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

ABSTRACT

Candidate:
Sara Milito

Supervisor:
Prof. Francesco Giordano

PhD Coordinator:
Prof. Alessandra Amendola

ACADEMIC YEAR 2018-2019

ABSTRACT

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. (2016), 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 (**?**). 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.

ABSTRACT in Italiano

Questa tesi propone un nuovo metodo per risolvere il problema della regressione non parametrica e non additiva in presenza di dati ad dimensionalità ultra-high. In questo contesto ci sono due aspetti rilevanti: la variable selection e la structure discovery, che riguardano, rispettivamente, l'identificazione delle variabili che influenzano la variabile di risposta e il tipo di effetto di tali variabili selezionate (lineare o non lineare).

In questa tesi proponiamo un metodo non parametrico di variable selection che lavora in due stadi. Al primo stadio, viene eseguita una procedura di screening: si seleziona un sottoinsieme di variabili che contiene le covariate vere con probabilità 1. Al secondo, la screening viene trasformata in variable selection utilizzando un approccio non penalizzato. In questo modo, sfruttando la semplicità della screening, riusciamo a superare il problema della stima di parametri di penalità. Inoltre, il nostro approccio di screening ha la potenzialità di poter distinguere le covariate lineari da quelle non lineari, ottenendo quindi anche la structure discovery.

Chang et al. (2016), senza imporre assunzioni sulla forma funzionale del modello sottostante, hanno proposto un metodo di screening che utilizza l'empirical likelihood e la stima dei polinomi locali. Una volta ottenuta la stima della funzione marginale tra una particolare variabile e la risposta, hanno utilizzato l'empirical likelihood per verificare se questa funzione sia significativamente diversa da zero. Nonostante gli eccellenti risultati raggiunti in termini di dimensionalità, gli autori non hanno costruito un approccio che faccia anche variable selection e structure discovery. Per risolvere tali limitazioni, in questa tesi proponiamo di complicare il loro approccio stimando la derivata prima marginale anziché la funzione marginale. In questo modo, abbiamo ottenuto un nuovo metodo di screening completamente non parametrico, chiamato *Derivative Empirical Likelihood Sure Independence Screening* (D-ELSIS). Per trasformare la nostra procedura di screening selection in un approccio di variable selection, abbiamo utilizzato la tecnica della subsample. In particolare, proponiamo di applicare l'idea della subsample non sui risultati di una procedura di variable selection, come in Meinshausen and Bühlmann (2010), ma dopo una procedura di screening. Con questo strumento, le variabili identificate tramite il D-ELSIS vengono quindi ulteriormente valutate per indagare la loro probabilità, in termini di frequenza relativa, di essere selezionate quando i dati vengono

campionati casualmente. Inoltre, sebbene in questo approccio vengano utilizzate threshold, non è necessario stimarle.

In sintesi, il potenziale dell'approccio proposto è triplice. Innanzitutto, otteniamo un metodo di screening totalmente non parametrico e che funziona nel contesto della regressione non parametrica e non additiva. In secondo luogo, trasformiamo la screening in variable selection senza stima dei parametri di penalità. Infine, stimando le derivate prime marginali, siamo in grado di distinguere l'effetto lineare o non lineare delle covariate selezionate sulla variabile di risposta.

In questa tesi verranno analizzate le proprietà teoriche del nostro nuovo approccio D-ELSIS come metodo di screening. In particolare, la proprietà di consistenza di tale procedura viene ottenuta con una dimensionalità esponenziale, ma paghiamo qualcosa in quanto stimiamo la derivata prima marginale. Per valutare le prestazioni della nostra proposta, verranno descritti i risultati di alcune simulazioni e dell'applicazione empirica su un set di dati reali. I risultati teorici saranno inoltre presentati per supportare la trasformazione da un metodo di screening ad un metodo di variable selection. Le proprietà teoriche di structure discovery della nostra procedura saranno oggetto di futuro lavoro di ricerca.

In aggiunta, estendiamo la nostra proposta all'analisi delle time-to-event variable. Grazie alla rapida crescita della tecnologia, sono sempre più disponibili dati ad alta dimensionalità. Negli ultimi anni, la tecnologia ha registrato una forte crescita anche in campo medico e genetico. La selezione delle variabili è fondamentale in survival analysis, in cui troviamo le variabili time-to-event, che rappresentano un diverso tipo di variabile perché l'oggetto di interesse non è solo il verificarsi o meno dell'evento di interesse, ma anche quando questo evento si è verificato. La maggior parte dei metodi in letteratura considera una stima condizionata della funzione di sopravvivenza, usando lo stimatore Kaplan e Meier (**?**). Poiché ciò non tiene conto dell'effetto diretto della covariata sulla probabilità di sopravvivenza, questo stimatore presenta alcuni svantaggi. Siamo riusciti ad evidenziare e giustificare la possibilità di applicare il metodo D-ELSIS anche in questo contesto. Con il nostro D-ELSIS, otteniamo una procedura di screening completamente non parametrica senza l'uso della stima di Kaplan e Meier della funzione di sopravvivenza. Questa rappresenta la differenza fondamentale tra il nostro metodo e le altre tecniche di screening. Inoltre, in base alle nostre conoscenze, nel contesto della survival analysis, non è mai stato utilizzato un metodo di screening che combini l'empirical likelihood e la regressione con i polinomi locali.

La tesi è divisa in due parti. Nella prima parte,verrà affrontato il problema della regressione con dati ad alta dimensionalità, la nostra proposta sarà esaminata da un punto di vista teorico, verranno presentati i risultati di alcune simulazioni e verrà presentato uno studio empirico. Nella seconda parte la nostra proposta sarà applicata nel contesto della survival analysis. Anche in questo caso verranno riportati i risultati dell'applicazione del nostro nuovo approccio su dati simulati.