UNIVERSITÀ DEGLI STUDI DI SALERNO

dottorato in

ECONOMIA E POLITICHE DEI MERCATI E DELLE IMPRESE

Curriculum

METODI STATISTICI

XXX ciclo

(A.A. 2016/2017)

tesi in

High-dimensional statistics for complex data

Tutor

Prof. Francesco GIORDANO

Autore

Massimo PACELLA

Coordinatore

Prof. Sergio Pietro DESTEFANIS

DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

Abstract

High dimensional data analysis has become a popular research topic in the recent years, due to the emergence of various new applications in several fields of sciences underscoring the need for analysing massive data sets.
One of the main challenge in analysing high dimensional data regards the interpretability of estimated models as well as the computational efficiency of procedures adopted. Such a purpose can be achieved through the identification of relevant variables that really affect the phenomenon of interest, so that effective models can be subsequently constructed and applied to solve practical problems. The first two chapters of the thesis are devoted in studying high dimensional statistics for variable selection. We firstly introduce a short but exhaustive review on the main developed techniques for the general problem of variable selection using nonparametric statistics. Lastly in chapter 3 we will present our proposal regarding a feature screening approach for non additive models developed by using of conditional information in the estimation procedure.
Differently, the second part of the thesis focuses on the spatio-temporal models in high dimensional contexts. Over the last decade, a particular class of spatio-temporal models has been rapidly developed, the spatial dynamic panel data models (SDPD). Several versions of the SDPD model have been proposed, based on different assumptions on the spatial parameters and different properties of the estimators. The standard version of the model assumes the spatial parameters constant over location, meanwhile another recently proposed version assumes the spatial parameters are adaptive over location. The assumption of different scalar coefficients is motivated by practical situations, in which empirical evidence shows how considering constant effect for each location can be limiting. While chapter 4 is devoted to introduce principal elements of spatio-temporal models in statistical and econometric frameworks, in chapter 5 we propose a strategy for testing the particular structure of SDPD model, by means of a multiple testing procedure that allows choosing between the version of the model with adaptive spatial parameters and some specific versions derived from the general one by imposing particular constraints on the

parameters. The multiple test is made in high dimensional setting by the Bonferroni technique and the distribution of the multiple test statistic is derived by a residual bootstrap resampling scheme.

Contents

List of Figures

List of Tables

Part I

Nonparametric variable selection in high-dimensions

Introduction

Nowadays high dimensional data analysis has become increasingly important, due to the emergence of various new applications in several fields of sciences, such as genomics, bioinformatics, economics and finance. Many of them underscore the need for analyzing massive data sets. For example, in genomics, using microarray data set, there could be hundreds or thousands genes potential predictors of a particular disease. Otherwise, in a portfolio allocation problem the amount of stocks to be included could be large enough to involve an intractable parametrization of the variance covariance matrix. Other situations involving high dimensional data sets are, for example, in high resolution image analysis, e-commerce and behavioural finance studies, among others. One of the main challenge in analyzing high dimensional data regards the interpretability of estimated models as well as the computational efficiency of procedures adopted. A fundamental objective of statistical analysis with high dimensional data is to identify relevant features, so that effective models can be subsequently constructed and applied to solve practical problems. Traditionally, there was two major types of variable selection methods. The first one is known as the best subset selection, which selects the best model among all possible combinations of the predictors based on some specific selection criterion. Example of well-known selection criteria include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). All of these criteria consider a trade-off between the goodness of fit of the model and its complexity. The second class of methods is known to include such procedures that employed the selection of a subset of predictors in a sequential order. One well-known example includes the forward, backward and stepwise selection. The forward selection starts from the model with no variables included, then adds variables sequentially according to the most significant if its p-values is below some predetermined level. Variables are added until none of the remaining variables are significant.

In contrast, the backward selection works at the opposite direction, thus it begins with the full model and least significant variables are excluded sequentially. Finally, the stepwise method is a combination of the two above, as it allows movement in either direction by adding or dropping variables at sequential steps.
More recently, a third class, closely related to the class of sub-sequential procedures, has been developed for linear and other parametric models. This class includes the well-known LASSO of Tibshirani (1996) and Adaptive LASSO (Zou (2006)), SCAD (Fan and Li 2001), Dantzig selector (Candes and Tao 2007), among others. These procedures are commonly defined *Regularization methods* because of the presence of one or more regularization parameters in estimation. They allow selection of variables and estimates of parameters simultaneously by solving high dimensional optimization problems. Most of them are based on the well-known l_1-penalization and for this reason are defined *LASSO-type* (or LASSO variants) estimators. For a complete review on parametric statistics in High dimensional see Bühlmann and Van De Geer (2011). Usually, in parametric regression the assumption of linearity could be very stringent in many practical analysis. It has been found from our knowledge that the traditional parametric linear model might not work well for detecting pattern when the underlying true relationships is nonlinear. Nevertheless, parametric techniques as the LASSO are limited in handling problems with a very large number of covariates and can fail in presence of highly correlated structure of covariates. For such reasons extension to nonparametric seems natural.
This part of the thesis is devoted to the context of nonparametric statistics for variable selection, then first of all we will introduce a short but exhaustive review on the main techniques developed for the general problem of variable selection in high dimensions (or ultra-high dimension) using nonparametric statistics, before presenting our research proposal. Particularly, in chapter 1 we will focus on the major techniques developed for variable selection in contexts of additive models. Most of the discussed methods adopt some type of regularization to reach the general aim of variable selection, other are developed in a different way and require, for instance, sequential estimation procedure. We will concentrate on such methods developed for regression problems. Then, in chapter 2 we will show the main results and technical details of variable

selection methods for non additive models. Non additivity assumption is the primary focus in our research purpose, thus we have regarded a full chapter in order to present as briefly as possible the strengths and weaknesses of the main techniques that allow for non additivity. Finally, in chapter 3 we will present our proposal regarding a feature screening approach for non additive models developed by using of conditional information in the estimation procedure.

Chapter 1

Nonparametric variable selection for additive models

1.1 Introduction

High dimensional data analysis has become a popular research topic in the recent years. However, the nature of high dimensional data makes many traditional statistical methods fail. Consequently, variable selection or dimensionality reduction is often the first step in high dimensional data analysis. In this chapter we examine the main nonparametric approaches of features selection adopted in contexts of high dimensionality regression.
The chapter is organized as follows. First, we define more precisely the context of interest, we give definitions of high dimensional and nonparametric settings, and in addition, we briefly explain the meaning of sparsity. In section 1.3 we present the main selection techniques developed for additive models and involving penalization in the procedure. Section 1.5 is devoted to the *feature screening* methods firstly introduced by Fan and Lv (2008). The chapter ends with a brief review of some technical results from such methods.

1.1.1 High-dimensional setting

From a statistical perspective, when the number of parameters p is greater than the number of observations n, regression problems cannot be solved by classical estimation procedures like the method of ordinary least squares. The standard procedures rely on the assumption that $\mathbf{X}'\mathbf{X}$ is nonsingular, otherwise $\mathbf{X}'\mathbf{X}$ cannot be inverted and the solution of the optimization problem is not

unique. This, obviously, occurs when $p > n$, as the covariate matrix does not have full column rank and this highly influences the estimation problem. Thus, to perform regressions when $p > n$ (or $p \gg n$), some kind of preselection or regularization is needed. Typically, we define the situation where the number of parameters increases with the number of observations as a problem of *High dimensions*, while we refer to *Ultra-high dimensions* when p increases and is exponentially in n.

Sparsity assumption

It is well-known that in many practical situations it is generally believed that only a small number of features are related to the phenomenon of interest. Such a point is primary justification to direct the analysis towards the aim of dimensionality reduction. Moreover, dimensionality reduction allows both the goals of constructing well interpretable models as well as to gain insight into relationship between predictive variables and response variable for scientific purposes. Most of the known methods developed to address the problem of high dimensional regression via the dimensionality reduction, usually require an assumption of sparseness. Given the general model (1.1), according to the definition of *sparsity* only a few covariates are in the true model. For instance, we regard that only q are relevant in explaining response variable, and $q \ll p$. Sparsity is an important theoretical aspect to reduce the complexity and the number of effective variables in the model. The intention of producing more interpretable models is especially productive in the High dimensional context. It is obviously easier and more convenient to interpret results from an estimate which involves some preselection or regularization rather than a result involving hundreds or thousands of covariates.

1.1.2 Nonparametric framework

Let $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ be a set of $\mathbb{R}^{p+1}$ random vectors from the model

$$Y = m(\mathbf{X}) + \varepsilon. \tag{1.1}$$

where Y represents the depended variable (or response variable) and $\mathbf{X}$ is the $\mathbb{R}^p$-valued covariates matrix. The function $m : \mathbb{R}^p \to \mathbb{R}$ is the regression function, commonly defined as multivariate conditional mean $\mathbb{E}(Y|\mathbf{X})$. The term ε represent the usual additive noise.
There nowadays exist many methods for obtaining nonparametric regression estimate of m. In nonparametric statistics there exists universally consistent estimates, but it is impossible to obtain a non trivial rate of convergence for all distributions of $(\mathbf{X}, Y)$. In other words, the estimator can converge very slowly. In order to get non trivial rates of convergence, one has to restrict the class of distributions, defining classes of such distributions where the corresponding regression function satisfies some smoothness conditions (e.g. m is d continuously differentiable or m is Lipschitz continuous). For classes of functions where m is d times continuously differentiable, the optimal rate of convergence is $n^{-2d/(2d+p)}$ (Györfi et al. (2006)). Thus, nonparametric regressions require a sample size n exponential in p in order to approximate m.

1.2 Some definitions

Additive and non additive models

Model (1.1) is a pure *non additive* model as regression function is a joint function of all p regressors. Instead, when m is expressed as linear combination of marginal unknown functions

$$m(\mathbf{X}_i) = \sum_{j=1}^{p} m_j(X_{ij}). \tag{1.2}$$

the model becomes *additive*. Assume linear combination of univariate functions is less general than joint multivariate regression function in (1.1), but it is actually very simple to estimate.
Differently, m can be expressed as function of linear combination of marginal regression functions,

$$m(\mathbf{X}_i) = g\left(\sum_{j=1}^{p} m_j(X_{ij})\right), \tag{1.3}$$

according to the function $g(\cdot)$, the corresponding model can be additive or non additive. For instance, if $g(\cdot) = \exp$ we have a nondditive model with $m(X) = \exp\{\mathbf{X}\beta\}$. Such a class of models are defined *Generalized additive models* (Hastie and Tibshirani (1990)).
The distinction between additive and non additive models is useful for our purpose since it directly affects performance and theoretical aspects of selection procedures. From a theoretical point of view, assuming regression functions to be a linear combination of marginal functions makes consistency results in variable selection as simple to prove as linear models, even though the marginal functions are nonlinear. This is because any estimation involved into selection procedure can be performed with the same difficulty of a single-variable model, despite the dimension p is growing with n. Conversely when the regression function is expressed as fully joint function of p regressors there is no information about structure of regression in order to simplify the estimation required for variable selection, thus asymptotics for increasing p becomes more difficult to assessed.

Oracle property

Following the definition of Fan and Li (2001), as well as for the parametric case, we say that a nonparametric estimator has the oracle property if it estimates the regression function at the optimal nonparametric rate and, at the same time, also selects the nonzero components with probability tending to one.

Sure screening property

In the context of screening it is usually said that a procedure holds the *sure screening property* that is similar to the *oracle property* of variable selection techniques. Formally, it means that the probability of a set of screened variables that contains the true set of relevant variables, converges to one as n goes to infinity.

1.3 Variable selection in additive models

Additivity assumption is a primary way of relaxing the linear assumption in order to account for nonlinearity in the regression. It is able to retain the

interpretable additive form of linear regression models, even though it allows for nonlinear marginal regression functions. More importantly, the additive models are able to circumvent the problem of *curse of dimensionality* arising in high dimensional regression problems, since they can be estimated at the same optimal rate of convergence for univariate functions (Stone (1985)).
For such reasons, there have been a large part of literature involved in variable selection for nonparametric additive models. Many authors treated this purpose by spline estimators, as such a class of estimators is easy to analyse and to adapt for variable selection. In the following we give a brief review of the most common techniques in such a field.

1.3.1 Penalized spline estimators

A traditional smoothing splines estimator could be defined as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda\sum_{j=1}^{p}\theta_j^{-1}\|P^{(j)}m\|^2, \tag{1.4}$$

where $P^{(j)}m$ denotes the orthogonal projection of m onto the j-th orthogonal subspace of a reproducing kernel Hilbert space (RKHS). Lin and Zhang (2006) proposed a penalized variant of smoothing splines estimator in (1.4).

Instead of the squared functional norm in the penalty, Lin and Zhang (2006) proposed the following estimator (COSSO)

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda\sum_{j=1}^{p}\|P^{(j)}m\|, \tag{1.5}$$

the penalty term $\sum_{j=1}^{p}\|P^{(j)}m\|_{\mathcal{M}}$ involves the sum of the norms of order one of the function components, thus it allows for shrinkage similar to the traditional LASSO. The parameter λ is as well a tuning parameter that control the amount of shrinkage.
In a similar way, the penalization term in (1.5) can be constructed by L_2-norm (Ravikumar et al. (2009)). In such a case the penalized spline estimator can be

expressed as minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{y_i - m(x_i)\Big\}^2 + \lambda\sum_{j=1}^{p}\|P^{(j)}m\|_2. \tag{1.6}$$

From a computational point of view, it is not easy to estimate model in (1.5) ensuring sparsity, since it is not equivalent to the standard smoothing spline in (1.4). In order to perform a proper estimates as solution of a smoothing spline problem, a different formulation is needed

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda_0\sum_{j=1}^{p}\theta_j^{-1}\|P^{(j)}m\|^2 + \lambda\sum_{j=1}^{p}\theta_j. \tag{1.7}$$

The (1.7) is minimized subject $\theta_j > 0$, the constant λ_0 can be take as any fixed positive value, while parameter λ needs to be tuned.
When the penalization is as in (1.6), each marginal regression function is approximated by orthonormal basis with respect to $L_2[0,1]$.
Differently from (1.5), the sparsity is taken into account by imposing penalty on the L_2 norm of the spline approximation to ensure identifiability. Thus the estimation reduces to a standard optimisation problem as for spline estimator. If we rewrite (1.6) in terms of basis function, we have

$$\frac{1}{2n}\left\|Y - \sum_{j=1}^{p}\Phi_j\beta_j\right\|_2^2 + \frac{\lambda}{\sqrt{n}}\sum_{j=1}^{p}\sqrt{\left(\beta_j'\Phi_j'\Phi_j\beta_j\right)}, \tag{1.8}$$

where Φ_j is the $n \times d$ matrix of the orthonormal basis. Lin and Zhang (2006) gave an accurate analysis in the special case of a tensor product design with an SS-ANOVA model built from the second order Sobolev spaces of periodic functions.
They assumed the space $\mathcal{M}$ of functions m to be constructed by the tensor product of p orthogonal subspaces where each subspace is itself a RKHS

$$\mathcal{M} = \bigotimes_{j=1}^{p}\mathcal{H}_j = \{1\} \oplus \Big\{\sum_{j=1}^{p}\bar{\mathcal{H}}_j\Big\} \oplus \Big\{\sum_{j=1}^{p}(\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k)\Big\} \oplus ... \tag{1.9}$$

where each $\mathcal{H}_j = 1 \oplus \bar{\mathcal{H}}_j$, and $\bar{\mathcal{H}}_j$ is the RKHS of functions m_j such that $\langle m_j, 1\rangle =$

0. The right side of (1.9) shows the decomposition into the space of constant functions, the main effects spaces $\bar{\mathcal{H}}_j$, the two-way interaction spaces $\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k$ and so on and usually it is cut off to an order d. Instead, the second-order Sobolev space of periodic functions can be expressed as

$$\begin{aligned}\bar{\mathcal{H}}_j =& \{m_j : \ m_j(x_j) = \sum_{k=1}^{\infty} a_k \sqrt{2} \cos 2\pi k x_j + \sum_{k=1}^{\infty} b_k \sqrt{2} \sin 2\pi k x_j, \\ & \sum_{k=1}^{\infty} (a_k^2 + b_k^2)(2\pi k)^4 < \infty\}.\end{aligned} \tag{1.10}$$

When the regression function is estimated by splines approximation, each nonparametric component is expressed as linear combination of a specified splines basis function, as in (1.10). Generally, it can be assumed that each regression function belongs to

$$\bar{\mathcal{H}}_j = \{m_j : \ m_j(x_j) = \sum_{k=0}^{\infty} \beta_{jk} \eta_{jk}(x_j), \ \sum_{k=0}^{\infty} \beta_{jk}^2 k^4 \leq \infty\}, \tag{1.11}$$

where $\{\eta_{jk}, \ k = 0, 1, ...\}$ defines a uniformly bounded orthonormal basis function.

Different from Lin and Zhang (2006), estimator of Ravikumar et al. (2009) allows for the problem of component selection adopting penalization inside the L_2 norm, thus it does not make use of Hilbert norm in order to obtain sparsity among projections. In other words, the coefficients in the polynomial approximation of splines are shrunk, the idea of grouped LASSO of Yuan and Lin (2006) is borrowed to take into account the grouping structure in the approximation, thus the work of Ravikumar et al. (2009) can be thought as a functional version of the grouped LASSO.

Adaptive variants

To achieve optimality in both estimation and selection, Storlie et al. (2011) proposed a variant of the COSSO characterized by a two-step estimation procedure where each individual norm in the penalty of equation (1.5) is adaptively

weighted. Thus, the model at the second-stage becomes

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda\sum_{j=1}^{p} w_j \|P^{(j)}m\|, \tag{1.12}$$

each weight is constructed as $w_j = \|P^{(j)}\tilde{m}\|_2^{-\gamma}$, where $\tilde{m}$ is the first-stage estimate.
The proposal in Storlie et al. (2011) shares all features of the adaptive LASSO (Zou (2006)). Naturally, it does not allow for overdetermined scenario with $p > n$, where one cannot be able to do first stage estimates, unless using one estimator that also allows for dimensionality reduction. In a similar way, Huang, Horowitz, and Wei (2010) proposed an adaptive version of penalized smoothing spline of Ravikumar et al. (2009).

Spline with interaction terms

Despite the models in Lin and Zhang (2006) and Storlie et al. (2011) assume the covariates space with interaction effects, they do not perform model selection in the sense of considering interactions among the selectable features. In fact interaction terms are provided in the approximation by the splines estimator, but the whole estimation procedure is not able to distinguish between main effects and interaction effects. A penalty function that account for interaction terms was proposed by Radchenko and James (2010). It can be expressed as

$$\lambda\left(\sum_{j=1}^{p}\|P^{(j)}m\|_2 + \sum_{j=1}^{p}\sum_{k=j+1}^{p}\|P^{(j,k)}m\|_2\right) \tag{1.13}$$

where $P^{(j,k)}m$ denotes the orthogonal projection of m onto the (j,k)-th two dimensional subspace of RKHS. In (1.13) the interactions are not only considered in the spline approximation, rather they act as covariates and allow in recognising covariates that jointly contribute to the response.
Nevertheless, single effects and interaction terms are treated similarly, in fact an entry of an interaction generally adds more predictors and it is difficult to interpret as well as to compute due to p^2 different terms. For that reason the

penalty term can be reformulated as

$$\lambda_1 \sum_{j=1}^{p} \left(\|P^{(j)}m\|^2 + \sum_{k:k\neq j}^{p} \|P^{(j,k)}m\|^2 \right)^{1/2} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} \|P^{(j,k)}m\| \tag{1.14}$$

where the L_2 and L_1 norms are considered separately for a better interpretation. The main problem of such a variant is that when the covariates dimension is high, the estimation still suffers from introduction of interactions. Just think that, for one-to-one interactions without consider interactions between the same covariate, the procedure involves $p(p-1)$ variables. When p is very large, i.g. in ultra-high dimensional contexts, the procedure becomes computationally hard.

1.4 Variable selection in nonadditive models

When we adopt nonparametric regression it is obviously to recall how the bias-variance trade-off affects the estimation procedure. Generally speaking, since linear regression model is more restrictive, it guarantees a lower variance of estimates with respect to the nonparametric model (1.1). Extending linear model to nonparametric one is in general a real problem, since nonparametric methods usually suffer from variance that grows exponentially with the dimension p. On the other hand, linear regression results in variance that increases linearly with p, but its bias decreases quickly. This is why extension to nonparametric by assuming additivity is someway straightforward, in fact it is able to mitigate between strength of fully nonparametric model and easiness of parametric one. That said, without any restriction, estimation becomes complicate, nevertheless when the main purpose is variable selection, *curse of dimensionality* involved in estimation could be treated in a different way. For instance, in *Feature screening* techniques selection requires marginal estimation, in such a case flexibility and generality of nonparametric model can be retained without involving very complicated and hard estimation procedure.

1.5 Feature Screening

We know that, when the number of explanatory variables grows with the number of observations and its order is exponential, we are facing up to *ultra-high dimensional* estimation problem. To be more specific, in such a situations standard variable selection procedures are not unaffected due to the so called *curse of dimensionality*. Motivated by this concerns, it has recently developed a new research field, called *Feature screening*. It is a class of computational approaches useful in preliminary analysis for preprocessing data to reduce the scale of dimension to a less order.

1.5.1 Sure Independence Screening (SIS)

Originally, Fan and Lv (2008) introduced a simple screening method (*Sure Independence Screening* or SIS). It relies on ranking estimations by measuring the marginal contributions of explanatory variables in explaining the response.
Let the linear model

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i \ , \tag{1.15}$$

where X_{ij} is the i-th row of the regressors matrix $\mathbf{X}$ corresponding to the j-th covariate.
SIS uses componentwise regression to rank the importance of features according to their marginal correlation with the response variable.
Marginal correlation of j-th predictor with the response variable can be computed as

$$\omega_j = \sum_{i=1}^{n} X_{ij}Y_i \ , \quad j = 1, ..., p. \tag{1.16}$$

The p marginal correlations obtained by (1.16) are sorted in a decreasing order and, for a given $\gamma_n \in (0, 1)$, the sub-model is defined by

$$\hat{M}_{\gamma_n} = \{1 \leq j \leq p : |\omega_j| \text{ is among the first } [\gamma_n n] \text{ largest of all}\}, \tag{1.17}$$

where $[\gamma_n n]$ denotes the integer part of $\gamma_n n$. By doing that, one is able to filter out the features that have weak correlation with the response. Broadly speaking, this method shrinks the full model down to a sub-model of less order by

sorting the p marginal magnitudes in a decreasing order and then discarding the components up to a fixed threshold.
As linear model with more than $p > n$ parameters are not identifiable, SIS appears a suitable methods primarily useful when p is exponential in n. Although it is proposed to reduce dimensionality to be below the sample size, it can be applied for reduction to a dimension $\tilde{p} > n$. It is obvious that larger $\tilde{p}$ assures larger probability of including the true model in the final sub-model $\hat{M}$.
Once the screening is carried out and the dimension is reduced, one of the well-developed variable selection technique is adopted to estimate the vector β in equation (1.15). For instance, LASSO type procedure or other variable selection techniques (i.e. SCAD in Fan and Li (2001), Adaptive LASSO by Zou (2006) and Dantzig selector in Candes and Tao (2007)) can be used.

1.5.2 Marginal spline estimator

Extension of SIS to generalized linear models was developed in Fan and Song (2010). The conditional expectation in the regression is assumed from an exponential family. SIS procedure for both linear models and GLM focuses on studying marginal contributions when the problem is strictly linear or at most generalized linear. They can address some methodological challenges regarding the missed joint information by adopting multistage or iterative versions, nonetheless, they can be crude in reducing ultra-high dimensions even if the linear model holds jointly. We could have that, when the joint distribution is normal, the marginal contribution can be highly nonlinear. Such a situation can be overcome throughout extensions to nonparametric statistics.
One way of extending ranking measures for screening to nonparametric framework is to adopt additive nonparametric regression function in (1.2) and apply some nonparametric estimator marginally.
Fan, Feng, and Song (2011) propose to estimate marginal nonparametric regressions by spline estimators. Let S_n be a polynomial space of degree $l \geq 1$, and let $\{\eta_{jk},\, k = 1, ..., d_n\}$ be a normalised B-spline basis with $\left\|\eta_{jk}\right\|_\infty \leq 1$. Under some proper conditions, each marginal regression function can be approximated as

$$m_{nj}(x) = \sum_{k=1}^{d_n} \beta_{jk}\eta_{jk}(x),\ 1 \leq j \leq p \quad \text{for some coefficients } \beta_{jk}.$$

The resulting marginal estimator can be expressed by

$$\hat{m}_{nj} = \min_{m_{nj} \in S_n} \mathbb{E}_n (Y - m_{nj}(X_j))^2$$

and is equivalent to minimize $\min_{\beta_j \in \mathbb{R}^{d_n}} \mathbb{E}_n (Y - \eta_j' \beta_j)^2$.
Then, as for standard SIS, the set of selected variables can be obtained as

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

for a predefined threshold value γ_n.

1.6 Some technical results

Technically, a marginal screening procedure requires that (i) if X_j contributes in explaining the response variable, then the marginal measure used for ranking takes non-negligible value; (ii) if X_j is not a relevant covariate, then the marginal measure takes negligible value. Such a condition, named *identification condition* (IC), ensures *sure screening property* of the method.
In the papers of Fan and Lv (2008) and Chang, Tang, and Wu (2013) for linear model, IC is viewed as a requirement for a minimal signal strength, thus it is ensured by the following inequality

$$\min_{j \in M} |\mathbb{E}(X_j Y)| \geq c n^{-k}, \quad \text{for } 0 \leq k < 1/2,$$

assuming that Y has finite variance. Similarly, in Fan, Feng, and Song (2011) it is ensured by

$$\min_{j \in M} \mathbb{E}\left[m_j(X_j)^2\right] \geq d n^{-2k}, \quad \text{for } 0 < k < r/(2r+1), \tag{1.18}$$

where d is the number of basis in the truncation of spline approximation and r is a non-negative constant related to the assumption about continuity of m_j. Such a type of assumptions are needed in order to guarantee a separation between the set of relevant variables and the set of irrelevant ones. When the separation is sufficiently large the two sets of variables can be easily identified.

As concern the high dimensions, it is well understood that in parametric linear regression when the design matrix satisfies some kind of irrepresentable condition, consistent estimation of the true relevant variables set, also called the sparsity pattern, is possible under the condition $q\log(p/q) = o(n)$ as $n \to \infty$, where q is the cardinality of the set of relevant covariates and p the number of all the covariates. Furthermore, if the quantity $(q\log(p/q))/n$ does not converge to zero when $n \to \infty$, but for instance is fixed or it remains bounded, then it is impossible to consistently estimate the sparsity pattern (see Bühlmann and Van De Geer (2011)).
Theoretical properties of Penalized Spline estimators of Lin and Zhang (2006) are assessed only for fixed p and q. Thus, the estimator properties are not provided for increasing dimensions, but its consistency in estimation of the sparsity pattern is guaranteed for fixed scenarios. Similar in Ravikumar et al. (2009), apart from some kind of incoherence condition on the design matrix in term of orthogonal basis, consistency in sparsity estimation is guaranteed also when $\log p = o(n)$, providing the number q of relevant variables remain at least bounded and the tuning and truncation parameters increase according to some specified orders (see Theorem 2 of Ravikumar et al. (2009) for details).
Otherwise, results on variable selection consistency in Lin and Zhang (2006) require detailed investigation on the eigen-properties of the RKHS, which in general is not straightforward. However, assumptions of m belonging to the class of periodic function and tensor product design make the derivation more tractable. Assuming observations from tensor product design means that the design points are of the form $x_{jk} = j/n_k$, $j = 1, ..., n_k$; $k = 1, ..., d$, for n_k such that the sample size is $n = n_k^p$. In such a case the estimator is shown to converge at the optimal rate $n^{-d/(2d+1)}$ (where d is the order truncation in spline approximation) if $\lambda = O(n^{-2d/(2d+1)})$. Thus for $d = 2$, λ should be of order $O(n^{-4/5})$ to ensure proper selection.
Despite Lin and Zhang (2006) show that penalized spline for additive model leads to consistent estimation of the true predictors and correct selection with probability tending to one, they do not provide accurate demonstration of *oracle property*. In fact, it is not clear how the order of tuning parameter λ should be, taking in consideration results of Theorem 2, in order to ensure the optimal rate estimation and correct selection, simultaneously.

Meanwhile, Storlie et al. (2011) show that, if both λ's in the first and second step estimation are of the same order of $n^{-4/5}$, the adaptive variant is *oracle*. In particular, for $\tilde{m}$ being the traditional smoothing spline in equation (1.4) with $\lambda_0 \sim n^{-4/5}$, suppose that $w_j^{-1} = O_p(1)$, for all $j = 1, ..., p$, and further that the weights of the irrelevant variables do not vanish, i.e. $w_j = O_p(1)$, $\gamma \geq 3/4$ and the second-stage tuning parameter $\lambda \sim O_p(n^{-4/5})$, then procedure in Storlie et al. (2011) is consistent at rate $O_p(n^{-2/5})$ and correctly selects the set of relevant variables with probability tending to one.
Another point to deal with is the choice of tuning parameters. Both Lin and Zhang (2006) and Ravikumar et al. (2009) proposed to choose tuning parameters by estimated risk through generalized cross-validation (GCV) or a criterion C_p based on the concept of generalized degree of freedom. Given a smoothing matrix A of the smoothing spline, GCV can be defined as

$$GCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_j \hat{m}_j(X_{ij})\}^2}{\{tr(I - A)/n\}^2}$$

while, the estimated risk C_p is given by

$$C_p = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_j \hat{m}_j(X_{ij})\}^2 + \frac{2\hat{\sigma}^2}{n}\sum_j tr(A)I(\|\hat{m}\| \neq 0).$$

In general, in both procedures tuning parameters are estimating through validation and it can lead to overfitting.

1.7 Conclusion

All the works presented in the chapter treat several techniques for variable selection and feature screening making use of nonparametric estimators for additive models. Some of them have been developed for linear models (see for instance feature screening methods in Fan and Lv (2008) and Chang, Tang, and Wu (2013)), others are suitable for treating nonlinearity through the less restrictive assumption of nonlinear additivity (Lin and Zhang (2006) and Fan, Feng, and Song (2011) among others). Estimation in additive models is actually very simple since it results in the same optimal rate of convergence for

univariate case. Most of the literature for additive model has been focused on smoothing splines estimator as primarily adaptation to Nonparametric since it can well fulfil the purpose of such a context. Otherwise, extension of splines to non additive models is not nearly as obvious, instead Kernel smoothing theory provides good tools for non additive regression models. This is why most of the proposed methods for non additive variable selection have been developed by using Kernel regression. In the next chapter we will present some of the most popular procedures belonging to this particular class.

Chapter 2

Nonadditive variable selection methods

Non additivity assumption is quite different and more general with respect to the additive one, in fact it does not require any kind of definition of the regression structure, but it affects the estimator properties as well as the maximum allowed number of covariates. Most of the main techniques for non additive models are developed in the framework of local polynomial estimators (LPE). Compared with the splines, LPE are however easier to program and to analyse mathematically, meanwhile advantages of spline are their computational speed and simplicity. Most of the bias-variance analysis for Kernel regression can be done with basic calculus, instead the corresponding analysis for splines requires working with Hilbert spaces, that are infinite-dimensional functional spaces. The following chapter is organised as follows. In the first part we will introduce some of the main techniques developed for non additive variable selection that make use of LPE, the second part will be devoted to feature screening methods that allows for non additivity and at the end we will discuss about the main problems encountered in this specific context.

2.1 Local polynomial estimators (LPE)

Local Polynomial Estimator is a very useful and widespread nonparametric tool whose properties have been deeply studied (see for instance Ruppert and Wand (1994)). It corresponds to a locally weighted least squares fit of a polynomial

function of order k. Thus estimate of regression function m can be expressed as

$$\hat{m}(x) = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - B_{\beta}^{(k)} \right\}^2 K_H(X_i - x) \tag{2.1}$$

where $B_{\beta}^{(k)}$ is the k-order polynomial of the form $\beta_0(x) + \beta_1'(x)(X_i - x) + ... + \beta_k'(x)(X_i - x)^k$, the weights are given by function $K_H(x) = |H^{-1}|\mathcal{K}(H^{-1}x)$, where $\mathcal{K}(x)$ represents a p-variate Kernel function and H is a $p \times p$ matrix of smoothing parameters. Usually smoothing parameters are also called *bandwidths* and their calibration is crucial in the estimation procedure.
For sake of simplicity we do not report more details on the matrix representation of LPE and minimization solution of (2.1) (for more details we refer to Györfi et al. (2006) among others). We only argue that when $k = 1$, the approximation reduces to a linear function, thus Y is assumed locally linear. Instead, when $k = 0$ we have a constant local approximation of response and the estimator reduces to the so-called Nadaraya-Watson estimator in (2.11).
Over the last years, there have been many development on LPE for variable selection, following we briefly introduce some of them.

2.1.1 A greedy method: RODEO

Lafferty and Wasserman (2008) proposed an innovative procedure named RODEO. Instead of defining the whole fitting into a global convex optimization problem, as in the lasso-type estimation, the greedy methods adopt iterative algorithms locally. During each iteration, only a small number of variables are actually involved in the model fitting so that the whole estimation only involves low dimensional models. This is why they naturally arise suitable for high dimensional regression problems.
The method proposed in Lafferty and Wasserman (2008) performs feature selection and estimation using LPE, thus without assuming any particular model structure and allowing for non additivity. The main idea underlying their approach can be summed as follows.
Let x be a fixed point and $m_h(x)$ be a local estimator of the function m based on a p-dimensional vector of smoothing parameters h. Let $M(h) = \mathbb{E}(m_h(x))$ denote its expected value and assume that $m_0(x) = Y$, this leads to $m(x) = M(0) =$

$E(Y)$. The authors assume h to be in a smooth path, say $P_h = \{h(t), 0 < t < 1\}$ with $h(0) = 0$ and $h(1) = 1$. In such a formulation, the regression function can be express as

$$\begin{aligned} m(x) &= \mathbb{E}(m_1(x)) - \int_0^1 \frac{d\, M(h(s))}{d\, s}\, ds \\ &= M(1) - \int_0^1 < Z(h(s)), h^{'}(s) > \, ds, \end{aligned} \tag{2.2}$$

where $Z(h)$ is the gradient of $M(h)$, $h^{'}(s)$ the derivative of $h(s)$ along the path and the first equality is obtained by some algebra.

The key factor in the RODEO is that if the true regression function is sparse in the sense that only some covariates are relevant in explaining the response, there should be paths of h for which also $Z(h)$ is sparse. Thus one can perform variable selection taking advantage of sparseness into derivatives estimation. The functional $Z(h)$ can be estimated through LPE and the sparsity is then reached by some threshold rule.

The idea of RODEO effectively helps in discriminating which covariates are locally relevant, that is if x is irrelevant one should expect that a change in the bandwidth causes only a small change in the estimation of functional $Z(h)$, while the opposite situation should happen when the covariate is locally relevant. Such a method is similar to that of Ruppert (1997) where bandwidth selection is performed by a greedy approach using nondecreasing sequence of bandwidths and the optimal h is estimated by minimizing the mean square error. Compared with Ruppert (1997) RODEO takes into account sparseness, thus it implicitly performs variable selection in addition to the bandwidth selection. Practically, it is done by replacing the continuum of bandwidths in P_h by a discrete set $B_h = \{h_0, \beta h_0, \beta^2 h_0, \dots\}$, for some $0 < \beta < 1$. Thus, the smoothing parameters are firstly inflated, then the estimation of $Z(h)$ is done sequentially for $h \in B_h$ setting $Z_j(h) = 0$ when $h_j < \hat{h}$, with $\hat{h}$ the first h such that $|\hat{Z}| < \lambda$. This threshold implementation takes into account sparsity in the derivative estimation. Such a sequential procedure results in shrunk bandwidths for relevant variables while those corresponding to irrelevant variables are left relatively large.

2.1.2 Penalized LPE

Bertin and Lecué (2008) suggested a penalized version of LPE. Their proposal mixes the idea of l_1-penalization with local polynomial estimation. The estimator can be expressed as

$$\hat{m}_1(x) = \underset{\theta \in \mathbb{R}^{p+1}}{\arg\min} \left[\frac{1}{nh^p} \sum_{i=1}^{n} \left(Y_i - m\left(\frac{X_i - x}{h} \right) \theta \right)^2 K\left(\frac{X_i - x}{h} \right) + 2\lambda \left\| \theta \right\|_1 \right]. \tag{2.3}$$

Any minimizer of the above equation is a L_1 penalized version of the classical LPE. Assuming that there exist m_{max} such that $|m(x)| < m_{max}$, they also consider another form of estimator

$$\hat{m}_2(x) = \underset{\theta \in \mathbb{R}^{p+1}}{\arg\min} \left[\frac{1}{nh^p} \sum_{i=1}^{n} \left(Y_i + m_{\max} + Ch - m\left(\frac{X_i - x}{h} \right) \theta \right)^2 K\left(\frac{X_i - x}{h} \right) + 2\lambda \left\| \theta \right\|_1 \right]. \tag{2.4}$$

where constant C is as in Assumption 6 of Bertin and Lecué (2008).
Through (2.3) and (2.4), minimizing a localised version of the penalized L_2-risk by LPE, one should detect the set of relevant variables by the corresponding local approximations. More precisely, this method provides LASSO-type penalization to the local polynomial approximation, in such a way the method performs a LASSO selection locally. Once the set of relevant variable has been selected, a standard LPE is constructed on this set to estimate the true regression function.
Comparing with Lafferty and Wasserman (2008), weaker assumptions on the regression function is required in Bertin and Lecué (2008), merely assuming m to belong to the Holder class with smoother order strictly greater than one, while in the RODEO the regression function is required continuously differentiable and its derivatives to be bounded.

2.2 Feature screening for non additive models

In the first work on feature screening by Fan and Lv (2008), marginal magnitude of each covariate was measured by marginal linear correlation. Although

correlation measures a linear relationship, there have been some extensions to nonparametric context that also adopt marginal correlation functions in order to rank the variables in nonlinear models. Usually, such a measures are defined *marginal functional correlations*.

2.2.1 Marginal functional correlation measures

Zhu et al. (2011) defined a screening procedure for a general model framework where the conditional distribution function of Y given $\mathbf{X}$ satisfies

$$F(Y|\mathbf{X}) = F_0(Y|\mathbf{X}_{\mathcal{M}}\beta) \tag{2.5}$$

where $\mathbf{X}_{\mathcal{M}}$ is the matrix composed by relevant regressors and $F_0(\cdot|\mathbf{X}_{\mathcal{M}}\beta)$ an unknown distribution function. Such a framework includes models with regression function of the form (1.3) where Y depends on the relevant predictors through some linear combinations given by $q \times d$ matrix β. Thus, depending on the shape of the inverse function of $F_0(\cdot|\mathbf{X}_{\mathcal{M}}\beta)$, the corresponding model can be of non additive type.
Given the set of relevant predictors

$$\mathcal{A} = \{k \ : \ F(Y|\mathbf{X}) \text{ functionally depends on } X_k \text{ for some } Y \in \mathcal{Y}\}, \tag{2.6}$$

marginal magnitude for screening is measured by a standard functional correlation between the conditional distribution of Y given $\mathbf{X}$ and each predictor X_k. Functional correlation in Zhu et al. (2011) is expressed as

$$\Omega_k = \mathbb{E}(X_k F(Y|\mathbf{X}))$$

and the marginal utility measured by $\omega_k = \mathbb{E}(\Omega_k^2)$. Such a correlation is able to describe the relation between response and predictors without assuming any model structure.
An estimates of Ω_k is given by

$$\hat{\Omega}_k(y) = \frac{1}{n}\sum_{i=1}^{n} X_{ik} I(Y_i < y) \tag{2.7}$$

Consequently, an estimates of ω_k is can be obtained as $\hat{\omega}_k = 1/n \sum_{j=1}^{n} \hat{\Omega}_k(y_j)$. Lin, Sun, and Zhu (2013) modified such an idea in order to project the marginal measure of correlation into the local information flows "$X_k < x_k$" (for $a_k < x_k < b_k$). Their estimator is computationally more expensive since it includes sample counterpart of distribution of X_k, it is given by

$$\hat{\Omega}_k(y) = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_{ik} I(X_{ik} < x_{ik}) I(Y_i < y) - \frac{1}{n} \sum_{i=1}^{n} X_{ik} I(X_{ik} < x_{ik}) \frac{1}{n} \sum_{i=1}^{n} I(Y_i < y) \right\} \tag{2.8}$$

Through functional correlation measures one is able to describe the relation between response and predictors in a very general nonparametric regression framework.

In a closed manner, Li, Zhong, and Zhu (2012) suggested a screening procedure without assuming any regression structure. As in Zhu et al. (2011) they define only existence of a functional dependence of $F(\cdot|\mathbf{X})$ on the set of relevant regressors. Differently, they use distance correlations for measuring marginal magnitudes. Assuming finite first moments, distance correlation between two random vector is defined as

$$w(u, v) = \frac{\zeta(u, v)}{\sqrt{\zeta(u, u)\zeta(v, v)}} \tag{2.9}$$

where $\zeta(u, v)$ is the distance covariance given by

$$\zeta(u, v)^2 = \int_{\mathbb{R}^{p_u + p_v}} \|\phi_{u,v}(t, s) - \phi_u(t)\phi_v(s)\|^2 \{c_{p_u} c_{p_v} \|t\|_{p_u}^{1+p_u} \|s\|_{p_v}^{1+p_v}\}^{-1} dt ds.$$

with $c_p = \pi^{(1+p)/2} / \Gamma\{(1 + p)/2\}$, p_u and p_v the dimensions of vectors u and v respectively; while $\phi_u(t)$ and $\phi_{u,v}(t, s)$ are the characteristic function of u and the joint characteristic function of (u, v), respectively. Estimates of $w(\cdot, \cdot)$ is used as marginal utility in order to rank the important of each predictor in explaining the response. Then, the set of relevant predictors is defined through a threshold rule, as usually.

Distance correlation is quite different from the functional correlation in Zhu et al. (2011), nevertheless such a screening method encloses nonparametric non additive models. Differently, it has some thigh requirement, in fact because the

quantities involved in the marginal measurement are based on the moments estimates, such a method needs response and predictors to be sub-exponential tail in order to guarantee an increasing dimension exponential in n. Moreover, it is required conditional independence of response with irrelevant variables, that could be too stringent for feature screening problems.

2.2.2 Marginal Empirical Likelihood

Differently from correlation measures, marginal magnitudes in screening procedure can be constructed also by empirical likelihood (EL). EL is a statistical inference tool whose scope recently has been extended to high dimensional problems. Generally, EL encounters substantial difficulty when data dimensionality is high. More specifically, data dimension p cannot exceed the sample size n in the conventional construction.
In the feature screening purpose, Chang, Tang, and Wu (2013) and Chang, Tang, and Wu (2016) proposed a novel idea. Since marginal contribution is assessed one at time individually, thus only involving univariate optimization for each regressor, EL approach may seem particularly useful to treat very general contexts of non additive model regression.
Let $(\mathbf{X}_i, Y_i)$ be i.i.d. collected data from model (1.15). Chang, Tang, and Wu (2013) define a marginal EL problem as

$$EL_j(\beta) = \sup\left(\prod_{i=1}^{n}\omega_i \ : \ \omega_i \geq 0,\ \sum_{i=1}^{n}\omega_i = 1,\ \sum_{i=1}^{n}\omega_i g_{ij}(\beta) = 0\right). \tag{2.10}$$

where, in their context (i.e. linear model) $g_{ij}(\beta) = X_{ij}Y_i - \beta$.
As we are interested in methods for non additive models, we report some details of Chang, Tang, and Wu (2016) since their work can be thought as extension of Chang, Tang, and Wu (2013) to nonparametric non additive regression models. Use of marginal EL in nonparametric regression problems requires a nonparametric estimator. Chang, Tang, and Wu (2016) consider LPE with polynomials

of order zero, thus $m_j(x)$ is estimated by the Nadaraya-Watson (NW) estimator

$$\hat{m}_j(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(X_{ij} - x)Y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(X_{ij} - x)}, \tag{2.11}$$

where, as usual K_h is a weight function depending on bandwidth parameter h and a Kernel function. For assessing $m_j(x) \equiv 0$ at a given x, they suggest the following EL problem

$$EL_j(x, 0) = \sup\left(\prod_{i=1}^{n} \omega_i \ : \ \omega_i \geq 0, \ \sum_{i=1}^{n} \omega_i = 1, \ \sum_{i=1}^{n} \omega_i K_h(X_{ij} - x)Y_i = 0\right). \tag{2.12}$$

Equation (2.12) is solved by Lagrange multiplier method and leads to the empirical likelihood ratio

$$l_j(x, 0) = 2\sum_{i=1}^{n} \log\left(1 + \lambda K_h(X_{ij} - x)Y_i\right), \tag{2.13}$$

where λ here is the Lagrange multiplier solving $\sum_{i=1}^{n} \frac{K_h(X_{ij}-x)Y_i}{1+\lambda K_h(X_{ij}-x)Y_i} = 0$.
Because the denominator in the NW estimator converges to the density of the j-th covariate evaluated in x, (2.13) can be used as a proxy of the local contribution to the response. Large values of $l_j(x, 0)$ are taken as evidence of significant contribution for testing locally whether or not the numerator in the NW equation has zero mean.
The statistics adopted is self-studentized, and hence it incorporates the uncertainties level that usually are taken into account by standard errors when the ranking method is based on magnitudes of parametric estimators. This clearly is a different way of considering marginal statistics for screening. Notice that the mean constraint in the empirical likelihood (2.12) is nothing more that the local correlation in x, between the smoothed version of X_j and Y. Such a screening method does not require strict distributional assumptions such as normally distributed error as in linear models or exponential family distributed response in GLM.
Finally, it can be viewed as an interesting extension of the original proposal in Fan and Lv (2008) and the additive version in Chang, Tang, and Wu (2013), in fact it results appealing in general nonparametric contexts, as it uses an

innovative way of measuring marginal magnitudes.

2.3 High-dimensional results for regression models

We are interested in a brief review about order of dimensions for non additive procedures introduced in the previous sections. In particular we will distinguish the case of fixed from increasing dimensions, with respect to the total number of regressors and the number of the relevant ones.
We know that if the number of relevant predictors q is fixed, then the condition that guarantees consistent estimation of the sparsity pattern is $(\log p)/n \to 0$ in linear regression, whereas it is $p = O(\log n)$ in the general nonparametric case. Probably, a justification to such an important gap between two conditions above resides in the fact that nonparametric regression is much more complex than the linear one. Results in Comminges and Dalalyan (2012) give a good benchmark for our purpose. They show existence of consistent estimators for recovering sparsity pattern in context of nonparametric regression models, under two regimes. The first is when the sample size and the dimension p of all covariates tend to infinity, but the dimension q of relevant ones is fixed; the second is when also q diverges with the sample size n. Interesting results from Comminges and Dalalyan (2012) are about second regime. They show that, in the fixed regime case and for nonparametric models, it is possible to estimate the sparsity pattern for q at the same order of the sample size (i.e. $q = O(n^{1-\varepsilon})$, for some $\varepsilon > 0$) if p is at most polynomially. In particular, it is possible under condition

$$\log p = O\left(\frac{c\, n}{q}\right), \tag{2.14}$$

for some constant $c > 0$.
The situation becomes worse when $q \to \infty$. Consistent recovery of sparsity pattern under regime of increasing dimensions requires together

$$\begin{aligned} & q = o(\log n) \\ & \text{and} \\ & \log\log\left(\frac{p}{q}\right) = o(\log n) \end{aligned} \tag{2.15}$$

Work of Comminges and Dalalyan (2012) is devoted to show the possibility of consistent estimation, rather than to provide a practical procedure for recovering the sparsity pattern, for that reason we have not included it among methods reviewed in previous sections.
Assumption of non additive regression function strongly affects the order of high dimensions in variable selection procedures. Lafferty and Wasserman (2008) and Bertin and Lecué (2008) proved the consistency of the proposed procedures under some more or less restrictive assumptions, moreover the former does not allow the dimension of covariates increases too fast and thus turns out to be sub-optimal, the latter is optimal in term of increasing dimensions but it results unfeasible. Estimation in Lafferty and Wasserman (2008) assumes the unknown regression function four times continuously differentiable with bounded derivatives, and it seems to be a bit stringent. In spite of this, the algorithm is shown to have good converge properties when the number of covariates p is at most $O(\log n / \log\log n)$ meanwhile the number of relevant variables q does not increase with n.
On the contrary, Bertin and Lecué (2008) redefine the same problem of RODEO in order to achieve a best treatable dimensionality. They show consistency when q is still fixed, but p is allowed to be $O(\log n)$, up to a constant, giving a little improvement in terms of high dimension orders. Apart from that, it remains impractical due to unfeasible calibration procedure of penalization balance that ensure a consistent selection.

Regarding the class of feature screening methods, we know that it has originally proposed to treat situations where the high dimensions order is non-polynomial in the sample size. All reviewed work for non additive models can handle non-polynomial dimensionality. Differently from variable selection techniques, in feature screening the required order of dimensionality is directly affected by several quantities related to several assumptions (i.e. smoothness, minimum signal, and so on) or by the class of adopted estimator. For instance, in Fan, Feng, and Song (2011) the order of high dimension is $\log p = o(n^{1-4k}d^{-3} + nd^{-3})$, it is affected by truncation order d of spline estimator; also in Chang, Tang, and Wu (2016) highest handleable dimension is exponential in n, nevertheless it actually

depends on k that controls the order of the signal strength, on the smoothness order r of regression functions and on tail probability distribution of the response. If we compare the two results we notice that the former provides a handleable faster diverging rate of p compared with the latter. Clearly, it is the price paid in Chang, Tang, and Wu (2016) by allowing weaker requirement on the continuity of projections m_j as well as on the absence of bounds on regression function. On the other hand, if we compare it with the work of Chang, Tang, and Wu (2013), slow diverging rate of p is offset by a better performance in keeping relevant covariates with weaker marginal contribution. Nevertheless, the idea of Chang, Tang, and Wu (2016) shares some practical challenges. In fact, it can be quite difficult to ensure proper rate of smoothing and other quantities that control the probabilistic behaviour of the empirical likelihood ratio, in order to achieve consistent screening. As to regard, it is sufficient to notice that good performance in distinguish between the true contributing variables from the false ones is achieved only under a correct estimation of the bandwidth parameter h that has to be selected in the estimation procedure.

2.4 Estimation of tuning parameters

In feature screening a threshold is adopted in order to reduce the full model to a sub-model of less order. The importance of the threshold is similar to the task of tuning parameters in regularization methods for variable selection, thus it plays a crucial rule in practical implementation. However, choosing the threshold is difficult in practice. Usually, this issue is overcome by defining a prespecified number of variables to be selected. This is one of the most important limits of screening methods comparing with variable selection purpose. Clearly, it is because it works marginally in order to perform reduction of large-scale dimension to a less order.

In a different way, variable selection procedures involving tuning or smoothing parameters require some data-driven criteria in order to select directly such parameters.

In Lafferty and Wasserman (2008) bandwidth estimation and variable selection can be performed simultaneously. Thus no calibration is needed since the procedure requires computing infinitesimal change of estimator as function

of the smoothing parameter h, and at the end selected variables automatically result from estimation procedure. From a theoretical point of view procedure in Lafferty and Wasserman (2008) does not require the smoothing parameter h goes to zero when n tends to infinity. This unusual behaviour appears since LPE are used for variable selection rather than estimation.
Similarly, Bertin and Lecué (2008) provide to use bandwidth h as in Lafferty and Wasserman (2008). Since their estimator uses LASSO-type penalization, it introduces a further regularization parameter λ that needs to be calibrated. Here a proper selection is based on a sort of balance between λ and the smoothing parameter h. In particular, it is required

$$0 < h < \frac{\mu_m}{32(q_0+1)L_\mu M_k} \wedge \eta \text{ and } \lambda = 8\sqrt{3M_k}\mu_M L h.$$

where q_0 is an integer such that $q < q_0$ and μ_m, μ_M, L, M_k are quantities connected to the assumption on bounded derivative of the regression function (see assumption 6 in Bertin and Lecué (2008) for more details). Since λ controls the amount of shrinkage, this sort of connection between the two parameters is crucial to guarantee, on one hand orthogonality of the design through the restriction about h, on the other hand consistency in selection through λ. An important aspect to remind is that such a method is inoperable in practice as no calibration criteria is able to guarantee a proper balance for consistent selection.
Apart from threshold parameter, procedure in Chang, Tang, and Wu (2016) also requires estimation of bandwidth h as it adopts LPE. From a theoretical point of view, it is assumed $h = O(n^{-\delta})$ for some positive constant δ that is directly connected to the order of smoothing of the regression function, Kernel requirements and identification assumption. Such a requirement on the order of h seems to be not compelling as the authors show that it can be satisfied by the conventional optimal bandwidth $h = O(n^{-1/5})$. In particular, if the first derivative of the regression functions exists, then the optimal bandwidth is shown to be $O(n^{-k/r})$, where k controls the order of the minimum signal strength and $r \geq 1$ is the degree of smoothness of m. Conversely, if m is infinitely differentiable, then $r = \infty$ and the order of optimal bandwidth for sure screening is given by $\delta \in (0, 1)$. As concern the effects of choice of h in practical implementation, differently from Lafferty and Wasserman (2008) here

h needs to be estimated. It could be done by using the most common validation criteria or plug-in estimation. However, a bad estimation of the bandwidth parameter strongly affects the selection procedure, this is because selection method is based on EL local testing of constrained means and, since h directly control the bias-variance trade-off in LPE, inaccurate estimation can affect the result of the test.

2.5 Selection with correlated predictors

Variable selection as well as feature screening are valid under some regularity assumptions and some of them directly affect the magnitude of correlation among regressors. As concern the main variable selection methods for non additive models, we have that RODEO of Lafferty and Wasserman (2008) requires the joint density function $m(x)$, valued at local point x, to be uniform. Such a condition makes theoretical proofs simpler, but clearly rules out any situation of correlated regressors.

Bertin and Lecué (2008) use all the theory from L_1 penalization in linear models to show theoretical properties of their method. Thus, a proper selection is ensured under some kind of irrepresentable condition on the design matrix in the Kernel of $\mathbf{X}$, and as usual, it involves bounds on the eigenvalues of the matrix. For that reason, particular correlation structures in the regressors matrix can affect the selection procedure in a similar way as for the standard LASSO.

As concern feature screening, linearity condition in Zhu et al. (2011) states that

$$\mathbb{E}(\mathbf{X}|\mathbf{X}^\top\beta) = cov(\mathbf{X},\mathbf{X}^\top)\beta\{cov(\mathbf{X}^\top\beta)\}^{-1}\beta^\top\mathbf{X} \tag{2.16}$$

Such a condition is always satisfied when the regressors are normal distributed or, more generally, have an elliptical distribution. Nonetheless, it is weaker than the assumption of normality since it is only required to hold for the true value of $q \times d$ coefficients β. Furthermore, linearity condition holds asymptotically if the number of regressors p diverges while the spanning dimension remains fixed.

In addition, requirements in Zhu et al. (2011) also involve a bound on correlation among the predictors (see condition C1). Let $\mathbf{X}_{\mathcal{M}}$ and $\mathbf{X}_{\bar{\mathcal{M}}}$ be the matrices of

true relevant and irrelevant predictors, respectively. For sure screening it is required

$$\frac{d^2\lambda_{max}\{cov(\mathbf{X}_{\mathcal{M}}, \mathbf{X}'_{\bar{\mathcal{M}}})cov(\mathbf{X}_{\bar{\mathcal{M}}}, \mathbf{X}'_{\mathcal{M}})\}}{\lambda^2_{min}\{cov(\mathbf{X}_{\mathcal{M}}, \mathbf{X}'_{\mathcal{M}})\}} < \frac{\min_{k\in\mathcal{M}} w_k}{\lambda_{max}\{\mathbf{!}_{\mathcal{M}}\}}. \tag{2.17}$$

where here $\lambda_{min}\{\mathbf{X}\}$ and $\lambda_{max}\{\mathbf{X}\}$ denote minimum and maximum eigenvalues of matrix $\mathbf{X}$, respectively; d is the spanning order as in (2.5). Under such assumption, the method is suitable for characterising the conditional distribution of the response given the full set of predictors $\mathbf{X}$ through a projection of $\mathbf{X}$ onto a space of dimension q less than p and spanned by $p \times d$ matrix of coefficients. Thus, (2.17) is the key assumption to ensure that the screening procedure works properly. First, as the dimension of spanning coefficients β increases, (2.17) becomes more stringent. Therefore, a model with a small matrix β is favoured by such a procedure. Second, the numerator in the left hand side of (2.17) measures the correlation between relevant predictors and the irrelevant ones, while the denominator measures the correlation among relevant predictors themselves. When the relevant and irrelevant groups are uncorrelated, the assumption holds automatically. This condition rules out the case in which there is strong collinearity between the two groups, or among relevant predictors themselves.

Also in Lin, Sun, and Zhu (2013), similar to Zhu et al. (2011), a moment condition for predictors is required in order to guarantee a correct separation between predictors. It should be ensured that the set of relevant predictors is weakly correlated with irrelevant ones. Such a condition states that

$$\max_{k\in\bar{\mathcal{M}}} E^2[C_k(\mathbf{X}_{\mathcal{M}})] < \frac{1}{4}\min_{k\in\mathcal{M}} \omega_k \tag{2.18}$$

where $C_k(\mathbf{X}_{\mathcal{M}}) = \sup_{x_k} |\mathbb{E}[X_k I(X_k < x_k)|\mathbf{X}_{\mathcal{M}}] - \mathbb{E}[X_k I(X_k < x_k)]|$ for $k \in \bar{\mathcal{M}}$. Since the absolute value of $\mathbb{E}[C_k(\mathbf{X}_{\mathcal{M}})]$ for $k \in \bar{M}$ can measure the correlation between the relevant and irrelevant predictors, it is representative of constraint on magnitude of correlation among covariates. Then, it effectively rules out some cases of strong correlation.

Correlation problem in Chang, Tang, and Wu (2016) is not directly determinate by some kind of assumption. This is why by EL one would marginally test

constrained mean given by some function of nonparametric smoother where bandwidth h controls bias-variance trade-off. Therefore h plays a key rule in the selection procedure since its wrong choice can directly affect testing result, but its calibration is not directly affected by correlation problems. Conversely, identification condition on minimal signal strength of relevant variables can be difficult to ensure in presence of highly correlated variables. When it is required that the smallest uniform norm of each relevant predictors is not too weak, i.e. $\min_{j \in M} \left\| m_j \right\|_\infty \geq c\, n^{-k}$, for some $k > 0$, technically it is to guarantee *sure screening* property. On the contrary, maximum signal strength of irrelevant variable should vanish at a faster rate. Since the probability of the size of recruited set of variables is normally affected by the order at which irrelevant signal strength vanishes, proper result in distinguish between true contributing variables from false ones can be ensured when there is a proper separation between the group of relevant variables and the group of irrelevant ones. In spite of this, when the irrelevant covariates are strongly correlated with the relevant ones such a decreasing rate of signal strength could not be faster enough to ensure consistent screening. That is because screening method, although working marginally, suffers in recruiting the true set of variables in such a situation of strong correlation, as it can retain irrelevant variables confounded as relevant.

Chapter 3

Conditional local independence feature screening for nonadditive models

3.1 Introduction

Screening methods are appealing and innovative, nevertheless, success of any screening procedure depends on how well the marginal utility, correlation coefficient between the response and each individual predictor, captures the importance of the predictors in a joint model. A variable may be retained by the screening procedure when it is marginally important but not jointly important, resulting in false positive, or a variable that is jointly important but not marginally important can be screened out, resulting in a false negative. False negatives have two potentially serious consequences. First, important covariates may be screened out and they will not be reinstated by the second-stage analysis. Second, the false negatives can lead to bias in subsequent inference. As an illustrative example, we consider the following model from Chang, Tang, and Wu (2016), $Y = 2X_1 + 2X_2 + 2X_3 - 3\sqrt{2}X_4 + \varepsilon$, where ε has standard normal distribution. The covariates vector $(X_1, ..., X_p)'$ is jointly normal distributed such that $\mathbb{E}(X_j) = 0$ and $var(X_j) = 1$ for all j, and its covariance matrix is constructed such that the relevant variable X_4 results marginally uncorrelated with the response even if it has largest coefficient. In such a situation we expect that screening procedure will give little priority to X_4, as also confirmed by simulation study in paragraph 3.4.

Fan and Lv (2008) tried to partial overcome these issues by proposing an iterative version of the SIS, which allows using more quite the joint information rather than just the marginal information. An iterative version basically works as follows. First, perform feature screening procedure resulting in a subset $\mathcal{A}_1$ of retained variables, then one effectively selects among them a subset of $\mathcal{M}_1$ of variables. As said, such a selection step can be done by one of the standard selection procedures (i.e. LASSO, COSSO, etc.). Then, a vector of residuals is obtained regressing the response over the $\mathcal{M}_1$ variables. In the next step, the residuals vector is treated as response variable and the same screening method is applied to the $\mathcal{X}/\{\mathcal{M}_1\}$ variables, which results in a set of $\mathcal{A}_2$ variables. Considering that the residuals from the first stage are uncorrelated with the selected variables, the priority associated to one irrelevant predictor highly correlated with the response can significantly reduced. The screening is doing iteratively until k disjoint subsets of variables are obtained and such that their union has prespecified dimension $\tilde{p} < n$. It should be noted that iterations cannot help to avoid the issue of discarding relevant predictors marginally unrelated.

In addition to the false positive and false negative issues, measuring the importance of features marginally can also yield wrong results in situations of high correlated covariates. In fact, due to the correlation among the covariates, marginal screening can recruit those variables who have strong marginal relevance but are jointly independent with the response variable. Such a situation, concerning with collinearity among predictors, introduces a further issue in feature screening.

To well understand how correlation affects a selection procedure, consider that in linear model one of the real difficulties of high dimensional data is that the matrix of regressors $\mathbf{X}$ is rectangular and the corresponding $\mathbf{X}'\mathbf{X}$ matrix is singular. Even in nonlinear or more general contexts, this means that maximum spurious correlation between a covariate and the response can be very large because of irrelevant predictors that are highly correlated with the response owing to the presence of relevant predictors associated with them. This is the situation in which, for instance, LASSO type methods or marginal screening techniques fail to correctly select the true set of relevant variables. In fact, some irrelevant predictors that are highly correlated with relevant predictors can

have higher priority of being selected than other relevant predictors that are relatively weakly marginal related to the response. Such situations can add difficulties in performing a proper selection.

3.2 Conditional screening

It is well-known that in many applications researchers often know in advance a set of certain predictors are related to the response variable. This is because of previous investigation and/or experience about phenomenon of study or, for instance it is the result of some previous screening procedure that returns a set of some relevant variables. Information from a known set of relevant variables can be used in order to reduce the correlation among predictors, solve problems of false positive/negative and, thus improve results from screening. This is why, instead of adopting iterative residual-based approach to circumvent the issue of classical screening methods, Barut, Fan, and Verhasselt (2016) proposed a conditional version in order to handle the situation in which relevant variables are marginally unrelated to the response and there is strong correlation among predictors. The idea is that conditioning upon a set $\mathcal{C}$ of variables regarded as significant by a prior knowledge on the problem, one can disclose variables that were hidden by collinearity or because marginally unrelated. Through conditional screening marginal magnitude of each variable X_j $(j \notin \mathcal{C})$ is evaluated by taking in consideration predictors in $\mathcal{C}$. The work of Barut, Fan, and Verhasselt (2016) is the original proposal for conditional screening and it is assessed for linear models with normal distributed errors. Hu and Lin (2017) extended the idea to the generalized linear models. Moreover, the properties of both methods in Barut, Fan, and Verhasselt (2016) and Hu and Lin (2017) are assessed in the frameworks of linear and generalized linear models only. In the following, we introduce a conditional screening procedure for non additive models.

3.3 Conditional local marginal empirical likelihood

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

$$Y = m(\mathbf{X}) + \varepsilon, \tag{3.1}$$

where $\mathbf{X} = (X_1, ..., X_p)'$ and ε is i.i.d. error term with $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$.
We assume that the predictors X_j are standardized such that $\mathbb{E}(X_j) = 0$ and $\mathbb{E}(X_j^2) = 1$, for all $j = 1, ..., p$.
No specification of regression function $m(\mathbf{X})$ is required, apart from that the true model is sparse in the sense that only a small subset of covariates is contributing to the response variable Y. We call this set of relevant covariates $\mathcal{M}^* = \{1 \leq j \leq p : \mathbb{E}(Y|X_j) \neq 0\}$. As mentioned in introduction, in many practical application, researchers have already known certain predictors are important for the response by some previous investigations and experiences, which means that a set of relevant predictors has been determined in advance. For instance, such a situation can happen in presence of hidden relevant predictors that have no marginal strength and could be missed by standard feature screening techniques.
There exist several screening methods for the ranking of marginal magnitudes in non additive models, see Li, Zhong, and Zhu (2012), Zhu et al. (2011), Lin, Sun, and Zhu (2013), Hu and Lin (2017) and Chang, Tang, and Wu (2016), among others. Comparing to the other methods, Chang, Tang, and Wu (2016) adopt the idea of marginal hypothesis testing to handle feature screening problem, while the other methods all deal such a problem by marginal estimation. In particular, they used marginal empirical likelihood approach with local polynomial estimation (LPE), this makes the method appealing since it requires a less restrictive distributional assumption. Empirical likelihood ratio evaluated at zero can be used to against the null hypothesis that the marginal effect is negligible. Moreover, it only involves univariate optimisation problem, thus it provides an appealing device for both theoretical analysis and practical implementation.
In order to investigate the marginal contribution from each covariate in explaining Y, it can be adopted the standard marginal nonparametric regression

problem

$$\min_{m_j \in \mathcal{M}_2} \mathbb{E}\big\{ \big[Y - m_j(X_j) \big]^2 \big\}, \quad j = 1, ..., p. \tag{3.2}$$

where $\mathcal{M}_2$ denotes the class of square integrable functions. The solution to the above minimization problem is $\mathbb{E}(Y|X_j)$ and Chang, Tang, and Wu (2016) use Nadaraya-Watson estimate of $m_j(x) = \mathbb{E}(Y|X_j = x)$ as building block to test marginal contribution of jth covariate locally. This is because $m_j(x) = 0$, for all $x \in \mathcal{X}$, if X_j is not relevant to explain Y.

Since our purpose is to introduce conditional information into such a marginal regression problem, let us to consider, without loss of generality, that the conditioning set of relevant known predictors is composed by the first c components $X_1, ..., X_c$ of $\mathbf{X}$. Thus let us partition the regressors matrix as $\mathbf{X} = \{\mathbf{X}_\mathcal{C}, \mathbf{X}_\mathcal{D}\}$ with $\mathbf{X}_\mathcal{C} = (X_1, ..., X_c)'$ and $\mathbf{X}_\mathcal{D} = (X_{c+1}, ..., X_p)'$.

To include conditional information into the local marginal regression, we first consider, as in Hu and Lin (2017), the following key quantity

$$\mathbb{E}\big\{ \big[X_j - \mathbb{E}(X_j|\mathbf{X}_\mathcal{C}) \big] \big[Y - \mathbb{E}(Y|\mathbf{X}_\mathcal{C}, X_j) \big] \big\}, \quad \forall j \in \mathcal{D}. \tag{3.3}$$

where $\mathbf{X}_\mathcal{C}$ is the matrix of conditioning variables, meanwhile $\mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ and $\mathbb{E}(Y|\mathbf{X}_\mathcal{C}, X_j)$ are the conditional expectations of X_j given $\mathbf{X}_\mathcal{C}$, and Y given $(\mathbf{X}_\mathcal{C}, X_j)$, respectively. Moment condition (3.3) allows to include conditional information about $\mathbf{X}_\mathcal{C}$ into marginal evaluation of covariate X_j. In particular $\mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ incorporates the conditional information from $\mathbf{X}_\mathcal{C}$ with respect to X_j, while $\mathbb{E}(Y|\mathbf{X}_\mathcal{C}, X_j)$ measures the strength of conditional contribution to Y of X_j given the conditional set.

If we thought of $X_j - \mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ as a centralized version of the jth variable, we can define the following nonparametric regression problem

$$\min_{m_{\mathcal{C},j} \in \mathcal{M}_2} \mathbb{E}\Big\{ \omega_{(\mathcal{C},j)} \big[Y - m_{\mathcal{C},j}(\mathbf{X}_\mathcal{C}, X_j) \big]^2 \Big\}, \quad j \in \mathcal{D}. \tag{3.4}$$

where $\omega_{(\mathcal{C},j)} = X_j - \mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$. Function $m_{\mathcal{C},j}(\mathbf{X}_\mathcal{C}, X_j) = \mathbb{E}(Y|X_j, \mathbf{X}_\mathcal{C})$ summarises the impact of jth variable jointly with the conditioning set $\mathcal{C}$.

In order to evaluate locally the magnitude for the jth variable in a conditional

screening problem, we solve (3.4) by Kernel smoothing estimation. Thus, assuming constant locally approximation for the regression function $m_{\mathcal{C},j}$, minimization problem (3.4) is equivalent to

$$\min_{m_c} \sum_{i=1}^{n} \{Y_i - m_c\}^2 \omega_{i(\mathcal{C},j)} K_H\big(X_{i(\mathcal{C},j)} - \mathbf{x}\big). \tag{3.5}$$

where $X_{i(\mathcal{C},j)}$ is the ith observations of matrix $\mathbf{X}_{\mathcal{C},j} = \{\mathbf{X}_{\mathcal{C}}, X_j\}$, meanwhile function $K_H(\cdot)$ is such that $K_H(\mathbf{x}) = |H|^{-1}\mathcal{K}(H^{-1}\mathbf{x})$ with $\mathcal{K}(\cdot)$ a $(c+1)$-variate Kernel and H is a $(c+1) \times (c+1)$ matrix of bandwidths.
A solution of the above equation is given by the weighted Nadaraya-Watson estimator of the form

$$\hat{m}_{\mathcal{C},j}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \omega_{i(\mathcal{C},j)} K_H\big(X_{i(\mathcal{C},j)} - \mathbf{x}\big) Y_i}{\sum_{i=1}^{n} \omega_{i(\mathcal{C},j)} K_H\big(X_{i(\mathcal{C},j)} - \mathbf{x}\big)}. \tag{3.6}$$

where each weight $\omega_{i(\mathcal{C},j)}$ needs to be estimated.
This setting results similar to the one in Chang, Tang, and Wu (2016) but now conditional information is introduced in the estimation procedure. Unfortunately, statistics (3.6) cannot help in discriminating relevant predictors when some conditional information is introduced in the estimation. Thus it is not useful to evaluate if the jth covariate is marginal contributing in explaining Y conditionally to the known set $\mathcal{C}$. In fact, it happens that, also when the jth covariate is conditionally irrelevant, $m_{\mathcal{C},j}(\mathbf{x})$ could not be equal to zero.
Furthermore, notice that also in the simplest case of conditioning set $\mathcal{C}$ composed by one variable (i.e. $|\mathcal{C}| = 1$), a bivariate Kernel estimation is involved with 2-dimensional Kernel function and the bandwidth H a 2×2 matrix that can be taken by imposing some restriction, for instance diagonal positive definite restriction or by assuming a single common bandwidth.
In order to avoid multivariate estimations even in the simplest case of univariate conditioning set, instead of (3.3) we consider the following quantities from Barut, Fan, and Verhasselt (2016)

$$X_j^* = X_j - \mathbb{E}(X_j|\mathbf{X}_{\mathcal{C}}) \,, \quad Y^* = Y - \mathbb{E}(Y|\mathbf{X}_{\mathcal{C}}) \qquad \forall j \in \mathcal{D}. \tag{3.7}$$

Define $\hat{\mathbb{E}}(X_j|\mathbf{X}_\mathcal{C})$ as an estimator for the conditional expectation $\mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ in (3.7), moreover we estimate the conditional expectation $\mathbb{E}(Y|\mathbf{X}_\mathcal{C}) = m_\mathcal{C}(\mathbf{x})$ by local polynomial estimator $\hat{m}_\mathcal{C}(\mathbf{x})$. It can be obtained as

$$\arg\min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - \beta_0(\mathbf{x}) + \beta_1'(\mathbf{x})(X_{i\mathcal{C}} - \mathbf{x}) + ... + \beta_k'(\mathbf{x})(X_{i\mathcal{C}} - \mathbf{x})^k \right\}^2 K_H\big(X_{i\mathcal{C}} - \mathbf{x}\big)$$

where the weights are given by function $K_H(x) = |H^{-1}|\mathcal{K}(H^{-1}x)$, $\mathcal{K}(x)$ represents a c-variate Kernel function and H is a $c \times c$ matrix of smoothing parameters.
If we assume constant local approximation ($k = 0$) for the regression function, $\hat{m}_j$ reduces to the marginal Nadaraya-Watson estimator

$$\hat{m}_\mathcal{C}(\mathbf{x}) = \frac{\sum_{i=1}^{n} K_H\big(X_{i\mathcal{C}} - \mathbf{x}\big) Y_i}{\sum_{i=1}^{n} K_H\big(X_{i\mathcal{C}} - \mathbf{x}\big)}. \tag{3.8}$$

Considering a conditioning set $\mathcal{C}$ composed by only one known relevant covariate, $\hat{m}_\mathcal{C}(x)$ can be obtained as solution of marginal regression problem in (3.2) with the covariate replaced by the conditioning variable, that is

$$\hat{m}_\mathcal{C}(x) = \frac{\sum_{i=1}^{n} K_h(X_{i\mathcal{C}} - x) Y_i}{\sum_{i=1}^{n} K_h(X_{i\mathcal{C}} - x)}, \tag{3.9}$$

where $K_h(x) = h^{-1}\mathcal{K}(x/h)$ and $\mathcal{K}(\cdot)$ is a univariate Kernel function and h is the bandwidth.
By estimators in (3.8) or (3.9) we can build the transformation $Y^* = Y - \hat{m}_\mathcal{C}(\mathbf{x})$ of the response variable that takes into account conditional information. We use this new response variable to investigate the marginal contribution from each centralized jth covariate $X_j^* = X_j - \hat{\mathbb{E}}(X_j|\mathbf{X}_\mathcal{C})$.
Notice that in order to use both the rebuilded variables for constructing a statistics, we first need to estimate the conditional expectation $\mathbb{E}(X_j|\mathbf{X}_\mathcal{C})$ for obtaining variables X_j^*.
We follow Hu and Lin (2017) that propose to estimate $\mathbb{E}(X_j|X_{i\mathcal{C}})$ in the following way. Let $\sigma_{\mathcal{C}j} = cov(X_j, \mathbf{X}_\mathcal{C})$ be the covariance between conditioning variables and jth covariate. Estimators for $\sigma_{\mathcal{C}j}$ and $\mathbb{E}(\mathbf{X}_\mathcal{C}\mathbf{X}_\mathcal{C}')$ can be obtained respectively

as

$$\begin{aligned} \hat{\sigma}_{\mathcal{C}j} &= \frac{1}{n}\sum_{i=1}^{n} X_{ij}X'_{i\mathcal{C}}\,, \\ \hat{\mathbb{E}}(\mathbf{X}_{\mathcal{C}}\mathbf{X}'_{\mathcal{C}}) &= \frac{1}{n}\sum_{i=1}^{n} X_{i\mathcal{C}}X'_{i\mathcal{C}}. \end{aligned} \tag{3.10}$$

Hence, an estimator for $\mathbb{E}(X_j|X_{i\mathcal{C}})$ is given by

$$\hat{\mathbb{E}}(X_j|X_{i\mathcal{C}}) = \frac{1}{n}\sum_{k=1}^{n} X_{kj}X'_{k\mathcal{C}}\left\{\frac{1}{n}\sum_{k=1}^{n} X_{k\mathcal{C}}X'_{k\mathcal{C}}\right\}^{-1} X_{i\mathcal{C}}, \tag{3.11}$$

and the centralized X_j^* variable can be obtained as

$$X_{ij}^* = X_{ij} - \sum_{k=1}^{n} X_{kj}X'_{k\mathcal{C}}\left\{\sum_{k=1}^{n} X_{k\mathcal{C}}X'_{k\mathcal{C}}\right\}^{-1} X_{i\mathcal{C}}, \quad i = 1, ..., n \tag{3.12}$$

If $\mathbb{E}(\mathbf{X}_{\mathcal{C}}\mathbf{X}'_{\mathcal{C}}) = I_{\mathcal{C}}$ then X_j^* can be rewritten as

$$X_{ij}^* = X_{ij} - n^{-1}\big(\sum_{k=1}^{n} X_{kj}X'_{k\mathcal{C}}\big)X_{i\mathcal{C}}, \quad i = 1, ..., n. \tag{3.13}$$

To apply local marginal screening with conditional information we need to define some discriminating rule in order to assessing whether the jth covariate is marginal relevant or irrelevant without any distributional assumption. For such a purpose we will adopt marginal empirical likelihood as in Chang, Tang, and Wu (2016) and define it for conditional screening problem.
Second stage estimation involves marginal nonparametric regression problem

$$\min_{m_j^* \in \mathcal{M}_2} \mathbb{E}\big\{\big[Y^* - m_j^*(x)\big]^2\big\}, \quad j \in \mathcal{D}. \tag{3.14}$$

As in Chang, Tang, and Wu (2016) we consider the Nadaraya-Watson estimator for m_j^*

$$\hat{m}_j^*(x) = \frac{\sum_{i=1}^{n} K_h(X_{ij}^* - x)Y_i^*}{\sum_{i=1}^{n} K_h(X_{ij}^* - x)}, \tag{3.15}$$

For assessing $m_j^*(x) \equiv 0$ at a given x without distributional assumptions, we adopt marginal local empirical likelihood problem defined as

$$EL_j(\mathbf{x}, 0) = \sup\left(\prod_{i=1}^{n} \pi_i \ : \ \pi_i \geq 0, \ \sum_{i=1}^{n} \pi_i = 1, \ \sum_{i=1}^{n} \pi_i \, g_{ij}(x) = 0\right). \tag{3.16}$$

estimator (3.15) takes into account conditional information in empirical likelihood. We set the constrain function $g_{ij}(x)$ as

$$g_{ij}(x) = K_h(X_{ij}^* - x)Y_i^* \tag{3.17}$$

since $\hat{m}_j^*(x) \equiv 0$ means that jth covariate is not relevant. The corresponding empirical likelihood ratio is given by

$$l_j(x; 0) = 2\sum_{i=1}^{n} \log\left(1 + \lambda \, g_{ij}(x)\right), \tag{3.18}$$

where λ is the Lagrange multiplier satisfying $\sum_{i=1}^{n} \frac{g_{ij}(x)}{1+\lambda \, g_{ij}(x)} = 0$.
Intuitively, $l_j(x; 0)$ should be small for all x in the corresponding support, if given the conditioning set $\mathcal{C}$, X_j does not contribute to Y.
For such a reason, conditional empirical likelihood ratio (3.18) can be viewed as device for feature screening in the context of nonparametric nonadditive models. More specifically, large value of $l_j(x; 0)$ is taken as evidence of relevant contribution of jth covariates given the conditional set $\mathcal{C}$. We select the set $\hat{\mathcal{M}}$ of explanatory variables as usual by

$$\mathcal{D} \cap \mathcal{M}_{\gamma_n} = \{j \in \mathcal{D} \ : \ l_j(0) \geq \gamma_n\}, \tag{3.19}$$

where $\mathcal{M}_{\gamma_n}$ is the set of selected predictors depending on threshold γ_n and $l_j(0) = \sup_{x \in supp(X)} l_j(x; 0)$.
We call this empirical likelihood problem for feature screening as *conditional local empirical likelihood* (CEL).
Since generally feature screening serves as a preliminary dimensionality reduction procedure, and it is often followed by a conventional variable selection procedure, feature screening is more concerned with recruiting all the truly

important covariates. Furthermore, conditional screening approach is non-iterative and has much less computational cost compared with the iterative one.

3.3.1 Theoretical properties

Following, we assume some regular conditions for Kernel regression.

(A.0) Suppose that $\mathbf{X}_{\mathcal{C}}$ is additive with respect to $\mathbf{X}_R$. Then, $Y = m_1(\mathbf{X}_{\mathcal{C}}) + m_2(\mathbf{X}_R) + \varepsilon$, where $\mathbf{X}_R$ is the set of relevant covariates without the set of the conditional ones, $\mathbf{X}_{\mathcal{C}}$. Without loss of generality, we assume that $E[m_2(\mathbf{X}_R)] = 0$.

(A.1) The marginal projections $\{m_j\}_{j=1,...,p}$ belong to $C^r(\mathcal{X})$. Where $C^r(\mathcal{X})$ denotes the class of all continuous functions defined over $\mathcal{X}$ that are r times differentiable.
If $r = 0$, m_j's satisfy the Lipschitz condition with order $\alpha \in (0, 1]$, that is $|m_j(x) - m_j(z)| \leq C_1|x - z|^\alpha$ for any $x, z \in \mathcal{X}$, where C_1 is a positive constant uniformly for any $j = 1, ..., p$.
In addiction, there exists a constant C_2 such that $|m_j^{(r)}(x)| \leq C_2$ for any $x \in \mathcal{X}$ and $j = 1, ..., p$.

(A.2) The marginal density function f_j of X_j satisfies $0 < C_3 \leq f_j(x) < C_4 < \infty$, $\forall x \in \mathcal{X}$ and $j = 1, ..., p$. In addiction, we assume that each f_j belongs to $C^r(\mathcal{X})$ for the r given in (A.1) and $|f_j^{(r)}(x)| \leq C_5$ for any $x \in \mathcal{X}$ and $j = 1, ..., p$.

(A.3) For r specified in (A.1), if $r \geq 1$, the Kernel function $\mathcal{K}(\cdot)$ is of order r, that is, $\int \mathcal{K}(u)du = 1$, $\int u^k \mathcal{K}(u)du = 0$ for $k = 1, ..., r-1$ and $\int u^r \mathcal{K}(u)du > 0$. If $r = 0$, the Kernel function satisfies $\mathcal{K}(u) > 0$ and $\int \mathcal{K}(u)du = 1$.

(A.4) The marginal projections $\{m_j^*\}_{j \in \mathcal{D}}$ satisfy condition (A.1) and there exist non negative constants $c_1 > 0$ and $k \in \left[0, \min\left(\frac{\max(r,\alpha)}{2\max(r,\alpha)+2}, \frac{\max(r,\alpha)}{2\max(r,\alpha)+d_{\mathcal{C}}}\right)\right)$ such that
$\min_{j \in \{\mathcal{D} \cap \mathcal{M}^*\}} \|m_j^*\|_\infty \geq c_1 n^{-k}$, where r and α are specified in (A.1) and $d_{\mathcal{C}} = |\mathcal{C}|$.

(A.5) Let $\|\mathcal{X}_n\|$ be the largest length of the intervals in the partition $\mathcal{X}_n$, there exists some positive constant η such that $\|\mathcal{X}_n\| = n^{-\eta}$.

(A.6) There exist positive constants C_6, C_7 and γ_1 such that $\mathbb{P}(|Y^*| \geq z) \leq C_6 \exp(-C_7 z^{\gamma_1})$ for any $z > 0$, where $\tilde{Y}^* = m_2(\mathbf{X}_R) + \varepsilon$.

Assumption (A.0) is only made to simplify the proof of the following Proposition. Assumption (A.1) is about smoothness of functions m in the regression problem. Assumption (A.2) is standard in nonparametric regression and ensures that density of each covariate is bounded. Assumption (A.3) is a standard requirement for the Kernel function that ensures the bias of smoothing is not dominating. Assumptions (A.4) and (A.5) are from Chang, Tang, and Wu (2016). The former is a sort of identification condition that ensure correct separation for relevant covariates, while the latter requires that the partition of the support of each covariate has size at least $O(n^{\eta})$. For sake of simplicity we take the same support $\mathcal{X} = [a, b]$ for all the covariates. Assumption (A.6) is again from Chang, Tang, and Wu (2016).

Proposition 1. *Suppose that the assumptions (A.0) - (A.6) hold. If X_j and $\mathbf{X}_{\mathcal{C}}$ are independent then the results of Chang, Tang, and Wu (2016), for the screening procedure with marginal Empirical Likelihood, still hold using $\hat{X}_j^*$ and $\hat{Y}^*$, $\forall j \in \mathcal{D}$.*

Proof. First, we consider the true quantities X_j^* and Y^*, $\forall j \in \mathcal{D}$. Since X_j and $\mathbf{X}_{\mathcal{C}}$ are independent, it follows that $X_j^* = X_j$, $\forall j \in \mathcal{D}$. Moreover, by assumption (A.0) we have that $Y^* = m_2(\mathbf{X}_R) + \varepsilon$. So, the results in Chang, Tang, and Wu (2016) hold.

Now, we consider $\hat{X}_j^* = X_j - \hat{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right) = X_j + O_p(n^{-1/2})$ by using the consistency results for $\hat{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right)$ as in Hu and Lin (2017). Therefore, we can find n^* such that $\forall n > n^*$, the marginal density function of X_j^*, say $f_j^*(\cdot)$, satisfies assumption (A.2) since assumption (A.2) is true for X_j, $\forall j \in \mathcal{D}$.

Now we write $\hat{Y} = Y - \hat{m}_1(\mathbf{X}_{\mathcal{C}}) = Y^* + [m_1(\mathbf{X}_{\mathcal{C}}) - \hat{m}_1(\mathbf{X}_{\mathcal{C}})]$. By using the optimal Kernel smoothing estimation, we have that

$$E|\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O(n^{-k_1}),$$

where $k_1 = \frac{\max(r,\alpha)}{2\max(r,\alpha)+d_{\mathcal{C}}}$. By assumption (A.4) $k < \min(k_1, \frac{\max(r,\alpha)}{2\max(r,\alpha)+2})$ is sufficient to assure that Proposition 2 of Chang, Tang, and Wu (2016) holds. So, the proof is complete. □

Remark. The assumption that X_j and $\mathbf{X}_{\mathcal{C}}$ are independent, $\forall j \in \mathbf{X}_{\mathcal{D}}$, is made to simplify the proof of Proposition 1. In fact, we always have that $|\hat{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right) - E\left(X_j|\mathbf{X}_{\mathcal{C}}\right)| = O_p(n^{-1/2})$.

3.4 Simulation study

In the following we study the problems related to feature screening by some simulation examples. Since we are interested in procedures that allow for non additive models, we compare our proposal (CEL) with method in Chang, Tang, and Wu (2016) (EL) and its iterative version. The results are obtained on $R = 100$ simulation replications, and we vary the sample size from 100 to 300 for different scenarios and the number of variables from 500 to 1000.
The bandwidth h is selected by cross-validation and the spherical Epanechnikov Kernel $\mathcal{K}(\mathbf{x}) \propto (1 - \mathbf{x}'\mathbf{x})I(|\mathbf{x}'\mathbf{x}| \leq 1)$ is used. If the conditioning set has cardinality one the Kernel function reduces to the univariate $\mathcal{K}(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$.
We adopt the first four examples in Chang, Tang, and Wu (2016) and also define some alternative cases of equicorrelated covariates with common ρ varying from 0.5 to 0.8.
We compare the different alternatives according to the number of times when each relevant predictor is included by the screening and when the true model is retained (all the active variables are retained apart from conditioning ones) on 100 replications. The data from first example are generated from nonlinear additive model with the first four covariates being relevant and the rest irrelevant. The marginal regression functions for the active predictors are $m_1(x) = x$, $m_2(x) = (2x-1)^2$, $m_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$ and $m_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. All covariates are generate from $U(0,1)$ and the error term is standard normal distributed.
Table 3.1 reports the results for example 1 with $\{n, p\} = \{200, 500\}$ and the

Example 1

S.1	X_1	X_2	X_3	X_4	all
EL	90	98	68	100	59
iterative	100	100	93	100	93
CEL ($\mathcal{C}$={1})	x	100	100	100	100
S.2	X_1	X_2	X_3	X_4	all
EL	98	1	53	2	0
iterative	100	100	100	100	100
CEL ($\mathcal{C}$={1})	x	100	100	100	100

TABLE 3.1: Simulation results on $R = 100$ replications from Example 1 with $\{n, p\} = \{200, 500\}$. In **S.1** results for example 1; meanwhile in **S.2** the covariates are all equicorrelated with $\rho = 0.8$.

covariates generated from independent uniform distribution. Each column shows the number of times when each relevant predictor is retained on 100 replications, while the last column ("*all*") shows the number of times when the true model is retained (in the conditional case the true model is taken without considering the conditioning variables). The second part of table (S.2) refers to the case of high correlated covariates. It is to illustrate that conditional screening procedure has much better performance when there are irrelevant variables that are correlated with relevant ones, in such a case all the covariates are equicorrelated with $\rho = 0.8$.

Results from first example show how conditional version of proposed screening has good performances when the conditioning set is composed by the first of relevant covariates, it still works well when there is high correlation among predictors. In example 2 the model is again nonlinear and additive with first four covariates being relevant with marginal regression functions, $m_1(x) = (2x-1)^2$, $m_2(x) = \cos(2\pi x)/(2+\sin(2\pi x))$, $m_3(x) = \cos(2\pi x)/(2-\cos(2\pi x))$ and $m_4(x) = \cos(\pi(2x-1))$. Covariates are generated from uniform over $[0, 1]$, while we use normal distributed heteroskedastic error term with heterogeneous conditional variance generated as $var(\varepsilon) = 4/(x_1^2 + x_2^2 + x_3^2 + x_4^2)$. As for the first example we consider the first relevant covariate known in advance as conditioning information. The goal of Example 3 is to show how our conditional screening procedure can make it possible to retain hidden relevant predictors. In such a particular case the fourth relevant variables shows small marginal

Example 2

	Homogeneous variance					*Heterogeneous variance*				
S.1	X_1	X_2	X_3	X_4	all	X_1	X_2	X_3	X_4	all
EL	16	100	12	47	3	40	94	44	40	13
CEL($\mathcal{C}$={1})	x	13	86	88	1	x	14	80	89	11
S.2	X_1	X_2	X_3	X_4	all	X_1	X_2	X_3	X_4	all
EL	9	62	83	100	4	44	67	88	88	24
CEL($\mathcal{C}$={1})	x	13	66	65	8	x	30	57	66	15

TABLE 3.2: Simulation results on $R = 100$ replications from Example 2 with $\{n, p\} = \{100, 1000\}$. In **S.2** the covariates are all equicorrelated with $\rho = 0.6$.

Example 3

S.1	X_1	X_2	X_3	X_4	all
EL	84	80	91	0	0
iterative	100	100	100	98	98
CEL ($\mathcal{C}$={1})	x	100	85	86	76
CEL ($\mathcal{C}$={1,2})	x	x	100	100	100
S.2	X_1	X_2	X_3	X_4	all
EL	89	28	55	0	0
iterative	100	100	100	36	36
CEL ($\mathcal{C}$={1})	x	56	88	43	37
CEL ($\mathcal{C}$={1,2,3})	x	x	x	100	100

TABLE 3.3: Simulation results on $R = 100$ replications from Example 3 with $\{n, p\} = \{100, 500\}$. In **S.1** $\beta_4 = -3\sqrt{2}$ is greater than the correlation between X_4 and other covariates; meanwhile in **S.2** all the covariates are equicorrelated with $\rho = 0.6$ and $\beta_4 = 1/3$, so that coefficient of hidden covariate is less than the correlation.

relevance with the response, thus it is not recruited by canonical screening procedures.

Data are generated from a linear model with independent normal distributed errors and true coefficients $\beta_1 = \beta_2 = \beta_3 = 2$, $\beta_4 = -3\sqrt{2}$ and $\beta_j = 0$ for $j > 4$. The covariates are such that $cov(X_j, X_k) = 0.5$ for $j \neq k \in \{1, ..., p\}/\{4\}$ and $cov(X_j, X_4) = 1/\sqrt{2}$ for $j \neq 4$. In such a case covariate X_4 shows null marginal correlation with the response through although it has coefficient β_4 greater than other relevant covariates.

First part (S.1) of Table 3.3 reports the simulation results for example 3 in Chang, Tang, and Wu (2016) while second part (S.2) is setting such that all covariates are equicorrelated with $\rho = 0.6$ and $\beta_4 = 1/3$, so that the coefficient of the hidden variable X_4 is smaller than the correlation exhibits by the the variable with the rest of covariates. Example 4 shows the case of non additive model. Regression function $m(\mathbf{X})$ is generated from $\exp\{-0.5(\beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2)\}$ with $\beta = (1/0.8^2, 1/0.9^2, 1, 1/1.1^2)'$ and $\beta_j = 0$ for $j > 4$. Corresponding results are shown in table 3.4. Simulation results show how standard screening per-

Example 4

	X_1	X_2	X_3	X_4	all
EL	97	75	77	64	39
iterative	87	76	83	91	58
CEL ($\mathcal{C}$={1})	x	97	66	67	46
CEL ($\mathcal{C}$={1,2})	x	x	100	100	100

TABLE 3.4: Simulation results on $R = 100$ replications from Example 4 with $\{n, p\} = \{100, 500\}$.

forms poorly when there exist a hidden predictor in the model. However both the iterative procedure and CEL have excellent performances, moreover CEL has less computational cost. Analysing the number of iterations on the total of simulation runs for the iterative version, we have detected that it variates from 2 to a maximum of 55 with the median at 4 and the third percentile at 9. That is, in 50% of the cases the iterative screening requires at least 4 iterations while our CEL is always computed in two steps.

3.5 Discussion

In this chapter we proposed a variant of screening procedure for non additive models, making use of conditional information in estimation process. Conditional feature screening for linear models was firstly proposed by Barut, Fan, and Verhasselt (2016) and, compared with the iterative screening approaches, it is able to circumvent the issue of hidden relevant predictors. Furthermore, it has the advantage of less computational cost since can be performed in few steps. Our proposal makes use of LPE for measuring the marginal contribution,

similarly to the work of Chang, Tang, and Wu (2016). In addition, we introduced information from a known set of relevant predictors and provided an extension to the standard procedure by two-steps estimation. Through a simulation study we showed how such a strategy works as well as the iterative-based approaches, and it results useful in situations where the standard screening approaches fail. Moreover, it also has the advantage of a computational cost comparable to the one of standard feature screening methods.

Part II

High Dimensional Spatio-temporal Models

Chapter 1

Spatio-temporal Models

The main purpose of first chapter is to briefly introduce principal elements of spatio-temporal models in statistical and econometric frameworks. At beginning of the chapter, we report some notion in spatial econometrics analysis, while in the second part, we widely review spatio-temporal models and in particular the class of spatio-temporal panel data models, the focus of interest of this part. Together with the presentation of the models and the distinction between static and dynamic types, also the main estimation techniques are discussed. At the end of the chapter, we argue about the theoretical advantage and disadvantage of considering such models in high dimensional settings.

1.1 Spatial Econometrics

The Spatio-temporal models belongs to the more general class of spatial econometrics techniques and represent one of the most recent developments in such topic. Following Anselin's definition, we can consider the *Spatial Econometric* as 'the field of spatial econometrics to consist of those methods and techniques that, based on a formal representation of the structure of spatial dependence and spatial heterogeneity, provide the means to carry out the proper specification, estimation, hypothesis testing and prediction for models in regional science' (see Anselin (2013), pag.10). The definition of *regional science models* has to be understood as those specifications that incorporate explicitly spatial interactions among individuals. The term *spatial* is not only referred to geographical space, it may have several understanding (i.e. economics meaning, among others). The spatial econometrics has been recently extend to panel data modelling. The advantage of panel data is that by using information about inter-temporal and

individual dynamics, it is possible to control for the effects of unobserved or missing variables. In spatial panel data models, the data analysis has to take into account of spatial dependence among different locations, but also that the observations at each location typically are not independent but present serial dependence. Broadly speaking, one must take into account of temporal (auto-) correlations as well as spatial (cross-) correlations.
Following, we shortly give some useful definitions in spatial econometrics, particularly we focus on the main definition of spatial effects, spatial weights, lags and spatial errors. Section 1.2 is devoted to introduce spatio-temporal models, while in section 1.3 the main estimation procedures for dynamic spatio-temporal models are presented.

1.1.1 Spatial effects

When the data are collected and organized by spatial units, the definition of some spatial dependence represent a crucial problem. Spatial dependence can be viewed as the functional relationship existing between one point in the space and everything elsewhere. Instead, spatial heterogeneity can be though as the 'set' of different aspects of each unity in the space coming from several factors related to the concept of space. That said, spatial econometrics deals with the incorporation of effects (spatial effects) that result from the above concepts in econometric modelling. Thus, spatial effects may results from spatial dependence or spatial heterogeneity, alternatively. In the first case the dependence structure is somehow related to the concept of location and distance, both in geographical sense or in a more general sense (economics, social, etc.). The second case, spatial heterogeneity, is meaning as a special case of the observed or unobserved heterogeneity among unities treated in panel data econometrics. It became *spatial* when the variability across two distinct units i and j is driven by spatial variables, such as distance or region (Anselin (2013)).

1.1.2 Neighbourhood and nearest neighbours

The notion of spatial dependence implies the determination of some functional relationship between units in the space, namely which unity has influence on

another one and vice versa. Formally, this is expressed through the definition of *Neighbourhood* and *Neighbour* in topological field.
The original notion of set of *neighbours* comes from the literature on *Geostatistics* and *lattice* processes. Historically, the former is considered the traditional approach to spatial analysis and it pertains with theory of stochastic processes indexed over continuous space of spatial locations, the latter assumes the stochastic process being indexed over sets of countable collection of locations. Since a *lattice* recalls a regularly spaced points set, it is thought as the spatial analogue of time series.
Considering a spatial process of the form $\{Z(s) : s \in D\}$, where D is the set of all the locations, a set J of neighbours for a spatial unit i can be thought as collection of those locations that are in the conditional probability of the process at i, formally

$$J = \{j : \mathbb{P}(z(i)) \neq \mathbb{P}(z(i)|z(j))\}, \quad \forall i \in D. \tag{1.1}$$

Moreover, considering some distance metric d, the set of neighbours can be expressed more generally as

$$J = \{j : \mathbb{P}(z(i)) \neq \mathbb{P}(z(i)|z(j)),\, d_{ij} < \varepsilon_i\}, \quad \forall i \in D \tag{1.2}$$

where d_{ij} measures the distance between i and j in a properly structured space and ε_i is a cut-off point for the spatial unit i. This definition of *neighbourhood* combines the notion of statistical dependence, shared by conditional probability, and notion of spatial dependence thorough the distance measure d_{ij}. More precisely, when a location j meets the distance criterion and the conditional influence together, it is said to be *nearest neighbour*. Conversely, when some location j does not meet the distance criterion it is considered as higher order neighbour.
The resulting set of neighbours of each location i can be represented in several ways, for instance graph or network structure. For our purpose it is useful to refer to the concept of spatial weights matrix as representation of neighbourhood.

1.1.3 Spatial weights

The spatial weights matrix $\mathbf{W}$ is a $p \times p$ positive matrix in which each element w_{ij} represent the interaction and/or relationship between cross-sectional units i and j. Originally, spatial dependencies were measured by binary contiguity between units. It is refer to the simplest case where the weights in the matrix are binary, so $w_{ij} = 1$ when i and j are neighbours, and $w_{ij} = 0$ when they are not. By convention, the diagonal elements are 0 to exclude self-neighbours. To increase the interpretation of the spatial variables and for computational purpose, the weights are almost always standardized such that row-elements of the matrix sum to 1. Often the weights matrix is symmetric, while in the so called *row-standardized* form is no longer, gives further computational complications. To extend their use in panel data econometric, the weights are assumed to remain constant over time.

Estimation of spatial weights matrix

The choice of spatial matrix $\mathbf{W}$ is crucial in spatial modelling as the weights share part of the structure of spatial dependence. The literature about adequate formulation and choice of the spatial weights shows a huge amount of works where the spatial weights matrix is assumed exogenous. In such a situation, the choice is typically driven by distance or geographic criteria while, generalizations of distance that include economic notions are only recently and increasingly used. Under exogeneity of spatial dependence, consistency and asymptotic normality of the estimation methods for the spatial models are well established in literature (see paragraph 1.3). When the spatial structure is assumed endogenous, $\mathbf{W}$ need to be estimated in some way.
As concern the estimation of matrix $\mathbf{W}$, Meen (1996) proposed a procedure consisting of two-stage regression where the second stage provides that the residuals of ordinary least squares (OLS) for each location are regressed on residuals of all other locations. Formally

$$\begin{aligned} \text{(1st)} \quad & \mathbf{y}_t = \mathbf{X}_t\beta_0 + \varepsilon_t, \\ & \hat{\varepsilon}_t = \mathbf{y}_t - \mathbf{X}_t\hat{\beta}, \\ \text{(2nd)} \quad & \hat{\varepsilon}_{tk} = \rho_k \sum_{j \neq k} w_{jk}\hat{\varepsilon}_{tk} + \eta_{tk} \end{aligned} \tag{1.3}$$

In first-stage estimation $\hat{\varepsilon}_t$ are computed, then they are used for estimating the weights. Second-stage regression produces inconsistent estimates of the regression coefficients w_{jk}, which are then used to construct the spatial weights matrix. Moreover, if the dimension p exceed the length T of time series there are no degree of freedom to estimate $\mathbf{W}$.
Bhattacharjee and Jensen-Butler (2013) proposed a method for estimating $\mathbf{W}$ from the autocovariance of the process and showed that the matrix $\mathbf{W}$ is only partially identified by the autocovariance matrix $\mathbb{E}(\varepsilon\varepsilon')$, thus additional structural assumption are needed to uniquely identified $\mathbf{W}$ by covariance matrix. In particular, the spatial weights matrix is fully identified if it is assumed symmetric.
Kelejian and Piras (2014) adopt the IV technique to estimate endogenous weights matrix. Qu and Lee (2015) try to explicitly model the source of endogeneity of spatial structure by two sets of equation, one that define the outcome for each unit linked to the outcome of all other units, and one that models random variable from which the spatial weights depend. The estimation is a two-stage procedure and it makes using of IVs.
Ahrens and Bhattacharjee (2015) propose a lasso-type procedure again developed in two-stages for dimension reduction in IV estimation. It results useful when the number of endogenous regressors and the number of IVs are larger than the number of observations.

1.1.4 Spatial lags

In spatial econometrics, spatial lags are constructed by weighted average of the neighbouring observation with the specified weights matrix $\mathbf{W}$. Using spatial lags produces new variables that can be included into a model specification. Spatial lag operator can be applied to dependent variable, explanatory variables or error terms. It corresponds to matrix operation $\mathbf{Wy}$, in the case of dependent variable, in which the $p \times p$ weights matrix is post-multiplied by the $p \times 1$ vector of cross-sectional observations. Introducing this type of lags into the model results in a spatial lag model

$$\mathbf{y}_t = \lambda \mathbf{W}\mathbf{y}_t + \varepsilon_t, \quad t = 1, \ldots, n \tag{1.4}$$

where the spatial lag term typically gives the interpretation of the spatial interaction process in which each value of the dependent variable is jointly determined by that of the neighbouring units.
The main problem encountered in estimation of models including spatial lags is endogeneity. The presence of spatial lags introduces joint dependence between **Wy** and ε at each cross-section. In model estimation, this simultaneity must be accounted, usually through instrumental variables (IV) or by assuming distributional model, as in maximum likelihood estimation. Another way of dealing with spatial lag is to detrend data by some spatial filter, thus a new dependent variable is obtained and the effect of spatial autocorrelation has been eliminated.

1.1.5 Spatial errors

A spatial error specification does not require any theoretical representation of spatial interaction among units, instead, it consists of specifying a parsimonious covariance structure in order to account for spatiality in the residuals. A standard representation of the spatial error model is

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\beta_0 + \varepsilon_t, \\ \varepsilon_t &= \rho\mathbf{W}\varepsilon_t + \eta_t. \end{aligned} \tag{1.5}$$

Cliff and Ord (1973) defined the above model as spatial autoregressive (SAR), since it is a spatial specification analogous to the well-known Box-Jenkins' approach for time series analysis. The spatial structure is driven by the spatial weights matrix and the autocorrelation parameter ρ.
Alternatively, the error can be modelled as spatial moving average process SMA (Anselin (1988))

$$\varepsilon_t = \mathbf{W}\zeta_t + \zeta_t \tag{1.6}$$

where ζ_t is a vector of independently distributed random terms.
As an alternative to the SAR and SMA, Kelejian and Robinson (1995) suggest a spatial error components specification in which the error term is decomposed into a local and a spillover effect.

1.1.6 Temporal and Spatial Heterogeneity

Both temporal and spatial heterogeneities are treated in a familiar way as in panel data analysis, by considering fixed or random effects. In the general case, the fixed effects treatment is attained by introduction of time-specific intercept and/or slopes, while the random effects treatment provides incorporation of a random type factor. Usually temporal heterogeneity as well as spatial heterogeneity is evidence for lack of uniformity in the effects space. While the temporal one consist in non constant error variances through time, spatial heterogeneity is related to the lack of uniformity of the effects for the different locations due to spatial dependence. Usually, in econometric analysis this lack can be carries out by considering explicitly varying parameters, random coefficients of other various form of structural change. Apart from the lack of structural stability over the space, generally in spatial phenomena all units of observation are far from homogeneity.

1.2 Spatio-temporal models

Spatio-temporal models allows for dependencies in both time and space dimensions. Under such specification, there is additional order of difficulty, e.g. in identification of the $pn \times (pn-1)/2$ elements of covariance matrix. The class of spatial panel data models has affected a very rapid development of research over the last decade. Classical representations of spatio-temporal models can be fund in Baltagi, Song, and Koh (2003) and Baltagi et al. (2007) and Kapoor, Kelejian, and Prucha (2007).

1.2.1 Spatial Static Panel Data models

Spatial panel data models can be specified in both static and dynamic cases. The former does not incorporate any temporal dependency of the dependent variable, meanwhile, the latter usually provides the lags of the dependent variables as explanatory variables in order to take into account serial dependence of the dependent variable. Following Lee and Yu (2010c) we can define a static spatial panel data model (SSPD) as

$$\begin{aligned} \mathbf{y}_t &= \lambda_1 \mathbf{W}_1 \mathbf{y}_t + \mathbf{X}_t \beta_0 + \mu + \varepsilon_t, \quad t = 1, \ldots, n \\ \mu &= \lambda_3 \mathbf{W}_3 \mu + c_0, \\ \varepsilon_t &= \lambda_2 \mathbf{W}_2 \varepsilon_t + \eta_t \quad t = 1, \ldots, n \end{aligned} \tag{1.7}$$

where $\mathbf{y}_t$ is a $p \times 1$ column vector of observations of each locations at time t, ε_t is $p \times 1$ vector of error terms and η_t has elements *i.i.d.* across i and t, zero mean and finite variance. $\mathbf{W}_j$, $j = 1, 2, 3$ represent the spatial weights matrices, $\mathbf{X}_t$ is an $p \times k$ matrix of regressors, μ denotes the unobserved individual time-invariant effects. The spatial weights matrices $\mathbf{W}_j$ are $p \times p$ and positive. Each nonzero element of such matrices indicates whether two locations are neighbours. Hence, each element indicates the intensity of the relationship between cross sectional units, meanwhile diagonal elements are set to zero to exclude self-neighbours. The model in (1.7) is a general representation of SSPD that accounts for spatial lags and regression through the terms $\lambda_1 \mathbf{W}_1 \mathbf{y}_t$ and $\mathbf{X}_t \beta_0$, respectively. In practice, panel data analysis deals with the choice between fixed effects or random effects specification. In the former, each location would have its own functional specification, whereas in the latter, all locations are assumed to obey the same encompassing model and individual characteristics are specified as random deviations from the overall mean. The general model in (1.7) can assume both of the above specifications of spatial effects, accordingly if the elements of μ are treated as fixed or random. In practice, fixed effects can be estimated by a direct approach or indirectly by eliminate individual effects before estimation. The first approach yields consistent estimations except for the residual variance, when the time series length n is small. The second approach uses methods of conditional likelihood when a sufficient statistic of the fixed effects can be found and follows a data transformation in order to eliminate fixed effects before estimation of the model.

Models in Baltagi, Song, and Koh (2003) and Baltagi et al. (2007) are special cases of (1.7) with $\lambda_1 = 0$, i.e. without spatial lags. They formulate a representation which take into account spatial correlations in both individual (μ) and error terms (ε) assuming different parametrization (i.e. $\mathbf{W}_3 \neq \mathbf{W}_2$). In particular, the model in Baltagi, Song, and Koh (2003) is a pure spatial error model allowing

for spatial dependence in the error term

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\beta_0 + \varepsilon_t, \\ \varepsilon_t &= \mu + \lambda\mathbf{W}\varepsilon_t + \eta_t, \quad t = 1, \ldots, n \end{aligned} \tag{1.8}$$

while Baltagi et al. (2007), in addition, allows for serial correlation in the error terms

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\beta_0 + \varepsilon_t, \\ \varepsilon_t &= \mu + \lambda\mathbf{W}\varepsilon_t + \eta_t \\ \eta_t &= \rho\eta_{t-1} + e_t, \quad t = 1, \ldots, n \end{aligned} \tag{1.9}$$

Kapoor, Kelejian, and Prucha (2007) consider a different specification where the individual components are accounted into the error term. This specification results in the same parametrization of spatial correlations for both individual and error components. It is equivalent to model (1.7) with $\mathbf{W}_3 = \mathbf{W}_2$ and $\lambda_3 = \lambda_2$.

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X}_t\beta_0 + \varepsilon_t, \\ \varepsilon_t &= \lambda\mathbf{W}\varepsilon_t + \eta_t, \quad t = 1, \ldots, n \end{aligned} \tag{1.10}$$

Under fixed effects specification, both of the above models reduce to the same representation.

As concern the random effects specification, the estimation will be more efficient if the individual (random) effects are independent of the exogenous regressors. In fact, assumption of independence between individual effects and the regressors in $\mathbf{X}_t$ is crucial to achieve efficient estimation of regression parameters.

In addition to the individual effects, SSPD can also include the time fixed effects from panel data. For instance, when n is short, they can be treated as regressors in the model. Similar to the individual effects, the time effects can be directly estimated or indirectly treated before estimation, when they are assumed fixed. As will be discussed later, even when both p and n are large such that both individual and time fixed effects can be consistently estimated, a permanent bias occurs in the MLE of common parameters. Hence, it is desirable to treat and eliminate the effects before estimation.

1.2.2 Spatial Dynamic Panel Data models (SDPD)

Differently from SSPD, Spatial Dynamic Panel Data models (SDPD) can take into account dynamic (time) effects. In other words, this kind of models are useful when the dependent variable depends on its own past realizations. One way to take into account dynamic effects is to use time lag terms of dependent variables as explanatory variable. Following Anselin (2013), SDPD can be divided in four categories. If the time dependence pertains only to neighbouring locations we obtain the *pure space recursive model*

$$\mathbf{y}_t = \lambda_0 \mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\beta_0 + \varepsilon_t, \quad t = 1, \ldots, n \tag{1.11}$$

where $\mathbf{y}_t = (y_{1t}, y_{2t}, ...y_{pt},)'$, $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, ..., \varepsilon_{pt})'$ are $p \times 1$ column vector of response observations and error terms at time t, respectively, λ_0 is the regression parameter capturing the time-spatial effect, β_0 is a $(k \times 1)$ vector of parameters and $\mathbf{X}_t$ a $(p \times k)$ matrix of exogenous regressors.
If the time dependence is related to both the location itself, as well as its neighbours at previous periods, we have the *time-space recursive model*

$$\mathbf{y}_t = \lambda_0 \mathbf{y}_{t-1} + \lambda_1 \mathbf{W}\mathbf{y}_{t-1} + \mathbf{X}_t\beta_0 + \varepsilon_t, \tag{1.12}$$

with λ_0 capturing the relation between different locations over time.
When an individual time lag and a contemporaneous spatial lag are included we obtain the *time-space simultaneous model*

$$\mathbf{y}_t = \lambda_0 \mathbf{y}_{t-1} + \lambda_1 \mathbf{W}\mathbf{y}_t + \mathbf{X}_t\beta_0 + \varepsilon_t. \tag{1.13}$$

Instead, the general SDPD model is obtained when all type of lags are included (Lee and Yu (2010c))

$$\begin{aligned} \mathbf{y}_t &= \lambda_0 \mathbf{W}\mathbf{y}_t + u_{t-1} + \mathbf{X}_t\beta_0 + \mu + \alpha_t I + \varepsilon_t \\ u_{t-1} &= \lambda_1 \mathbf{y}_{t-1} + \lambda_2 \mathbf{W}\mathbf{y}_{t-1}. \end{aligned} \tag{1.14}$$

The ε_{it} are i.i.d. across i and t with zero mean and constant variance σ^2 (hypothesis of homoskedasticity and independence across time and cross-section in the error term). The spatial weight matrix $\mathbf{W}$ is known and determines spatial

dependence between units, $\mu = (\mu_1, \mu_2, ..., \mu_p)'$ is the vector of fixed individual effects and $\alpha_t I$, with I $p \times p$ identity matrix, accounts for time effects. The model in 1.14 includes spatial lags and exogenous regressors, as in the SSPD model, and in addition takes into account dynamic effects throughout components $\lambda_1 \mathbf{y}_{t-1}$ and $\lambda_2 \mathbf{W}\mathbf{y}_{t-1}$. More precisely, λ_1 captures the pure dynamic effect, while λ_2 captures the spatial-time effect.

Lee and Yu (2010c) classify the above SDPD model into different cases depending on the structure of eigenvalues of matrix $\mathbf{A}$ from the following reduced form

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{S}^{-1}\mathbf{X}_t\beta_0 + \mathbf{S}^{-1}c_0 + \alpha_t\mathbf{S}^{-1}I + \mathbf{S}^{-1}\varepsilon_t \tag{1.15}$$

with $\mathbf{S} = (I - \lambda_0\mathbf{W})$ and $\mathbf{A} = \mathbf{S}^{-1}(\lambda_1 I + \lambda_2\mathbf{W})$.
The *stable case* occurs when all the eigenvalues of $\mathbf{A}$ are smaller than 1, when some of the eigenvalues of $\mathbf{A}$ are equal to 1 we have the *spatial cointegration case*, meanwhile the *explosive case* occurs when some of them are greater than 1.

1.3 Estimation of SDPD

Estimation of spatio-temporal models in the dynamic case, can be done by maximum likelihood (ML) or generalized method of moments (GMM). Since the model incorporate lags of dependent variable, the individual effects μ and the dependent variable $\mathbf{y}_t$ are not independent. Furthermore, due to the lag of dependent variable endogeneity must be deal with. Direct estimation of fixed effects introduces a problem of incidental parameters, this is because the introduction of fixed effects increases the number of parameters to be estimated. Thus, direct estimation of individual effects produces biased and inconsistent QML estimate when the temporal dimension n is fixed. As an alternative, individual effects can be ruled out by data transformation or GMM estimation by instrumental variables can be adopted in order to achieve consistent but biased estimation. The incidental parameters issue becomes less severe in MLE when both p and n go to infinity, in fact in scenarios of time increasing, MLE are consistent but remain biased.

1.3.1 GMM and IV

Probably the most common used estimation approach in practice is the GMM with instrumental variables (IV). Such a method yields consistent estimation for the SDPD model. Basically it uses lagged values and exogenous variables in order to construct the moment conditions. Once the entire transformed system is obtained, the error terms result uncorrelated across i and t. However, as the SDPD model allows for lags of the response variable, the lagged terms on the right side of (1.14) ($\mathbf{y}_{t-1}$ and $\mathbf{W}\mathbf{y}_{t-1}$ transformed) still remain correlated with disturbance. Thus, a set of IV is needed. After introduction of instruments, the moment conditions are used for a generalized two-stage least squares estimation.

Anderson and Hsiao (1981) showed that IV method can improve the efficiency of estimator in dynamic models. They first eliminate fixed effects by data transformation and then construct a new system by instrumental variables. When n is fixed, IV technique improves the efficiency of estimation. However, if the number of instruments is too large, the asymptotic bias, due to IV introduction, increases. Moreover, contrary to the ML estimation that requires larger n, GMM can be adopted when n is small, because after elimination of individual effects it does not suffer from bias of order $O(1/n)$ (Arellano and Bond (1991)).

1.3.2 MLE

Consider the simple dynamic model (1.14) as in Lee and Yu (2010c), the log-likelihood function can be expressed as

$$l(\theta,\mu) = -\frac{pn}{2}\log 2\pi - \frac{pn}{2}\log\sigma^2 + n\log|\mathbf{S}| - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\mathbf{V}_t'\mathbf{V}_t, \tag{1.16}$$

where $\theta = (\delta', \lambda_0, \sigma^2)$, $\delta = (\lambda_1, \lambda_2)$, $\mathbf{S} = I - \lambda_0\mathbf{W}$ and $\mathbf{V}_t = (\mathbf{S}\mathbf{y}_t - \mathbf{z}_t\delta - \mu)$, with $\mathbf{z}_t = (\mathbf{y}_{t-1}, \mathbf{W}\mathbf{y}_{t-1})$.
For the stable case (i.e. when $\lambda_0 + \lambda_1 + \lambda_2 < 1$), QMLE $\hat{\theta}$ and $\hat{\mu}$ are derived from the optimization of the above equation.
When the number of parameters tends to infinity, it is convenient to concentrate μ using the first order condition and focus on the likelihood function for the parameter θ. In this way, the parameters space does not change with p and n,

as in the standard model the regression parameters are assumed constant. Given the first order condition $\frac{l(\theta,\mu)}{\mu} = \frac{1}{\sigma^2}\sum_t \mathbf{V}_t$ and $\hat{\mu} = \frac{1}{n}\sum_t(\mathbf{S}\mathbf{y}_t - \mathbf{z}\delta)$, the concentrate estimator of μ given θ, the concentrate log-likelihood function for θ is

$$l(\theta) = -\frac{pn}{2}\log 2\pi - \frac{pn}{2}\log\sigma^2 + n\log|\mathbf{S}| - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\tilde{\mathbf{V}}_t'\tilde{\mathbf{V}}_t, \tag{1.17}$$

where $\tilde{\mathbf{V}} = \mathbf{S}\tilde{\mathbf{y}}_t - \tilde{\mathbf{z}}_t\delta$. The asymptotic analysis in Yu, De Jong, and Lee (2008) requires the following assumptions

1. $\mathbf{W}$ is constant with $w_i i = 0,\ i = 1, ..., p$;
2. ε_{it} are i.i.d. across i and t, zero mean, variance σ^2 and $\mathbb{E}|\varepsilon_{it}|^{4+\eta} < \infty$ for some $\eta > 0$;
3. $\mathbf{S}$ is invertible for all $\lambda_0 \in \Lambda$, and Λ is compact and the true parameter is an interior point of it;
4. the regressors in $\mathbf{X}$ are nonstochastic, uniformly bounded and the limit of $1/np\sum_t \mathbf{X}'\mathbf{X}$ exists and is nonsingular;
5. $\mathbf{W}$ and $\mathbf{S}^{-1}$ are uniformly bounded in row and column sums in absolute value (UB);
6. $\sum_{h=1}^{\infty}|\mathbf{A}^h|$ is UB, where $\mathbf{A}$ is as in the reduced form (1.15);
7. p is nondecreasing in n and n goes to infinity.

The first four assumptions are quite standard in spatial modelling. Assumption 5 limits the spatial correlations through the spatial weights in order to make the asymptotics handleable. Assumption 6 is of more interest, it relies on the absolute summability in row and column of the matrix $\mathbf{A}$ in the reduced equation (1.15). Essentially, this assumption limits the dependence between serial correlation and cross sectional units. In practice, imposing limited dependence across locations or time series effectively rules out some special cases. For instance, the required assumption is not checked in the aforementioned *spatial cointegration case*, instead it only occurs in *stable case*. Motivated by this, Yu, Jong, and Lee (2012) extend the previous results to the *spatial cointegration case*.

Yu, Jong, and Lee (2012) also show the GMM estimator and its properties for dynamic models. GMM estimates for SDPD are consistent and asymptotically

normal. Compared with MLE, GMM does not suffer from asymptotic bias. However, when n is large so that the bias of MLE vanishes, the asymptotic variance of MLE is smaller that that of GMM.

1.3.3 ML and individual effects estimation

Lee and Yu (2010b) investigate the asymptotic properties of QMLE for models with spatial lags and fixed effects specification. The model is (1.14) with u_{t-1} and α_t equal to zero and ε_t following a SAR process in (1.5). They compare the asymptotics of the estimation including fixed effects (direct approach) and those of estimation after elimination of the fixed effects (transformation approach). The estimates of the parameters λ_0 and ρ, are properly centred at their true values, but the estimate of variance in direct approach may not be centred at 0, even though both n and p tend to infinity, unless p/n goes to zero. For the transformation approach, the estimated variance is unbiased even with finite n and growing p.
In addition, they study the properties when also time effects are included ($\alpha_t \neq 0$). The direct and transformation approaches do not yield the same estimate for λ_0 and γ, even though they are both consistent. For the direct approach, consistency is achieved even when p is fixed, while the estimate of variance requires both p and n goes to infinity. For the transformation approach, all the estimates are consistent when n is small. Also Elhorst (2005) considers estimation of model as in Lee and Yu (2010b), but only focusing on the case of n fixed and treating the fixed effects before estimation by first difference elimination. On the contrary, Su and Yang (2015) derive asymptotics even with random effects specification. They showed that, when n is finite and fixed effects are treated directly, MLE is biased and inconsistent.

1.3.4 Generalized Yule-Walker estimator

Dou, Parrella, and Yao (2016) proposed a new estimator for the following SDPD model

$$\mathbf{y}_t = D(\lambda_0)\mathbf{W}\mathbf{y}_t + D(\lambda_1)\mathbf{y}_{t-1} + D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1} + \varepsilon_t \tag{1.18}$$

Compared with the model in Yu, De Jong, and Lee (2008) it allows the scalar coefficients λ_i to be different for each location. In fact $D(\lambda_i) = diag(\lambda_{i1}, \lambda_{i2}, .., \lambda_{ip})$ for $i = 0, 1, 2$ are the matrices of the unknown coefficient parameters, differently from the model (1.14) where they are considered equal for each location. As before, $D(\lambda_0)$ captures the pure spatial effect, $D(\lambda_1)$ captures the pure dynamic effect and $D(\lambda_2)$ the time spatial-time effect. No endogenous regression components are included in the model.
The extension to a model with different scalar coefficients is motivated by practical situations, in which empirical evidence shows how considering constant effect for each location can be limiting. Nevertheless, considering more parameters in the model leads to inconsistent MLE. Thus, the authors proposed a simple and new method of estimation based on Yule-Walker estimator. They propose to apply least squares to each of individual rows of a Yule-Walker equation system obtained by the model (1.18) as

$$(\mathbf{I} - D(\lambda_0)\mathbf{W})\Sigma_1 = (D(\lambda_1) + D(\lambda_2)\mathbf{W})\Sigma_0 \tag{1.19}$$

where $\Sigma_k = Cov(\mathbf{y}_{t+k}, \mathbf{y}_t)$, $k \geq 0$. The matrices Σ_0 and Σ_1 are replaced by the respective sample counterparts. Each row of the above equation is a system of p linear equations and the parameters $(\lambda_{0j}, \lambda_{1j}, \lambda_{2j})$, $j = 1, .., p$ are estimated by least squares. The resulting estimator has the closed form

$$(\hat{\lambda}_{0j}, \hat{\lambda}_{1j}, \hat{\lambda}_{2j}) = (\hat{\mathbf{X}}_j'\hat{\mathbf{X}}_j)^{-1}\hat{\mathbf{X}}_j'\hat{\mathbf{Y}}_j, \tag{1.20}$$

where $\hat{\mathbf{X}}_j = (\hat{\Sigma}_1'\omega_j, \hat{\Sigma}_0 e_j, \hat{\Sigma}_0\omega_j)$ and $\hat{\mathbf{Y}}_j = \hat{\Sigma}_1 e_j$. e_j is a p-dimensional null vector with jth element equal to 1 and w_j corresponds to the jth row of matrix $\mathbf{W}$. Unfortunately, this estimator suffers of increasing parameters, in fact when $n/\sqrt{p} \to \infty$ it admits nonstandard convergence rate. The authors propose to restrict the number of equations to be estimate in order to restore the usual rate of convergence. Such a restriction is perform keep only those rows for which the contribution in estimation is not too small. The contribution is measured in term of correlation between $y_{k,t-1}$ and $(\omega_j'\mathbf{y}_t, y_{j,t-1}, \omega_j'\mathbf{y}_{t-1})$ by $\rho_j^{(i)} = |e_k'\Sigma_1'\omega_j| + |e_k'\Sigma_1 e_j| + |e_k'\Sigma_0\omega_j|$. The idea is that, when $\rho_k^{(j)}$ is small, say close to zero, the k-th equation carries little information in the estimation and can be ruled out. This thresholding results in transformation of the starting overdetermined scenario

in a new one where the estimation can be done consistently. The new system will have a restricted number of equations.
The asymptotic analysis in Dou, Parrella, and Yao (2016) requires the following assumptions

1. **W** is known with $w_{jj} = 0,\ j = 1, ..., p$;
2. the error term ε_t satisfies $Cov(\mathbf{y}_{t-1}, \varepsilon_t) = 0$;
3. the process $\mathbf{y}_t$ is α-mixing with the mixing coefficient $\alpha(k)$ satisfying $\sum_{k=1}^{\infty} \alpha(k)^{\frac{\gamma}{4+\gamma}} < \infty$ for some constant $\gamma > 0$;
4. for $\gamma > 0$ in the above assumption, it holds $\sup_p \mathbb{E}|w_i'\Sigma_0\mathbf{y}_t|^{4+\gamma} < \infty$, $\sup_p \mathbb{E}|w_i'\Sigma_1\mathbf{y}_t|^{4+\gamma} < \infty$, $\sup_p \mathbb{E}|e_i'\Sigma_0\mathbf{y}_t|^{4+\gamma} < \infty$, $\sup_p \mathbb{E}|w_i'\mathbf{y}_t|^{4+\gamma} < \infty$, $\sup_p \mathbb{E}|e_i'\mathbf{y}_t|^{4+\gamma} < \infty$;
5. the rank of matrix $(\Sigma_1' w_i, \Sigma_0 e_i, \Sigma_0 w_i)$ is equal to 3;

Dou, Parrella, and Yao (2016) show the asymptotics of their estimator in both case of p fixed and growing. When p is fixed the estimator is $\sqrt{n}$ consistent. When p is diverging, the standard $\sqrt{n}$ rate is achieved as long as p grows less than $\sqrt{n}$, while convergence rate may be slower if p is of higher order of $\sqrt{n}$. The estimator is asymptotically normal as long as the number of restricted equations in the estimation is $o(\sqrt{n})$.

1.3.5 Estimation with increasing dimensions

When both p and n go to infinity, the issue of increasing parameters in MLE becomes less stringent and efficiency of estimator improves. For the stable case Yu, De Jong, and Lee (2008) show the properties of estimator in setting both p and n going to infinity. The spatial cointegration case is elegantly treated in Yu, Jong, and Lee (2012). Meanwhile, the explosive case is more difficult to handle. In fact, when some eigenvalues of matrix **A** are greater than 1, it might be difficult to obtain good estimates, as asymptotic properties of ML in such a case are unknown. Moreover, such situations can be treated by data transformation in order to eliminate the unstable components.
For panel data models, Alvarez and Arellano (2003) showed that the bias of MLE of autoregressive parameters is of order $O(1/n)$ when p and n grow proportionally. Hahn and Kuersteiner (2002) extended the work of Alvarez and

Arellano (2003) introducing a bias corrected estimator. Meanwhile in the context of spatio-temporal models, Yu, De Jong, and Lee (2008) provided a rigorous asymptotic theory that covers several high-dimensional scenarios, as the one where p may grow faster than n, and vice versa. They consider model (1.14) with no temporal effects ($\alpha_t = 0$), fixed individual effects and i.i.d error term. Under some specific assumptions, reported in the section 1.3.2, they prove that the MLE of common parameters is consistent but has an asymptotic bias of order $O(1/n)$, as in panel data model of Alvarez and Arellano (2003). More specific, when n grows both proportional to p or faster than p, the estimators are $\sqrt{pn}$ consistent, but in the case of n increasing proportional to p, the limit distribution is not centre and bias occurs; instead when p is relatively large with respect to n, the estimator are n consistent and have a degenerate limit distribution. They also prove how to obtain bias-corrected estimator achieving $\sqrt{pn}$ consistency even when p/n goes to infinity.
Note that, the model in (1.14) can be regarded as a vector autoregressive (VAR(1)) process with existence of cointegration relationship among locations. Thus, the SDPD in Yu, De Jong, and Lee (2008) assumes this configuration and its cointegrating space is completely known. In fact, it is determined by the spatial weight matrix $\mathbf{W}$ in equation (1.14). This assumption is rather enforced compared with classical cointegration time series analysis where the primary purpose is inference about $\mathbf{W}$. However, here the dimension of process in (1.14) can be large and asymptotically tends to infinity, while in the canonical VAR(1) it is fixed and relatively small, hence it can be of particular interest in studying high dimensional contexts.

Chapter 2

Testing different structures of Spatial Dynamic Panel Data models

2.1 Introduction

The spatio-temporal models have affected a very rapid development of research in econometric field. Classical representation of spatio-temporal model can be found in Baltagi et al. (2007), Kapoor, Kelejian, and Prucha (2007), Lee and Yu (2010b), Lee and Yu (2014), and Yu, De Jong, and Lee (2008).
Over the last decade, it has been developed a particular class of models for spatio-temporal data analysis, the spatial dynamic panel data models (SDPD). In particular, several versions of the SDPD model have been proposed, based on different assumptions on the spatial parameters and different properties of the estimators. The standard version of the model assumes that the spatial parameters are constant over location. Two are the most common methods developed to estimate standard SDPD. One of them is by maximum likelihood (MLE) or quasi-maximum likelihood (QML) estimators, the other method is based on instrumental variables and generalized method of moments (IV/GMM). Yu, De Jong, and Lee (2008) constructed a bias-corrected estimator for dynamic model with spatial fixed effects, Lee and Yu (2010a) extend this study to include time-period fixed effects. Another recently proposed version, called *generalized SDPD*, assumes that the spatial parameters are adaptive over location.
The assumption of different scalar coefficients is motivated by practical situations, in which empirical evidence shows how considering constant effect for each location can be limiting. Nevertheless, considering an increasing number of parameters in the model leads to inconsistent MLE. Therefore, generalized

model is usually estimated by different method based on Yule-Walker estimator. In this work we propose a strategy for testing the particular structure of the spatial dynamic panel data model, by means of a multiple testing procedure that allows choosing between the generalized version of the model and some specific versions derived from the general one by imposing particular constraints on the parameters. The multiple test is made by the Bonferroni technique and the distribution of the multiple test statistic is derived by a residual bootstrap resampling scheme.

2.2 Model

Consider a multivariate stationary process $\{\mathbf{y}_t\}$ of order p generating the observations at time t from p different locations. The following model

$$\mathbf{y}_t = D(\lambda_0)\mathbf{W}\mathbf{y}_t + D(\lambda_1)\mathbf{y}_{t-1} + D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1} + \varepsilon_t, \quad t = 1, \ldots, n. \tag{2.1}$$

where $\mathbf{y}_t = (y_{1t}, y_{2t}, ...y_{pt},)'$, is the *generalized* SDPD and has been proposed by Dou, Parrella, and Yao (2016) as generalization of the spatial dynamic panel data model of Yu, De Jong, and Lee (2008). The errors ε_t are serially uncorrelated, they have zero mean value and may show cross-sectional correlation and heteroskedasticity, which means that ε_t have a full variance/covariance matrix Σ_ε; the *spatial matrix* $\mathbf{W}$ is a weight matrix with zero main diagonal; the matrices $D(\lambda_i)$ are diagonal and λ_i are the vectors with the coefficients λ_{ij} for $i = 0, 1, 2$ and $j = 1, \ldots, p$. Model (2.1) guarantees adaptivity by means of its $3p$ parameters and it is characterized by the sum of three terms: the *spatial component*, driven by matrix $\mathbf{W}$ and the vector parameter λ_0; the *dynamic component*, driven by λ_1; and the *spatial–dynamic component*, driven by $\mathbf{W}$ and λ_2.
Model (2.1) allows for non constant regression coefficients among locations, starting from it we can derive different models as special cases by considering some constraints on the parameters. The first one is the *standard* SDPD of Yu, De Jong, and Lee (2008), that has constant spatial coefficients for all locations

$$\mathbf{y}_t = \lambda_0\mathbf{W}\mathbf{y}_t + \lambda_1\mathbf{y}_{t-1} + \lambda_2\mathbf{W}\mathbf{y}_{t-1} + \varepsilon_t. \tag{2.2}$$

with homoskedastic and uncorrelated errors. Other special cases (available in the literature) of the model can be derived from the *standard* SDPD by testing the significance of specific λ_{ij} coefficients. Both (2.1) and (2.2) can be expressed in VAR representation. For the *generalized* model, let $\mathbf{S}_0 = (\mathbf{I} - D(\lambda_0)\mathbf{W})$, thus we have the reduced form

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \eta_t \tag{2.3}$$

where $\mathbf{A} = \mathbf{S}_0^{-1}(D(\lambda_1) + D(\lambda_2)\mathbf{W})$ is the parameters matrix and $\eta_t = \mathbf{S}_0^{-1}\varepsilon_t$. Model (2.3) is a p-dimensional VAR(1) with coefficient matrix given by $\mathbf{A}$, its equivalent for model (2.2) can be found in Yu, De Jong, and Lee (2008).

2.3 Estimation of the SDPD models

In the sequel, we assume that $\mathbf{y}_1, \cdots, \mathbf{y}_n$ are n observations from a stationary process defined by (2.1) or (2.2). We assume that the process has mean zero and denote with $\Sigma_h = Cov(\mathbf{y}_t, \mathbf{y}_{t-h}) = E(\mathbf{y}_t\mathbf{y}'_{t-h})$ the autocovariance matrix of the process at lag h, where the prime subscript denotes the transpose operator.

From literature of spatio-temporal models we know that as $\mathbf{y}_t$ occurs on both sides of (2.2), $\mathbf{W}\mathbf{y}_t$ and ε_t are correlated. Applying least squares method directly based on regressing $\mathbf{y}_t$ on $(\mathbf{W}\mathbf{y}_t, \mathbf{y}_{t-1}, \mathbf{W}\mathbf{y}_{t-1})$ leads to inconsistent estimators. Usually the estimation of *standard* SDPD is performed by MLE, meanwhile ML applied to the *generalized* SDPD requires at least $3p$ parameters estimation and can result in inconsistent estimates. A proper estimator of the parameters for the *generalized* SDPD model (2.1) has been proposed and analysed by Dou, Parrella, and Yao (2016). They suggested applying least square estimator to the Yule-Walker equations to avoid endogeneity problem and propose a particular estimator, called *Generalized Yule-Walker* estimator, that allows for heterogeneous coefficients among locations.
In the following we define the Yule-Walker equation system as in Dou, Parrella, and Yao (2016)

$$(\mathbf{I} - D(\lambda_{0\cdot})\mathbf{W})\Sigma_1 = (D(\lambda_{1\cdot}) + D(\lambda_{2\cdot})\mathbf{W})\Sigma_0 \tag{2.4}$$

where Σ_h are the autocovariance matrices of the process at lags $h = 0, 1$. The coefficients matrices $D(\lambda_{k\cdot})$, with $k = 0, 1, 2$, are $p \times p$ diagonal matrices with

main diagonal given by coefficients λ_{kj}, $j = 1, ..., p$.
The matrices Σ_0 and Σ_1 are replaced by the respective sample counterparts

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_t', \; \hat{\Sigma}_1 = \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_{t-1}' \tag{2.5}$$

Generalized Yule-Walker estimator of Dou, Parrella, and Yao (2016) involves least square estimation for each row of the equation system in (2.4), that is

$$(\mathbf{e}_j' - \lambda_{0j}\omega_j')\hat{\Sigma}_1 = (\lambda_{1j}\mathbf{e}_j' + \lambda_{2j}\omega_j')\hat{\Sigma}_0 \; , \quad j = 1, ..., p$$

where $\mathbf{e}_j$ denotes the p-variate null vector with 1 in the jth position and ω_j is the jth row of matrix $\mathbf{W}$. Parameters vector $(\hat{\lambda}_{0j}, \hat{\lambda}_{1j}, \hat{\lambda}_{2j})'$ is estimated as solution of the problem

$$\min \left\| \hat{\Sigma}_1'(\mathbf{e}_j - \lambda_{0j}\omega_j) - \hat{\Sigma}_0(\lambda_{1j}\mathbf{e}_j + \lambda_{2j}\omega_j) \right\|_2^2 \; , \quad j =, ..., p$$

The resulting estimator has closed form

$$\begin{aligned} &\underline{\hat{\lambda}}_j = (\hat{\lambda}_{0j}, \hat{\lambda}_{1j}, \hat{\lambda}_{2j})' = (\hat{\mathbf{X}}_j'\hat{\mathbf{X}}_j)^{-1}\hat{\mathbf{X}}_j'\hat{\mathbf{Y}}_j, \quad j = 1, ..., p \\ &\text{where} \\ &\hat{\mathbf{X}}_j = (\hat{\Sigma}_1'\omega_j, \hat{\Sigma}_0\mathbf{e}_j, \hat{\Sigma}_0\omega_j) \;\text{ and }\; \hat{\mathbf{Y}}_j = \hat{\Sigma}_1\mathbf{e}_j \end{aligned} \tag{2.6}$$

Following we will introduce our proposal for testing the structure of SDPD models. In particular, we will define a test statistics for a multiple test procedure that allows choosing between the *generalized* model or its *standard* version.

2.4 The test statistics

In order to test the structure of the SDPD model, we define the test statistics

$$\hat{D}_{ij} = \sqrt{n}\Big(\hat{\lambda}_{ij} - \frac{1}{p}\sum_{k=1}^{p}\hat{\lambda}_{ik}\Big), \quad j = 1, \ldots, p, \text{ and } i = 0, 1, 2; \tag{2.7}$$

In (2.7) we are comparing the estimator under the generalized model, $\hat{\lambda}_{ij}$, with a proxy of the estimator under the standard model with constant coefficients.

When the true model is the *standard* SDPD, it has not been proved that estimator in Dou, Parrella, and Yao (2016) is consistent for parameters in the restricted model. We can expect that it is bound to converge to some constant equal for all the locations.
Since our purpose is not to evaluate the properties of *generalized* estimator under the *standard* model, we consider to evaluate estimation under the *standard* model by the mean value of the estimates over different locations

$$\bar{\lambda} = \frac{1}{p}\sum_{j=1}^{p}\hat{\underline{\lambda}}_j = \frac{1}{p}\sum_{j=1}^{p}(\hat{\mathbf{X}}_j'\hat{\mathbf{X}}_j)^{-1}\hat{\mathbf{X}}_j'\hat{\mathbf{Y}}_j. \tag{2.8}$$

Considering (2.8) as a proxy of constant estimated coefficients is sufficient for our purpose. Our proposal is about a strategy for testing the particular structure of the spatial dynamic panel data model, therefore, first of all we are interested in studying the behaviour of the test statistics when the true model is the *generalized* one (with nonconstant coefficients) and when the model is the *standard* one (with constant coefficients) as in (2.2).
Notice that large values of the statistics in (2.7) denote a preference for the *generalized* SDPD model. Instead, when the true model has constant parameters, as in the SDPD model of Yu, De Jong, and Lee (2008), the statistics in (2.7) are expected to be around zero. By Theorem 2 of Dou, Parrella, and Yao (2016), it can be shown that under the null hypothesis $\hat{D}_{ij} = O_p(1)$ since $\bar{\lambda} \xrightarrow{p} \lambda^*$, even that $\lambda^* \neq \lambda$. Conversely, if the true model is the *generalized* one, $\hat{D}_{ij}$ does not converge to zero inasmuch as $\bar{\lambda}$ is not consistent for the heterogeneous parameters. In such a case, it can be shown that $\hat{D}_{ij} = O_p(\sqrt{n})$.
In order to give an empirical evidence of this, Figure 2.1 shows the estimated density (based on $N = 250$ replications of the model) of the statistic (2.7), for $i = 2$, $j = 1$ and dimension $p = 50$, with different time series lengths (going from $n = 100$ to $n = 1000$ and denoted by the line width, as indicated in the legend). The left side of the figure refers to a case where the true model is the *standard* SDPD model, with constant parameters, therefore this is a case generated under the null hypothesis. In such a case, as expected, the distribution of the statistic is centred around zero. The right side of the figure refers to a case where the true model is a *generalized* SDPD, with non-constant parameters, therefore this is a case generated under the alternative hypothesis. In such a

case, as expected, the statistic $\hat{D}_{ij}$ is far away from zero. Moreover, as required for consistency, the value of the statistic increases for increasing time series length. Similar results for other values of j, i and p will be shown later in the simulation study.

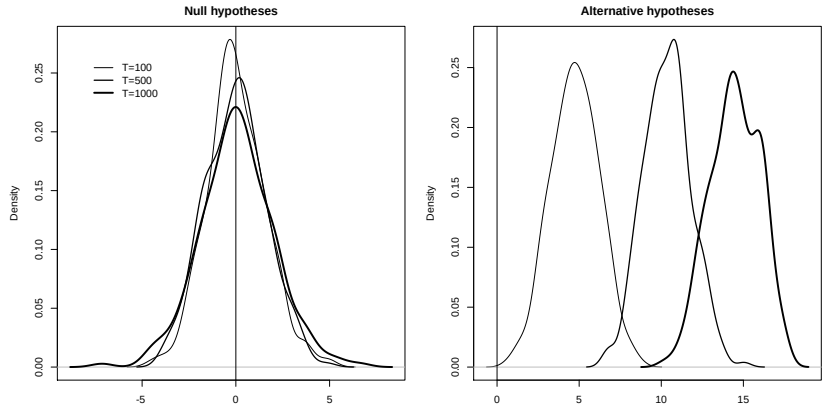

FIGURE 2.1: Estimated densities (based on $N = 250$ replications) of the statistic $\hat{D}_{ij} = \sqrt{n}\left(\hat{\lambda}_{ij} - \bar{\lambda}_i\right)$, for $i = 2$, $j = 1$ and dimension $p = 50$, with different time series lengths denoted by the line width. The left side refers to the case generated under the null hypothesis of true *standard* SDPD model. The right side refers to the case generated under the alternative hypothesis of true *generalized* SDPD model.

Before explain in detail the strategy of the test by using statistics (2.7), we state some basic assumptions and definitions useful for the next results.

Let $\mathcal{F}_a^b$ be the σ-algebra generated by $\{\mathbf{y}_t, a \leq t \leq b\}$ and let

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |P(A)P(B) - P(AB)| \tag{2.9}$$

denote the strong mixing coefficient for two σ-algebras $\mathcal{A}$ and $\mathcal{B}$ on the same probability space.

Definition 1 (strong mixing coefficient)**.** *The strong mixing coefficients α_n for the process $\{\mathbf{y}\}_t$ are defined by*

$$\alpha_n = \sup_{i \in \mathbb{N}} \alpha(\mathcal{F}_{-\infty}^i, \mathcal{F}_{n+i}^{\infty}) \tag{2.10}$$

moreover, for the sequence α_n we define a function $\alpha(t) = \alpha_{[t]}$ that is right continuous and has a left limit, while we denote by α^{-1} its inverse.

Definition 2 (double array)**.** *Let $(Y_{i,j})_{i=1,...,n;\, j=1,...,p_n}$ be a double array of dimensions (p_n, n)*

$$\begin{aligned} &Y_{1,1}\, Y_{1,2} \dots Y_{1,j} \dots Y_{1,p_n} \\ &Y_{2,1}\, Y_{1,2} \dots Y_{2,j} \dots Y_{2,p_n} \\ &\vdots \\ &Y_{n,1}\, Y_{n,2} \dots Y_{n,j} \dots Y_{n,p_n} \end{aligned} \tag{2.11}$$

where we denote p_n in order to considering the high-dimensional case with growing p as $n \to \infty$.
We also set $Y_{n,p_n} = \sum_{j=1}^{p_n} Y_{n,j}$ and $V_{n,p_n} = Var(Y_{n,p_n})$.

Definition 1 is standard definition of alpha mixing coefficient for multivariate time series and can be found in Rio (1995), meanwhile definition 2 gives us a way of rearranging our data process in order to state some theoretical results. Let $\{\mathbf{y}_t\}$ be the multivariate process considered in the previous sections, we denote by F its distribution function and by Q_F the inverse function of $u \to \mathbb{P}(|\mathbf{y}| > u)$. We arrange it in a double array according to definition 2 and state the following assumption

(A.1) The process $\{\mathbf{y}_t\}$ is strictly stationary and α-mixing with strong mixing coefficients α_n satisfying

$$V_{n,p_n}^{3/2} \sum_{i=1}^{n} \int_0^1 \alpha^{-1}(x/2)\; Q_F^2(x)\; \inf\left(\alpha^{-1}(x/2) Q_F(x), \sqrt{V_{n,p_n}}\right) dx \to 0 \tag{2.12}$$

We need some further regularity assumptions, most of them are from Dou, Parrella, and Yao (2016). In particular we require that

(A.2) The spatial weight matrix $\mathbf{W}$ is known with zero main diagonal elements and is uniformly bounded in row and column sums in absolute value (UB) (i.e. we have, $\sup_{p\geq 1} \|W\|_\infty$ and $\sup_{p\geq 1} \|W\|_1$, where $\|W\|_\infty := \sup_{1\leq i,j\leq p} \sum_j |w_{ij}|$ is the row sum norm and $\|W\|_1 := \sup_{1\leq i,j\leq p} \sum_i |w_{ij}|$ is the column sum norm);

(A.3) the matrix $S_0 = I - \lambda_0 \mathbf{W}$ is invertible and UB;

(A.4) ε_{it} are i.i.d. across i and t, zero mean, variance σ^2 and $\mathbb{E}|\varepsilon_{it}|^{4+\gamma} < \infty$ for some $\gamma > 0$;

(A.5) $\sum_{h=1}^{\infty} |\mathbf{A}^h|$ is UB, where $\mathbf{A}$ is as in the reduced form (2.3);

(A.6) the rank of the matrix $\hat{\mathbf{X}}_j$ is equal to 3, moreover the matrix $\hat{\mathbf{V}}_j = (\hat{\mathbf{X}}_j' \hat{\mathbf{X}}_j) \xrightarrow{p} \mathbf{V}_j$ positive definite matrix, for $j = 1, ..., p$.

Conditions A.1 is a Lindeberg condition for multivariate strong mixing processes. Assumptions A.2-A.5 are standard in SDPD modelling. Essentially, conditions A.2 and A.3 limits the spatial correlation through spatial weights, condition A.5 limits the dependence across serial and cross-sectional units, condition A.6 ensures identifiability in (2.4).

2.4.1 Some theoretical results

Following, we give some theoretical results about the behaviour of the test statistics (2.7). In detail, we state a Lemma for assessing asymptotic distribution of the statistics under the null hypothesis.

Lemma 1. *Let the model 2.1, under assumptions (A.1)-(A.6) as $n \to \infty$, $p \to \infty$ and $p = o(\sqrt{n})$, we have that*

$$\frac{\sqrt{n}}{p} \sum_{j=1}^{p} (\hat{\lambda}_j - \lambda_j) \xrightarrow{d} N(\mathbf{0}, \Gamma),$$

for some positive definite variance-covariance matrix Γ.

Proof of Lemma 1. By Corollary 1 of Dou, Parrella, and Yao (2016) we have that

$$\left\| \frac{1}{p} \sum_{j=1}^{p} (\hat{\lambda}_j - \lambda_j) \right\|_1 \leq \frac{1}{p} \sum_{j=1}^{p} \|\hat{\lambda}_j - \lambda_j\|_1 = O_p\Big(\frac{1}{\sqrt{n}}\Big), \tag{2.13}$$

similar to Dou, Parrella, and Yao (2016), we have to show that

$$\frac{\sqrt{n}}{p} \Gamma^{1/2} \sum_{j=1}^{p} \begin{pmatrix} \mathbf{w}_j^\top \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_{t-1}^\top \times \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_{t-1} \varepsilon_{j,t} \\ \mathbf{e}_j^\top \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_t^\top \times \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_{t-1} \varepsilon_{j,t} \\ \mathbf{w}_j^\top \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_t^\top \times \frac{1}{n} \sum_{t=1}^{n} \mathbf{y}_{t-1} \varepsilon_{j,t} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \mathbf{I}) \tag{2.14}$$

By the proof of Theorem 2 of Dou, Parrella, and Yao (2016), it is sufficient to show that

$$\mathbf{a}^\top \begin{pmatrix} \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \mathbf{w}_j^\top \Sigma_1 \mathbf{y}_{t-1} \varepsilon_{j,t} \\ \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \mathbf{e}_j^\top \Sigma_0 \mathbf{y}_{t-1} \varepsilon_{j,t} \\ \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \mathbf{w}_j^\top \Sigma_0 \mathbf{y}_{t-1} \varepsilon_{j,t} \end{pmatrix} \tag{2.15}$$

is asymptotic normal, where $\mathbf{a} = (a_1, a_2, a_3)'$ is any nonzero vector.
Let

$$\begin{aligned} S_{n,p} =& \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \left(a_1 \, \mathbf{w}_j^\top \Sigma_1 \mathbf{y}_{t-1} \varepsilon_{j,t} + a_2 \, \mathbf{e}_j^\top \Sigma_0 \mathbf{y}_{t-1} \varepsilon_{j,t} + a_3 \, \mathbf{w}_j^\top \Sigma_0 \mathbf{y}_{t-1} \varepsilon_{j,t} \right) = \\ =& \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \left(a_1 \, X_1^{(j,t)} + a_2 \, X_2^{(j,t)} + a_3 \, X_3^{(j,t)} \right) = \frac{1}{p\sqrt{n}} \sum_{j=1}^{p} \sum_{t=1}^{n} \left(\mathbf{a}' \mathbf{X}^{(j,t)} \right)'. \end{aligned}$$

where $\mathbf{X}^{(j,t)} = \left(X_1^{(j,t)}, X_2^{(j,t)}, X_3^{(j,t)} \right)$
Since by assumption (A.5) it holds that

$$\limsup_{n \to \infty} \max_{1 \le j \le p} V_{j,n}^{(X)} / V_{p,n}^{(X)} < \infty$$

where $V_{j,n}^{(X)} = Var(\mathbf{X}^{(j,n)})$ and $V_{p,n}^{(X)} = Var(\mathbf{X}^{(p,n)})$, then by Theorem 1 in Rio (1995), 2.15 is asymptotically normal for any nonzero vector $\mathbf{a} = (a_1, a_2, a_3)'$. Substituting $\mathbf{a}$ with $\Gamma^{1/2}$, (2.14) holds and the proof follows. □

2.4.2 Simulation study

We conduct a Monte Carlo simulation in order to examine the performance of the statistic $\hat{D}_{ij}$. We replicate the estimation procedures 1000 times for two DGP, the *standard* SDPD model and its *generalized* version, respectively, and compute the statistic $\hat{D}_{ij}$. We expect that statistic (2.7) takes value far from zero when the true model is the *generalized* SDPD (assumed in the alternative hypothesis). Meanwhile, when the true model is *standard* SDPD with constant parameters, we expect it lies around zero. This behaviour can make the statistics useful in testing spatial dependencies by the proposed test, in such a case it is able to correctly discriminate both the hypothesis.
The spatial matrix $\mathbf{W}$ has been randomly generated as a rook matrix and has

been row-normalized. The parameters have all been randomly generated in the interval $[-0.8, 0.8]$, assuring that the stationarity condition of the model is guaranteed. Figures 2.2 and 2.3 report box-plots of the statistics under different settings. Each panel shows the box-plots of the statistics $\hat{D}_{ij}$ for the first 10 locations, the plots in column correspond to each of the 3 regression parameters in the model λ_{0j}, λ_{1j} and λ_{2j}. The behaviours of the statistics is similar under the null hypothesis, at the same way, box-plots of statistics under the alternative hypothesis are not centred at zero for the most part of locations. Instead, figure 2.4 shows the density of the statistics under both the DGPs and compare it with the densities of $\hat{\lambda}_{i1}$ and mean value of $\hat{\bar{\lambda}}_1$. The upper part is referred to the case low dimensional case $\{n, p\} = \{100, 10\}$, while the bottom to $\{n, p\} = \{500, 1000\}$. In each of them, first row refers to the case generated under the null hypothesis of true *standard* SDPD model, while the second row refers to the case generated under the alternative hypothesis of true *generalized* SDPD model. To make the plot more clear, under the Null we have preferred to plot the mean values of $\bar{\lambda}_1$ instead of the densities since the latters appeared very flattened already for low dimensional case. As expected, the distribution of the statistic is centred around zero under the null hypothesis, while the *generalized* estimator is not centred at zero; under the alternative hypothesis, although the statistic and the *generalized* estimator show similar behaviours, we can see how the former is away from zero, due to the effect of $\bar{\lambda}_1$.

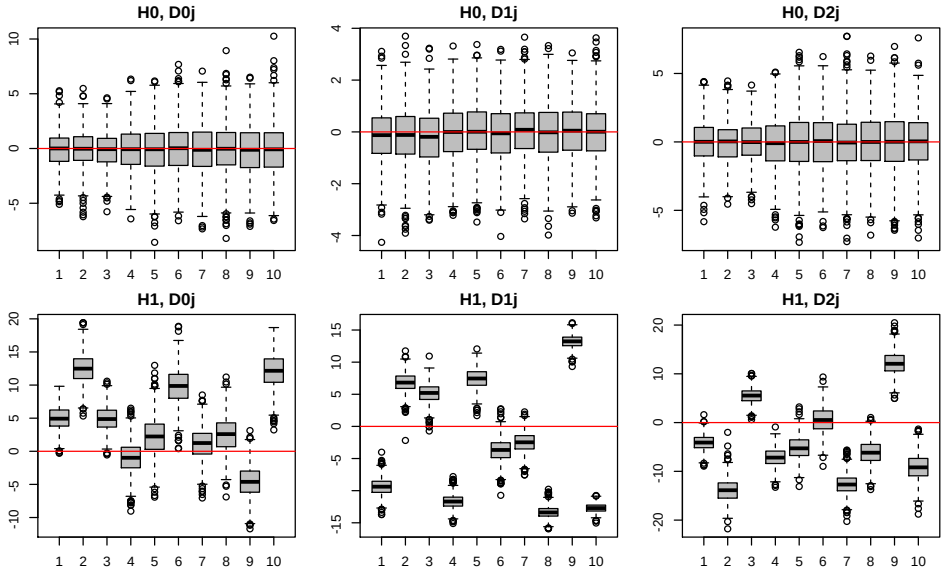

FIGURE 2.2: Box-plot of the statistics D_{ij}, for $i = 0, 1, 2$ (by column) and first 10 locations; $R = 1000$ simulation runs. On the upper part the DGP is the *standard* SDPD, on the bottom it is the *generalized* SDPD with $\{n, p\} = \{300, 500\}$.

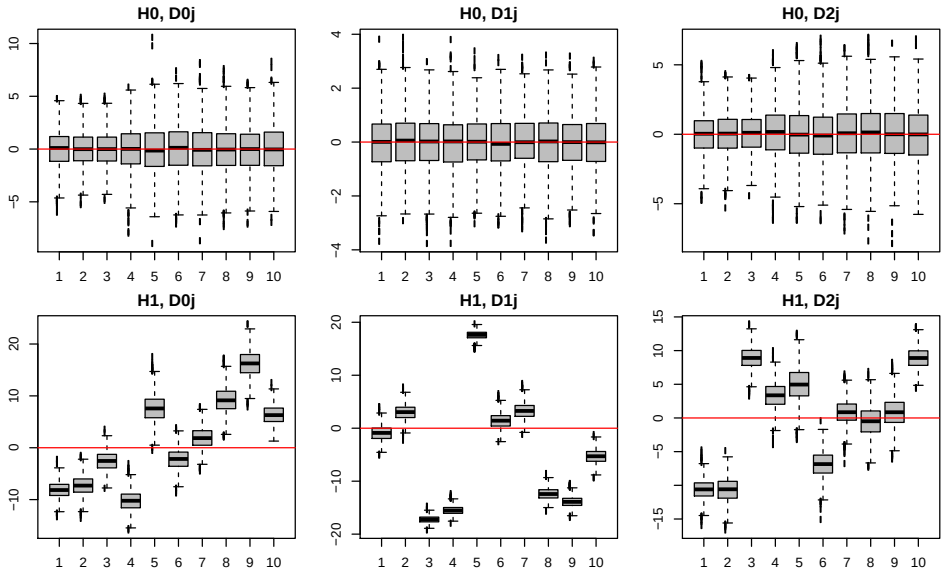

FIGURE 2.3: Box-plot of the statistics D_{ij} for $i = 0, 1, 2$ (by column) and first 10 locations; $R = 1000$ simulation runs. On the upper part the DGP is the *standard* SDPD, on the bottom it is the *generalized* SDPD with $\{n, p\} = \{500, 1000\}$.

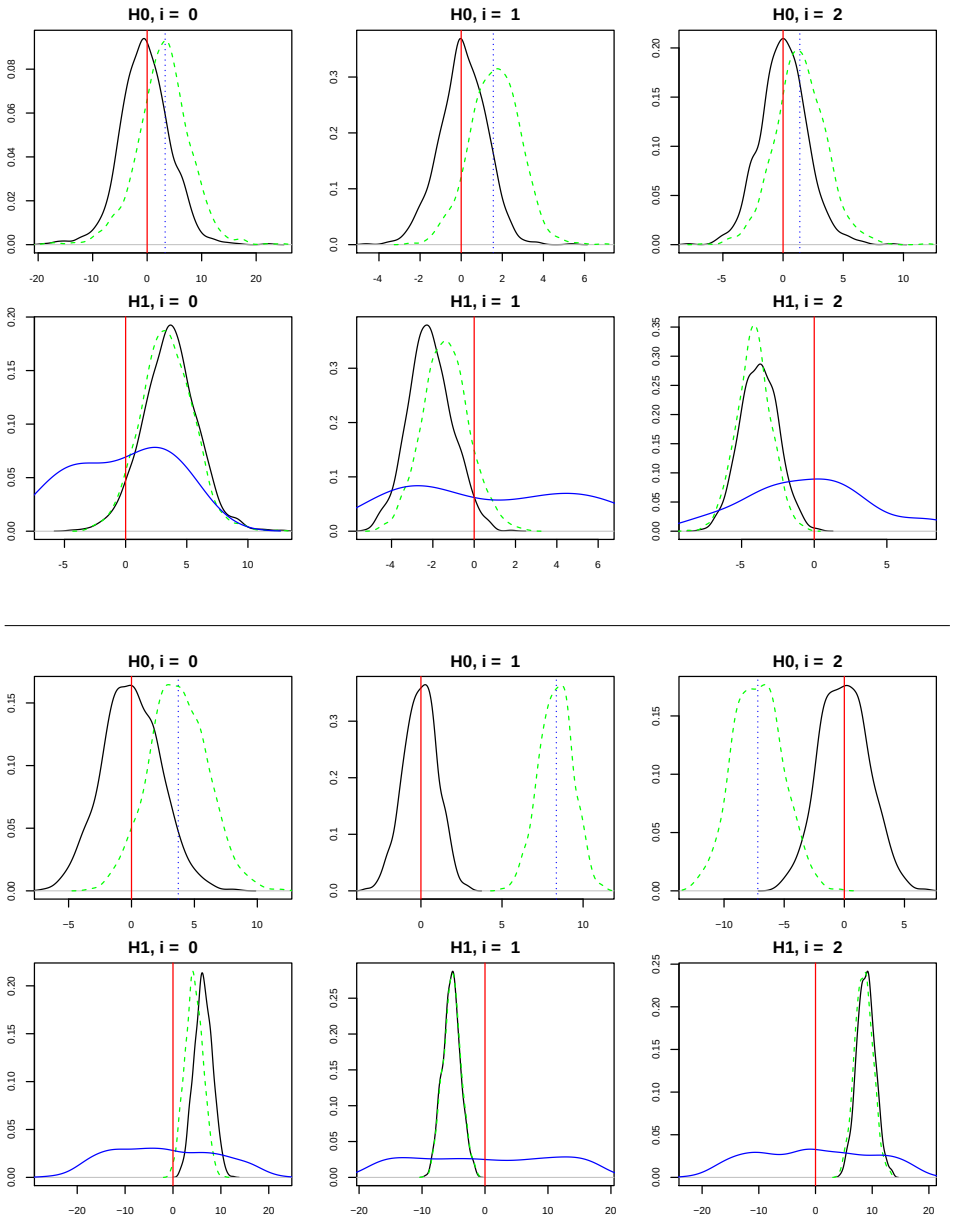

FIGURE 2.4: Densities (derived by $R = 1000$ replications of the model) of the estimators $\hat{\lambda}_{i1}$ (dashed green line), $\bar{\lambda}_1$ (blue solid line) and statistic $\hat{D}_{i1}$ (solid black line). Under the Null we have reported the mean of $\bar{\lambda}_1$ (dotted blue line). Upper side refers to the case $\{n, p\} = \{100, 10\}$, bottom side to $\{n, p\} = \{500, 1000\}$.

2.5 A strategy for the test

We are proposing a test strategy to identify the specific structure of the spatial dynamic model. In particular we are interested in classifying between the *standard* SDPD, with constant parameters, and the its *generalized* version with heterogeneous parameters among location.
Thus, for each term in (2.1) we need to test if $\lambda_{i1} = ... = \lambda_{ip}$ or $\lambda_{ij} \neq \lambda_{ik}$ for at least one $k \neq j$ and $j = 1, ..., p$.
Results from previous section show that the statistics in (2.7) can be used as building blocks of such a testing purpose, thus the hypothesis we need to test is

$$\begin{cases} H_{0j} \ : \ D_{ij} = 0; \\ H_{j1} \ : \ D_{ij} \neq 0 \quad \text{for at leat one } j = 1, ..., p \end{cases} \tag{2.16}$$

where i denotes the specific spatial parameter, with $i = 0, 1, 2$. Test (2.16) has clearly a multiple testing structure and the problem then becomes how to decide which hypotheses to reject, taking into account the multitude of tests.

2.5.1 Multiple hypothesis testing

When many hypotheses are tested jointly, some are bound to appear as significant by chance alone, even if in reality they are not relevant. If we follow the same rejection rule independently for each test, the resulting probability of making at least one type I error is substantially higher than the nominal level used for each test, particularly when the number of total tests is large. To address this issue, multiple testing procedures seek to make the individual tests more conservative so as to minimize the number of type I errors while maintaining an overall error rate. The type I error rates most discussed in the literature are FWER (family-wise error rate) and FDR (false discovery rate). The FWER is defined as the probability of making at least one false rejection when all the null hypotheses are true, meanwhile the FDR is defined as the expected percentage of rejected hypotheses that have been wrongly rejected. Instead of controlling the probability of a type I error at a set level for each test, these methods control the overall FWER or FDR at level of the overall error rate.
To prevent us from declaring true null hypotheses to be false, we seek control

(at least asymptotically) of the FWER or FDR. We choose to control FWER by applying the well-known Bonferroni method. It is the most familiar scheme for controlling the FWER : for each null hypothesis H_{0j}, individual p-values p_js are computed and the hypothesis H_{0j} is rejected at global level α if $p_j \leq \alpha/m$.

2.5.2 Bootstrap approach

We know by Lemma 1 that statistics (2.7) has normal asymptotic distribution. Thus, if we would use the asymptotic distribution for our test purpose, we have to consider studentized form, say

$$\hat{Z}_j = \Gamma_{D_j}^{-1/2} \hat{D}_j \sim N(\mathbf{0}, \mathbf{I}), \tag{2.17}$$

where Γ_{D_j} is a 3×3 positive definite matrix. Then, under the Null we can bound the probability of event $\{|\hat{Z}_j| < c_n\}$ as follows

$$\mathbb{P}(|\hat{Z}_j| < c_n) < 1 - \frac{1}{\sqrt{2\pi}} \int_{c_n}^{\infty} \exp\{-u^2/2\}\, du < 1 - \frac{1}{c_n\sqrt{2\pi}} \exp\{-c_n^2/2\}$$

for a given c_n growing with n. Thus, a primary way to proceed with the test could be find condition that ensures $\mathbb{P}(|\hat{Z}_j| < c_n)$ goes to one under the null hypothesis, and at the same time it vanishes under the alternative hypothesis, for the same threshold c_n. Under such a condition the statistics has a proper behaviour that makes it useful for infinite p hypothesis testing in a multiple scheme like this one. Unfortunately, this strategy is no easy to adopt, first of all we need to take into account for the variance of $\hat{D}_{ij}$ in order to obtain the studentized version (2.17), second we need some procedure in order to find threshold c_n. Therefore, for testing significance of $\hat{D}_j$, we must account for some estimate $\hat{\Gamma}_{D_j}$ of the variance-covariance matrix Γ_{D_j}. Unfortunately, direct estimation of variance-covariance matrix is not straightforward. Moreover, we have a further problem, if the order of increasing dimension p is not $o(\sqrt{n})$ we encounter a non-vanishing bias that affects the estimate, thus we also need to take into account for bias estimates. For that reason, in order to provide a good approximation for our test, we apply a bootstrap method useful for finite samples.

As concern the derivation of the individual p-values p_is, we use a resampling procedure based on the residual bootstrap approach. In particular, we follow a procedure that allows us to approximate the distribution of the test statistics $\hat{D}_{ij}$, then by such a bootstrap distribution we compute the p-values for the test. This procedure runs as follows.

1. First obtain the bootstrap errors $\{\varepsilon_t^*\}$ by drawing $B = 999$ replicates independently from the residuals $\hat{\varepsilon}_t = \mathbf{y}_t - \hat{\mathbf{y}}_t$, where $\hat{\mathbf{y}}_t = \bar{\lambda}_0 \mathbf{W}\mathbf{y}_t + \bar{\lambda}_1 \mathbf{y}_{t-1} + \bar{\lambda}_2 \mathbf{W}\mathbf{y}_{t-1}$ and $\bar{\lambda}_i = \frac{1}{p}\sum_{j=1}^{p} \hat{\lambda}_{ij}$.

2. Generate the bootstrap series, under the null hypothesis, as

$$\hat{\mathbf{y}}_t^* = \hat{\mathbf{A}}\mathbf{y}_{t-1}^* + \hat{\mathbf{S}}_0^{-1}\varepsilon_t^*,$$

   where $\hat{\mathbf{S}}_0 = (\mathbf{I}_p - \bar{\lambda}_0 \mathbf{W})$ and $\hat{\mathbf{A}} = \hat{\mathbf{S}}_0^{-1}(\bar{\lambda}_1 \mathbf{I}_p + \bar{\lambda}_2 \mathbf{W})$.

3. Compute the bootstrap statistics $\hat{D}_{ij}^* = \sqrt{n}\Big(\hat{\lambda}_{ij}^* - \bar{\lambda}_i^*\Big)$, as in (2.7), with $\hat{\lambda}_{ij}^*$ and $\bar{\lambda}_i^*$ estimated from the bootstrap data $\hat{\mathbf{y}}_t^*$.

4. For a given $j = 1, \ldots, p$, the individual p-value p_j for testing H_{0j} is defined as the probability $\mathbb{P}(|D_{ij}^*| > |\hat{D}_{ij}| \Big| \mathbf{y}_1, ..., \mathbf{y}_n)$, which is approximated by the relative frequency of the event $\{|D_{ij}^*| > |\hat{D}_{ij}|\}$ over the 999 bootstrap replications.

Bootstrap estimates consistency for VAR models has been studied, from empirical and theoretical point of view, in Kim (1999), Kim (2004), Staszewska-Bystrova (2011), and Paparoditis (1996). In our case, VAR representation (2.3) is stationary and the estimator satisfies proper conditions to assure consistency of the corresponding bootstrap quantities.
In particular, given the (2.4), we obtain the following reparametrization

$$\Sigma_1' \Sigma_0^{-1} = (\mathbf{I} - D(\lambda_{0\cdot})\mathbf{W})^{-1}(D(\lambda_{1\cdot}) + D(\lambda_{2\cdot})\mathbf{W}). \tag{2.18}$$

Notice that the left side of the above equation corresponds to the Yule-Walker equation for the reduced form of the model, meanwhile the right part corresponds to the matrix $\mathbf{A}$ in the VAR representation (2.3).

Given the *generalized* Yule-Walker estimator (2.6), we define

$$\hat{\mathbf{A}} = (\mathbf{I} - D(\hat{\lambda}_{0\cdot})\mathbf{W})^{-1}(D(\hat{\lambda}_{1\cdot}) + D(\hat{\lambda}_{2\cdot})\mathbf{W}) \tag{2.19}$$

the *generalized* Yule-Walker estimator of $\mathbf{A}$ obtained by estimators $\hat{\lambda}_{ij}$. Now, given the error term ε_t, let F_ε be its distribution function and denote by $\hat{F}_\varepsilon$ the empirical distribution of $\hat{\varepsilon}_t$.
The following proposition shows the asymptotic validity of bootstrap procedure for estimator (2.19).

Proposition 1. *Suppose that the assumptions (A.1) - (A.6) hold. If $p = o(\sqrt{n})$, we have that*

$$d_2(\hat{F}_\varepsilon\,,\, F_\varepsilon) \to 0 \quad \textit{in probability}, \tag{2.20}$$

where $d_2(\cdot,\cdot)$ is the mallow's metric (see Paparoditis (1996), pag.282).

Proof. By results of Theorem 2 in Dou, Parrella, and Yao (2016) and the continuous mapping theorem, $\hat{\mathbf{A}}$ is a consistent estimator for $\mathbf{A}$. Moreover, we know that model (2.3) is a particular VAR(1) process. Then, by delta method (Serfling (1980)), it can be shown that

$$\|\hat{\mathbf{A}} - \mathbf{A}\| = O_p\Big(\sqrt{\frac{p}{n}}\Big). \tag{2.21}$$

where $\|\mathbf{M}\|^2 = tr(\mathbf{M}'\mathbf{M})$ denotes the Schur's matrix norm and p is the order for the matrix $\mathbf{A}$.

Since Theorem 2.3 of Paparoditis (1996) does not require a specific estimator for $\mathbf{A}$ but, it is valid for any estimator and furthermore, it only needs $\|\hat{\mathbf{A}} - \mathbf{A}\| = o_p(1)$, we can use result (2.21). Then, by Theorem 2.4 of Paparoditis (1996) the result follows since $p = o(\sqrt{n})$. □

Remark. After some algebra and using the same approach as in Paparoditis (1996), we can show that $(\hat{\lambda}^*_{0j}, \hat{\lambda}^*_{1j}, \hat{\lambda}^*_{2j})'$ are consistent. The main difference with respect to the paper of Paparoditis (1996) is that we have a dimension p which goes to infinity when $n \to \infty$.

2.5.3 Simulation results

In the following, we evaluate the performance of the test by a simulation study. We perform 250 simulation runs, and generate data under the Null and Alternative hypothesis, respectively. At each step we run 999 bootstrap replication in order to compute p-values for the test statistics, we set the nominal size α equal to 0.1.

Table 2.1 reports values of the size of the test and the power over the whole simulation replications. Similarly, figure 2.5 shows the same results for different combinations of number of locations p and time series length n. Each plot of the figure shows size (black points) and power (green points) from multiple test corresponding to one of the $i = 0, 1, 2$ index and for a given value of dimension p, meanwhile the sample size n is from 100 to 1000. We also report horizontal lines at level of the nominal size and maximum power 1. As we can see, apart from some cases, p-values converge to the nominal size and the power to 1 as the sample size n grows.

Under Null	for $i = 0$			for $i = 1$			for $i = 2$		
(=size)	n=100	500	1000	100	500	1000	100	500	1000
$p = 10$	0.124	0.1	0.072	0.184	0.144	0.148	0.136	0.112	0.108
$p = 50$	0.024	0.12	0.092	0.156	0.144	0.172	0.144	0.164	0.24
$p = 100$	0.888	0.2	0.204	0.82	0.216	0.244	0.884	0.128	0.136
Under Altern.	for $i = 0$			for $i = 1$			for $i = 2$		
(=power)	n=100	500	1000	100	500	1000	100	500	1000
$p = 10$	0.204	1	1	1	1	1	0.988	1	1
$p = 50$	0.056	0.108	0.148	1	1	1	1	1	1
$p = 100$	0.44	0.968	1	1	1	1	1	1	1

TABLE 2.1: Values of size and power of the test for different settings of sample size n and number of parameters p.

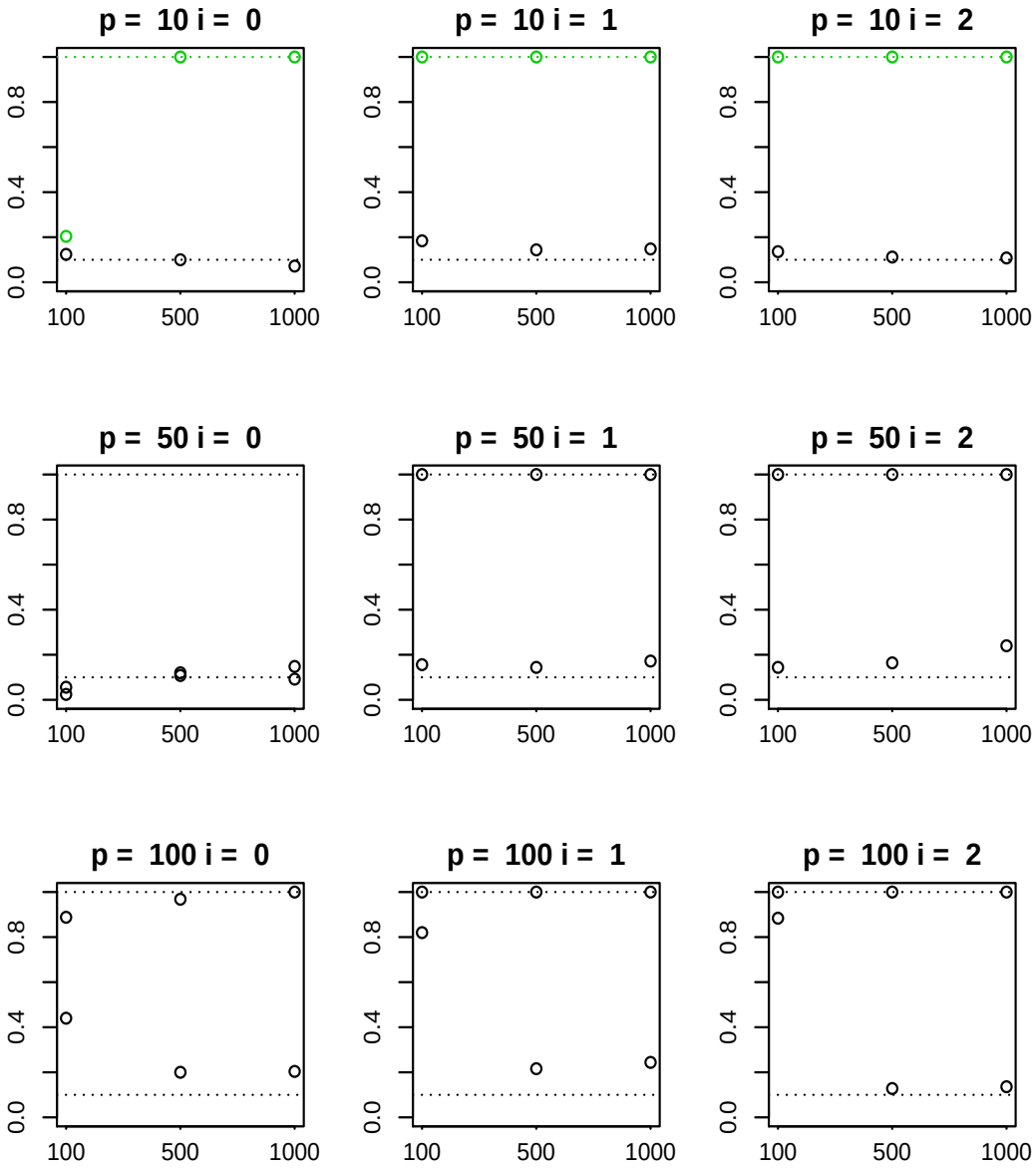

FIGURE 2.5: Power (green points) and size (black points) of the test for different combinations of p and n. The green dotted line is at level 1, while the black one at nominal size 0.1.

2.6 Discussion

In this chapter we proposed a strategy for testing the particular structure of the SDPD, by means of a multiple testing procedure based on bootstrap. Our procedure allows choosing between a generalized version of SDPD that makes assumption of different scalar coefficients, and the standard SDPD with equal coefficients among locations. While standard SDPD is highly regarded among spatio-temporal models, the generalized version has been recently introduced by Dou, Parrella, and Yao (2016). Motivation for different scalar coefficients can be found in practical situations in which empirical evidence shows how considering constant effects for each location can be too restrictive. This suggested us to propose a test for such a class of models. Some basic theoretical results for the statistic and the bootstrap strategy have been showed. The multiple scheme of the test was handled by the Bonferroni technique and the distribution of the statistic derived by a residual resampling scheme. Results from a simulation study show that the proposed procedure for testing SDPD models works well.

Bibliography

Ahrens, Achim and Arnab Bhattacharjee (2015). "Two-step Lasso estimation of the spatial weights matrix". In: *Econometrics* 3.1, pp. 128–155.

Alvarez, Javier and Manuel Arellano (2003). "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators". In: *Econometrica* 71.4, pp. 1121–1159.

Anderson, Theodore Wilbur and Cheng Hsiao (1981). "Estimation of dynamic models with error components". In: *Journal of the American statistical Association* 76.375, pp. 598–606.

Anselin, Luc (2013). *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media.

Arellano, Manuel and Stephen Bond (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations". In: *The review of economic studies* 58.2, pp. 277–297.

Baltagi, Badi H, Seuck Heun Song, and Won Koh (2003). "Testing panel data regression models with spatial error correlation". In: *Journal of econometrics* 117.1, pp. 123–150.

Baltagi, Badi H et al. (2007). "Testing for serial correlation, spatial autocorrelation and random effects using panel data". In: *Journal of Econometrics* 140.1, pp. 5–51.

Barut, Emre, Jianqing Fan, and Anneleen Verhasselt (2016). "Conditional sure independence screening". In: *Journal of the American Statistical Association* 111.515, pp. 1266–1277.

Bertin, Karine and Guillaume Lecué (2008). "Selection of variables and dimension reduction in high-dimensional non-parametric regression". In: *Electronic Journal of Statistics* 2, pp. 1224–1241.

Bhattacharjee, Arnab and Chris Jensen-Butler (2013). "Estimation of the spatial weights matrix under structural constraints". In: *Regional Science and Urban Economics* 43.4, pp. 617–634.

Bühlmann, Peter and Sara Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Candes, Emmanuel and Terence Tao (2007). "The Dantzig selector: Statistical estimation when p is much larger than n". In: *The Annals of Statistics*, pp. 2313–2351.

Chang, Jinyuan, Cheng Yong Tang, and Yichao Wu (2013). "Marginal empirical likelihood and sure independence feature screening". In: *Annals of statistics* 41.4.

Chang, Jinyuan, Cheng Yong Tang, Yichao Wu, et al. (2016). "Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood". In: *The Annals of Statistics* 44.2, pp. 515–539.

Cliff, ADADC and John Keith Ord (1973). *Spatial autocorrelation*. Tech. rep.

Comminges, Laëtitia and Arnak S Dalalyan (2012). "Tight conditions for consistency of variable selection in the context of high dimensionality". In: *The Annals of Statistics* 40.5, pp. 2667–2696.

Dou, Baojun, Maria Lucia Parrella, and Qiwei Yao (2016). "Generalized Yule–Walker estimation for spatio-temporal models with unknown diagonal coefficients". In: *Journal of Econometrics*.

Elhorst, J Paul (2005). "Unconditional Maximum Likelihood Estimation of Linear and Log-Linear Dynamic Models for Spatial Panels". In: *Geographical analysis* 37.1, pp. 85–106.

Fan, Jianqing, Yang Feng, and Rui Song (2011). "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models". In: *Journal of the American Statistical Association* 106.494. PMID: 22279246, pp. 544–557. DOI: 10.1198/jasa.2011.tm09779. eprint: http://dx.doi.org/10.1198/jasa.2011.tm09779. URL: http://dx.doi.org/10.1198/jasa.2011.tm09779.

Fan, Jianqing and Runze Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American statistical Association* 96.456, pp. 1348–1360.

Fan, Jianqing and Jinchi Lv (2008). "Sure independence screening for ultrahigh dimensional feature space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5, pp. 849–911.

Fan, Jianqing, Rui Song, et al. (2010). "Sure independence screening in generalized linear models with NP-dimensionality". In: *The Annals of Statistics* 38.6, pp. 3567–3604.

Györfi, László et al. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hahn, Jinyong and Guido Kuersteiner (2002). "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large". In: *Econometrica* 70.4, pp. 1639–1657.

Hastie, Trevor and Robert Tibshirani (1990). *Generalized additive models*. Wiley Online Library.

Hu, Qinqin and Lu Lin (2017). "Conditional sure independence screening by conditional marginal empirical likelihood". In: *Annals of the Institute of Statistical Mathematics* 1.69, pp. 63–96.

Huang, Jian, Joel L Horowitz, and Fengrong Wei (2010). "Variable selection in nonparametric additive models". In: *Annals of statistics* 38.4, p. 2282.

Kapoor, Mudit, Harry H Kelejian, and Ingmar R Prucha (2007). "Panel data models with spatially correlated error components". In: *Journal of econometrics* 140.1, pp. 97–130.

Kelejian, H and D Robinson (1995). "Spatial correlation: a suggested alternative to the autocorrelation Model". In: *New Directions in Spatial Econometrics*.

Kelejian, Harry H and Gianfranco Piras (2014). "Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes". In: *Regional Science and Urban Economics* 46, pp. 140–149.

Kim, Jae H (1999). "Asymptotic and bootstrap prediction regions for vector autoregression". In: *International Journal of Forecasting* 15.4, pp. 393–403.

— (2004). "Bias-corrected bootstrap prediction regions for vector autoregression". In: *Journal of Forecasting* 23.2, pp. 141–154.

Lafferty, John and Larry Wasserman (2008). "Rodeo: sparse, greedy nonparametric regression". In: *The Annals of Statistics*, pp. 28–63.

Lee, Lung-fei and Jihai Yu (2010a). "A spatial dynamic panel data model with both time and individual fixed effects". In: *Econometric Theory* 26.2, pp. 564–597.

— (2010b). "Estimation of spatial autoregressive panel data models with fixed effects". In: *Journal of Econometrics* 154.2, pp. 165–185.

Lee, Lung-fei and Jihai Yu (2010c). "Some recent developments in spatial panel data models". In: *Regional Science and Urban Economics* 40.5, pp. 255–271.

— (2014). "Efficient GMM estimation of spatial dynamic panel data models with fixed effects". In: *Journal of Econometrics* 180.2, pp. 174–197.

Li, Runze, Wei Zhong, and Liping Zhu (2012). "Feature screening via distance correlation learning". In: *Journal of the American Statistical Association* 107.499, pp. 1129–1139.

Lin, Lu, Jing Sun, and Lixing Zhu (2013). "Nonparametric feature screening". In: *Computational Statistics & Data Analysis* 67, pp. 162–174.

Lin, Yi, Hao Helen Zhang, et al. (2006). "Component selection and smoothing in multivariate nonparametric regression". In: *The Annals of Statistics* 34.5, pp. 2272–2297.

Paparoditis, Efstathios (1996). "Bootstrapping autoregressive and moving average parameter estimates of infinite order vector autoregressive processes". In: *Journal of Multivariate Analysis* 57.2, pp. 277–296.

Qu, Xi and Lung-fei Lee (2015). "Estimating a spatial autoregressive model with an endogenous spatial weight matrix". In: *Journal of Econometrics* 184.2, pp. 209–232.

Radchenko, Peter and Gareth M James (2010). "Variable selection using adaptive nonlinear interaction structures in high dimensions". In: *Journal of the American Statistical Association* 105.492, pp. 1541–1553.

Ravikumar, Pradeep et al. (2009). "Sparse additive models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.5, pp. 1009–1030.

Rio, Emmanuel (1995). "About the Lindeberg method for strongly mixing sequences". In: *ESAIM: Probability and Statistics* 1, pp. 35–61.

Ruppert, David (1997). "Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation". In: *Journal of the American Statistical Association* 92.439, pp. 1049–1062.

Ruppert, David and Matthew P Wand (1994). "Multivariate locally weighted least squares regression". In: *The annals of statistics*, pp. 1346–1370.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley. ISBN: 9780471024033. URL: https://books.google.it/books?id=eIXGaQP6qLsC.

Staszewska-Bystrova, Anna (2011). "Bootstrap prediction bands for forecast paths from vector autoregressive models". In: *Journal of Forecasting* 30.8, pp. 721–735.

Stone, Charles J. (1985). "Additive Regression and Other Nonparametric Models". In: *The Annals of Statistics* 13.2, pp. 689–705. ISSN: 00905364.

Storlie, Curtis B et al. (2011). "Surface estimation, variable selection, and the nonparametric oracle property". In: *Statistica Sinica* 21.2, p. 679.

Su, Liangjun and Zhenlin Yang (2015). "QML estimation of dynamic panel data models with spatial errors". In: *Journal of Econometrics* 185.1, pp. 230–258.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Yu, Jihai, Robert De Jong, and Lung-fei Lee (2008). "Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large". In: *Journal of Econometrics* 146.1, pp. 118–134.

Yu, Jihai, Robert de Jong, and Lung-fei Lee (2012). "Estimation for spatial dynamic panel data with fixed effects: the case of spatial cointegration". In: *Journal of Econometrics* 167.1, pp. 16–37.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zhu, Li-Ping et al. (2011). "Model-free feature screening for ultrahigh-dimensional data". In: *Journal of the American Statistical Association* 106.496, pp. 1464–1475.

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". English. In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. ISSN: 01621459. URL: `http://www.jstor.org/stable/27639762`.