the use of subsample tool, as we explained in Section 2.5 of this thesis.

Chapter 8

Simulations for screening in Survival analysis

Simulation studies are conducted to investigate the performance of our D-ELSIS method, proposed in Chapter 2 of this thesis and extended to survival analysis in Chapter 7, in terms of the following three criteria: (i) the median of the minimum model size (MMSs, i.e., the smallest number of the selected covariates including all the active explanatory variables) for 100 repetitions; (ii) the IQR divided by 1.34 (SD), that is the robust measure of the standard error of MMS; (iii) the percentage of true positive rate (TPR) that controls the precision measuring the proportion of actual relevant variables that are correctly identified as such. To calculate the TPR we consider that the predicted relevant variables are the first 20. In order to have a very good method, the MMS should be equal to the number of the true active variable, with small SD and high TPR. We set $n = (500, 750, 1000)$ and $p = (100, n/2, 2n)$.

For comparison, we also consider other two existing screening methods for nonparametric models: the Integrated Powered Density (IPOD) of Hong et al. (2018) and fused Kolmogorov –Smirnov statistic-based (KS-SIS) of Liu et al. (2018), presented in Chapter 6. For the implementation of the best bandwidth in the kernel regression estimation for D-ELSIS, we use the R package **NonpModelCheck** of Zambom et al. (2017). Among the various options of the package, we choose the cross-validation leave-one-out, which performs satisfactorily. Instead, as regards the likelihood estimation for empirical likelihood, we use the R package **emplik** of Zhou (2018). Furthermore, for the IPOD method, we use the code that Hong et al. (2018) provided in their paper and we consider $\gamma = (0.8, 1.0, 1.2)$ as in their study.

We consider the following experiments in the simulation study.

Example 5 : Cox model

The Cox model represents the most widely used model in survival analysis. In this case we want to verify how much the increase of censoring rate affects the performance of our

estimator. The survival time is generated from a Cox model

$$h(t|\boldsymbol{X}) = \exp(\beta^T \boldsymbol{X}).$$

Here predictors X_j's are generated from a multivariate normal distribution with mean $\mu = 0$, variance $\sigma^2 = 1$, correlation $\rho = \{0, 0.5\}$ and $\beta = (1_5^T, 0_{p-5}^T)$. The censoring time is generated from a uniform distribution $U(0, c)$, where c is chosen to achieve censoring proportion of 20% and 40%. The results are shown in Table 8.1.

Example 6 : Non-linear covariate-response relationship

In this case we are interested in evaluating the effect of the increased correlation between the covariates on the results of the screening. The survival time is generated from

$$\log(T) = 5X_1 - 4X_2(1-X_2) + 10\left[\exp\{-3(X_3-1)^2\} + \exp\{-4(X_3-3)^2\}\right] - 1.5 + 4\sin(2\pi X_4) + \epsilon$$

Here predictors X_j's are generated from a multivariate normal distribution with mean $\mu = 0$, variance $\sigma^2 = 1$ and correlation $\rho = \{0, 0.5\}$. The error $\epsilon \sim N(0, 1)$ is independent from $\boldsymbol{X}$. The four true main effects were initially generated as

$$f_1(x) = 5x, \quad f_2(x) = -4x(1-x)$$

$$f_3(x) = 10\left[\exp\{-3(x-1)^2\} + \exp\{-4(x-3)^2\}\right] - 1.5, \quad f_4(x) = 4\sin(2\pi x)$$

Then, each f_j is standardized by subtracting $E[f_j(x)]$ and dividing by $SD[f_j(x)]$ to have zero mean and unit variance. The censoring time C is generated from a 3-component normal mixture distribution $N(0, 4) - N(5, 1) + 0.5N(25, 1)$. This example is adopted from Li et al. (2016). The results are displayed in Table 8.2

8.1 Simulation results

The results shown in Table 1 reports the values of MMS and TPR for the first scenario where the data are generated from the Cox model. In order to check how the D-ELSIS method works, we consider different settings to analyze the effects of correlation and censoring on the performance of the D-ELSIS. In particular, the correlation is fixed to be equal to $\rho = c(0.00, 0.50)$, and the censoring percentage is set to be equal to $c(20\%, 40\%)$. Moreover, the D-ELSIS performance is compared with that of IPOD (with different values of γ) and KS-SIS. Looking at the TPR, if correlation between covariates does not exist, i.e. $\rho = 0$, D-ELSIS has a better performance when the censoring is low and when $n > p$. Then, when $n < p$, our proposed screening technique is less efficient for smaller sample sizes, while its

performance increases when the sample size goes up. In absence of correlation, its performance is substantially equal in terms of TPR. Moreover, the censoring affects the TPR especially when $n < p$. In fact in order to make the results stable, we have to increase the sample size at $n = 1000$. In terms of MMS, D-ELSIS fails only when $n = 500$ and $p = c(100, 250)$. In all the other cases, it performs well. Now, if we assume that the covariates are linearly correlated, D-ELSIS works very well for most of the settings, in terms of both TPR and MMS. Moreover, the SD of MMS decreases when the sample size increases. Compared with the IPOD and KS-SIS, D-ELSIS performs particularly well when there is correlation, independently of the censoring percentage. Instead, in terms of MMS, it works better because it is able to capture the number of relevant covariates when both $n > p$ and $n < p$.

For the second scenario, where a non-linear relationship between covariates and time of interest is assumed, the results are shown in Table 2. It is possible to observe that D-ELSIS works as better as the other methods when there is not correlation between covariates. Instead, it outperforms them when there is a linear relationship in terms of both TPR and MMS, whatever the sample size and number of covariates are. It is particular evident that IPOD and KS-SIS fail in presence of correlation, and their performance is very poor when $n < p$, in terms of MMS and SD, while D-ELSIS is quite stable in selecting the true set of covariates.

Table 8.1: Simulation results from Example 5

$s=5$	$n=500$						$n=750$						$n=1000$					
	$p=100$		$p=250$		$p=1000$		$p=100$		$p=375$		$p=1500$		$p=100$		$p=500$		$p=2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method																		
	$\rho=0$ CR=20 %																	
D-ELSIS	5 (0.75)	100.00	6 (0.93)	99.60	8 (3.73)	98.40	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	99.80	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=0.8$)	6 (1.49)	100.00	8 (4.66)	97.60	15 (5.11)	90.00	5 (0.00)	100.00	5 (0.00)	100.00	6 (1.49)	99.60	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	99.80
IPOD ($\gamma=1.0$)	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=1.2$)	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
KS-SIS	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
	$\rho=0$ CR=40 %																	
D-ELSIS	5 (0.75)	99.60	6 (2.99)	98.60	8 (6.71)	96.20	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.19)	99.80	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=0.8$)	6 (0.93)	99.80	6 (2.99)	99.00	11 (9.14)	95.20	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.75)	99.60	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	99.80
IPOD ($\gamma=1.0$)	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=1.2$)	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
KS-SIS	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
	$\rho=0.5$ CR=20 %																	
D-ELSIS	5 (0.75)	100.00	6 (1.49)	100.00	9 (6.43)	97.60	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.75)	99.40	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=0.8$)	9 (6.72)	97.40	15 (12.87)	91.40	44 (47.95)	76.80	5 (1.49)	99.40	6 (3.73)	98.40	13 (19.96)	91.20	5 (0.00)	100.00	6 (2.24)	99.40	7 (5.97)	98.00
IPOD ($\gamma=1.0$)	8 (5.22)	98.40	11 (11.38)	93.80	30 (48.13)	82.60	5 (1.49)	99.60	7 (4.48)	98.80	11 (15.30)	92.80	5 (0.00)	100.00	5 (1.49)	99.00	7 (4.48)	98.40
IPOD ($\gamma=1.2$)	8 (3.92)	99.00	11 (8.40)	95.20	25 (34.33)	84.20	5 (0.75)	99.60	6 (3.73)	99.00	10 (13.06)	93.80	5 (0.00)	100.00	5 (0.75)	99.00	6 (2.43)	99.00
KS-SIS	5 (0.75)	99.80	6 (2.24)	99.40	8 (6.16)	96.80	5 (0.00)	100.00	5 (0.00)	100.00	5 (1.49)	98.60	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
	$\rho=0.5$ CR=40 %																	
D-ELSIS	5 (0.75)	100.00	6 (1.49)	99.60	10 (7.65)	99.60	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.75)	100.00	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00
IPOD ($\gamma=0.8$)	15 (11.20)	92.80	28 (26.12)	82.80	91 (96.45)	59.60	8 (4.48)	98.40	21 (25.75)	86.80	65 (82.28)	71.40	6 (2.24)	100.00	11 (11.94)	93.40	27 (40.49)	83.80
IPOD ($\gamma=1.0$)	9 (5.97)	97.60	13 (9.70)	93.80	35 (44.96)	80.60	6 (2.24)	100.00	8 (6.16)	97.00	19 (19.22)	89.00	5 (0.75)	100.00	5 (2.24)	99.00	7 (5.97)	97.20
IPOD ($\gamma=1.2$)	7 (4.48)	98.20	10 (7.46)	96.20	19 (30.41)	87.20	5 (1.49)	100.00	6 (2.99)	98.40	10 (12.13)	93.20	5 (0.00)	100.00	5 (0.75)	99.80	6 (3.17)	99.20
KS-SIS	5 (1.49)	99.20	6 (2.24)	99.80	11 (13.43)	93.40	5 (0.00)	100.00	5 (0.75)	100.00	5 (1.49)	98.80	5 (0.00)	100.00	5 (0.00)	100.00	5 (0.00)	100.00

Table 8.2: Simulation results from Example 6

$s = 4$	$n = 500$						$n = 750$						$n = 1000$					
	$p = 100$		$p = 250$		$p = 1000$		$p = 100$		$p = 375$		$p = 1500$		$p = 100$		$p = 500$		$p = 2000$	
	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR	MMS (SD)	TPR
Method																		
	$\rho = 0$																	
D-ELSIS	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
IPOD ($\gamma = 0.8$)	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
IPOD ($\gamma = 1.0$)	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.75)	99.50	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
IPOD ($\gamma = 1.2$)	4 (0.00)	99.75	4 (0.75)	100.00	4 (2.42)	99.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
KS-SIS	4 (0.00)	100.00	4 (0.75)	100.00	5 (3.92)	97.25	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00	4 (0.00)	100.00
	$\rho = 0.5$																	
D-ELSIS	4 (1.49)	98.75	5 (3.73)	97.75	11 (17.35)	90.75	4 (0.00)	100.00	4 (0.75)	99.00	5 (1.49)	98.50	4 (0.00)	99.75	4 (0.00)	100.00	4 (0.75)	98.75
IPOD ($\gamma = 0.8$)	19 (24.07)	87.25	24 (71.08)	84.50	190 (229.85)	71.25	13 (16.60)	90.25	62 (102.43)	82.00	127 (305.60)	77.50	13 (21.08)	89.25	37 (86.38)	84.50	138 (384.13)	77.25
IPOD ($\gamma = 1.0$)	18 (21.83)	88.75	21 (50.19)	86.00	183 (255.97)	70.50	13 (16.04)	91.00	57 (97.76)	81.25	103 (318.84)	76.50	13 (20.15)	90.75	32 (80.78)	85.00	139 (308.58)	77.50
IPOD ($\gamma = 1.2$)	18 (22.20)	88.50	25 (47.95)	84.25	190 (242.72)	68.50	13 (17.35)	89.00	42 (89.55)	81.75	105 (308.77)	74.75	12 (17.91)	90.25	35 (77.98)	84.75	153 (286.01)	76.00
KS-SIS	20 (21.27)	85.50	46 (78.92)	77.50	217 (279.85)	65.00	10 (14.18)	93.00	27 (51.49)	85.50	206 (336.01)	74.00	11 (26.31)	90.50	33 (79.85)	84.25	306 (530.22)	74.25

Chapter 9

Conclusions

In the second part of this thesis, after a review on the recent developments of variable and screening selection for survival data analysis in ultra-high dimensions, we proposed to use D-ELSIS screening procedure with time-to-event data. Many approaches developed for uncensored data have been adapted to time-to-event data. Following the same idea introduced in the first part of the thesis, we have shown that it is possible to use our new proposal also in this context.

D-ELSIS differs from other model-based screening methods in survival analysis since it selects the relevant covariates without imposing assumptions on the underlying model. Furthermore, it does not use the KM estimator and the fused technique, as some model-free screening methods do. In particular, we focused our attention on handling the conditional survival function, without using the KM estimator, which has numerous disadvantages in the presence of continuous covariates. In fact, in order to find the relevant variables for the survival function we used the general regression model, in which the response variable is a function of time. Under some regularity conditions, we have shown that the relevant covariates for the regression model are relevant also for the survival probability.

The simulations results show that our approach is able to select the relevant covariates, especially in the presence of correlation between the relevant and non-relevant ones. In fact, compared to model-free competitors based on fused technique, when the effect of the covariate on the event of interest is non-linear and in the presence of correlation, D-ELSIS has significantly better results.

Since the results obtained from the simulations are very promising, as future works we will demonstrate that D-ELSIS has the screening property with survival data in ultra-high dimensions. Furthermore, we will test that the subsample technique of Chapter 2, that is very general, transforms screening into variable selection also in this particular context. Finally, we will apply our proposed method on some real data sets, in order to assess its performance.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Amendola, A., Giordano, F., Parrella, M. L., and Restaino, M. (2017). Variable selection in high-dimensional regression: a nonparametric procedure for business failure prediction. *Applied Stochastic Models in Business and Industry*, 33(4):355–368.

Augugliaro, L., Mineo, A. M., and Wit, E. C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):471–498.

Barut, E., Fan, J., and Verhasselt, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bradburn, M. J., Clark, T. G., Love, S., and Altman, D. (2003). Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431.

Bradic, J., Fan, J., and Jiang, J. (2011). Regularization for cox's proportional hazards model with np-dimensionality. *Annals of statistics*, 39(6):3092.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Chang, J., Tang, C. Y., and Wu, Y. (2013a). Marginal empirical likelihood and sure independence feature screening. *Annals of statistics*, 41(4).

Chang, J., Tang, C. Y., and Wu, Y. (2013b). Supplement to "marginal empirical likelihood and sure independence feature screening.". *Ann Stat. https://doi. org/10.1214/13-AOS1139SUPP*.

Chang, J., Tang, C. Y., and Wu, Y. (2016a). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Annals of statistics*, 44(2):515.

Chang, J., Tang, C. Y., and Wu, Y. (2016b). Supplement to "local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood.". *Ann Stat. https://doi. org/10.1214/15-AOS1374SUPP*.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chu, Y. and Lin, L. (2018). Conditional sirs for nonparametric and semiparametric models by marginal empirical likelihood. *Statistical Papers*, pages 1–18.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Annals of Statistics*, 40(5):2667–2696.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641.

Dicker, L. and Lin, X. (2013). Parallelism, uniqueness, and large-sample asymptotics for the dantzig selector. *Canadian Journal of Statistics*, 41(1):23–35.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Fan, J. (2018). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J., Li, R., et al. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Fox, J. (2002). Cox proportional-hazards regression for survival data. *An R and S-PLUS companion to applied regression*, 2002.

Gai, Y., Zhu, L., and Lin, L. (2013). Model selection consistency of dantzig selector. *Statistica Sinica*, pages 615–634.

Gorst-Rasmussen, A. and Scheike, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):217–245.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

He, X., Wang, L., Hong, H. G., et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369.

Hong, H. G., Chen, X., Christiani, D. C., and Li, Y. (2018). Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics*, 74(2):421–429.

Hong, H. G. and Li, Y. (2017). Feature selection of ultrahigh-dimensional covariates with survival outcomes: a selective review. *Applied Mathematics-A Journal of Chinese Universities*, 32(4):379–396.

Hu, Q. and Lin, L. (2017). Conditional sure independence screening by conditional marginal empirical likelihood. *Annals of the Institute of Statistical Mathematics*, 69(1):63–96.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the cox model. *Annals of statistics*, 41(3):1142.

Huang, J., Wei, F., and Ma, S. (2012). Semiparametric regression pursuit. *Statistica Sinica*, 22(4):1403.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when p¿¿ n. *Statistica Sinica*, pages 595–611.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.

Kleinbaum, D. and Klein, M. (1996). Survival analysis: a self-learning text springer. *New York*.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63.

Li, G., Peng, H., Zhang, J., Zhu, L., et al. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

Li, J., Zheng, Q., Peng, L., and Huang, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154.

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Lian, H., Du, P., Li, Y., and Liang, H. (2014). Partially linear structure identification in generalized additive models with np-dimensionality. *Computational Statistics & Data Analysis*, 80:197–208.

Liang, H., Wang, H., and Tsai, C.-L. (2012). Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models. *Statistica Sinica*, pages 531–554.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Liu, Y., Zhang, J., and Zhao, X. (2018). A new nonparametric screening method for ultrahigh-dimensional survival data. *Computational Statistics & Data Analysis*, 119:74–85.

Mai, Q. and Zou, H. (2013). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234.

Mai, Q. and Zou, H. (2015). The fused kolmogorov filter: a nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, 24(2):165–179.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Owen, A. B. (2001). *Empirical likelihood*. Chapman and Hall/CRC.

Peng, L. and Fine, J. P. (2009). Competing risks quantile regression. *Journal of the American Statistical Association*, 104(488):1440–1453.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A (General)*, 135(2):185–198.

Radchenko, P. and James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, pages 1346–1370.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Song, R., Lu, W., Ma, S., and Jessie Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.

Storlie, C. B., Bondell, H. D., Reich, B. J., and Zhang, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica*, 21(2):679.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

Walters, S. J. (1999). *What is a Cox model?* Citeseer.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

Wit, E., Augugliaro, L., Abegaz, F., and Gonzalez, J. (2014). Dgcox: a differential geometric approach for high-dimensional cox proportional hazard models. In *Eleventh international Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB 2014*.

Yan, X., Tang, N., Xie, J., Ding, X., and Wang, Z. (2018). Fused mean–variance filter for feature screening. *Computational Statistics & Data Analysis*, 122:18–32.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zambom, A. Z., Akritas, M. G., et al. (2017). Nonpmodelcheck: An r package for nonparametric lack-of-fit testing and variable selection. *Journal of Statistical Software*, 77(10):1–28.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495):1099–1112.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411.

Zhou, M. (2018). Package 'emplik'.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.