Non-parametric Methods under Cross-sectional Dependence

Jungyoon Lee

Dissertation submitted for the degree of

Doctor of Philosophy in Economics at

The London School of Economics and Political Science

June 2012

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others.

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgment is made. This thesis may not be reproduced without the prior written consent of the author.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Jungyoon Lee

Abstract

The possible presence of cross-sectional dependence in economic panel or cross-sectional data needs to be taken into consideration when developing econometric theory for data analysis. This thesis consists of three works that either allow for or estimate cross-sectional dependence in the disturbance terms of a regression model, each addressing different problems, models and methods in the areas of non- and semi-parametric estimation.

Chapter 1 provides an overview of the motivations for, and contributions of, the three topics of this thesis. A review of relevant literature is given, followed by a summary of main results obtained in order to help place the present thesis in perspective. Chapter 2 develops asymptotic theory for series estimation under a general setting of spatial dependence in regressors and error term, including cases analogous to those known as long-range dependence in the time series literature. A data-driven studentization, new to non-parametric and cross-sectional contexts, is theoretically justified, then used to develop asymptotically correct inference. Chapter 3 discusses identification and kernel estimation of a non-parametric common regression with additive individual fixed effects in panel data, with weak temporal dependence and arbitrarily strong cross-sectional dependence. An efficiency improvement is obtained by using estimated cross-sectional covariance matrix in a manner similar to generalised least-squares, achieving a Gauss-Markov type efficiency bound. Feasible optimal bandwidths and feasible optimal non-parametric regression estimation are established and asymptotically justified. Chapter 4 deals with efficiency improvement in the estimation of pure Spatial Autoregressive model. We construct a two-stage estimator, which adapts to the unknown error distribution of non-parametric form and achieves the Cramer-Rao bound of the correctly specified maximum likelihood estimator. In establishing feasibility of such adaptive estimation, we find that the gain in efficiency from adaptive estimation is typically smaller than in the relevant time series context, but could be also greater under certain asymptotic behaviour of the weight matrix of the model.

To my parents

Acknowledgments

First of all, I am greatly indebted to my advisor, Professor Peter Robinson, for his generous advice, guidance and encouragement.

Many thanks are due to Abhisek Banerjee, Ziad Daoud, Liudas Giraitis, Abhimanyu Gupta, Javier Hidalgo, Tatiana Komarova, Oliver Linton, Francesca Rossi, Myung Hwan Seo, Marcia Schafgans and Sorawoot Srisuma for plenty of helpful comments and discussions. I also thank all my friends and fellow researchers at the LSE, in particular the participants at the work in progress seminars, for discussions and comments on my work.

I owe a lot to my family. I deeply thank my parents for their unending support and encouragement over the years, my two brothers, Jungjoon and Jungho, for their joyful company. Special thanks are due to my husband, Waki, for brightening up my days with his humour and love. All this would not have been possible without them. I also thank all the friends who made me feel at home in London.

Financial support from Samsung Scholarship Foundation and ESRC Grant RES-062-23-0036 is most gratefully acknowledged.

Contents

List of Tables

1 Introduction

This chapter provides an overview of the motivations for, and contributions of, the present thesis. A review of relevant literature on the topics of cross-sectional dependence and non- and semi-parametric methods is provided in detail, in order to help place the thesis in perspective. We then summarise the main contributions of each of the three topics of this thesis in relation to the existing studies.

1.1 Cross-sectional dependence

Three types of data are encountered in economics, namely, time series, cross-section and panel data. This thesis focuses on estimation and inference for the latter two types and consists of three chapters developing non- and semi-parametric methods that either allow for or estimate cross-sectional dependence in the disturbance terms of a regression model. Implications of possible dependence between cross-sectional units on econometric methods has been less studied in the literature than that of dependence across time periods. Unfortunately, the nature of cross-sectional dependence and heterogeneity observed in economic data hinders a simple extension of the time series literature to cross-section or panel data.

In economic datasets, cross-sectional units naturally correspond to economic entities, such as individuals, households, firms, industries, cities, regions or countries. A typical type of dataset involving smaller units like individuals or households consists of survey data collected by governments or firms using various sampling schemes. The most prevalent sampling schemes encountered in economics are simple random sampling where each unit has the same probability of being sampled, cluster sampling where clusters consisting of individual units are sampled, or stratified sampling where units in the sample are represented with different frequencies than they are in the population, see Wooldridge (2002, pp. 132-135) for a good exposition. When the cross-section units are larger entities such as firms within an industry, regions or countries, the sampling may be exhaustive, i.e. all population units are observed in the data. It is obvious that the need to allow for dependence and heterogeneity across cross sectional units is even more compelling when the sample coincides with the population.

A standard practice in the econometric literature, particularly with survey data, has been to assume that cross-sectional observations are independent and identically distributed (*i.i.d.*). An exception to this is the literature on data collected using cluster sampling, where accounting for possible group effects via cluster-robust standard errors of Liang and Zeger (1986) is widely available. This method allows for arbitrary dependence within clusters but assumes independence between clusters and works well when the number of clusters is large relative to the sample size. There seems to be a common misconception that the simple random sampling scheme leads to the *i.i.d* property of the collected data. This, along with the difficulty of dealing with cross-sectional dependence and heterogeneity of economic data in theoretical development, partly explains the relative lack of econometric literature concerning cross-sectional dependence in survey data. In the case of larger cross-sectional units, there has been relatively more literature that allows for cross-sectional dependence, as will be discussed below.

In the case of survey data, it is important to appreciate that when there is dependence and heterogeneity between underlying cross-sectional units in the population, the *i.i.d* assumption on the sampled units is *at best* an approximation, that needs to be carefully weighed against the specific data setting under consideration. Even the simple random sampling scheme does *not* warrant that the sampled units are *i.i.d.*, as clearly exposited in Andrews (2005) and Conley (1999). They offer probabilistic frameworks which first define random vectors for all units in the population, not just the observed units, and then consider drawing sampled units from the population.

There are two possible sources of dependence in the error terms of cross sectional units that have been discussed in the econometric literature. Firstly, there may be common shocks that affect all or some of individual units. Andrews (2003) gives a comprehensive discussion on possible common shocks that may arise in economic contexts, such as macroeconomic, technological, legal/institutional, political, environmental, health and sociological shocks. Such shocks could have either global or local effects, influencing individual units in a possibly heterogeneous manner, that may depend on the unit's characteristics. Secondly, there may be dependence between individual units' unobservables due to their economic interactions. Conley (1999) provides an example where insurance contracts are made by risk-averse agents in order to smooth individual idiosyncratic shocks. This inevitably leads to dependence in consumption across those individuals. Another example arises due to spill-over between agents: an idiosyncratic productivity shock to a firm/industry, such as technological innovation, may subsequently affect the productivity of other related firms/industries. Yet another example arises in hedonic pricing model of houses: neighbouring houses may share similar unobservable characteristics resulting in spatial dependence in the disturbance terms, although this example does not arise from economic interaction as such. In these three examples, it is clear that such dependence will be governed by the degree of interaction/proximity between units. Dependence arising from economic interaction is likely to be local in nature, in contrast to that generated by the presence of common shocks, which can produce either global or local effects.

1.1.1 Models of cross-sectional dependence in disturbance terms

This subsection discusses three existing classes of models for cross-sectional dependence in disturbances.

Models with common shocks

For the case of common shocks, recent works by Bai (2009) and Pesaran (2006) consider the linear regression model with large N, large T panel data. They model the error term of the i-th cross sectional unit's t-th time period observation as, $U_{it} = \lambda_i' F_t + \varepsilon_{it}$, where F_t is the vector of unobserved common factors, λ_i the vector of individual-specific factor loadings, giving rise to cross-sectional dependence, and ε_{it} the idiosyncratic error. Both papers allows the unobserved factors, F_t, to also affect the regressors linearly, which seems plausible especially in macroeconomic settings such as cross-country data, for which the large N and large T asymptotic framework of the papers is particularly relevant as N and T may be of similar magnitude. This however results in the component $\lambda_i' F_t$ in the error terms that are correlated with the regressors, which can be seen as individual-specific and time-varying "fixed effects" that cannot be purged by simple data transformation like first differencing. The two papers provide estimation methods that lead to consistent estimates of the linear parameters of the regression model despite the presence of such fixed effects. The estimation methods of the above papers are unfortunately not applicable to cross-section data and it is not straightforward to extend similar methods to nonlinear or non- and semi-parametric regression models or to relax the linear specification in which the unobserved factors affect the disturbance term and/or regressor.

Andrews (2005) looks at the linear regression model with cross-section data when there is arbitrary dependence and heterogeneity in the error terms between underlying units in population generated by the presence of common shocks, and observations are collected using random sampling. He derives asymptotic properties of the least squares (LS) estimates of the linear parameters of the regression and establishes a necessary and sufficient condition for consistency. This condition requires the regressors and errors to be uncorrelated conditional on the σ-field, $\mathcal{C}$, generated by the common shocks. The random sampling assumption implies that observations are *i.i.d.* conditional on $\mathcal{C}$, needed for law of large numbers (LLN) and central limit theorem (CLT) results that are used to show asymptotic properties of the LS estimate. The asymptotic framework offered by Andrews (2005) is indeed very useful for survey data collected using random sampling schemes but not when random sampling does not hold.

Spatial models

For cross-sectional dependence in the unobservables arising from economic agents' interdependence, two classes of models of dependence have been prominent in recent literature, involving a concept of "economic location". As mentioned above, cross-sectional units in economic data correspond to economic agents such as individuals or firms. One could envisage that these agents are positioned in some socio-economic (even geographical) space, whereby their relative locations in this space underpin the strength of dependence between them. For a detailed discussion and examples of such proximity, see e.g. Conley (1999) and Pinkse, Slade and Brett (2002). These models are often called "spatial", reflecting the inclusion of space in the set-up.

The first type of model includes the pure Spatial Autoregressive (SAR) and related models, which form a part of a more general class of models, first suggested by Cliff and Ord (1968). This class of models is characterized by the use of exogenously given weight matrices, which capture the structure of spatial dependence between units up to a finite number of unknown parameters. In the case of modelling dependence in the error terms, the spatial dependence is simply modelled parametrically as a linear transformation of underlying shocks. Let $U = (U_1, \cdots, U_n)'$ be a vector of observations having zero mean, with the prime denoting transposition. The class of spatial dependence models is given by

$$Q\left(\lambda_0\right) U = \sigma_0 \varepsilon, \tag{1.1.1}$$

where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$ is a vector of *i.i.d.* random variables with zero mean and unit variance, σ_0 is a scalar, λ_0 is a finite-dimensional vector of parameters, and $Q(\lambda_0)$ is a known, non-singular $n \times n$ matrix function of its argument. In general λ_0, μ_0 and σ_0 are unknown and Q depends on one or more known spatial weight matrices. Denote by W a generic $n \times n$ matrix,with real-valued elements w_{ij} such that

$$w_{ii} = 0, \qquad \sum_{j=1}^{n} w_{ij} = 1, \quad i = 1, \cdots, n. \tag{1.1.2}$$

The latter condition, called row normalization restriction, is not always imposed in the literature, but some normalization on W is required in order to identify λ_0. The quantities w_{ij} are typically interpreted as inverse economic distances, see e.g. Arbia (2006), and may form triangular arrays. The following are three examples of Q in which λ_0 is scalar:

1. Pure SAR(1) (spatial autoregression of degree 1)

$$Q(\lambda_0) = I - \lambda_0 W, \tag{1.1.3}$$

   where I is the $n \times n$ identity matrix and $\lambda_0 \in (-1, 1)$.

2. Pure SMA(1) (spatial moving average of degree 1)

$$Q(\lambda_0) = \left(I - \lambda_0 W\right)^{-1},$$

   for $\lambda_0 \in (-1, 1)$.

3. MESS (matrix spatial exponential, see LeSage and Pace (2009)):

$$Q(\lambda_0) = \exp\left(-\lambda_0 W\right).$$

The clear limitation of these models is the presumption that the spatial dependence is known to the practitioner up to a small number of parameters (λ_0). Nonetheless, these models, the pure SAR model in particular, have gained popularity in empirical works, see Arbia (2006) for examples. In these models where the spatial dependence is parsimoniously captured by the unknown λ_0, the estimation of λ_0 is often of interest, possibly for the purpose of testing for lack of spatial dependence. In chapter 3 of the thesis, efficient estimation of λ_0 in a generalised version of (1.1.3) is considered.

The second class of models involves the use of mixing coefficients familiar from the time series literature. Suppose unit i is endowed with a vector of characteristics z_i, the economic distance between units i and j is defined as the distance between z_i and z_j, e.g. the Euclidean norm $\|z_i - z_j\|$. Conley (1999) approximates the locations z_i by regularly spaced lattice points and applies strong mixing conditions in deriving asymptotic theory for generalized method of moments (GMM) estimates. An alternative mixing condition in spatial setting was proposed in Pinkse, Shen and Slade (2007). Mixing conditions, in contrast to the SAR and related models mentioned above, are essentially non-parametric, desirably avoiding a specific parametric description of dependence.

It is notable that in the models with common factors, cross-sectional dependence is allowed to be "strong" as well as "weak", in the sense that common shocks are allowed to affect all units in the sample (and population) significantly. In contrast, the afore-mentioned spatial models require spatial dependence to fall as the economic distance between units increases, sufficiently fast that the strength of spatial-dependence satisfies weak dependence conditions analogous to ones in time series literature. For this thesis, the "weak" dependence in U_i is defined by the condition $\sum_{i,j=1}^{n} |Cov(U_i, U_j)| = O(n)$, which is analogous to the concept of weak dependence in stationary time series: $\sum_{k=-\infty}^{\infty} |Cov(U_1, U_{1+k})| < \infty$. "Strong" dependence in U_i, on the other hand, is defined by the condition $\sum_{i,j=1}^{n} |Cov(U_i, U_j)|/n \to \infty$ as $n \to \infty$. In Chapter 4, it is explained how the existing SAR literature imposes weak dependence restriction. In the case of weak dependence, the common factor models and spatial models may produce similar patterns of dependence, although the motivation and specification of the disturbance terms may be rather different.

Model of Robinson (2011)

Robinson (2011) provides an alternative way of modeling cross-sectional dependence, which can produce strong as well as weak dependence, and need not involve known economic distances although can readily accommodate them. The following general, possibly non-stationary, linear process is used to describe the disturbance of a regression model:

$$U_i = \sigma_i(X_i)e_i, \quad e_i = \sum_{j=1}^{\infty} b_{ij}\varepsilon_j, \quad \sum_{j=1}^{\infty} b_{ij}^2 = 1, \quad 1 \le i \le n, \quad n = 1, 2, \cdots, \tag{1.1.4}$$

where U_i is the scalar disturbance term, X_i a finite dimensional vector of regressors in the regression model, ε_j's are independent random variables with zero mean and unit variance that are independent of $\{X_i, i = 1, \cdots, n, n \ge 1\}$, σ_i's are scalar unknown functions and b_{ij}'s are unknown fixed weights. These weights b_{ij}'s, and hence U_i's, may form triangular arrays, and the reference to n is suppressed for ease of notation. Notice that e_i's are generated by summation over $j = 1$ to infinity, letting the sampled units be also affected by unsampled units, in contrast to the pure SAR and related models. This specification allows both unconditional and conditional heteroscedasticity. The triangular array structure also accommodates the panel data case where some relabeling of observations would be required if both T and N are allowed to grow as $n = NT \to \infty$. As the unknown weights b_{ij}'s may vary across i and j, the above specification offers a general model of spatial dependence.

An important question to ask when specifying the model for the disturbance in a regression with stochastic regressors is the extent to which the disturbance term is dependent with the regressors. Pesaran (2006) and Bai (2009) allow the same set of unobserved factors to enter both regressors and error terms, and Andrews (2005) requires that they are uncorrelated conditional on the σ-algebra of common shocks. In comparison, the above specification is relatively more restrictive in that e_i is independent of the regressors X_i's. In particular, one may be concerned that the dependence patterns between units i and k in their disturbance terms and the regressors may be similar, especially in the spatial setting where they may be governed by the same distance measure between the units. The specification (1.1.4) does allow the dependence between units i and k in regressors and disturbances to be related. For example, one could let the joint density function f_{ik} of X_i and X_k, which reflects the dependence between two units' regressors, be a function of a distance between unit i and k, denoted d_{ik}, i.e. $f_{ik}(x, y) = f(x, y; d_{ik})$ and at the same time also allow the weights b_{ik}'s to be governed by the same distance measure, $b_{ik} = b(d_{ik})$.

1.1.2 Some implications of cross-sectional dependence

The consequence of cross-sectional dependence in estimation of a regression model varies according to the strength of dependence. It has been shown in the time series literature that weak dependence typically does not affect consistency or asymptotic normality results of parameter estimates, but does alter their variances relative to the *i.i.d.* setting. Therefore disturbance variance structures need to be suitably estimated in order to carry out valid inference. In case of strong dependence, depending on the specification, even consistency may sometimes break down, although in the afore-mentioned papers that allow for strong dependence (Andrews (2005) and Robinson (2011)), consistency and asymptotic normality of estimates were shown under suitable conditions. The issues discussed in Pesaran (2006) and Bai (2009) are rather different as there is the additional problem of correlation between regressors and disturbance terms and the two papers offer new methods of estimation that achieve consistency of regression parameter estimates.

Developing standard errors that are robust to dependence and heterogeneity is considerably more difficult in the cross-sectional setting than in time series, where so called heteroscedasticity autocorrelation consistent (HAC) estimation is facilitated by the information carried by the time index. The dependence between observations at times t and s is modelled in terms of $|t-s|$. In the spatial context, an extension of HAC estimation is feasible if additional information which may take the role of the time indices is available e.g. the socio-economic or geographical distance between units which underpin the structure of the spatial dependence. Conley (1999) has considered HAC estimation under a stationary random field with measurement error in distance measures, Kelejian and Prucha (2007) for models of Cliff and Ord (1968) and Robinson and Thawornkaiwong (2010) for a more general set-up than Cliff-Ord type models. Chapter 1 of this thesis offers an alternative method of robust inference to that based on HAC estimation.

1.2 Non- and semi-parametric methods in economics

In the afore-mentioned papers, regression models take a parametric form, with the exception of Robinson (2011). However economic theory usually does not imply a particular functional form and there may be little confidence that a linear or specific nonlinear regression model is correctly specified. Non-parametric estimation allow researchers to drop the presumption of known functional form, instead requiring non-parametric restrictions such as smoothness and existence of certain moments, that may be less restrictive. In some contexts, specifying some components of the model to be parametric while keeping the others non-parametric may be more appealing than a fully non-parametric specification. This could be either due to practical reasons, such as avoiding "curse of dimensionality" in non-parametric regression with many regressors, or due to the practitioner having confidence and interest in parameterization of some components while not in the others. Such models are called semi-parametric. Chapter 2 consider estimation of *known* functionals of the non-parametric regression function, an example of which is estimation of slope parameters in partly linear regression model, in which some regressors enter linearly and others non-parametrically. In Chapter 4, the regression model is parametric but the unknown error density is left to be non-parametric. A non-parametric estimate of the "score" function, the ratio of the first derivative of the error density to the density itself, is used in order to construct an estimate of the parameter of the regression model that achieves the

Cramer-Rao bound of the correctly specified MLE.

There are some differences in the type of theoretical results we typically obtain for non-parametric and semi-parametric estimates. Non-parametric estimates achieve a slower rate of convergence to the true value, compared to the correctly specified parametric estimates, which is a natural consequence of the parsimony of the latter. In contrast, some semi-parametric estimates have been shown to achieve the same rate of convergence as its parametric counterpart under suitable conditions, which is remarkable since semi-parametric estimates rely on the first stage non-parametric estimates that exhibit a slower rate of convergence. This type of results has received wide interest in econometric literature, starting from Robinson (1988) and Powell, Stock and Stocker (1989), which deal with the two well-known semi-parametric models, partly linear regression model and single index model, respectively. Chapter 2 of this thesis provides a set of sufficient conditions, including those on the strength of dependence and heterogeneity in the data, for semi-parametric estimates to obtain the parametric rate of convergence.

There are two main methods of non- and semi-parametric estimation used in econometrics. The first method is the kernel approach, which uses local smoothing/averaging with a chosen kernel function and bandwidth parameter. The second is the sieve estimation method, which uses an increasing number of base functions as the sample size increases, to approximate the non-parametric function of interest. In this thesis, the focus is on estimation of the conditional expectation, i.e. regression function, which is the topic of the first two chapters, and the score function used in the third chapter is estimated using the same method as regression estimation. In the sieve estimation literature, regression estimation has been developed under the name of series estimation and in the context of this thesis, the terms "sieve" and "series" are used interchangeably.

There are many theoretical results on kernel estimation of regression and density under temporal dependence, see e.g. Roussas (1969), Rosenblatt (1971), Robinson (1983) for weak dependence and Robinson (1991), Robinson (1997), Hidalgo (1997) for strong dependence. A notable difference from parametric estimation is that in the case of weak dependence, the asymptotic variance of non-parametric kernel estimates are the same as in the *i.i.d.* setting, which arises from the local nature of estimation. Robinson (2011) and Robinson and Thawornkaiwong (2010) have considered kernel estimation in non-parametric regression and partly linear regression, respectively, under strong and weak cross-sectional dependence and found similar results to the time series literature.

The main advantages of series estimation over kernel estimation are four-fold. When economic theory generates certain restrictions on the non-parametric function of interest, such as monotonicity, convexity and additive separability, series estimation offers a more natural way of using such information in estimation by reflecting it in the choice of series functions. Secondly, it is computationally convenient, because the data is summarized by a relatively few estimated coefficients. Thirdly, from a theoretical point of view, theories can be developed in a unified way to include both non-parametric regression and general semi-parametric quantities, as will be made clear in Chapter 2. This is in contrast to kernel estimation where an asymptotic theory for each semi-parametric model needs to be developed separately. Finally, semi-parametric estimation with kernel methods typically involves ”trimming” out some observations, if the density estimates at their values are smaller than a certain trimming parameter. This is because of the form of many kernel estimates which have as their denominator, the random density estimate that can be very close to zero. Introduction of a user-chosen trimming-parameter could be in itself a disincentive to the practitioner, while generating complications in development of econometric theory behind estimation. Semi-parametric estimation of series method is free from the need for trimming.

The asymptotic behaviour of series estimation under independence has been studied in Andrews (1991) and Newey (1997). For weakly dependent time series data, Chen and Shen (1998) and Chen, Liao and Sun (2011) together offer asymptotic theory and robust inference of general sieve M estimation, which includes series estimation as a special case. Importantly, Chen, Liao and Sun (2011) found that certain non-parametric sieve estimates under weak temporal dependence also have the same asymptotic variance as in the *i.i.d.* setting, analogous to the result reported for kernel estimation in Robinson (1983). Chapter 2 of this thesis establishes asymptotic theory for a spatial setting similar to Robinson (2011), which covers strong, as well as weak, dependence. In future work, it would be of interest to extend Chen *et al.* (2007)’s finding to the spatial setting and compare asymptotic results on series estimates of non-parametric regression to those of kernel estimates, reported in Robinson (2011).

1.3 Summary of main contributions

This section highlights the contributions of each of the three chapters of this thesis, and place them in relation to the existing studies.

1.3.1 Chapter 2

In Chapter 2 ”Series estimation under cross-sectional dependence”, the following model is studied,

$$\begin{aligned} &Y_i = m(X_i) + U_i, \quad \text{where} \quad U_i = \sigma(X_i)e_i, \\ &e_i = \sum_{j=1}^{\infty} b_{ij}\varepsilon_j, \quad \sup_i \sum_{j=1}^{\infty} b_{ij}^2 < \infty, \quad 1 \le i \le n, \quad n = 1, 2, \cdots, \end{aligned}$$

where $Y_i, U_i \in \mathbb{R}$ and $X_i \in \mathcal{X} \subset \mathbb{R}^q$ are random variables, $m : \mathcal{X} \to \mathbb{R}$ is an unknown function of interest, $\sigma(\cdot)$ is a real bounded function, $\{b_{ij}, i, j \ge 1\}$ are unknown constants and $\{\varepsilon_j, j \geq 1\}$ are independent random variables with zero mean and unit variance. Processes $\{X_i\}$ and $\{\varepsilon_j\}$ are assumed to be independent of each other. The specification of U_i has been slightly modified from Robinson (2011), but both conditional and unconditional heteroscedasticities are still allowed.

The quantity of interest is a $d \times 1$ functional of m denoted $\theta_0 = a(m)$, which is estimated by plugging in the series estimate $\hat{m}$ of m in the known functional operator $a(\cdot)$. Theorem 2.1 reports a uniform rate of convergence of $\hat{m}$ to m. A uniform rate of convergence for non-parametric regression estimates is useful for many semi-parametric problems and has been extensively studied in the context of kernel estimation, see e.g. Masry (1996) and Hansen (2008). For series estimation, Newey (1997) provided a rate for the *i.i.d.* setting, which was subsequently improved by de Jong (2002), under the additional assumption of compact $\mathcal{X}$. The rate result obtained in Theorem 2.1 reduces to that of Newey (1997) under the *i.i.d.* setting and contains a variance contribution term that reflects the collective dependence in the U_i's and X_i's.

In this work, dependence in $\{X_i\}$ is allowed to be strong, as well as weak, and two measures of dependence are introduced. The first is used in showing the consistency and CLT results, while the second is needed for the variance matrix of the estimate $\hat{\theta} = a(\hat{m})$ conditional on $\{X_i\}$ to be well-behaved so that the unconditional asymptotic variance matrix can be obtained. The first measure of dependence is defined in terms of departure of bivariate density function from the product of marginals. Denote by f_{ij} the joint density function of X_i and X_j and define,

$$\triangle_n := \sum_{i,j=1, i \neq j}^{n} \int_{\mathcal{X}^2} |f_{ij}(x,y) - f(x)f(y)| dxdy.$$

The rate of growth of $\triangle_n$ is a measure of bi-variate dependence in the X_i's and has an upper bound of $2n^2$. The quantity $\triangle_n$ is zero in case of independence across i and we may view the condition $\triangle_n = O(n)$ as an analogue to short-range/weak dependence in the time series literature. We find an upper bound on $\triangle_n$ for the case of Gaussian X_i's to be $\sum_{i,j=1, i \neq j}^{n} |Cov(X_i, X_j)|$, which is the quantity used in the definition of weak dependence earlier.

Similar measures of dependence have been used in Robinson (2011), where the local nature of kernel estimation meant conditions were confined to the neighbourhoods of the points at which the function m was estimated. For establishing asymptotic normality result for kernel estimates of m, Robinson (2011) also imposed conditions on the third and fourth order joint probability densities of X_i.

For the partly linear regression model, Robinson and Thawornkaiwong (2010) also offered a global measure of dependence which involves the supremum over the entire support of X_i. However, the measures of dependence used in their Assumption B6 are more complicated and less tractable than $\triangle_n$ and they impose uniform boundedness of marginal and joint densities of order up to four. We avoid imposing restrictions on the third and fourth order joint densities of X_i's, which are considerably harder to verify than that involving bivariate density even in the simple case of Gaussian random variables. Instead we formulate our second measure of dependence in $\{X_i\}$ in terms of the fourth order cumulants of quantities combining U_i and X_i. Our restrictions on their collective dependence allow strong, as well as weak dependence in both X_i's and U_i's. Cumulants have been often used in the time series literature as a measure of dependence, see e.g. Brillinger (1981). Chapter 2 also provides sufficient conditions for $\hat{\theta} = a(\hat{m})$ of certain smooth functionals $a(\cdot)$ to be $\sqrt{n}$-consistent, extending Newey (1997)'s results obtained in the *i.i.d* setting. It is interesting that some strong dependence in X_i is allowed although the U_i's need to be weakly dependent for the $\sqrt{n}$-consistency result.

Chapter 2 of this thesis also establishes theoretical justification for the use of the studentization method of Kiefer, Vogelsang and Bunzel (2000), which is new to the cross-sectional setting and non- and semi-parametric methods. In time series, HAC estimation of the asymptotic variance matrix is widely known to perform poorly in small samples, see e.g. Andrews and Monahan (1992) and Den Haan and Levin (1997). Kiefer *et al.* (2000) offer a data-driven studentization method that can produce better small sample performance than HAC-based inference. Kiefer *et al.* (2000)'s assumption A1 requires a functional central limit theorem (FCLT) result on a data-driven quantity that forms a basis of the studentizing matrix. They provide conditions of Phillips and Durlauf (1986) as an example of a set of sufficient conditions for this FCLT result to hold. Phillips and Durlauf (1986)'s conditions require weakly stationary α-mixing sequences. The contributions of the extensions offered by the current thesis are as follows.

1) This is the first work to apply Kiefer *et al.* (2000)'s studentization method in non- and/or semi-parametric contexts to the best of our knowledge.

2) We relax the conditions of homogeneity, regular spacing and ordering of U_i's and X_i's, which are exhibited by stationary time series, but not by cross-sectional data.

Indeed, a notable contribution of Chapter 2 is evaluating the extent to which the dependence and heterogeneity of the U_i's can depart from stationary mixing and still achieve the FCLT result required in order to apply Kiefer *et al.* (2000)'s studentization method. The degree of relaxation of the regularity conditions is summarised by Assumption C3 of Chapter 2, where a detailed discussion can also be found.

The main results of Chapter 2 are, Theorem 2.1 which states the uniform rate of convergence for the series regression estimate, Theorem 2.2 that presents the asymptotic distribution of the estimate of a functional of the regression function, and Theorem 2.5 which establishes the validity of the studentization method.

1.3.2 Chapter 3

Chapter 3, "Panel data model with non-parametric common regression and individual fixed effects", considers the following model for a balanced panel data set of size $N \times T$. Below Y_{it} denotes a one dimensional dependent variable, λ_i an additive individual-specific fixed effect of individual i, Z_t is a vector of time-varying regressor common to all individuals, whereas $m(\cdot)$ is the non-parametric regression function of interest, and U_{it} denotes the error term:

$$Y_{it} = \lambda_i + m(Z_t) + U_{it}, \quad E(U_t U_t' | Z_t = z) = \Omega(z), \quad i = 1, \cdots, N, \quad t = 1, \cdots, T,$$

where $\Omega(\cdot)$ is a $N \times N$ matrix of smooth functions. Importantly, an arbitrary form and strength of cross-sectional dependence are allowed in U_{it}, while Z_t and U_{it}'s are required to satisfy a β-mixing condition over time. Nadaraya-Watson (N-W) kernel estimate is used to estimate $m(\cdot)$. The setting in mind is with larger T than N, such as in regional data with long time series. Cross-sectional units are typically large entities like regions or countries in such settings, where the need to allow for strong dependence and heterogeneity across the cross-section may be compelling. Therefore we do not impose restrictions on Ω, other than some smoothness and boundedness conditions.

A similar model was considered in Robinson (2011) in the context of common trend estimation, with Z_t replaced by the deterministic argument t/T where U_{it}'s are assumed to be *i.i.d.* across time. He clarified the issue of joint identification of $m(\cdot)$ and λ_i's and showed how to incorporate the knowledge of cross-sectional dependence in U_{it}'s into estimating $m(t/T)$ in order to obtain an efficiency gain. In particular, a generalised least squares (GLS) type estimate under the full knowledge of cross-sectional dependence was shown to be superior in the mean square error (MSE) sense, to the one that does not incorporate such information. Asymptotic equivalence between the infeasible and feasible GLS type estimates was also established.

Chapter 3 of this thesis essentially extends the results of Robinson (2011) to the case of multivariate non-stochastic regressor, which leads to the need for conditional heteroscedasticity captured in $\Omega(z)$. We also allow Z_t and U_{it}'s to be jointly weakly dependent over time, rather than using the *i.i.d.* condition imposed on U_{it}'s in Robinson (2011). We first establish asymptotic MSE, the consequent optimal bandwidth choice and asymptotic distribution of a simple N-W estimate of $m(\cdot)$ based on the simple cross-sectional average of Y_{it}'s, $\sum_{i=1}^{n} Y_{it}/n$. We then obtain similar results for the optimal N-W estimate, based on the knowledge of the cross-sectional error covariance structure Ω. This optimal estimate is analogous to the GLS estimate in linear regression model where a data transformation based on Ω produces transformed error terms that are homoscedastic and uncorrelated. A similar principle is used here, leading to the optimal estimate achieving a Gauss-Markov type bound. We then construct a feasible version of such optimal estimate using an estimate of Ω based on the fitted residuals from the simple N-W estimation. We establish asymptotic equivalence between the feasible and infeasible versions of the optimal N-W estimate and also between their optimal bandwidth choices. Unlike in Robinson (2011), the conditional heteroscedasticity here implies that the optimal weight for the GLS-type estimation now varies over the point z at which the function $m(\cdot)$ is estimated.

In obtaining the theoretical results, we prove a useful result, Lemma 3.6, that represents an additional and significant contribution of the chapter. Many sample quantities in econometrics take the form of a U-statistic, whose properties in the *i.i.d.* setting are well understood. Similar results on the behaviour of U-statistics under dependence is often obtained by showing the negligibility of the departure of the U-statistic under dependent process from its counterpart under the *i.i.d.* setting. Dehling (2006) offers an excellent review of both the literature under the *i.i.d.* setting and that covering the dependent case. Fan and Li (1999) utilized Yoshihara (1976)'s lemma, to obtain an upper bound on the difference in expectations of a third order U-statistic under independence and the β-mixing process. Lemma 3.6 of this chapter extends Fan and Li (1999)'s result to U-statistics with an asymmetric kernel and of order up to four. This lemma would be useful in many applications, especially when finding the asymptotic order of magnitude for moments of various estimates with time series data.

The main results of Chapter 3 are, Theorem 3.7 which establishes how good an estimate of the cross-sectional error covariance matrix we have, and Lemma 3.6 that presents useful decomposition of the expectation of U-statistics based on β-mixing processes.

1.3.3 Chapter 4

While Chapters 2 and 3 deal with non-parametric regression models, in Chapter 4 "Efficiency improvement in the semi-parametric pure Spatial Autoregressive (SAR) model", we consider the parametric pure SAR model, presented in (1.1.1) and (1.1.3). We state the model again for ease of reference:

$$(I - \lambda_0 W)U = \sigma_0 \varepsilon. \tag{1.3.1}$$

For the weight matrix W, it is assumed that $\max_{1\le i,j\le n} |w_{ij}| = O(1/h)$, with a sequence $h = h_n$ that is either fixed or divergent as $n \to \infty$. This is a typical condition imposed in the SAR literature, e.g. in Lee (2002, 2004), Kelejian and Prucha (1998), and it has been shown that the behaviour of the sequence h has implications on asymptotic theory of parameter estimates. In Chapter 4, U_i will be allowed to have a non-zero mean, providing a model for observables, as well as disturbances. However for clarity of exposition, we stick to the simpler version (1.3.1) in the present description.

Estimation of λ_0 in pure SAR model was considered in Lee (2002, 2004). In the former, the ordinary least squares (OLS) estimate was shown to be inconsistent while in the latter the Gaussian pseudo maximum likelihood estimate (PMLE) was shown to be consistent, at the usual rate $\sqrt{n}$ when h is fixed, and at a slower rate, $\sqrt{n/h}$, if h is divergent. When ε_j's, therefore U_i's, are Gaussian, the Gaussian PMLE is of course the MLE itself, attaining the Cramer-Rao bound. This property is lost once the true likelihood seizes to be Gaussian. The lack of efficiency property of the Gaussian PMLE of λ_0 when U_i's are not Gaussian, in addition to the possibly slower rate of convergence in case of divergent h, gives rise to an interest in improved estimation of λ_0. This is the focus of Chapter 4 and we treat μ_0 and σ_0 as nuisance parameters.

We let ε_i's be *i.i.d.* with *unknown* density function of non-parametric form, thus avoiding possible parametric misspecification of density function, which could lead to inconsistency of the corresponding MLE. This is what makes the model semi-parametric, as it contains a non-parametric error density function along with a parametric regression model. There is a large literature addressing whether the Cramer-Rao bound of the correctly specified MLE can be achieved in the absence of knowledge of the density function, with only non-parametric assumptions. This is attained by an estimate that takes an approximate Newton-step from an initial consistent and inefficient estimate of λ_0, which is the Gaussian PMLE in our case. In Beran (1976), Newey (1988), Robinson (1995) and Robinson (2011), the Newton-step is constructed from a non-parametric series estimate of the score function of the error density. Chapter 4 also uses this estimate and establishes that it indeed achieves the Cramer-Rao bound of the correctly specified MLE.

Another notable and interesting finding of Chapter 4 is that the relative efficiency of the adaptive estimate $\hat{\lambda}$ to the PMLE can be either less or more than ones in the classical outcome, which includes the results under mixed regressive SAR model considered in Robinson (2011) and the time series setting. As mentioned earlier, pure SAR model is a particularly popular model in the more general class of models and it is hoped that the results of this chapter may be extended in future to other models.

The main results of Chapter 4 are, Lemma 4.1 which establishes feasibility of efficiency improvement from the (Gaussian) PMLE of λ_0, and Theorem 4.1 which presents the asymptotic distribution of the improved estimate, which coincides with that of the true MLE.

2 Series Estimation under Cross-sectional Dependence

2.1 Introduction

Economic agents are typically interdependent, due for example to externalities, spill-overs or the presence of common shocks. Such dependence is often overlooked in cross-sectional or panel data analysis, in part due to a lack of econometric literature that deals with the issue at hand. Implications of dependence on econometric analysis have long been studied in the context of time series data, where the temporal dependence is naturally modeled in terms of the distance between observations along the time axis. Unfortunately, the nature of cross-sectional dependence observed in economic data hinders a simple multi-dimensional extension of time series literature to spatial data. For example, the index of observations in economic cross-sectional data cannot be used to describe the dependence between units in the way that the time index can be. This is because there is often no natural ordering of cross-sectional data and the indices do not represent relative positioning of the units sampled.

In order to start accounting for possible cross-sectional dependence, one needs first to establish a framework under which the structure of such dependence can be suitably formalised. Three classes of models of cross-sectional dependence have been prominent in recent literature. The first class of models deal with the presence of unobserved common factors that may affect some/all of individual units, see Andrews (2005), Pesaran (2006) and Bai (2005). These models could give rise to cross-sectional dependence that are persistent throughout units, analogous to "strong" or "long-range" dependence in the time series literature.

The other two classes of models involve a concept of "economic location". In economic data, cross-sectional units correspond to economic agents such as individuals or firms. One could envisage that these agents are positioned in some socio-economic (even geographical) space, whereby their relative locations in this space underpin the strength of dependence between them. For a detailed discussion and examples of such proximity, see e.g. Conley (1999) and Pinkse, Slade and Brett (2002).

The second class of models is the Spatial Autoregressive (SAR) model of Cliff and Ord (1968, 1981), see e.g. Lee (2002, 2004), Kelejian and Prucha (1998, 1999), Robinson (2010a), Rossi (2010). In this approach, the dependent variable (or disturbance) of a given unit is assumed to be affected by a weighted average of the dependent variables (or disturbances) of the other sampled units. The weights used in the averaging are presumed to be known and reflect the degree of proximity between agents, leaving a finite number of parameters (often scalar) to be estimated to explain the spatial dependence. The SAR model has gained popularity in empirical works, see e.g. Arbia (2006).

The third class of models involve the use of mixing coefficients familiar from the time series literature. Conley (1999) and the related papers develop spatial mixing conditions in terms of economic distance between agents, under a suitable stationarity assumption. An alternative mixing condition in spatial setting was proposed in Pinkse, Shen and Slade (2007).

Robinson (2011) has offered a new way of modeling cross-sectional dependence, which does *not* hinge on the idea of economic distance although can certainly accommodate it. A general, possibly non-stationary, linear process is assumed for disturbances, which, unlike a mixing framework, allows possible strong dependence. The dependence in the regressors is phrased in terms of the departure of joint densities from the product of marginals, allowing possible heterogeneity across units. The model's ability to cover both weak and strong dependence in the error term and regressors allows the development of a general set of theory. While the model accommodates many spatial settings plausible in economic data, no new concepts, other than those familiar from standard econometric literature, need to be introduced.

On the other hand, non-parametric and semi-parametric estimation have become an established method in econometric analysis. Such methods allow researchers to drop the assumption of known parametric functional form that is often not warranted by economic theory. There are many theoretical results on non-parametric kernel estimation under temporal dependence, see e.g. Robinson (1983) and Hidalgo (1997). Robinson (2011) and Robinson and Thawornkaiwong (2010) have considered kernel estimation in the non-parametric regression model and the partly linear regression model, respectively, under cross-sectional dependence.

The asymptotic behaviour of the series estimation under independence has been studied in Andrews (1991) and Newey (1997). For weakly dependent time series data, Chen and Shen (1998) and Chen, Liao and Sun (2011) offer a rather complete treatment of asymptotic theory and robust inference of the general sieve M estimation, which includes series estimation as a special case. This chapter produces an asymptotic theory that covers general cross-sectional heterogeneity and dependence, including weak and strong dependence. The conditions of the chapter, while designed for spatial setting, readily lend themselves to time series and panel data, expanding the applicability of the results to those settings. They follow the framework of Robinson (2011), however the nature of series estimation necessitated some modifications. This chapter offers alternative conditions in terms of the fourth order cumulants familiar from time series literature, enabling to avoid conditions on joint densities which may be difficult to verify for some processes. Due to a number of similarities of series estimation to OLS in linear regression, asymptotic results derived here easily extend to the linear regression.

The other main contribution of this chapter is establishing a theoretical background for the use of a studentization method that offers an alternative to the existing variance estimation literature in spatial setting. In the spatial context, an extension of HAC estimation familiar from the time series literature, see e.g. Hannan (1957), Newey and West (1987), is possible if additional information such as the socio-economic distance between units which underpin the structure of the spatial dependence is available. Conley (1999) has considered HAC estimation under a stationary random field with measurement error in distance measures, Kelejian and Prucha (2007) for models of Cliff and Ord (1981) and Robinson and Thawornkaiwong (2010) for a more general set-up than Cliff-Ord type models. Bester, Conley, Hansen and Vogelsang (2008) consider the asymptotic theory of the HAC estimation when a fixed, rather than a vanishing, proportion of the sample is used in the variance estimation. However, the small sample performance of HAC estimation is known to be poor even in time series setting and an alternative method that achieves better finite sample performance was suggested by Kiefer, Vogelsang and Bunzel (2000) in the linear regression in time series context. This chapter provides theoretical justification for extending the use of Kiefer *et al.*'s studentization to spatial or spatio-temporal data.

The chapter is structured as follows. In Section 2.2, the setting of the model is outlined. In Section 2.3, the series estimation is introduced and a uniform rate of convergence for the non-parametric component is established. Section 2.4 contains asymptotic normality results. Section 2.5 presents sufficient conditions for the $\sqrt{n}$ rate of convergence of certain semi-parametric estimators, with data-driven studentization. Section 2.6 presents a small Monte Carlo study of finite sample performance. Section 2.7 discusses some empirical examples and Section 2.8 concludes. The Appendix contains the proofs.

2.2 Setting of the model

This chapter discusses inference on the following non-parametric regression model,

$$Y_i = m(X_i) + U_i, \quad i = 1, 2, \cdots, n, \tag{2.2.1}$$

where $Y_i, U_i \in \mathbb{R}$ and $X_i \in \mathcal{X} \subset \mathbb{R}^q$ are random variables and $m : \mathcal{X} \to \mathbb{R}$ is an unknown function of interest. The error term U_i of the model is assumed to follow

$$U_i = \sigma(X_i)e_i, \quad e_i = \sum_{j=1}^{\infty} b_{ij}\varepsilon_j, \quad \sum_{j=1}^{\infty} b_{ij}^2 < \infty \quad i = 1, 2, \cdots, n, \tag{2.2.2}$$

where $\sigma(\cdot)$ is a real function, $\{b_{ij}, i, j \geq 1\}$ are unknown constants and $\{\varepsilon_j, j \geq 1\}$ are independent random variables with zero mean and unit variance. Processes $\{X_i\}$ and $\{\varepsilon_j\}$ are assumed to be independent of each other. The linear process e_i in U_i was also used in Robinson (2011) and Robinson and Thawornkaiwong (2010). Quantities $Y_i, X_i, U_i, e_i, \varepsilon_j, b_{ij}$ are allowed to admit a triangular structure throughout this work, accommodated by the proofs of later theorems. The additional n subscript in e.g. $b_{ijn} = b_{ij}$ is suppressed for the ease of notation. Triangular array structure takes into account the possible need to re-label observations as n increases in panel-data or multi-dimensional lattice data as this work uses a single index i for observations, see Robinson (2011) for discussion. Allowing coefficients $b_{ij} = b_{ijn}$ to vary with n is important in making (2.2.2) to cover the popular SAR model, whereby $e_i = \sum_{j=1}^{n} b_{ijn}\varepsilon_j$ will include summation only up to n.

The structure (2.2.2) is designed to encompass various forms of spatial dependence and heterogeneity in the unobserved errors U_i, that could arise in economic applications. Conditional and unconditional heteroscedasticity of the errors U_i is allowed, while the restrictions later imposed on b_{ij}'s are rather mild, affording an ample scope for possible non-stationarity/heterogeneity across i. For example, b_{ij}'s need not exhibit any form of conformity across i and j, in particular, be affected by $|i-j|$. The specification (2.2.2) also accommodates the idea of "economic distance", in which case b_{ij} will be determined by distances between units. Restrictions on dependence and heterogeneity in X_i are stated and discussed in Section 3.

Errors (2.2.2) obviously cover equally-spaced stationary time series, where b_{ij} is of the form $b_{|i-j|}$. An alternative to (2.2.2) is a mixing framework, which would allow us to relax the condition of independence between $\{X_i\}_{i=1}^n$ and $\{e_i\}_{i=1}^n$. However, a mixing framework necessitates the introduction of some distance measures and the notion of stationarity, which are not always justifiable in economic applications. More importantly, long-range dependence is not covered by a mixing-framework.

Regarding the function $m(x) = E(Y_i|X_i = x)$ in (2.2.1), it denotes the conditional expectation of Y_i at $X_i = x$. For any given function $g(\cdot) : \mathcal{X} \to \mathbb{R}$, let $a(g)$ denote a $d \times 1$ vector-valued functional of $g(\cdot)$, i.e. a mapping from a possible conditional expectation to a real vector. There are many applications where a (known) functional $a(m)$ of the conditional expectation m is of interest. It can be estimated by $a(\hat{m})$, where $\hat{m}(\cdot)$ denotes a series estimator of $m(\cdot)$, constructed as a linear combination of pre-specified approximating functions. Simple examples of $a(g)$ include the value of the function at multiple fixed points, $a(g) = (g(x_1), \cdots, g(x_d))'$, $(x_1, \cdots, x_d) \in \mathcal{X}^d$, which is of interest in non-parametric regression estimation, and the value of partial derivative of the function with respect to the ℓ^{th} argument at fixed points,

$$a(g) = \Big(\frac{\partial g(\mathrm{x})}{\partial \mathrm{x}_\ell}\big|_{x_1}, \cdots, \frac{\partial g(\mathrm{x})}{\partial \mathrm{x}_\ell}\big|_{x_d}\Big)',$$

which is of interest in the case of non-parametric derivative estimation. An example of nonlinear functional $a(\cdot)$ given in Newey (1997) is the consumer surplus. Letting Y_i be the log consumption and $X_i = (\log p_i, \log I_i)'$, a 2×1 vector of log price and log income, the estimated demand function at a fixed point $X_i = x$ is given by $\exp(\hat{m}(x))$, whereas the approximate consumer surplus is equal to the integral of the demand function over a range of prices. For a fixed income $\bar{I}$, an estimator of this functional, when $\underline{p}$ and $\bar{p}$

represent the lower and upper bounds on the price, is

$$a(\hat{m}) = \int_{\underline{p}}^{\bar{p}} \exp\big(\hat{m}(\log t, \log \bar{I})\big) dt.$$

If one was interested in the approximate consumer surplus at multiple fixed values of income, $a(\hat{m})$ would take a vector form. Further example of $a(\cdot)$ arises in the context of the partly linear regression model and will be discussed in detail in Section 5.

Previously, Andrews (1991) showed asymptotic normality for a vector-valued linear functional $a(\hat{m})$, using for $\{X_i\}_{i=1}^n$ and $\{U_i\}_{i=1}^n$ independent and non-identically distributed $(i.n.i.d)$ setting and, in addition, indicating that "the proof can be extended to cover strong mixing regressors without too much difficulty." Newey (1997) has established uniform rate of convergence for $|\hat{m}(x) - m(x)|$ and asymptotic normality result for $a(\hat{m}) - a(m)$ when $\{X_i\}_{i=1}^n$ and $\{U_i\}_{i=1}^n$ are both *i.i.d.* and $a(g)$ is a general (possibly nonlinear) scalar functional. Newey (1997) has also offered a set of conditions on the functional $a(\cdot)$, under which $a(\hat{m})$ converges to $a(m)$ at the parametric rate. Chen and Shen (1998) and Chen, Liao and Sun (2011) consider the problem of sieve extreme estimation for weakly dependent time series setting. In the context of series estimation, Chen and Shen (1998)'s results yield a convergence rate for the non-parametric regression estimate $\hat{m}(\cdot)$ and asymptotic normality for $a(\hat{m})$ in the case of $\sqrt{n}$-rate of convergece. Chen, Liao and Sun (2011) offer an asymptotic normality result that also covers the case of slower-than-$\sqrt{n}$ rate of convergence and provide methods of inference robust to time series weak dependence. They also unveil a rather striking fact that for certain cases of slower-than-$\sqrt{n}$ rate of convergence, the asymptotic variance of the estimate $a(\hat{m})$ coincides with that obtained under independence. An important example is the case of non-parametric regression function evaluated at a finite number of fixed points, for which a similar observation was made by Robinson (1983) for kernel estimation.

2.3 Estimation of m and uniform consistency rate

Estimation of m is based on the use of approximating functions. Denote by $p_s(\cdot), s = 1, 2, \cdots$ a set of approximating functions from $\mathcal{X}$ to $\mathbb{R}$:

$$p^k(\cdot) = (p_1(\cdot), \cdots, p_k(\cdot))'.$$

Next, introduce a deterministic sequence of positive integers $K = K_n$, nondecreasing in n, which denotes the number of approximating functions used in the series estimation where n stands for the sample size. The integer K can be regarded as a bandwidth parameter, analogous to the window length in kernel estimation, and its choice gives rise to a bias/variance trade-off as seen below. Under a suitable choice of approximating functions, larger values of K will reduce the bias while increasing the variance of the estimate $\hat{m}$. A number of assumptions introduced in the following two sections reflect the reliance of the theory on a suitable choice of K.

Let $\hat{\beta} = (\mathbf{p}'\mathbf{p})^{-}\mathbf{p}'Y \in \mathbb{R}^K$, where $\mathbf{p} = \mathbf{p_n} = [p^K(X_1), \cdots, p^K(X_n)]' \in \mathbb{R}^{n\times K}$, $Y = (Y_1, \cdots, Y_n)' \in \mathbb{R}^n$, and A^- denotes the Moore-Penrose inverse for a matrix A.

Definition 1. A series estimator of $m(x)$, at a fixed point $x \in \mathcal{X}$, based on K approximating functions, $p^K(\cdot) = (p_1(\cdot), \cdots, p_K(\cdot))'$, is given by

$$\hat{m}(x) = p^K(x)'\hat{\beta}. \tag{2.3.1}$$

In the remainder of this section, we establish a uniform consistency rate of the estimate $\hat{m}(x)$.

Assumption A1. *The random variables $\{X_i\}_{i=1}^n$, $n = 1, 2, \cdots$, are independent of $\{\varepsilon_j\}_{j=1}^\infty$ and identically distributed with the probability density function $f(x)$, $x \in \mathcal{X}$. The joint density of X_i and X_j, $f_{ij}(x, y)$, $x, y \in \mathcal{X}$, exists for all i and j.*

The assumption of identity of distribution on $\{X_i\}_{i=1}^n$ later facilitates proofs of theorems by offering some algebraic simplification, allowing us to afford more heterogeneity in the unobserved $\{U_i\}_{i=1}^n$, which was deemed more crucial than allowing for non-identical distribution of the observed $\{X_i\}_{i=1}^n$. Heterogeneity/non-stationarity in X_i across i is still allowed as the dependence in X_i may vary across the index i.

Assumption A2. *The random variables $\{U_i\}_{i=1}^n$, $n = 1, 2, \cdots$, follows the linear specification (2.2.2) with some bounded positive function $\sigma(x) : \mathcal{X} \to \mathbb{R}$, and innovations $\{\varepsilon_j\}$ are independent across j, satisfying $E(\varepsilon_j^2) = 1$ and $\max_{j\geq 1} E|\varepsilon_j|^{2+\nu} < \infty$, for some $\nu > 0$.*

For any $k \geq 1$, define a $k \times k$ matrix

$$B_k := E(p^k(X_i)p^k(X_i)'), \quad k = 1, 2, \cdots . \tag{2.3.2}$$

Let $\underline{\lambda}(A)$ and $\bar{\lambda}(A)$ denote the minimal and maximal eigenvalues of a square matrix A. In this work, Euclidean norm is used for vectors: $\|a\|^2 = a'a$. For matrices, we use spectral norm, induced by Euclidean vector norm: $\|A\| = \max_{\|x\|=1} \|Ax\| = \bar{\lambda}^{1/2}(A'A)$. For functions, the uniform norm $|g|_\infty = \sup_{x\in\mathcal{X}} |g(x)|$ is used.

Define a sequence of scalar constants $\xi(k)$ as

$$\xi(k) := \sup_{x\in\mathcal{X}} \|p^k(x)\|, \quad k = 1, 2, \cdots .$$

Quantities similar $\xi(k)$ were also used in Andrews (1991) and Newey (1997). If it is known that m is a bounded function, one may choose bounded and non-vanishing series functions, in which case $\xi(k)$ increases at the rate of $\sqrt{k}$: $\sup_{x\in\mathcal{X}} \|p^k(x)\| = \sup_{x\in\mathcal{X}} \big(\sum_{i=1}^k p_i^2(x)\big)^{1/2} \leq C\sqrt{k}$. It is sometimes possible to obtain the rate of $\xi(K)$ explicitly in terms of K. Newey (1997) provides examples where under suitable conditions, $\xi(K) = K$ when series functions are orthogonal polynomials, and $\xi(K) = K^{1/2}$ when they are B-splines.

Assumption A3. (i) *There exists $c > 0$ such that $\underline{\lambda}(B_k) \geq c, \quad \forall k \geq 1$.*

(ii) *K and $p^K(\cdot)$ are such that $K^2\xi^4(K) = o(n)$.*

Condition $\underline{\lambda}(B_k) \geq c$ of Assumption A3(i) requires B_k to be nonsingular for all values of k and is also assumed in Andrews (1991) and Newey (1997). When this assumption fails, some series functions in $p_s(\cdot), s \geq 1$ may be redundant and need to be eliminated to make it hold. Assumption 3(ii) imposes an upper bound on the rate of increase of $\xi(K)$ as $K \to \infty$. Using the explicit bounds $\xi(K)$ mentioned in the previous paragraph, A3 (ii) boils down to $K = o(n^{1/4})$ for the case of B-splines and $K = o(n^{1/6})$ in the case of orthonormal polynomials under the suitable conditions required for those expressions of $\xi(K)$.

Assumption A4. *The function $m(\cdot) : \mathcal{X} \to \mathbb{R}$ and series functions $p_s(\cdot), s \geq 1$, are such that there exist a sequence of vectors β_K and a number $\alpha > 0$ satisfying, as $K \to \infty$,*

$$|m - p^{K\prime}\beta_K|_\infty = O(K^{-\alpha}).$$

Assumption A4 is a standard condition used in the series estimation literature, and appears in Andrews (1991) and Newey (1997). It requires the uniform approximation error of $m(\cdot)$ by a linear combination of the chosen set of series functions to diminish fast enough. It can be seen as a smoothness condition imposed on $m(\cdot)$, if the functions $p_s(\cdot), s = 1, 2, \cdots$ are ordered so that higher values of s correspond to less smooth functions. In such case, the smoother the function $m(\cdot)$ is, the faster is the rate of decay in the coefficients of the vector β_K in the series expansion $p^{K\prime}\beta_K$ of $m(\cdot)$. Some further insights into Assumption A4 for certain choices of the approximating functions, including polynomials, trigonometric polynomials, splines and orthogonal wavelets, can be found in Chen (2007), pp. 5573. Assumption A4 will control the bias term of our estimate $\hat{m}$, and α is also related to the number of the regressors. Newey (1997) points out that for splines and power series, Assumption A4 is satisfied with $\alpha = s/q$ where s is the number of continuous derivatives of m and q is the dimension of x. Conditions imposing an upper bound on the rate of increase in K, such as A3 (ii), may necessitate a stronger assumption on the smoothness of the unknown m.

Now, we will state an assumption that is required to control the strength of dependence in X_i's across i. Introduce the quantity:

$$\triangle_n := \sum_{i,j=1,i\neq j}^{n} \int_{\mathcal{X}^2} |f_{ij}(x,y) - f(x)f(y)|dxdy. \tag{2.3.3}$$

The rate of growth of $\triangle_n$ is a measure of bi-variate dependence in X_i's and has an upper bound of $2n^2$, a useful property used in the proofs. The quantity $\triangle_n$ is zero in case of independence across i and we may view the condition $\triangle_n = O(n)$ as an analogue to the concept of short-range/weak dependence in time series literature. Quantities of similar nature were used in Robinson (2011) and Robinson and

Thawornkaiwong (2010).

In the case that X_i's are Gaussian random variables, $\triangle_n$ satisfies the following simple bound. Let $\sigma_{ij}^{(X)} := Cov(X_i, X_j)$. Assume for simplicity that $\sigma_{ii}^{(X)} = \sigma_i^{(X)} = 1$. If for some $c_0 < 1$, one has $|\sigma_{ik}^{(X)}| \le c_0, \forall i, k = 1, \cdots, n$; $i \ne k$, $n \ge 1$, then

$$\triangle_n \le C \sum_{i,k=1, i\ne k}^{n} |\sigma_{ik}^{(X)}|, \ n \ge 1,$$

see Proposition 2.1 in Appendix B. Clearly, if $\max_{1\le k\le n} \sum_{i=1}^{n} |\sigma_{ik}^{(X)}| \le Cn$, then $\triangle_n = O(n)$, whereas $\triangle_n = o(n^2)$ holds for a large class of covariances. Thus, in the Gaussian case, $\triangle_n$ can be replaced by the sum $\sum_{i,k=1:i\ne k}^{n} |\sigma_{ik}^{(X)}|$.

Assumption A5. As $n \to \infty$, $n^{-2}K^2\xi^4(K)\triangle_n = o(1)$.

Assumption A5 indicates that stronger dependence in X_i will require the use of smaller K. In the light of Assumption A4, this will necessitate a stronger assumption on the smoothness of the unknown function m. Under weak dependence, i.e. $\triangle_n = O(n)$, A5 reduces to $K^2\xi^4(K) = o(n)$ which is stated in A3(ii). Otherwise, A5 is a stronger condition than A3(ii) imposed on the upper bound of the growth in K and $\xi(K)$.

To state our first theorem, it is necessary to introduce some notation. Define normalised functions $P^k(x) := B_k^{-1/2}p^k(x)$ with B_k as in (2.3.2) such that $E(P^k(X_i)P^k(X_i)')$ $= I_k$. We shall write $P(x) = P^K(x)$ with $K = K_n$, suppressing the superscript K for the rest of the chapter for the ease of notation. Note that $P(\cdot) = [P_{1K}(\cdot), \cdots, P_{KK}(\cdot)]'$, with the double subscripts at $P_{sK}(\cdot)$ arising from the definition $P(\cdot) = B_K^{-1/2}p^K(\cdot)$. Such normalised functions were also used in Newey (1997). Let $\mathbf{P} = \mathbf{P}_n = (P(X_1), \cdots, P(X_n))' \in \mathbb{R}^{n\times K}$. For a given sequence $K = K_n$, define the following $K \times K$ variance-covariance matrix Σ_n of the $K \times 1$ vector sum $\sum_{i=1}^{n} P(X_i)U_i/\sqrt{n}$:

$$\begin{aligned} \Sigma_n &:= E(\mathbf{P}'UU'\mathbf{P}/n) = Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} P(X_i)U_i\right) \qquad (2.3.4) \\ &= \frac{1}{n}\sum_{i,k=1}^{n} E\big(P(X_i)U_iU_kP'(X_k)\big) \\ &= \frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik}E(\sigma(X_i)\sigma(X_k)P(X_i)P'(X_k)), \end{aligned}$$

where

$$\gamma_{ik} := Cov(\sum_{j=1}^{\infty} b_{ij}\varepsilon_j, \sum_{j=1}^{\infty} b_{kj}\varepsilon_j) = \sum_{j=1}^{\infty} b_{ij}b_{kj}.$$

The following theorem obtains the uniform rate of convergence of the estimator $\hat{m}(x)$.

Theorem 2.1 (Uniform Rate of Convergence). *Under Assumptions A1-A5,*

$$\sup_{x\in\mathcal{X}} |\hat{m}(x) - m(x)| = O_p\left(\xi(K)\left[\sqrt{\frac{tr(\Sigma_n)}{n}} + K^{-\alpha}\right]\right), \quad as \quad n \to \infty.$$

This result coincides with the rate obtained by Newey (1997) for *i.i.d.* $\{X_i\}$ and $\{U_i\}$. In the latter case $\Sigma_n = \sigma^2 E(P(X_i)P(X_i)') = \sigma^2 I_K$ leading to $tr(\Sigma_n) = O(K)$. The proof of the above Theorem is given in the Appendix A. The first term in the rate of Theorem 2.1 reflects the contribution of the variance of $\hat{m}$, while the second term arises from the bias component. The uniform rate of consistency highlights the bias/variance trade-off in the selection of K: the use of larger K reduces the bias and increases the variance.

The rates obtained in Theorem 2.1 need to be verified to be $o_p(1)$, to establish uniform consistency of the series estimate $\hat{m}$. The requirement $\xi(K)K^{-\alpha} = o(1)$ of negligible bias suggests that it may be favourable to choose series functions which are bounded. To evaluate the contribution of the variance, suppose for now that the original series functions and thus, the normalized functions $P_{1K}, \cdots, P_{KK}$, are uniformly bounded. Then, $tr(\Sigma_n) = K \cdot \sum_{i,k=1}^{n} \gamma_{ik}/n$, making the variance contribution $\xi(K)\sqrt{K}\big(\sum_{i,k=1}^{n} \gamma_{ik}/n^2\big)^{1/2}$. Under weak dependence on e_i's, $\sum_{i,k=1}^{n} \gamma_{ik} = O(n)$, meaning the rate becomes $\xi(K)\sqrt{K/n} = K/\sqrt{n}$ which is $o(1)$ by Assumption A3 (ii). Under strong dependence of e_i's, the rate is slower and further conditions restricting the increase of K and $\xi(K)$ may be needed to show uniform consistency.

For the *i.i.d.* setting, the uniform rate of convergence obtained by Newey (1997) was improved by de Jong (2002), under the additional assumption of compact $\mathcal{X}$. Under the presence of dependence, it was not possible to obtain improvement similar to that achieved in de Jong (2002), whose proof makes use of Hoeffding's inequality for a sum of *i.i.d.* random variables. It would be of interest for future work to sharpen the bound provided by Theorem 2.1.

2.4 Asymptotic normality

The previous section established the uniform rate of convergence for $\hat{m} - m$, whilst our ultimate interest lies in inference on the functional $a(m)$. Denote $\theta_0 = a(m)$ and $\hat{\theta} = a(\hat{m})$. In this section, we study the asymptotic distribution of $\hat{\theta} - \theta_0$. First, we provide some technical assumptions needed for establishing asymptotic normality. Recall that $a(\cdot)$ is a vector-valued functional operator.

Assumption B1. *One of the following two assumptions holds.*

(i) *$a(g)$ is a linear operator in g.*

(ii) *For some $\epsilon > 0$, there exists a linear operator $D(g)$ and a constant $C = C_\epsilon < \infty$ such that $\|a(g) - a(m) - D(g-m)\| \leq C(|g-m|_\infty)^2$, if $|g-m|_\infty \leq \epsilon$.*

Assumption B2. *For some $C < \infty$, $D(\cdot)$ of Assumption B1 satisfies $\|D(g)\| \leq C|g|_\infty$.*

Assumptions B1 and B2 are the same as in Newey (1997). Assumption B2 requires the linear functional $D(\cdot)$ to be continuous, which follows from the fact that $D(\cdot)$ is the Frechet-differential of $a(\cdot)$ at m. A functional $a(\cdot)$ is said to be Frechet-differentiable at m if there exists a bounded linear operator $D(\cdot)$ satisfying the following property: $\forall \delta > 0, \quad \exists \epsilon > 0$ such that $\|a(g) - a(m) - D(g-m)\| \leq \delta |g-m|_\infty$ if $|g-m|_\infty \leq \epsilon$. Assumption B1(ii) imposes a stronger smoothness condition on $a(\cdot)$ at m than Frechet differentiability. It is not restrictive, see e.g. its verification for some $a(\cdot)$ in Newey (1997, pp. 153). When $a(\cdot)$ is a linear operator, its Frechet-derivative is itself, $D(g) = a(g)$.

Define a $K \times d$ matrix A, with $D(\cdot)$ as in Assumption B1 and the $K \times 1$ vector of normalised functions $P(\cdot)$ as defined above, setting

$$A = (D(P_{1K}), \cdots, D(P_{KK}))' \in \mathbb{R}^{K \times d}.$$

Consider a linear operator $a(m) = (m(x_1), \cdots, m(x_d))'$, for some $(x_1, \cdots, x_d) \in \mathcal{X}^d$. The linearity of $a(m)$ yields $a(P_{sK}) = D(P_{sK}) = \big(P_{sK}(x_1), \cdots, P_{sK}(x_d)\big)'$, $s = 1, \cdots, K$.

Denote by $\bar{V}_n$ the $d \times d$ conditional variance-covariance matrix of the sum $\sum_{i=1}^{n} A'P(X_i)U_i/\sqrt{n}$,

$$\bar{V}_n := Var\left(\sum_{i=1}^{n} A'P(X_i)U_i/\sqrt{n}|X_1, \cdots, X_n\right) = \frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik}\sigma(X_i)\sigma(X_k)A'P(X_i)P'(X_k)A.$$

To gain an insight into the the matrix $\bar{V}_n$ and its role in the statement of the asymptotic distribution, note that one may alternatively write

$$\bar{V}_n = A^{*\prime}B_K^{-1}\left[\frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik}\sigma(X_i)\sigma(X_k)p^K(X_i)p^{K'}(X_k)\right]B_K^{-1}A^*,$$

where $A^* := (D(p_1), \cdots, D(p_K))' = B_K^{1/2}A \in \mathbb{R}^{K \times d}$, the matrix of Frechet-derivatives of the original series functions. One sees that the matrix $\bar{V}_n$ takes the form of the conditional variance-covariance matrix of a nonlinear function of least squares estimates, where the matrix A^* is the Jacobian term and

$$B_K^{-1}\left[\frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik}\sigma(X_i)\sigma(X_k)p^K(X_i)p^{K'}(X_k)\right]B_K^{-1}$$

is the conditional variance-covariance matrix of LS estimates for a possibly misspecified model. Assumption B3 below specifies the conditions under which $\bar{V}_n$ is the correct normalising matrix to be used in the statement of the asymptotic result of Theorem

2.2 below.

Two alternative representations of $\bar{V}_n$ in terms of $P(\cdot)$ or $p^K(\cdot)$ were given above. In the statement of assumptions and theorems, quantities will be written in terms of the vector of normalized functions $P(\cdot)$ to facilitate discussion of the quantity $\bar{V}_n$ in a more tractable manner. We shall need the following assumptions.

Assumption B3. *As $n \to \infty$,*

$$\begin{aligned}&\text{(i)}\quad \xi^2(K)tr(\Sigma_n) = o(n^{1/2}).\\ &\text{(ii)}\quad K^3\xi^6(K)tr(\Sigma_n)\left(\frac{1}{n} + \frac{\triangle_n}{n^2}\right) = o(1).\\ &\text{(iii)}\quad n\xi^2(K)K^{-2\alpha+1} = o(1).\end{aligned}$$

Assumption B3 combines various conditions on the rate of increase of K, $\xi(K)$, $tr(\Sigma_n)$ and $\triangle_n$ as $n \to \infty$. The rate of increase of $tr(\Sigma_n)$ depends on that of K and the strength of dependence in U_i and X_i. Uniform consistency of Theorem 2.1 required smoothness condition $\xi(K)K^{-\alpha} = o(1)$ on the unknown m, while deriving asymptotic normality in Theorem 2.2 needs a stronger smoothness condition of B3 (iii). Revisiting the case of bounded functions $P_{sK}(\cdot)$'s and weakly dependent e_i's leading to $tr(\Sigma_n) = O(K)$, note that B3 (i) is implied by A3 (ii), while B3 (ii) becomes $K^4\xi^6(K) = o(n)$ which implies A3 (ii).

Assumption B4. *K and functions $p^K(\cdot)$ are such that, as $n \to \infty$,*

$$\frac{\xi^2(K)}{\sqrt{n}} \max_{1 \le j \le n} \left\{ \sum_{i=1}^{n} |b_{ij}| \right\} = o(1).$$

Assumption B4 requires the influence of ε_j of any particular j on $U_i, i = 1, 2, \cdots$ to die off, more quickly if $\xi(K)$ grows faster.

Assumption B5. As $n \to \infty$, $\|\bar{V}_n^{-1}\| = O_p(1)$.

Assumption B5 trivially holds in the case when the random matrix $\bar{V}_n$ converges to a finite nonsingular matrix, considered in the next section of $\sqrt{n}$ rate of convergence. Validation of such convergence requires stronger restrictions both on the functional $a(\cdot)$ and the strength of dependence in X_i's and U_i's. Theorem 2.3 allows $\bar{V}_n$ to diverge with n as long as approximation $\|\hat{V}_n - V_n\| = o(1)$ holds for some sequence of deterministic nonsingular matrices V_n. Such approximation still requires certain, although weaker, restrictions to be placed on the strength of dependence in X_i's and U_i's. We present Theorem 2.2 separately from Theorem 2.3, to separate assumptions yielding asymptotic normality from those required for $\|\hat{V}_n - V_n\| = o_p(1)$. Assumption B5 certainly assumes the derivative matrix A to have rank d for all $K \ge d$. Throughout this work, denote by $A^{1/2}$ the unique positive definite square root of a positive definite matrix A.

Theorem 2.2 (Asymptotic Normality). *Under assumptions A1-A5 and B1-B5,*

$$\sqrt{n}\bar{V}_n^{-1/2}(\hat{\theta}-\theta_0) \to_d N(0, I_d), \quad \textit{as} \quad n \to \infty. \tag{2.4.1}$$

The proof of Theorem 2.2 is given in the Appendix A.

2.4.1 Properties of $\bar{V}_n$

The conditional covariance matrix $\bar{V}_n$ is a random quantity. In this section we study conditions, under which $\|\bar{V}_n - V_n\|$ converges to zero, where

$$V_n := E(\bar{V}_n) = Var\left(\sum_{i=1}^{n} A'P(X_i)U_i/\sqrt{n}\right) = \frac{1}{n}\sum_{i,k=1}^{n}\gamma_{ik}E\big[\sigma(X_i)\sigma(X_k)A'P(X_i)P'(X_k)A\big].$$

This will allow us to present the asymptotic distribution result (2.4.1) for $(\hat{\theta}-\theta_0)$ with normalisation V_n. In Theorem 2.3 below, the i^{th} element of the $d \times 1$ estimator $\hat{\theta}$ is shown to be $\sqrt{n}(V_n^{-1/2})_{ii}$-consistent, where $(V_n^{-1/2})_{ii}$ denotes the i^{th} diagonal element of $V_n^{-1/2}$.

To gain an intuition of implications of this rate, let's focus on the case of scalar $a(\cdot)$ in this paragraph. We rule out the possibility of shrinking V_n which corresponds to presence of negative dependence in X_i's or U_i's, as this is rather unlikely for real data. The above expression of V_n indicates that $V_n = O(1)$ would correspond to the case of short range dependence in the combined quantity $A'P(X_i)U_i$ if K were fixed. This may still allow for possibility of long range dependence in $A'P(X_i)$ or U_i to a certain degree. With increasing K, V_n may be increasing even under short-range dependence of $A'P(X_i)U_i$. The main contribution of this chapter is developing inference procedures when V_n is unknown and deriving asymptotic distribution results under additional generality in the strength of dependence in both $\{X_i\}$ and $\{U_i\}$.

The following two conditions state restrictions on the strength of dependence in X_i's and U_i's across i. Again, an upper bound is imposed on the rate of increase in the measure of bivariate dependence in X_i, $\triangle_n$.

Assumption B6. As $n \to \infty$,

$$\frac{\xi^8(K)(n+\triangle_n)}{n^2}\big(\max_{1\le j\le n}\sum_{i=1}^{n}\left|\gamma_{ij}\right|\big)^2 = o(1).$$

Assumption B6 indicates how the dependence in the data restricts the choice of the bandwidth parameter K and series functions. The stronger the dependence is, the slower the rate of increase in K and $\xi(K)$ is required to be, leading to further repercussions on the smoothness in Assumption B3 (iii), where a larger value of α would be needed to compensate for slower rate of growth in K.

Next we state an assumption on the strength of dependence in $\{X_i\}$ across i in terms of their 4th order joint cumulant. The following definition is required to do this.

Definition 2. Let Z_1, Z_2, Z_3, Z_4 be zero-mean random variables with finite fourth moments. Then, the joint cumulant of these four random variables is defined as

$$\begin{aligned}\kappa(Z_1, Z_2, Z_3, Z_4) &:= E(Z_1Z_2Z_3Z_4) - E(Z_1Z_2)E(Z_3Z_4) \\ &\qquad -E(Z_1Z_3)E(Z_2Z_4) - E(Z_1Z_4)E(Z_2Z_3) \\ &= Cov(Z_1Z_2, Z_3Z_4) - Cov(Z_1, Z_3)Cov(Z_2, Z_4) - Cov(Z_1, Z_4)Cov(Z_2, Z_3).\end{aligned}$$

Recalling $A = (A_1, \cdots, A_K)' \in \mathbb{R}^{K\times d}$, introduce the following notations:

$$\begin{aligned} h_i^{(\ell)} &:= \sigma(X_i)A_\ell'P(X_i), \qquad (2.4.2)\\ \bar{h}_i^{(\ell)} &:= \sigma(X_i)A_\ell'P(X_i) - E\left(\sigma(X_i)A_\ell'P(X_i)\right), \quad 1 \le i \le n, \quad 1 \le \ell \le d.\end{aligned}$$

The latter term is a de-meaned version of the former, introduced here so that we can make use of the definition of joint cumulant for mean-zero random variables.

In the time series literature, see e.g. Brillinger (1968), the weak dependence characterization in terms of cumulants typically implies that the 4^{th} order cumulant satisfies,

$$\left|\sum_{i_1,i_2,i_3,i_4=1}^{n} \kappa(Z_{i_1}, Z_{i_2}, Z_{i_3}, Z_{i_4})\right| = O(n). \qquad (2.4.3)$$

Assumption B7. $E\big[(\bar{h}_i^{(\ell)})^4\big] < \infty$ and $\kappa(\bar{h}_{i_1}^{(\ell)}, \bar{h}_{i_2}^{(p)}, \bar{h}_{i_3}^{(\ell)}, \bar{h}_{i_4}^{(p)})$ are such that

$$\max_{1\le \ell,p\le d} \frac{1}{n^2}\left|\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1i_2}\gamma_{i_3i_4}\kappa(\bar{h}_{i_1}^{(\ell)}, \bar{h}_{i_2}^{(p)}, \bar{h}_{i_3}^{(\ell)}, \bar{h}_{i_4}^{(p)})\right| = o(1).$$

Comparing Assumption B7 to (2.4.3), one observes that Assumption B7 is not restrictive and may allow strong dependence in both X_i and U_i. One can have arbitrarily strong dependence in U_i if $\{\bar{h}_i^{(\ell)}\}$ are weakly dependent, c.f. (2.4.3):

$$LHS \le C \max_{1\le \ell,p\le d} \frac{1}{n^2} \sum_{i_1,i_2,i_3,i_4=1}^{n} |\kappa(\bar{h}_{i_1}^{(\ell)}, \bar{h}_{i_2}^{(p)}, \bar{h}_{i_3}^{(\ell)}, \bar{h}_{i_4}^{(p)})| = o(1),$$

noting $|\gamma_{ik}| \le \sqrt{\gamma_{ii}\gamma_{kk}} \le C < \infty$, $i, k = 1, \cdots, n$, $n \ge 1$.

Assumption B8. As $n \to \infty$, $\|V_n^{-1}\| = O(1)$.

The following theorem establishes asymptotic normality if $\hat{\theta}$.

Theorem 2.3 *Under Assumptions B7-B8,*

$$\|\bar{V}_n^{-1}\| = O_p(1), \quad \text{and,} \qquad (2.4.4)$$

$$\|\bar{V}_n - V_n\| = o_p(1). \qquad (2.4.5)$$

Consequently, under assumptions A1-A5 and B1-B8,

$$\sqrt{n}V_n^{-1/2}(\hat{\theta} - \theta_0) \rightarrow_d N(0, I_d). \tag{2.4.6}$$

Theorem 2.3 covers non-parametric, as well as parametric rate of convergence of $\hat{\theta}$ to θ_0, and the $\sqrt{n}$ rate case will be the focus of the next section. An important example of slower-than-$\sqrt{n}$ rate of convergence is the non-parametric regression estimation at d number of fixed points, $a(m) = (m(x_1), \cdots, m(x_d))'$. Noting linearity of this functional, we have the following expression for A:

$$A = \begin{pmatrix} P_{1K}(x_1) & P_{1K}(x_2) & \cdots & P_{1K}(x_d) \\ P_{2K}(x_1) & P_{2K}(x_2) & \cdots & P_{2K}(x_d) \\ \vdots & \vdots & \ddots & \cdots \\ P_{KK}(x_1) & P_{KK}(x_2) & \cdots & P_{KK}(x_d) \end{pmatrix}.$$

The (ℓ, p)th element of V_n is therefore

$$(V_n)_{\ell p} := \sum_{j,m=1}^{K} P_{jK}(x_\ell)P_{mK}(x_p)\Big\{\frac{1}{n}\sum_{i,k=1}^{n}\gamma_{ik}E\big[\sigma(X_i)\sigma(X_k)P_{jK}(X_i)P_{mK}(X_k)\big]\Big\}.$$

In order to use Theorem 2.3 to carry out inference on $a(\hat{m}) = (\hat{m}(x_1), \cdots, \hat{m}(x_d))'$, we need to estimate the term in the curly bracket, which reflects dependence in X_i's and U_i's across i. Such estimation typically requires additional information like distance measure between units in spatial setting, as discussed in Section 2.1. In contrast, kernel non-parametric regression estimation literature found that the relevant asymptotic covariance matrix coincides with that under independence when X_i's and U_i's are weakly dependent across i, see e.g. Robinson (1983, 2011). This justifies the use of the covariance matrix under independence, that is easily estimated, for inference on kernel non-parametric regression estimates under weak dependence, at least for sample with a large n. It is notable that similar result has been recently obtained for series estimation by Chen, Liao and Sun (2011) in the context of weakly dependent time series data. They found that under certain conditions on the functional $a(\cdot)$, that include the case of $a(m) = (m(x_1), \cdots, m(x_d))'$ and preclude the $\sqrt{n}$ rate of convergence of $a(\hat{m})$, V_n reduces asymptotically to the same matrix as under independence, which in our setting is equal to $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\gamma_{ii}E\big[\sigma(X_i)^2A'P(X_i)P'(X_i)A\big]$. In future work, it is of great interest to extend Chen, Liao and Sun (2011)'s result to the spatial setting considered here, which may offer a method of inference for some cases of slower-than-$\sqrt{n}$ rate of convergence under weak dependence. Alternatively, devising a consistent estimation of V_n, that may even offer method of inference under strong dependence, is a challenging yet important task for future research.

2.5 $\sqrt{n}$ rate inference

Theorem 2.3 provides sufficient conditions for convergence $\sqrt{n}V_n^{-1/2}(a(\hat{m})-a(m)) \to_d N(0, I_d)$, where V_n is a $d \times d$ matrix that may grow with n. In this section, we establish sufficient conditions under which V_n converges to a finite limit V, as $n \to \infty$, which in turn implies the parametric $\sqrt{n}$ rate convergence of $\hat{\theta}$ to θ_0. Attainment of the parametric rate of convergence by some semi-parametric estimates have received wide interest in econometric literature, starting from Robinson (1988) and Powell, Stock and Stocker (1989). This type of results are available for the two well-known semi-parametric models: the single index model and partly linear model. While in the kernel estimation each semi-parametric model needs to be considered separately, Newey (1997) has shown that series estimation allows introducing a general semi-parametric estimate encompassing both afore-mentioned popular models, enabling attainment of a unified theory of $\sqrt{n}$-rate of convergence. Chen and Shen (1998) obtained similar results for weakly dependent time series case. It is of interest to extend these results to the setting of cross-sectional dependence, since semi-parametric estimates, such as in the partly linear regression model, are widely used in empirical works, generating a need for a method of inference robust against general spatial dependence and heterogeneity. This section provides a data-driven studentization method that overcomes certain limitations of the existing alternatives.

2.5.1 Partly linear regression model

Before starting the formal statement of theory, we discuss the partly linear regression model in some detail, as the semi-parametric estimate of this model satisfies the conditions of this section and will be used in the Monte Carlo study and empirical examples. This model is a popular alternative to the fully non-parametric regression model and imposes a restriction on the non-parametric function $m(\cdot)$ that a d-dimensional subset of the regressors enter $m(\cdot)$ linearly. For notational convenience, denote this subset by Z_i and the remaining regressors by X_i. Then the model can written as

$$Y_i = Z_i'\delta_0 + h_0(X_i) + U_i, \tag{2.5.1}$$

where $h_0(\cdot)$ is a function of unknown non-parametric form. The model is particularly suitable when Z_i are categorical variables, and is often used when the number of regressors is large since the fully non-parametric specification suffers from the curse of dimensionality. This model has received much attention in kernel estimation, see e.g. Robinson (1988) and Fan and Li (1999), where it has been noted that the parameter δ_0 can be estimated at the $\sqrt{n}$ rate despite the first stage non-parametric estimate having a slower-than-$\sqrt{n}$ rate of convergence.

Series estimation of (2.5.1) had been considered in Chamberlain (1986), where the choice of the series functions takes into account the partly linear regression form. The first d number of series functions are set to be Z_i, while the remaining $K-d$ number of series functions include only X_i in their arguments. The series estimate of δ_0 is then the first d elements of $\hat{\beta}$, and $\hat{h}(x) = \hat{m}(z,x) - z'\hat{\delta}$. At first glance, the form of series estimation of δ_0 may seem very different from kernel estimate, where first-stage non-parametric regression estimates of Y_i's and Z_i's in terms of X_i's are required. Contrary to what meets the eye, they are in fact very similar, as explained below.

Kernel and series estimates of δ_0 are both based on the following relation: subtracting $E(Y_i|X_i) = E(Z_i|X_i)'\delta_0 + h_0(X_i)$ from (2.5.1) yields

$$Y_i - E(Y_i|X_i) = [Z_i - E(Z_i|X_i)]'\delta_0 + U_i,$$

suggesting that δ_0 could be estimated by running a regression of $Y_i - E(Y_i|X_i)$ on $Z_i - E(Z_i|X_i)$. In Robinson (1988)'s kernel estimate denoted, $\tilde{\delta}$, the unknown quantities $E(Y_i|X_i)$ and $E(Z_i|X_i)$ are replaced by suitable kernel estimates,

$$\tilde{\delta} = [(Z - \tilde{E}(Z|X))'(Z - \tilde{E}(Z|X))]^{-1}(Z - \tilde{E}(Z|X))'(y - \tilde{E}(y|Z)),$$

with $Z = (Z_1, \cdots, Z_n)'$, $X = (X_1, \cdots, X_n)'$, and where $\tilde{E}(Z|X)$ and $\tilde{E}(y|Z)$ denote the first stage kernel estimates of the $n \times d$ matrix of conditional expectations $E(Z|X)$ and the $n \times 1$ vector $E(y|X)$.

In the series estimation, the same operation is being implemented by the property of $\hat{\beta} = (\mathbf{p}'\mathbf{p})^{-}\mathbf{p}'Y \in \mathbb{R}^K$, albeit implicitly. To see this, one may write down the following partitioned regression formula familiar from linear regression. Recall that $\hat{\delta}$ is the first d elements of $\hat{\beta}$ such that $\hat{\beta} = (\hat{\delta}', \hat{\lambda}')' = (\mathbf{p}'\mathbf{p})^{-}\mathbf{p}'Y \in \mathbb{R}^K$, with $p^K(Z_i, X_i) = (Z_i', q(X_i)')'$, where $q(\cdot)$ is the vector of $K-d$ series functions in terms of X_i. Define the $n \times n$ residual maker matrix $M := I - \mathcal{P}(\mathcal{P}'\mathcal{P})^{-}\mathcal{P}'$, using $n \times (K-d)$ matrix $\mathcal{P} = (q(X_1), \cdots, q(X_n))'$. Then, partitioned regression formula yields,

$$\hat{\delta} = (Z'MZ)^{-}Z'My.$$

The projections $\mathcal{P}(\mathcal{P}'\mathcal{P})^{-1}\mathcal{P}'Z$ and $\mathcal{P}(\mathcal{P}'\mathcal{P})^{-1}\mathcal{P}'y$ are series estimates of $E(Z|X)$ and $E(y|X)$. Therefore, the series estimate $\hat{\delta}$ of δ_0 effectively takes the same form as the kernel estimate $\tilde{\delta}$ of Robinson (1988), with series estimates of $E(Z|X)$ and $E(y|X)$ replacing corresponding kernel estimates.

Next, we clarify the functional $a(\cdot)$ used to represent the quantity of interest δ_0. There is more than one functional $a(\cdot)$ that yields $a(m) = \delta_0$. Andrews (1991) notes one could write $a(m) = \partial m(x,z)/\partial z = \delta_0$ for any values of x, z. In this work, we use the following functional as in Newey (1997), since this facilitates verification of conditions for $\sqrt{n}$-consistency. Denote $\mathtt{Z}^* = \mathtt{Z} - E(\mathtt{Z}|\mathtt{X})$, where $\mathtt{Z}$ and $\mathtt{X}$ are random variables independent of the data used to construct $\hat{\delta}$. Suppose $E(\mathtt{Z}^*\mathtt{Z}^{*\prime})$ is a non-singular matrix, which is an identification condition for δ_0, and consider the following functional of m:

$$\begin{aligned} a(m) &:= E\left\{[E(\mathtt{Z}^*\mathtt{Z}^{*\prime})]^{-1}\mathtt{Z}^* m(\mathtt{X},\mathtt{Z})\right\} \\ &= [E(\mathtt{Z}^*\mathtt{Z}^{*\prime})]^{-1}\left\{E(\mathtt{Z}^*\mathtt{Z}')\delta_0 + E[\mathtt{Z}^* h_0(\mathtt{X})]\right\} = \delta_0. \end{aligned} \tag{2.5.2}$$

The last equality follows from

$$\begin{aligned} E(\mathtt{Z}^*\mathtt{Z}^{*\prime}) &= E(\mathtt{Z}\mathtt{Z}') - E[E(\mathtt{Z}|\mathtt{X})\mathtt{Z}'] - E[\mathtt{Z}E(\mathtt{Z}'|\mathtt{X})] + E[E(\mathtt{Z}|\mathtt{X})E(\mathtt{Z}'|\mathtt{X})] \\ &= E(\mathtt{Z}\mathtt{Z}') - E[E(\mathtt{Z}|\mathtt{X})\mathtt{Z}'] = E(\mathtt{Z}^*\mathtt{Z}'), \end{aligned}$$

since $E[\mathtt{Z}E(\mathtt{Z}'|\mathtt{X})] = E[E(\mathtt{Z}|\mathtt{X})E(\mathtt{Z}'|\mathtt{X})]$ by the law of iterative expectation, and

$$E[\mathtt{Z}^* h_0(\mathtt{X})] = E[\mathtt{Z}h_0(\mathtt{X})] - E[E(\mathtt{Z}|\mathtt{X})h_0(\mathtt{X})] = 0.$$

To see how this functional can be used to characterise the series estimate of δ_0, recall $\hat{\beta} = (\hat{\delta}', \hat{\lambda}')'$ and $p^K(x,z) = (z', q(x)')'$. Then, $\hat{m}(x,z) = z'\hat{\delta} + q(x)'\hat{\lambda}$. Hence, conditioning on the data, and consequently on $\hat{\beta}$, we have,

$$\begin{aligned} a(\hat{m}) &= E\left\{[E(\mathtt{Z}^*\mathtt{Z}^{*\prime})]^{-1}\mathtt{Z}^*\hat{m}(\mathtt{X},\mathtt{Z})\right\} \\ &= [E(\mathtt{Z}^*\mathtt{Z}^{*\prime})]^{-1}\left[E(\mathtt{Z}^*\mathtt{Z}')\hat{\delta} + E[\mathtt{Z}^* q(\mathtt{X})']\hat{\lambda}\right] = \hat{\delta}. \end{aligned}$$

2.5.2 $\sqrt{n}$ rate of convergence

Returning to the discussion of the $\sqrt{n}$ rate of convergence, the following assumption states the key condition and is from Newey (1997).

Assumption C1. *There exists a $d\times 1$ vector-valued function $w(x) = (w_1(x), \cdots, w_d(x))'$ with the following properties.*

(i) $E[w(X_i)w'(X_i)]$ *is finite and nonsingular,*

(ii) $D(m) = E[w(X_i)m(X_i)]$, $D(P_{sK}) = E[w(X_i)P_{sK}(X_i)], 1 \le s \le K$ *for all* K,

(iii) $E[\|w(X_i) - \delta_K P(X_i)\|^2] \to 0$ *for some sequence of fixed* $d \times K$ *matrices* δ_K.

Discussion of sufficient conditions for Assumption C1 can be found in Newey (1997), pp. 155. The vector-valued function $w(\cdot)$ is the element of the domain of $D(\cdot)$ that is used in the Riesz representation of $D(\cdot)$. Assumption C1 (iii) requires such function $w(\cdot)$ to lie in the linear span of the series functions. Newey (1997) explicitly verifies that Assumption C1 holds for the semi-parametric estimands in the partly linear and single index models and also for the case of average consumer surplus estimation, where the quantity of interest is the approximate consumer surplus integrated over a range of income. The verification for the partly linear regression case is straightforward in the view of (2.5.2). Interested readers are referred to pp. 155 of Newey (1997).

By Assumption C1, $D(P_{sK}) = E[w(X_i)P_{sK}(X_i)]$, $1 \le s \le K$. Thus, one can write $A = E[P(X_i)w'(X_i)]$. Since the $K \times 1$ vector of normalized functions $P(\cdot)$ satisfies

$E[P(X_i)P'(X_i)] = I_K$, $A'P(x)$ can be written as the mean square projection of $w(x)$ on the $K \times 1$ vector $P(\cdot)$ of approximating functions:

$$A'P(x) = A'I_K^{-1}P(x) = E[w(X_i)P'(X_i)]E[P(X_i)P'(X_i)]^{-1}P(x).$$

Denote $d \times 1$ vector $A'P(x) =: v_K(x) = (v_{1K}(x), \cdots, v_{dK}(x))'$, with the subscript K indicating that v_K is a mean-square projection of w onto the linear space spanned by K series functions. Then V_n can be written as

$$V_n = \frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik} E[\sigma(X_i)\sigma(X_k)v_K(X_i)v_K'(X_k)].$$

Next, define $d \times d$ matrix W_n where $v_K(\cdot)$ is replaced by the function $w(\cdot)$:

$$W_n := \frac{1}{n}\sum_{i,k=1}^{n} \gamma_{ik} E[\sigma(X_i)\sigma(X_k)w(X_i)w'(X_k)].$$

The following assumption provides sufficient conditions for $\sqrt{n}$ rate of convergence of $a(\hat{m})$ to $a(m)$.

Assumption C2. (i) $V := \lim_{n\to\infty} W_n$ *exists;* (ii) $\sum_{i,k=1}^{n} |\gamma_{ik}| = O(n)$.

Existence of the limit V is a condition imposed on the collective strength of dependence in U_i and X_i, comparable to Assumption A4 of Robinson and Thawornkaiwong (2010). Assumption C2 (ii) is a weak dependence restriction for e_i's.

Theorem 2.4. ($\sqrt{n}$ rate of convergence). *Under assumptions C1 and C2,*

$$V_n \to V < \infty, \quad \text{as} \quad n \to \infty. \tag{2.5.3}$$

Consequently, under assumptions A1-A5, B1-B7, and C1-C2,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, V), \quad \text{as} \quad n \to \infty.$$

Theorem 2.4 has obtained the $\sqrt{n}$ rate of convergence for certain semi-parametric estimates under weak dependence. The asymptotic variance-covariance matrix V is unknown and needs to be estimated to construct a confidence interval or carry out hypothesis testing for the unknown θ_0. The next subsection considers the issues related to this.

2.5.3 Studentization

In the earlier Section 1, possible problems of using the HAC estimator in the cross-sectional setting have been discussed. Under the conditions for $\sqrt{n}$ rate of convergence of $a(\hat{m})$ to $a(m)$ given in this section, it is possible to construct a new studentization for $a(\hat{m}) - a(m)$ that does not require availability of economic distances. Theorem 2.4

states $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, V)$, where the matrix

$$V = \lim_{n\to\infty} A'\mathbf{P}'E(UU'|X)\mathbf{P}A/n = \lim_{n\to\infty} A^{*\prime}B_K^{-1}\mathbf{p}'E(UU'|X)\mathbf{p}B_K^{-1}A^*/n$$

is unknown. It is not possible to consistently estimate V, unless one resorts to additional information of suitable distance measures, as considered in Conely (1999), Kelejian and Prucha (2007) and Robinson and Thawornkaiwong (2010). Instead, we devise a matrix $\hat{C}_n$, defined in the subsequent discussion, such that the limit of $\sqrt{n}\hat{C}_n^{-1/2}(\hat{\theta}_n - \theta_0)$ is free from unknown parameters. Similar idea was used in a setting of linear OLS estimation in Kiefer, Vogelsang and Bunzel (2000).

Recall some notations: $B_K = E(p^K(X_i)p^K(X_i)')$, $P(x) = B_K^{-1/2}p^K(x)$, $A = (D(P_{1K}), \cdots, D(P_{KK}))' \in \mathbb{R}^{K\times d}$, $A^* = (D(p_1), \cdots, D(p_K))' = B_K^{1/2}A \in \mathbb{R}^{K\times d}$ with $D(\cdot)$ from Assumption B1(i). Denote by $\hat{A}^*$ and $\hat{B}_K$ the estimates of the corresponding true values A^* and B_K.

$$\hat{A}^* := \frac{\partial a(p^{K\prime}\beta)}{\partial\beta}\Big|_{\beta=\hat{\beta}}, \quad \hat{B}_K := \mathbf{p}'\mathbf{p}/n = \sum_{i=1}^{n} p^K(X_i)p^K(X_i)'/n. \qquad (2.5.4)$$

Given $\hat{A}^*$ and $\hat{B}_K$, we can construct the sample analogue of $A'\mathbf{P}'U/\sqrt{n}$ by $\hat{A}^{*\prime}\hat{B}_K^{-1}\mathbf{p}'\hat{U}/\sqrt{n}$, where $\hat{U} = Y - \hat{M}$, with $\hat{M} = \big(\hat{m}(X_1), \cdots, \hat{m}(X_n)\big)$, is the $n \times 1$ vector of corresponding residuals. To introduce $\hat{C}_n$, set

$$\hat{S}^*_{n,m} := \sum_{i=1}^{m} \hat{A}^{*\prime}\hat{B}_K^{-1}p^K(X_i)\hat{U}_i/\sqrt{n}, \quad 1 \le m \le n.$$

Now, define

$$\hat{C}_n := \frac{1}{n}\sum_{m=1}^{n} \hat{S}^*_{n,m}\hat{S}^{*\prime}_{n,m}, \quad \text{and} \quad \Psi_d := \int_0^1 [W_d(r) - rW_d(1)][W_d(r) - rW_d(1)]'dr,$$

where $W_d(\cdot)$ denotes a d-dimensional vector of independent Brownian motions and Ψ_d is the integral of the outer product of d-dimensional multivariate Brownian bridge. Recall that $EW_d(r)W_d(u)' = rI, \quad 0 \le r \le u \le 1$.

Assumption C3. (i) $\sum_{i=1}^{[rn]} \sum_{k=[rn]+1}^{n} |\gamma_{ik}| = o(n)$ uniformly in $r \in [0, 1]$;

(ii) $\max_{1\le i\le n} \sum_{k=1}^{n} |\gamma_{ik}| = O(1)$.

Previously, Assumption C2 (ii) of Theorem 2.4 required e_i's to be weakly dependent. C3 (ii) further rules out the presence of any "dominant" unit whose error covariances with new units added to the sample are persistently significant. Assumption C3 (i) requires some falling-off of dependence as $|i-k|$ increases, which inevitably necessitates the ordering of the data to carry at least some information of the structure of dependence, albeit with a significant relaxation from the time series case where dependence is a function of $|i-k|$. Both C3 (i) and (ii) are natural implications of weak dependence in the time series context where the dependence is a fast-decreasing function of the distance in time. The current setting differs from the time series in two ways; firstly it allows $\gamma_{ik}=\gamma_{ikn}$ to admit a triangular array structure, and secondly it relaxes the link between γ_{ik} and $|i-k|$. For example, Assumption C3(i) is satisfied if there exists a positive function, $\eta(\cdot)$, such that $|\gamma_{ik}|\le\eta(i-k), i,k=1,2,\cdots$ and $\sum_{j=-\infty}^{\infty}\eta(j)<\infty$. See Proposition 2.2, Appendix B. If γ_{ik} takes on a triangular array structure, as allowed in the pure SAR model, then Assumption C3 (i) is potentially more restrictive. In this setting, Assumption C3 (ii) allows a unit i to interact with infinitely many others as the sample increases, as long as the bilateral interaction $\gamma_{ikn}, k=1,2,\cdots$, falls suitably fast in n, whereas C3 (i) requires a faster uniform-in-n rate of reduction in γ_{ikn} as $|i-k|$ increases.

Therefore, in order to apply the studentization method to cross-sectional data, the ordering of data needs to carry some meaning. This rules out the case where the data was collected at random from the population without any record of how units may be related. Another issue is that in spatial settings it may not always be straightforward to order units along a single line, as their dependence structure may be based on e.g. a plane. Nonetheless, there are many economic applications where the data can be ordered to adhere to the requirements of Assumption C3. For example, with firm data, one may expect that firms using similar inputs or producing similar outputs would exhibit high correlation in disturbances, the knowledge of which can help put the data in order. This ordering may not be perfect because some perturbation may result from the imperfection of practitioner's knowledge of the underlying dependence and also because of the challenge of ordering along a single index, as when trying to order locations on a plane into a line. These considerations are dealt with in a simulation study later.

Assumption C4. (i) $\triangle_n=O(n)$; (ii) $tr(\Sigma_n)=O(K)$; (iii) $\bar{\lambda}(B_K)=O(1)$; (iv) $\sqrt{n}\xi^3(K)K^{-\alpha}=o(1)$.

Assumption C4 (i) can be seen as weak dependence condition on X_i's, whereas Assumption C4 (iii) is a restriction on the choice of the approximating functions, requiring their second moments to be bounded. Assumption C4 (ii) is a condition on the strength of dependence across i in the combined quantity $P(X_i)U_i$. Assumption C (iv) strengthens the smoothness condition of Assumption B3 (iii).

Assumption C5. *$E(\varepsilon_j^4)=\kappa<\infty$ for all $j=1,2,\cdots$.*

Recall the functional derivative $D(\cdot)$ from Assumptions B1 and B2. It is Frechet differential of the functional $a(\cdot)$, evaluated at m. Now, let $D(\cdot;g)$ denote the functional derivative of $a(\cdot)$ evaluated at g. Let $D(\cdot;g)=\big(D_1(\cdot;g),\cdots,D_d(\cdot;g)\big)'$.

Assumption C6. *For some $0<C,\epsilon<\infty$ and all $\tilde{g},\bar{g}$ such that $|\tilde{g}-m|_\infty\le\epsilon$ and $|\bar{g}-m|_\infty\le\epsilon$, $\|D_i(g;\tilde{g})-D_i(g;\bar{g})\|\le C|g|_\infty|\tilde{g}-\bar{g}|_\infty$, $i=1,\cdots,d$.*

Assumption C5 is the same as in Newey (1997) and requires the functional derivatives $D_i(\cdot; g)$ to exhibit continuity over g, the point at which the derivative is taken.

The following theorem shows that the asymptotic distribution for the estimation error $\hat{\theta} - \theta_0$, when studentized by the matrix $\hat{C}_n$, is free from the unknown variance matrix V and only depends on d, and is non-Gaussian.

Theorem 2.5. Under the assumptions of Theorem 2.4 and Assumptions C1-C6,

$$\hat{C}_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \Psi_d^{-1/2} W_d(1).$$

Now, suppose we are interested in testing the hypothesis $H_0 : a(m) = r$ against the alternative $H_1 : a(m) \neq r$ for a $d \times 1$ fixed vector r. Then the test statistic can be constructed as $t_n^* := n(\hat{\theta} - r)'\hat{C}_n^{-1}(\hat{\theta} - r)$. Since $t_n^* = \|\sqrt{n}(\hat{\theta} - r)'\hat{C}_n^{-1/2}\|^2$, Theorem 2.5 implies the following result.

Theorem 2.6. Under Assumptions of Theorem 2.5,

$$\begin{aligned} &t_n^* \Rightarrow W_d(1)'\Psi_d^{-1}W_d(1), \quad \text{under} \quad H_0, \\ &t_n^* \Rightarrow \infty, \quad \text{under} \quad H_1. \end{aligned}$$

The critical values c_α, satisfying $Pr(t_n \leq c_\alpha) \to 1-\alpha$, required to carry out hypothesis tests can be obtained from Table 2 of Kiefer *et al.* (2000) for $d = 1, \cdots, 30$. In particular, for $d = 1$, $c_{5\%} = 46.39$, and $c_{10\%} = 28.88$. Correspondingly, the 97.5^{th} and 95^{th} percentiles for $\Psi_1^{-1/2}W_1(1)$ in Theorem 2.5 are $\sqrt{46.39}$ and $\sqrt{28.88}$.

2.6 Monte Carlo Study of Finite-Sample Performance

In this section, we focus on the partly linear model of (2.5.1) where regressors X_i and Z_i are both one-dimensonal:

$$Y_i = \delta_0 Z_i + h(X_i) + U_i.$$

It was noted in Section 2.5 that the functional $a(m) = \delta_0$ satisfies the conditions of Theorem 2.4. Therefore the studentization devised in Theorem 2.5 and 2.6 applies. We set the true model at $\delta_0 = 0.3$ and $h(x) = \log(1 + x^2)$.

There are two issues we would like to address in this section, related to the difficulty of ordering data in line with the requirements of Assumption C3. Firstly, there may be noise in our information about the ordering. For example, in a spatial setting, one may correctly know which characteristic of individual units underpin the structure of dependence, but this characteristic may be observed with error. Secondly, it may not be straightforward to order the data with a single index as the underlying dependence structure is more complex. For instance, one may observe units residing on a plane, and there is no single obvious rule to order them with only a single index. In this simulation, we consider two set-ups that cover the two issues separately.

In the first set of simulations, we generate random locations for individual units along a line, which determines the underlying dependence structure. We then compare performance of studentization under the correct ordering of data to that under perturbed ordering, that arises when the original locations are observed with noise, then used to order the data. To be specific, the locations of the observations, denoted $s = (s_1, \cdots, s_n)'$, were generated by a random draw from the uniform distribution over $[0, n]$. Keeping these locations fixed across replications, U_i and Z_i were generated independently as scalar normal random variables with mean zero and covariances $Cov(U_i, U_j) = Cov(Z_i, Z_j) = \rho^{|s_i - s_j|}$. To construct X_i, we generate another scalar normal random variable V_i in the same way as U_i and Z_i and let $X_i = 1 + V_i + 0.5Z_i$. The dependent variable is then formed as $Y_i = \log(1 + X_i^2) + 0.3Z_i + U_i$.

For the studentization part of simulations, we add noise to the locations, to generate four sets of "perturbed" locations:

$$s_i' = s_i + \epsilon_i', \quad s_i'' = s_i + \epsilon_i'', \quad s_i''' = s_i + \epsilon_i''', \quad s_i'''' = s_i + \epsilon_i'''',$$

where the perturbations are independently drawn from

$$\begin{aligned} \epsilon' &= (\epsilon_1', \cdots, \epsilon_n')' \sim N(0, 4I_n), \quad \epsilon'' = (\epsilon_1'', \cdots, \epsilon_n'')' \sim N(0, 25I_n), \\ \epsilon''' &= (\epsilon_1''', \cdots, \epsilon_n''')' \sim N(0, 100I_n), \quad \epsilon'''' = (\epsilon_1'''', \cdots, \epsilon_n'''')' \sim N(0, 400I_n). \end{aligned}$$

These perturbations may be seen as the measurement error in observations of the locations. We use studentization with 5 different ordering of the data, according to the five sets of locations s, s', s'', s''', s''''.

We let $n = 100, 400$ and $\rho = 0, 0.2, 0.4, 0.6$, giving 8 combinations. For each combination, three values of $K = 4, 6, 9$ were tried and 1000 iterations carried out. For the series functions of X_i, the first $K - 1$ orthonormal Legendre polynomials were used.

The first objective of this simulation study is to analyse the finite sample performance of the series estimation for both the non-parametric function m and semi-parametric quantity $a(m)$ under differing sample sizes, strength of dependence and choices of K. We report in Table 2.1 the Monte Carlo MSE, bias and variance of the non-parametric regression estimate at a fixed point $(x, z) = (0.5, 0.5)$, and the Monte Carlo integrated MSE, defined as $E[(\hat{m}(X_i) - m(X_i))^2]$ conveying how the non-parametric estimation performs globally. Table 2.1 also contains the Monte Carlo MSE of the estimate $\hat{\delta}$ of δ_0. The Monte Carlo variance and bias of the non-parametric estimate at a fixed point are in line with the prediction that larger values of K reduce the bias while increasing variance. As for the Monte Carlo MSE for $\hat{m}(0.5, 0.5)$, under all four values of ρ, $K = 4$ or $K = 6$ led to the smallest MSE for $n = 100$, while $K = 6$ did so for $n = 400$. For the MISE, $K = 4$ for $n = 100$ always led to the smallest MISE, while $K = 6$ did so for $n = 400$. The Monte Carlo MSE of the semi-parametric estimate $\hat{\delta}$ shows remarkable invariance to the choice of K across all of the 8 settings, which is especially important as the optimal choice of the bandwidth parameter K for semi-parametric estimate is often more difficult than in the case of non-parametric estimation. See Robinson and Thawornkaiwong (2010) for a discussion.

The second objective is to investigate how the studentization of Section 2.5.3 performs in finite samples. Theorem 2.5 implies in this setting,

$$n(\hat{\delta}-\delta_0)'\hat{C}_n^{-1}(\hat{\delta}-\delta_0) \rightarrow_d W_1(1)'\sqrt{\Psi_1}^{-1}W_1(1),$$

$$\frac{\sqrt{n}(\hat{\delta}-\delta_0)}{\sqrt{\hat{C}_n}} \rightarrow_d \frac{W_1(1)}{\sqrt{\Psi_1}}.$$

Kiefer *et al.* (2000, Table 2) give simulated values of the percentiles of $W_1^2(1)/\Psi_1$, from which the corresponding percentiles of the square-rooted quantity $W_1(1)/\sqrt{\Psi_1}$ can be easily derived. The $99.5^{th}, 97.5^{th}$ and 95^{th} percentile of $W_1(1)/\sqrt{\Psi_1}$ are $\sqrt{101.2}$, $\sqrt{46.39}$ and $\sqrt{28.88}$, respectively. Based on this, we construct the asymptotic 95% confidence interval for δ_0:

$$\Pr\left(\delta_0 \in \left[\hat{\delta}-\sqrt{46.39\frac{\hat{C}_n}{n}}, \hat{\delta}+\sqrt{46.39\frac{\hat{C}_n}{n}}\right]\right) \rightarrow 0.95.$$

Table 2.2 reports the Monte Carlo average length of the 95% confidence intervals for studentization based on the correctly ordered data, i.e. ordered according to s. The length of confidence intervals decreases with the sample size, increases with ρ and does not report much variation over the choice of K. The same patterns are observed with results under perturbed ordering.

Table 2.3 reports the empirical coverage probabilities for the 99%, 95% and 90% asymptotic confidence intervals under the five different orderings of data, based on locations s, s', s'', s''', and s''''. When $\rho = 0$, studentizations with all orderings produce a rather precise coverage probabilities for both samples sizes. For $\rho = 0.2, 0.4, 0.6$ and correct ordering based on s, the coverage proabilities suffer slightly in the small sample $n = 100$, while being rather good for $n = 400$, at least for $\rho = 0.2$ and 0.4. As we perturb the ordering, a gradual deterioration in coverage probabilities is reported. Nevertheless, even with the perturbation caused by substantial noises $\epsilon_i''' \sim N(0, 100)$ and $\epsilon_i'''' \sim N(0, 400)$, the reported coverage probabilities are remarkably encouraging.

Table 2.4 reports empirical power of testing $H_0 : \delta_0 = \delta$ against $H_1 : \delta_0 \neq \delta$, for $\delta = 0.3, 0.4, 0.5, 0.7$. Since the true δ_0 used in data generation is 0.3, the columns corresponding to $\delta = 0.3$ report empirical size of the test. Not surprisingly for $\rho = 0$, empirical powers across different orderings are similar, while for $\rho = 0.2, 0.4$ and 0.6, power tends improve with larger perturbations to ordering.

The second set of simulations aims to investigate the implications of ordering spatial data with a single index, while their underlying dependence may be more complex. We generate random locations on a plane then order the data with a single index in an ascending order of the distance from the origin (0,0). To generate the data, we follow the random location setting of Robinson and Thawornkaiwong (2010), where the vector of locations of the observations, denoted $s_1, \cdots, s_n$, were generated by a random draw from the uniform distribution over $[0, 2n^{1/2}] \times [0, 2n^{1/2}]$. Again, keeping these locations fixed across replications, U_i and Z_i were generated independently as scalar normal random variables with mean zero and covariances $Cov(U_i, U_j) = Cov(Z_i, Z_j) = \rho^{\|s_i - s_j\|}$, where $\| \cdot \|$ denotes Euclidean norm. To construct X_i, we generate another scalar normal random variable V_i in the same way as U_i and Z_i and let $X_i = 1 + V_i + 0.5Z_i$. The dependent variable is then formed as $Y_i = \log(1 + X_i^2) + 0.3Z_i + U_i$. Again, we considered two sample sizes, $n = 100, 400$, three values of $K = 4, 6, 9$ and carry out 1000 iterations. For the series functions of X_i, the first $K - 1$ orthonormal Legendre polynomials were used. For the values of ρ's, we considered $\rho = 0, 0.2, 0.4, 0.52$ for $n = 100$ and $\rho = 0, 0.2, 0.35, 0.5$ for $n = 400$. The random location setting implies the degree of dependence is determined not only by the value of ρ, but also by the set of distances based on random locations. The fact that we are considering locations on a plane, rather than along a line, implies that the value of ρ produces differing strength of dependence compared to the stationary time series AR(1) model we are familiar with, making it difficult to get a sense of the degree of dependence in data generating models considered in our simulations. One way of comparing dependence between different settings is to measure it by the value $\sum_{i,j=1}^{n} |Cov(U_i, U_j)|$. The choices of ρ's were such that this summation is of similar magnitude to that in the time series AR(1) setting with $\rho = 0, 0.2, 0.4, 0.6$. For $n = 100$, the values of the above summation in our spatial simulations corresponding to $\rho = 0, 0.2, 0.4, 0.52$ were $100, 152, 255, 384$, respectively, which are comparable to $100, 150, 232, 396$ corresponding to time series AR(1) models with $\rho = 0, 0.2, 0.4, 0.6$. For $n = 400$, the values of the above summation in our simulation corresponding to $\rho = 0, 0.2, 0.4, 0.52$ were $400, 611, 949, 1602$, respectively, which are comparable to $400, 599, 930, 1590$ of time series AR(1) models with $\rho = 0, 0.2, 0.4, 0.6$.

We report in Table 2.5, the Monte Carlo MSE, bias and variance of the non-parametric regression estimate at a fixed point $(x, z) = (0.5, 0.5)$, Monte Carlo integrated MSE, Monte Carlo MSE of the estimate $\hat{\delta}$ of δ_0. Again, patterns of bias and variance of the non-parametric regression estimate with changing K is in line with the theory's predictions and the choice $K = 4$ generated the lowest MSE for all combinations for $n = 100$ and $K = 6$ did so for $n = 400$.

As mentioned before, we ordered data in an ascending order of Euclidean distance from the origin for the purpose of studentization. Table 2.6 reports the Monte Carlo average length of the 95% confidence intervals. As before, the length of confidence intervals decreases with the sample size, increases with ρ and does not show much variation over the choice of K. Table 2.7 reports the empirical coverage probabilities for the 99%, 95% and 90% asymptotic confidence intervals, which are reported to be highly satisfactory despite the difficulty of ordering and dependence. Table 2.8 reports empirical power of testing $H_0 : \delta_0 = \delta$ against $H_1 : \delta_0 \neq \delta$, for $\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1$ with 5% significance level. The asymptotic distribution of the test statistic is symmetric, and as expected, powers reported for $\delta = 0.1$ and 0.2 are similar to those reported for $\delta = 0.5$ and 0.4, respectively.

Table 2.1: Monte Carlo MSE, Variance and Bias

ρ	n	K	$MSE(\hat{g}_x)$	$Var(\hat{g}_x)$	$Bias(\hat{g}_x)$	$MISE(\hat{g})$	$MSE(\hat{\delta})$
0	100	4	0.0353	0.0283	0.0842	0.0595	0.0126
		6	0.035	0.0347	0.017	0.0701	0.0125
		9	0.0463	0.0463	0.0039	0.0989	0.0132
	400	4	0.0162	0.0071	0.0956	0.0265	0.0033
		6	0.0082	0.0079	0.0174	0.0199	0.0033
		9	0.0098	0.0098	-0.0024	0.025	0.0034
0.2	100	4	0.0526	0.0453	0.0855	0.0863	0.0216
		6	0.055	0.0546	0.0201	0.0992	0.022
		9	0.0671	0.067	0.0066	0.1261	0.0229
	400	4	0.0219	0.0121	0.099	0.033	0.005
		6	0.0141	0.0135	0.0254	0.0278	0.0051
		9	0.0151	0.0151	0.0041	0.0334	0.0051
0.4	100	4	0.0693	0.0647	0.0674	0.106	0.0268
		6	0.0757	0.0756	0.005	0.1207	0.0273
		9	0.0915	0.0915	-0.002	0.1493	0.0278
	400	4	0.025	0.0148	0.1014	0.0394	0.0065
		6	0.0175	0.0168	0.0265	0.0347	0.0065
		9	0.0193	0.0192	0.0058	0.0404	0.0065
0.6	100	4	0.0863	0.0809	0.0738	0.1326	0.0341
		6	0.0861	0.0859	0.0112	0.1465	0.0348
		9	0.1028	0.1028	-0.0013	0.1739	0.0358
	400	4	0.034	0.0253	0.0931	0.0517	0.0107
		6	0.0272	0.0267	0.0222	0.0481	0.0107
		9	0.0301	0.0301	-0.0006	0.0542	0.0107

Table 2.2: Monte Carlo average 95 % CI length

n	K	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$
100	4	0.5605	0.6746	0.7447	0.8328
	6	0.5608	0.6701	0.7401	0.8276
	9	0.5608	0.6736	0.7353	0.8224
400	4	0.2955	0.3519	0.4043	0.4889
	6	0.2933	0.3501	0.4039	0.4874
	9	0.2922	0.3489	0.402	0.4869

Table 2.3: Coverage Probabilities

			s			s’			s”’			s”’			s””		
ρ	n	K	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
0	100	4	0.894	0.951	0.987	0.891	0.942	0.987	0.897	0.936	0.991	0.889	0.937	0.989	0.892	0.94	0.987
		6	0.897	0.944	0.99	0.891	0.947	0.989	0.892	0.942	0.99	0.891	0.937	0.99	0.895	0.942	0.985
		9	0.896	0.938	0.989	0.891	0.94	0.989	0.889	0.937	0.984	0.878	0.933	0.985	0.882	0.94	0.984
	400	4	0.908	0.947	0.988	0.91	0.947	0.989	0.909	0.948	0.989	0.906	0.95	0.989	0.907	0.95	0.991
		6	0.897	0.947	0.991	0.897	0.947	0.991	0.891	0.944	0.992	0.897	0.955	0.99	0.902	0.951	0.992
		9	0.89	0.952	0.991	0.889	0.951	0.988	0.891	0.944	0.99	0.891	0.947	0.989	0.897	0.943	0.992
0.2	100	4	0.865	0.939	0.979	0.863	0.927	0.98	0.859	0.922	0.975	0.836	0.915	0.972	0.829	0.898	0.971
		6	0.862	0.926	0.984	0.855	0.919	0.984	0.849	0.916	0.979	0.837	0.913	0.978	0.828	0.899	0.97
		9	0.852	0.917	0.984	0.853	0.908	0.981	0.848	0.906	0.978	0.845	0.902	0.972	0.827	0.892	0.973
	400	4	0.89	0.951	0.989	0.889	0.951	0.988	0.881	0.946	0.986	0.876	0.941	0.983	0.868	0.936	0.986
		6	0.889	0.948	0.988	0.889	0.946	0.987	0.881	0.941	0.988	0.877	0.938	0.985	0.866	0.934	0.983
		9	0.883	0.943	0.986	0.885	0.94	0.986	0.876	0.934	0.986	0.872	0.934	0.984	0.863	0.929	0.983
0.4	100	4	0.863	0.918	0.976	0.858	0.916	0.97	0.837	0.904	0.971	0.825	0.898	0.968	0.807	0.881	0.959
		6	0.866	0.916	0.971	0.858	0.909	0.969	0.846	0.91	0.966	0.827	0.891	0.964	0.797	0.877	0.962
		9	0.864	0.913	0.973	0.86	0.907	0.972	0.841	0.901	0.96	0.84	0.901	0.969	0.796	0.868	0.967
	400	4	0.887	0.941	0.992	0.885	0.938	0.99	0.88	0.933	0.986	0.878	0.931	0.986	0.866	0.916	0.977
		6	0.893	0.935	0.992	0.889	0.933	0.99	0.885	0.93	0.99	0.88	0.929	0.991	0.868	0.912	0.975
		9	0.891	0.939	0.992	0.891	0.936	0.991	0.887	0.931	0.99	0.879	0.928	0.99	0.865	0.91	0.975
0.6	100	4	0.863	0.93	0.974	0.855	0.923	0.974	0.846	0.921	0.963	0.823	0.894	0.961	0.804	0.872	0.951
		6	0.865	0.919	0.978	0.864	0.915	0.976	0.84	0.912	0.967	0.821	0.893	0.956	0.799	0.874	0.948
		9	0.86	0.914	0.979	0.856	0.907	0.977	0.846	0.902	0.971	0.816	0.88	0.958	0.796	0.862	0.945
	400	4	0.866	0.923	0.977	0.861	0.921	0.977	0.858	0.915	0.975	0.846	0.913	0.973	0.835	0.905	0.96
		6	0.872	0.927	0.98	0.869	0.923	0.981	0.864	0.919	0.978	0.861	0.912	0.975	0.837	0.897	0.96
		9	0.869	0.93	0.981	0.867	0.928	0.978	0.864	0.922	0.977	0.856	0.913	0.975	0.84	0.901	0.966

Table 2.4: Empirical power of 95% test, $K = 6$

		s				s’				s”				s”’				s””			
	$\rho\backslash\delta$	**0.3**	0.4	0.5	0.7	**0.3**	0.4	0.5	0.7	**0.3**	0.4	0.5	0.7	**0.3**	0.4	0.5	0.7	**0.3**	0.4	0.5	0.7
$n = 100$	0	0.049	0.113	0.319	0.778	0.058	0.112	0.32	0.781	0.064	0.119	0.314	0.779	0.063	0.121	0.32	0.781	0.06	0.12	0.31	0.779
	0.2	0.061	0.131	0.275	0.656	0.073	0.141	0.28	0.667	0.078	0.142	0.287	0.679	0.085	0.139	0.314	0.695	0.102	0.163	0.313	0.726
	0.4	0.082	0.127	0.245	0.597	0.084	0.131	0.246	0.607	0.096	0.139	0.273	0.635	0.102	0.167	0.28	0.642	0.119	0.174	0.331	0.683
	0.6	0.07	0.108	0.229	0.542	0.077	0.115	0.236	0.55	0.079	0.136	0.263	0.587	0.106	0.158	0.293	0.601	0.128	0.187	0.304	0.639
$n = 400$	0	0.053	0.287	0.751	0.994	0.053	0.287	0.749	0.994	0.052	0.283	0.751	0.994	0.05	0.288	0.754	0.995	0.05	0.29	0.756	0.994
	0.2	0.049	0.219	0.624	0.971	0.049	0.22	0.622	0.971	0.054	0.228	0.633	0.969	0.059	0.224	0.632	0.968	0.064	0.243	0.648	0.971
	0.4	0.059	0.177	0.508	0.942	0.062	0.183	0.507	0.943	0.067	0.188	0.517	0.946	0.069	0.193	0.531	0.946	0.084	0.209	0.546	0.952
	0.6	0.077	0.164	0.387	0.846	0.079	0.169	0.394	0.848	0.085	0.17	0.401	0.85	0.087	0.174	0.406	0.857	0.095	0.191	0.441	0.872

Table 2.5: Monte Carlo MSE, Variance and Bias

ρ	n	K	$MSE(\hat{g}_x)$	$Var(\hat{g}_x)$	$Bias(\hat{g}_x)$	$MISE(\hat{g})$	$MSE(\hat{\delta})$
0	100	4	0.1884	0.024	0.4054	0.1965	0.0149
		6	0.0315	0.0258	0.0752	0.0587	0.0131
		9	0.0384	0.0384	-0.0029	0.0808	0.0136
	400	4	0.1717	0.006	0.407	0.1785	0.0037
		6	0.016	0.0064	0.098	0.0264	0.0031
		9	0.0081	0.0077	0.0211	0.0213	0.0031
0.2	100	4	0.1891	0.0316	0.3969	0.2009	0.0191
		6	0.0394	0.0334	0.0775	0.0676	0.017
		9	0.0433	0.0432	0.0097	0.0873	0.017
	400	4	0.1707	0.0083	0.403	0.1811	0.004
		6	0.0168	0.0083	0.0924	0.028	0.0035
		9	0.01	0.0099	0.0138	0.0233	0.0034
0.4	100	4	0.198	0.0473	0.3881	0.2107	0.0184
		6	0.0529	0.0475	0.0734	0.0815	0.0179
		9	0.0578	0.0578	0.0028	0.1009	0.018
0.35	400	4	0.177	0.0118	0.4064	0.1827	0.0046
		6	0.0214	0.0115	0.0996	0.0321	0.0042
		9	0.013	0.0126	0.0205	0.0272	0.0042
0.52	100	4	0.2089	0.0558	0.3913	0.2186	0.0225
		6	0.0614	0.0542	0.085	0.0942	0.021
		9	0.0654	0.065	0.0195	0.1146	0.0212
0.5	400	4	0.1778	0.0164	0.4018	0.1878	0.0062
		6	0.0264	0.017	0.0968	0.0387	0.0058
		9	0.0183	0.018	0.0171	0.0343	0.0058

Table 2.6: Monte Carlo average 95 % CI length

n	K	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.52$
100	4	0.6241	0.6653	0.6816	0.717
	6	0.5823	0.628	0.6471	0.6776
	9	0.5812	0.624	0.6484	0.6747
n	K	$\rho = 0$	$\rho = 0.2$	$\rho = 0.35$	$\rho = 0.5$
400	4	0.3099	0.3234	0.3483	0.3793
	6	0.2869	0.3026	0.3288	0.361
	9	0.2862	0.3016	0.3271	0.3584

Table 2.7: Coverage Probabilities

		$\rho = 0$			$\rho = 0.2$			$\rho = 0.4$			$\rho = 0.52$		
n	K	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
100	4	0.901	0.946	0.99	0.881	0.935	0.984	0.892	0.939	0.991	0.876	0.939	0.981
	6	0.904	0.951	0.987	0.889	0.932	0.985	0.874	0.938	0.986	0.879	0.93	0.989
	9	0.888	0.946	0.989	0.885	0.932	0.989	0.884	0.936	0.988	0.877	0.925	0.982
		$\rho = 0$			$\rho = 0.2$			$\rho = 0.35$			$\rho = 0.5$		
n	K	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
400	4	0.9	0.943	0.985	0.905	0.945	0.987	0.904	0.956	0.993	0.881	0.932	0.976
	6	0.895	0.947	0.985	0.912	0.957	0.992	0.896	0.944	0.988	0.884	0.934	0.975
	9	0.901	0.952	0.99	0.911	0.959	0.987	0.892	0.944	0.989	0.882	0.923	0.975

Table 2.8: Empirical power of 95% test

$n = 100$	$\rho \backslash \delta$	0	0.1	0.2	**0.3**	0.4	0.5	0.7	1
100	0	0.538	0.296	0.13	0.049	0.123	0.299	0.759	0.992
	0.2	0.481	0.291	0.111	0.068	0.108	0.293	0.716	0.966
	0.4	0.481	0.263	0.111	0.062	0.124	0.291	0.686	0.969
	0.52	0.446	0.267	0.115	0.07	0.12	0.273	0.649	0.941
400	0	0.964	0.789	0.284	0.3	0.4	0.775	0.995	1
	0.2	0.951	0.738	0.276	0.043	0.287	0.753	0.993	1
	0.35	0.912	0.676	0.259	0.056	0.248	0.664	0.988	1
	0.5	0.868	0.602	0.235	0.066	0.225	0.613	0.963	1

2.7 Empirical examples

This section presents two illustrative empirical examples in which the series estimation and studentization method of this chapter are applied. The examples are from Yatchew (2003) and are analysed by fitting the partly linear specification of

$$Y_i = Z_i'\delta_0 + h_0(X_i) + U_i.$$

The series estimation of δ_0 yields similar values of $\hat{\delta}$ to the kernel estimates reported in Yatchew (2003). To test the hypothesis $H_0 : \delta_{0\ell} = 0$ against $H_1 : \delta_{0\ell} \neq 0$, $\ell = 1, \cdots, d$, the test using the usual t-statistic derived under independence assumption is contrasted with that based on the test statistic $t_n^* := n(\hat{\theta} - r)'\hat{C}_n^{-1}(\hat{\theta} - r), r = 0$ of Theorem 2.6 of this chapter, which allows for spatial dependence.

The first example involves a hedonic pricing of housing attributes. The data consists of a relatively small sample of 92 detached homes in Ottawa that were sold during 1987. The dependent variable is the sale price of a given house (*price*), while the regressors contain various attributes of the house including the lot size (*lotarea*), square footage of housing (*usespc*), number of bedrooms (*nrbed*), average neighbourhood income (*acginc*), distance to highway (*dhwy*), presence of garage (*grge*), fireplace (*frplc*), and luxury bathroom (*lux*). In the non-parametric function enter two location coordinates denoted s and w (south and west) of the house:

$$\begin{aligned} price &= h(s,w) + \delta_1 frplc + \delta_2 grge + \delta_3 lux + \delta_4 acginc + \delta_5 dhwy \\ &\quad + \delta_6 lotarea + \delta_7 nrbed + \delta_8 usespc + u. \end{aligned}$$

The first set of columns of Table 2.9 recalls the results of kernel semi-parametric estimation reported in Yatchew (2003) based on the work of Robinson (1988). The second set of columns reports the corresponding results from series estimation. The estimates of coefficients, their standard errors and the t-statistics are broadly similar, reporting significance of many of the regressors at the 5% level. Series estimation was based on $(1, \quad s, \quad w, \quad sw)$ as approximating functions.

In applying the studentization of the previous section, the ordering of the data is important in the light of Assumption C3. We have ordered data in ascending order of the distance from the geographical coordinate $(s, \quad w) = (0,0)$, expecting spatial dependence in the error terms of neighbouring houses. SE reports standard error under assumption of independence, TS^* is test statistic t_n^* of Section 4.5 with critical values $46.39, 28.88$ at sizes 5% and 10%, respectively. Test statistics with * are significant at 5% level and those with $\triangle$ at 10% significance level. The test statistic t_n^* of this work, which accounts for dependence, reports that the presence of fire place and luxury bathroom are significant at the 5% significance level and square footage, presence of fire place, luxury bathroom, and garage at 10% level, which may be more informative, bearing in mind the small sample size of 92.

Table 2.9: Hedonic House Pricing

	kernel			**series**			
Variable	Coef	SE	t-stat	Coef	SE	t-stat	TS^*
frplc	12.6	5.8	2.17*	12.7	5.62	2.26*	126.23*
grge	12.9	4.9	2.63*	12.8	4.31	2.97*	29.98$^{\triangle}$
lux	57.6	10.6	5.43*	58.2	11.3	5.15*	177.10*
acginc	0.6	0.23	2.61*	0.61	0.2	3.08*	22.06
dhwy	1.5	21.4	0.07	-9.2	5.86	-1.57	10.38
lotarea	3.1	2.2	1.41	3.8	1.85	2.03*	22.12
nrbed	6.4	4.8	1.33	7.8	4.2	1.85$^{\triangle}$	14.57
usespc	24.7	10.6	2.33*	23.6	11.6	2.04*	37.67$^{\triangle}$

The contrasting conclusions on the significance of δ-coefficients between the t-test under independent errors and test t_n^* allowing for dependence may be due to a presence of cross-sectional dependence in the data. This seems to be natural as the dependent variable is price of houses of the same type, sold in the same year and city, which would have been subject to an overlapping set of demand and supply side factors, driven by the same macroeconomic fundamentals.

The second empirical example concerns the following cost function of distributing electricity, also from Yatchew (2003):

$$\begin{aligned} tc &= f(cust) + \delta_1 wage + \delta_2 pcap + \frac{\delta_3}{2} wage^2 + \frac{\delta_4}{2} pcap^2 + \delta_5 wage \cdot pcap \\ &\quad + \delta_6 PUC + \delta_7 kwh + \delta_8 life + \delta_9 lf + \delta_{10} kmwire + u. \end{aligned}$$

We have as the dependent variable, tc, the log of total cost per customer. As regressors, $cust$ is the log of the number of customers, $wage$ is the log wage rate, $pcap$ is the log price of capital, PUC is a dummy variable for public utility commissions that deliver additional services, therefore may benefit from economies of scope, $life$ is the log of the remaining life of distribution assets, lf is the log of the load factor (this measures capacity utilization relative to peak usage), and $kmwire$ is the log of kilometers of distribution wire per customer. In Yatchew (2003), it is of interest to non-parametrically estimate the conditional expectation of tc given $cust$, holding other regressors fixed, as the shape of this curve reveals whether there are increasing/decreasing returns to scale in electricity distribution. For the purpose of this chapter, we are interested in the estimates of the linear parameters δ's and test of their significance, $H_0 : \delta_l = 0, \quad H_1 : \delta_l \neq 0$ for $l = 1, \cdots, d$, when allowing for dependence in the error terms. The data consists of 81 municipal distributors in Ontario, Canada, during 1993.

The first set of columns of Table 2.10 replicates the kernel estimates of δ's and their standard errors assuming uncorrelatedness of error terms from Yatchew (2003). The second set of columns report the estimates using series estimation, where the first three Legendre polynomials were used as the series functions. The test statistics with

$*$ are those significant at 5% significance level, while those with $\triangle$ are significant at 10% significance level. In order to apply the studentization of section 2.5.3, the data needs to be ordered with Assumption C3 in mind. Two different orderings were tried. Firstly, the data was ordered in the ascending order of wage rate faced by the firm. The rationale behind this ordering is that firms may be subject to input shocks, and those with similar wage rate may use similar inputs, leading to dependence in their disturbance terms. Test statistics based on this stundentization is denoted TS_w^* in Table 2.10. Secondly, data was ordered according to the number of employees of the firm, which is a measure of the firm size. One may envisage that firms with similar sizes are subject to similar shocks, or alternatively, are interdependent due to e.g. competition. Test statistics based on this stundentization is denoted TS_e^* in Table 2.10.

Inference based on the assumption of uncorrelated error terms lead to $PUC, life, lf$ and *kmwire* being significant at 5% level with kernel estimation, while $PUC, life$ and *kmwire* are reported significant at 5% level with series estimation. When allowing for dependence in the error terms, $life$ and *kmwire* are still reported significant at 5% level with both orderings, while lf is now reported significant at 10% level. *pcap* and *wage* $\cdot$ *pcap* are reported significant at 10% level under both orderings, while PUC, which was significant at 5% level under uncorrelatedness assumption on error terms, is now significant at 10%, only with the ordering according to number of employees.

Table 2.10: Cost function in Electricity Distribution

	kernel			**series**				
	Coef	SE	t-stat	Coef	SE	t-stat	TS_w^*	TS_e^*
wage	-6.298	12.453	-0.506	-6.002	15.736	-0.381	0.426	0.261
pcap	-1.393	1.6	-0.872	-2.531	1.846	-1.371	44.08$^{\triangle}$	35.433$^{\triangle}$
$\frac{1}{2}wage^2$	0.72	2.13	0.3388	1.731	12.837	0.135	0.061	0.036
$\frac{1}{2}pcap^2$	0.032	0.066	0.485	0.148	0.318	0.466	1.593	1.491
$wage \cdot pcap$	0.534	0.599	0.891	2.044	1.553	1.317	43.155$^{\triangle}$	40.27$^{\triangle}$
PUC	-0.086	0.039	-2.205^*	-0.043	0.017	-2.6^*	11.042	28.893$^{\triangle}$
kwh	0.033	0.086	0.384	0.0828	0.102	0.8085	8.208	9.486
life	-0.634	0.115	-5.513^*	-0.613	0.124	-4.935^*	104.6*	92.7*
lf	1.249	0.436	2.865*	0.746	0.486	1.535	39.669$^{\triangle}$	36.587$^{\triangle}$
kmwire	0.399	0.087	4.586*	0.442	0.088	5.012*	202.65*	151.02*

2.8 Conclusion

This chapter has established the theoretical background for the series estimation of a vector-valued functional of the non-parametric regression function under cross-sectional dependence and nonstationarity. A uniform rate of consistency, asymptotic normality and sufficient conditions for the $\sqrt{n}$ rate of convergence were provided. Importantly, a data-driven studentization method that offers an alternative to exiting methods of inference was introduced for the $\sqrt{n}$-consistent semi-parametric estimates.

The problem of inference for non-parametric or semi-parametric estimates that do not achieve the $\sqrt{n}$ rate of convergence remains open and calls for further research.

The framework of cross-sectional dependence and non-stationarity of this chapter and its asymptotic arguments, e.g. application of the FCLT, may be used to establish asymptotic theory for other estimation methods under the cross-sectional setting. The robust inference offered by the studentization of this chapter provides a new tool for inference with cross-sectional data and needs to be extended to other commonly used methods such as GMM estimation of parametric models.

2.9 Appendix A. Proofs of Theorems 2.1-2.5.

The main matrix norm used in this work is *spectral* norm, $\|A\|^2 := \bar{\lambda}(A'A)$, defined as the largest eigenvalue of the matrix $A'A$. It is submultiplicative, i.e. $\|AB\| \leq \|A\|\|B\|$, and when A is positive semi-definite and symmetric, it satisfies $\|A^{1/2}\|^2 = \|A\|$ and $\|A^{-1}\|^{-1} = \|A\|$. When A is positive semi-definite, symmetric and random, one has that

$$\|A\| = O_p(E\|A\|) = O_p\big(E(\bar{\lambda}(A))\big) \leq O_p\big(E(tr(A))\big) = O_p\big(tr(E(A))\big).$$

In addition, three other matrix norms appear in the proofs. Let $\|\cdot\|_E$ denote Euclidean norm for matrix, $\|\cdot\|_C$ maximum column sum norm and $\|\cdot\|_R$ maximum row sum norm. Let $A = (a_{ij})$ be a $q \times q$ matrix. Then,

$$\|A\|_E^2 := \big(\sum_{i,j=1}^{q} a_{ij}^2\big), \quad \|A\|_C := \max_{1\leq j\leq q}\big(\sum_{i=1}^{q}|a_{ij}|\big), \quad \|A\|_R := \max_{1\leq i\leq q}\big(\sum_{j=1}^{q}|a_{ij}|\big).$$

The following inequalities hold:

$$\|A\| \leq \|A\|_E, \quad \|A\|^2 \leq \|A\|_R\|A\|_C, \quad |tr(AB)| \leq \|A\|_E\|B\|_E,$$

$$\|AB\|_E \leq \|A\|_E\|B\|, \quad \|AB\|_E \leq \|A\|_E\|B\|_E.$$

The above facts can be found in Searle (1982), Horn and Johnson (1990) and the appendix of Davies (1973).

Alternative representations of $\hat{m}$ in p^K and P

In Section 3, we introduced a $K \times 1$ vector of normalised functions $P(x) = P^K(x) = B_K^{-1/2}p^K(x)$ satisfying $E(P(X_i)P(X_i)') = I_K$. Given that the series estimator $\hat{m}(\cdot)$ is a projection of the unknown function $m(\cdot)$ onto the linear space spanned by $p_1(\cdot), \cdots, p_K(\cdot)$, the estimate $\hat{m}(\cdot)$ is invariant to any nonsingular linear transformation of approximating functions. Hence,

$$\hat{m}(x) = p^K(x)'\hat{\beta} = P(x)'\hat{\gamma}, \tag{2.9.1}$$

where $\hat{\beta} = (\mathbf{p}'\mathbf{p})^{-}\mathbf{p}'Y \in \mathbb{R}^K$ with

$$\mathbf{p} = \mathbf{p}_n = [p^K(X_1), \cdots, p^K(X_n)]' \in \mathbb{R}^{n\times K}, \qquad Y = Y_n = (Y_1, \cdots, Y_n)' \in \mathbb{R}^n$$

and $\hat{\gamma} = (\mathbf{P}'\mathbf{P})^{-}\mathbf{P}'Y \in \mathbb{R}^K$, where

$$\mathbf{P} = \mathbf{P}_n = [P(X_1), \cdots, P(X_n)]' \in \mathbb{R}^{n\times K}.$$

To show such invariance, one can use the equality $\mathbf{P} = \mathbf{p}B_K^{-1/2}$ to establish the following relation between $\hat{\gamma}$ and $\hat{\beta}$:

$$\hat{\gamma} = (\mathbf{P}'\mathbf{P})^{-}\mathbf{P}'Y = B_K^{1/2}(\mathbf{p}'\mathbf{p})^{-}B_K^{1/2}B_K^{-1/2}\mathbf{p}'Y = B_K^{1/2}\hat{\beta}.$$

Above equality holds, because the fact that $(\mathbf{p}'\mathbf{p})^{-}$ is the Moore-Penrose inverse of $\mathbf{p}'\mathbf{p}$ implies $(\mathbf{P}'\mathbf{P})^{-} = (B_K^{-1/2}\mathbf{p}'\mathbf{p}B_K^{-1/2})^{-} = B_K^{1/2}(\mathbf{p}'\mathbf{p})^{-}B_K^{1/2}$.[1]

Proof of Theorems 2.1-2.5 benefits from algebraic convenience of studying the representation $\hat{m}(x) = P(x)'\hat{\gamma}$ instead of $\hat{m}(x) = p^K(x)'\hat{\beta}$. Assumptions imposed on quantities involving $p^K(\cdot)$ such as $\xi(K)$ will continue to hold for their counterparts defined in terms of $P^K(\cdot)$. To show this fact for Assumption A4, note that

$$p^K(x)'\beta_K = P(x)'\gamma_K, \quad \text{where} \quad \gamma_K = B_K^{1/2}\beta_K.$$

Therefore, Assumption A4 implies

$$|m - P'\gamma_K|_\infty = O(K^{-\alpha}), \quad \text{as} \quad K \to \infty.$$

To verify that assumptions involving the upper bound $\xi(K)$ continue to hold for the corresponding quantity based on $P(\cdot)$, define:

$$\zeta(K) = \sup_{x\in\mathcal{X}} \|P^k(x)\|.$$

Then, for some $C < \infty$, $\zeta(k) \leq C\xi(k)$ for all $k \geq 1$, because

$$\zeta(k) = \sup_{x\in\mathcal{X}} \|B_k^{-1/2}p^k(x)\| \leq \|B_k^{-1/2}\| \sup_{x\in\mathcal{X}} \|p^k(x)\| \leq C\xi(k), \tag{2.9.2}$$

noting that by Assumption A3(i) and symmetry and positive semi-definiteness of B_K,

$$\|B_K^{-1/2}\| = \|B_K^{-1}\|^{1/2} = (\bar{\lambda}(B_K^{-1}))^{1/2} = (\underline{\lambda}(B_K))^{-1/2} \leq C.$$

The bound indicates that assumptions involving the upper bound $\xi(K)$ continue to hold also for $\zeta(K)$. The rest of the proof will be completed using $\hat{m}(x) = P(x)'\hat{\gamma}$. Wherever needed, translation to and from the two alternative representations of $\hat{m}$ given in (2.9.1) is clarified.

Proof of Theorem 2.1. Let $M := M_n = (m(X_1), \cdots, m(X_n))' \in \mathbb{R}^n$ and $\hat{Q} := \hat{Q}_n = \mathbf{P}'\mathbf{P}/n \in \mathbb{R}^{K\times K}$. We shall use these notations for the rest of the proof. To study the order of $|\hat{m} - m|_\infty$, we decompose the quantity $\hat{m}(x) - m(x)$ into the bias and stochastic terms. Let $\gamma_K = B_K^{1/2}\beta_K$ for β_K of Assumption A4. Write:

$$\hat{m}(x) - m(x) = \left[P(x)'(\hat{\gamma} - \gamma_K)\right] + \left[P(x)'\gamma_K - m(x)\right], \tag{2.9.3}$$

[1] Existence and uniqueness of Moore-Penrose inverse were established in Penrose (1955). The four Penrose conditions can be found in Searle (1982), pp. 212.

where $\hat{\gamma} = (\mathbf{P}'\mathbf{P})^{-}\mathbf{P}'Y = (\hat{Q})^{-}\mathbf{P}'Y/n$. Recall

$$\Sigma_n = E\left(\mathbf{P}'UU'\mathbf{P}/n\right)$$

the $K \times K$ variance-covariance matrix of the vector $\sum_{i=1}^{n} P(X_i)U_i/\sqrt{n}$. We shall show below that

$$\|\hat{\gamma} - \gamma_K\| = O_p\left(\frac{tr(\Sigma_n)^{1/2}}{n^{1/2}} + K^{-\alpha}\right). \tag{2.9.4}$$

Then, by the definition of $\zeta(K)$ and Assumption A4,

$$\begin{aligned} |\hat{m} - m|_\infty &\leq |P'(\hat{\gamma}_K - \gamma_K)|_\infty + |P'\gamma_K - m|_\infty \\ &\leq \zeta(K)\|\hat{\gamma}_K - \gamma_K\| + O(K^{-\alpha}) \\ &= O_p\left(\zeta(K)\left[\frac{tr(\Sigma_n)^{1/2}}{n^{1/2}} + K^{-\alpha}\right]\right). \end{aligned}$$

Therefore, we obtain the statement of Theorem 2.1:

$$|\hat{m} - m|_\infty = O_p\left(\xi(K)\left[\frac{tr(\Sigma_n)^{1/2}}{n^{1/2}} + K^{-\alpha}\right]\right).$$

Proof of (2.9.4). Observe that the matrix $\hat{Q}$ in $\hat{\gamma} = (\hat{Q})^{-}\mathbf{P}'Y/n$ depends on the sample $(X_1, \cdots, X_n)$ of random variables. Thus invertibility of $\hat{Q}$ for any given sample cannot be taken for granted. Let $1_n := I(\underline{\lambda}(\hat{Q}) \geq a)$ be the indicator function for the smallest eigenvalue of $\hat{Q}$, $\underline{\lambda}(\hat{Q})$, to be greater than some positive number $a < 1$. Then the inverse of $\hat{Q}$ exists when $1_n = 1$. It will be shown that $Pr(1_n = 1) \to 1$ as $n \to \infty$, so that $\hat{Q}^{-1}$ exists with probability tending to 1. First we study the quantity $1_n(\hat{\gamma} - \gamma_K)$, subsequently used to get the required result. Decompose $1_n(\hat{\gamma} - \gamma_K)$ as follows:

$$1_n(\hat{\gamma} - \gamma_K) = 1_n\left[\hat{Q}^{-1}\mathbf{P}'(Y - M)/n + \hat{Q}^{-1}\mathbf{P}'(M - \mathbf{P}\gamma_K)/n\right]. \tag{2.9.5}$$

Applying triangle inequality to (2.9.5) and the property $\|AB\| \leq \|A\|\|B\|$ of the spectral norm gives

$$\begin{aligned} \|1_n(\hat{\gamma} - \gamma_K)\| &\leq \|1_n\hat{Q}^{-1}\mathbf{P}'U/n\| + \|1_n\hat{Q}^{-1}\mathbf{P}'(M - \mathbf{P}\gamma_K)/n\| \\ &\leq \|1_n\hat{Q}^{-1}\|\|\mathbf{P}'U/n\| + \|1_n\hat{Q}^{-1}\mathbf{P}'/\sqrt{n}\|\|(M - \mathbf{P}\gamma_K)/\sqrt{n}\| \end{aligned} \tag{2.9.6}$$

Below we shall prove that

$$\begin{aligned} \|1_n\hat{Q}^{-1}\mathbf{P}'/\sqrt{n}\| &= O_p(1), & (2.9.7)\\ \|\mathbf{P}'U/n\| &= O_p\left(\frac{tr(\Sigma_n)^{1/2}}{\sqrt{n}}\right), & (2.9.8)\\ \|(M-\mathbf{P}\gamma_K)/\sqrt{n}\| &= O_p(K^{-\alpha}). & (2.9.9)\end{aligned}$$

These lead to

$$\|1_n(\hat{\gamma}-\gamma_K)\| = O_p\left(\frac{tr(\Sigma_n)^{1/2}}{n^{1/2}}+K^{-\alpha}\right),$$

which gives $\|\hat{\gamma}_K-\gamma_K\| = O_p\left(tr(\Sigma_n)^{1/2}/n^{1/2}+K^{-\alpha}\right)$. To see this, use the fact that $1-1_n = o_p(1)$ and the triangle inequality, to obtain

$$\begin{aligned} \|\hat{\gamma}-\gamma_K\| &\le \|1_n(\hat{\gamma}-\gamma_K)\| + \|(1-1_n)(\hat{\gamma}-\gamma_K)\| & (2.9.10)\\ &\le \|1_n(\hat{\gamma}-\gamma_K)\| + o_p(1)\|\hat{\gamma}-\gamma_K\|. \end{aligned}$$

Thus

$$\begin{aligned} &\|\hat{\gamma}-\gamma_K\|(1+o_p(1)) \le \|1_n(\hat{\gamma}-\gamma_K)\|,\\ &\|\hat{\gamma}-\gamma_K\| \le \|1_n(\hat{\gamma}-\gamma_K)\|/(1+o_p(1)) = O_p\left(tr(\Sigma_n)^{1/2}/n^{1/2}+K^{-\alpha}\right)(2.9.11)\end{aligned}$$

Proof of $1_n \to_p 1$. It suffices to show that $\underline{\lambda}(\hat{Q}) \to_p 1$, as $n\to\infty$.

First we derive $tr\left\{(\hat{Q}-I)^2\right\} = o_p(1)$. Recall the definition $P(x) := B_K^{-1/2}p^K(x) = [P_{1K}(x),\cdots,P_{KK}(x)]$. Observe that

$$\begin{aligned} E\left[tr\left\{(\hat{Q}-I)^2\right\}\right] &= \sum_{p,\ell=1}^{K} E[\{n^{-1}\sum_{i=1}^{n}P_{pK}(X_i)P_{\ell K}(X_i) - 1(\ell=p)\}^2]\\ &= n^{-2}\sum_{p,\ell=1}^{K} Var\left(\sum_{i=1}^{n}P_{pK}(X_i)P_{\ell K}(X_i)\right),\end{aligned}$$

noting that $E(\hat{Q}) = I$: $E(\hat{Q}) = n^{-1}\sum_{i=1}^{n}E(P(X_i)P'(X_i))$ where $E(P(X_i)P'(X_i)) = B_K^{-1/2}E[p^K(X_i)p^K(X_i)']B_K^{-1/2} = B_K^{-1/2}B_KB_K^{-1/2} = I$. For any pair $p,\ell = 1,\cdots,k$,

$$\begin{aligned} &Var\left(\sum_{i=1}^{n}P_{pK}(X_i)P_{\ell K}(X_i)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}Cov\left\{P_{pK}(X_i)P_{\ell K}(X_i), P_{pK}(X_j)P_{\ell K}(X_j)\right\}\\ &= \sum_{i=1}^{n}Var\left(P_{pK}(X_i)P_{\ell K}(X_i)\right) + \sum_{i,j=1,j\neq i}^{n} Cov\left\{P_{pK}(X_i)P_{\ell K}(X_i), P_{pK}(X_j)P_{\ell K}(X_j)\right\}\\ &=: V_{n,1}^{(p,\ell)} + V_{n,2}^{(p,\ell)}.\end{aligned}$$

Then $E[\|\hat{Q}-I\|^2] \leq n^{-2}\sum_{p,\ell=1}^{K}(V_{n,1}^{(p,\ell)}+V_{n,2}^{(p,\ell)})$. One has

$$\frac{1}{n^2}V_{n,1}^{(p,\ell)} = \frac{1}{n^2}\sum_{i=1}^{n} Var\big(P_{pK}(X_i)P_{\ell K}(X_i)\big) \leq \frac{\zeta^4(K)}{n}.$$

To bound $V_{n,2}^{(p,\ell)}$ we use Assumption A5:

$$\begin{aligned}\frac{1}{n^2}|V_{n,2}^{(p,\ell)}| &= |\int P_{pK}(x)P_{\ell K}(x)P_{pK}(y)P_{\ell K}(y)\big(\frac{1}{n^2}\sum_{i,j=1,j\neq i}^{n}\{f_{ij}(x,y)-f(x)f(y)\}\big)dxdy| \\ &\leq \zeta^4(K)\big(\frac{1}{n^2}\sum_{i,j=1,i\neq j}\int|f_{ij}(x,y)-f(x)f(y)|dxdy\big) = \zeta^4(K)n^{-2}\triangle_n.\end{aligned}$$

Therefore,

$$\begin{aligned} E\left[tr\left\{(\hat{Q}-I)^2\right\}\right] &= \sum_{p,\ell=1}^{K}(V_{n,1}^{(p,\ell)}+V_{n,2}^{(p,\ell)}) \\ &\leq \frac{K^2\zeta^4(K)}{n}+\frac{K^2\zeta^4(K)\triangle_n}{n^2} \\ &= K^2\zeta^4(K)\left(\frac{1}{n}+\frac{\triangle_n}{n^2}\right) = o(1), \qquad (2.9.12)\end{aligned}$$

by Assumptions A3(ii), and A5.

Hence to show $\underline{\lambda}(\hat{Q}) \to_p 1$, it suffices to verify that $|\underline{\lambda}(\hat{Q})-\underline{\lambda}(I)| \leq \left[tr\left\{(\hat{Q}-I)^2\right\}\right]^{1/2}$. The symmetric matrix $(\hat{Q}-I)$ can be written as $\hat{Q}-I = C(\hat{\Lambda}-I)C'$, where $C=(c_{ij})\in\mathbb{R}^{K\times K}$ is orthonormal eigenvector matrix such that $C'C=I$ and $\hat{\Lambda}$ is a diagonal matrix consisting of eigenvalues of $\hat{Q}$. Consequently, $(\hat{Q}-I)^2 = C(\hat{\Lambda}-I)^2C'$. Now, $tr\{(\hat{Q}-I)^2\} = tr\left(C(\hat{\Lambda}-I)^2C'\right) = \sum_{\ell=1}^{K}(\lambda_\ell(\hat{Q})-1)^2$, because

$$tr\left(C(\hat{\Lambda}-I)^2C'\right) = \sum_{i=1}^{K}\sum_{j=1}^{K}c_{ij}^2(\hat{\lambda}_j-1)^2 = \sum_{j=1}^{K}(\hat{\lambda}_j-1)^2\left(\sum_{i=1}^{K}c_{ij}^2\right) = \sum_{j=1}^{K}(\hat{\lambda}_j-1)^2,$$

because columns of C are orthonormal. Therefore,

$$(\underline{\lambda}(\hat{Q})-1)^2 \leq tr\{(\hat{Q}-I)^2\}, \qquad |\underline{\lambda}(\hat{Q})-1| \leq [tr\{(\hat{Q}-I)^2\}]^{1/2} = o_p(1),$$

as was concluded in (2.9.12). This completes the proof of $Pr(1_n=1)\to 1$ as $n\to\infty$.

Now we prove (2.9.7) -(2.9.9).

Proof of (2.9.7). Note that $\hat{Q}$ is symmetric and nonnegative definite. Thus, by the properties of the spectral norm,

$$\|1_n\hat{Q}^{-1}\| = 1_n\bar{\lambda}(\hat{Q}^{-1}) = 1_n(\underline{\lambda}(\hat{Q}))^{-1}.$$

The facts $1_n \to_p 1$ and $\underline{\lambda}(\hat{Q}) \to_p 1$ established above imply $1_n(\underline{\lambda}(\hat{Q}))^{-1} \to_p 1$. Hence, by Slutsky theorem, $\|1_n\hat{Q}^{-1}\| = O_p(1)$. Therefore,

$$\|1_n\hat{Q}^{-1}\mathbf{P}'/\sqrt{n}\|^2 = \|1_n\hat{Q}^{-1}\mathbf{P}'\mathbf{P}\hat{Q}^{-1}/n\| = \|1_n\hat{Q}^{-1}\| = O_p(1).$$

Proof of (2.9.8). One has

$$\|\mathbf{P}'U/n\| = \frac{1}{\sqrt{n}}\|\mathbf{P}'U/\sqrt{n}\| = \frac{1}{\sqrt{n}}\left[\bar{\lambda}\left(\frac{\mathbf{P}'UU'\mathbf{P}}{n}\right)\right]^{1/2} = O_p\left(\frac{tr\,(\Sigma_n)^{1/2}}{n^{1/2}}\right).$$

Proof of (2.9.9). We have,

$$\begin{aligned}\|(M-\mathbf{P}\gamma_K)/\sqrt{n}\|^2 &= (M-\mathbf{P}\gamma_K)'(M-\mathbf{P}\gamma_K)/n\\ &= \frac{1}{n}\sum_{i=1}^{n}(g(X_i)-P(X_i)\gamma_K)^2 = O_p(K^{-2\alpha}),\end{aligned}$$

by Assumption 4, which completes the proof of (2.9.4) and of the theorem. ■

Proof of Theorem 2.2. Let $T_n := A'\mathbf{P}'U/n$, where $\mathbf{P} = \mathbf{p}^K B_K^{-1/2} \in \mathbb{R}^n$, $A = \big(D(P_{1K}), D(P_{2K}), \cdots, D(P_{KK})\big)' \in \mathbb{R}^{K\times d}$ and $U = (U_1, \cdots, U_n)' \in \mathbb{R}^n$. Write

$$\hat{\theta}_n - \theta_0 = T_n + r_n, \quad r_n := \hat{\theta}_n - \theta_0 - T_n.$$

We shall show that

$$\sqrt{n}\bar{V}_n^{-1/2} r_n = o_p(1), \tag{2.9.13}$$

$$\sqrt{n}\bar{V}_n^{-1/2} T_n \to_d N(0, I_d), \tag{2.9.14}$$

which implies convergence (2.4.1) of Theorem 2.2.

Proof of (2.9.13). Again, let $1_n = I(\underline{\lambda}(\hat{Q}) \geq a)$ for some positive number $a < 1$ as in the proof of Theorem 2.1, hence $1_n = 1 + o_p(1)$. By the same argument as in proof of Theorem 2.1, (2.9.13) follows if we show that

$$1_n\sqrt{n}\bar{V}_n^{-1/2} r_n = o_p(1). \tag{2.9.15}$$

We shall use the bound $\|1_n\sqrt{n}\bar{V}_n^{-1/2} r_n\| \leq \sqrt{n}\|\bar{V}_n^{-1/2}\|\|1_n r_n\|$. To evaluate $\|1_n r_n\|$, recall $\bar{m} = P'\gamma_K$. Write

$$\begin{aligned}r_n = \hat{\theta}_n - \theta_0 - T_n &= \{a(\hat{m}) - a(m) - D(\hat{m}) + D(m)\}\\ &\quad + \{D(\hat{m}) - D(\bar{m}) - T_n\} + \{D(\bar{m}) - D(m)\}.\end{aligned}$$

Then

$$\begin{aligned} \|r_n\| &\leq \|a(\hat{m}) - a(m) - D(\hat{m}) + D(m)\| \\ &+ \|D(\hat{m}) - D(\bar{m}) - T_n\| + \|D(\bar{m}) - D(m)\| \\ &=: \|r_{n,1}\| + \|r_{n,2}\| + \|r_{n,3}\|. \end{aligned}$$

To show (2.9.13), note that by assumption of the theorem, $\|\bar{V}_n^{-1/2}\| = \|\bar{V}_n^{-1}\|^{1/2} = O_p(1)$. Thus, it suffices to prove that

$$1_n\sqrt{n}\|r_{n,i}\| = o_p(1), \quad i = 1, 2, 3. \tag{2.9.16}$$

For $i = 1$, by Assumption B1, $\|r_{n,1}\| = O_p(|\hat{m} - m|_\infty^2)$. Thus by Theorem 2.1 and Assumption B3(i), (iii)

$$\sqrt{n}\|r_{n,1}\| = O_p\left(\sqrt{n}\zeta(K)^2(\frac{tr(\Sigma_n)}{n} + K^{-2\alpha})\right) = o_p(1). \tag{2.9.17}$$

For $i = 2$, to bound $\|r_{n,2}\|$ recall the notation: $\hat{m}(x) = P(x)'\hat{\gamma}$, $\hat{Q} = \mathbf{P}'\mathbf{P}/n$, $\hat{\gamma} = (\mathbf{P}'\mathbf{P})^{-}\mathbf{P}'Y = \hat{Q}^{-}\mathbf{P}'Y/n$, $Y = M + U$ and $A = (D(P_{1K}), \cdots, D(P_{KK}))'$. Then,

$$\begin{aligned} D(\hat{m}) &= D(P'\hat{\gamma}) = A'\hat{\gamma} = A'\hat{Q}^{-}\mathbf{P}'(M + U)/n, & (2.9.18) \\ D(\bar{m}) &= D(P'\gamma_K) = A'\gamma_K. & (2.9.19) \end{aligned}$$

As in the proof of Theorem 2.1, one can replace $1_n\hat{Q}^{-}$ with $1_n\hat{Q}^{-1}$. Hence

$$\begin{aligned} \|1_n r_{n,2}\| &= \|1_n(A'\hat{Q}^{-1}\mathbf{P}'Y/n - A'\gamma_K - A'\mathbf{P}'U/n)\| \\ &= \|1_n A'\hat{Q}^{-1}\mathbf{P}'(M + U)/n - A'\gamma_K - A'\mathbf{P}'U/n\| \\ &= \|1_n A'(\hat{Q}^{-1} - I)\mathbf{P}'U/n + A'\hat{Q}^{-1}\mathbf{P}'(M - \mathbf{P}\gamma_K)/n\| \\ &\leq \|1_n A'(\hat{Q}^{-1} - I)\mathbf{P}'U/n\| + \|A'\hat{Q}^{-1}\mathbf{P}'(M - \mathbf{P}\gamma_K)/n\| \\ &\leq \|A'\|\|1_n(\hat{Q}^{-1} - I)\|\|\mathbf{P}'U/n\| + \|A'\|\|1_n\hat{Q}^{-1}\mathbf{P}'/\sqrt{n}\|\|(M - \mathbf{P}\gamma_K)/\sqrt{n}\|. \end{aligned}$$

Note that $\|A\|^2 \leq \zeta^2(K)$, $\|1_n\hat{Q}^{-1}\| = O_p(1)$, and by (2.9.7)- (2.9.9),

$$\|1_n\hat{Q}^{-1}\mathbf{P}'/\sqrt{n}\| = O_p(1), \quad \|(M - \mathbf{P}\gamma_K)/\sqrt{n}\| = O_p(K^{-\alpha}), \quad \|\mathbf{P}'U/n\| = O_p\left((tr(\Sigma_n)/n)^{1/2}\right).$$

Next, $\|1_n(\hat{Q}^{-1} - I)\| = \|1_n\hat{Q}^{-1}(I - \hat{Q})\| \leq \|1_n\hat{Q}^{-1}\|\|I - \hat{Q}\| = O_p(\|I - \hat{Q}\|)$. Thus,

$$\|r_{n,2}\| = O_p(1)\sqrt{K}\zeta(K)\left(\|I - \hat{Q}\|(tr(\Sigma_n)/n)^{1/2} + K^{-\alpha}\right).$$

To bound $\|I - \hat{Q}\|$ note that $E[\|\hat{Q} - I\|^2] = E\left[\bar{\lambda}\left\{(\hat{Q} - I)^2\right\}\right] \leq E\left[tr\left\{(\hat{Q} - I)^2\right\}\right]$.

From (2.9.12),

$$\sqrt{n}\|r_{n,2}\| \leq \left(nK\zeta^2(K)\right)^{1/2}\left\{\left[K^2\zeta^4(K)\left(\frac{1}{n}+\frac{\triangle_n}{n^2}\right)\frac{tr(\Sigma_n)}{n}\right]^{1/2}+K^{-\alpha}\right\} = o_p(1)$$

by Assumptions B3(ii) and (iii).

For $i = 3$, by linearity of $D(\cdot)$ and Assumption B2 and A4, $\|r_{n,3}\| = O(|\bar{m}-m|_\infty) = O(K^{-\alpha})$,

$$\sqrt{n}\|r_{n,3}\| = O_p(\sqrt{n}K^{-\alpha}) = o_p(1), \tag{2.9.20}$$

by Assumptions B3(iii), which implies $nK^{-2\alpha} = o(1)$.

Proof of (2.9.14). To show asymptotic normality of the main term $\sqrt{n}\bar{V}_n^{-1/2}T_n$, introduce the following representation

$$\begin{aligned}\sqrt{n}\bar{V}_n^{-1/2}T_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{V}_n^{-1/2}A'P(X_i)U_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{V}_n^{-1/2}A'P(X_i)\sigma(X_i)\sum_{j=1}^{\infty}b_{ij}\varepsilon_j \\ &= \sum_{j=1}^{\infty}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{V}_n^{-1/2}A'P(X_i)\sigma(X_i)b_{ij}\right)\varepsilon_j = \sum_{j=1}^{\infty}w_{jn}\varepsilon_j,\end{aligned}$$

letting

$$w_{jn} := \sum_{i=1}^{n}\bar{V}_n^{-1/2}A'P(X_i)\sigma(X_i)b_{ij}/\sqrt{n}. \tag{2.9.21}$$

Noting that w_{jn} is a function of $\{X_i\}_{i=1}^n$, we show asymptotic normality conditional on $\|\bar{V}_n^{-1}\| \leq C$ and $\{X_i\}_{i=1}^n$, treating w_{jn} as non-random. The key point here is to obtain the conditional asymptotic distribution to be $N(0, I_d)$, which is independent of $\{X_i\}_{i=1}^n$. This yields the required unconditional asymptotic normality result of Theorem 2.2. Such line of reasoning was used in Robinson (2011).

By Cramer-Wold device, to derive asymptotic normality of the vector $\sqrt{n}\bar{V}_n^{-1/2}T_n$, we focus on a scalar summation $\sum_{j=1}^{\infty}c'w_{jn}\varepsilon_j$ with any fixed vector $c \in \mathbb{R}^d$ such that $c'c = 1$. Consider splitting $\sqrt{n}c'\bar{V}_n^{-1/2}T_n$ into two sums,

$$\sqrt{n}c'\bar{V}_n^{-1/2}T_n = \sum_{j=1}^{N(n)}c'w_{jn}\varepsilon_j + \sum_{j=N(n)+1}^{\infty}c'w_{jn}\varepsilon_j,$$

where the integer $N(n)$ is chosen to be the smallest satisfying

$$\sum_{j=N(n)+1}^{\infty}(c'w_{jn})^2 \leq 1/\log n.$$

The choice of $N(n)$ is deterministic once we condition on $\{X_i\}_{i=1}^n$. The purpose of this truncation is to make the contribution from the second summation negligible:

$$\begin{aligned}(\sum_{j=N(n)+1}^{\infty} c'w_{jn}\varepsilon_j)^2 &= O_p(E(\sum_{j=N(n)+1}^{\infty} c'w_{jn}\varepsilon_j)^2) \\ &= O_p(\sum_{j=N(n)+1}^{\infty} (c'w_{jn})^2) = O_p\left(\frac{1}{\log n}\right) = o_p(1).\end{aligned}$$

Since $\{c'w_j\varepsilon_j\}$ are martingale differences under assumption A2, asymptotic normality of the first summation is established by verifying the following two sufficient conditions for asymptotic normality from Scott (1973), adapted for our setting.

$$\sum_{j=1}^{N(n)} E\big((c'w_j\varepsilon_j)^2\big) \to_p 1, \tag{2.9.22}$$

$$\sum_{j=1}^{N(n)} E\big((c'w_{jn}\varepsilon_j)^2 1(|c'w_{jn}\varepsilon_j| > \delta)\big) \to_p 0, \quad \forall\delta > 0. \tag{2.9.23}$$

By Assumption A2, we have

$$\sum_{j=1}^{N(n)} E\big((c'w_j\varepsilon_j)^2\big) = \sum_{j=1}^{N(n)} (c'w_{jn})^2.$$

By the choice of $N(n)$,

$$\sum_{j=1}^{N(n)} (c'w_{jn})^2 = \sum_{j=1}^{\infty}(c'w_{jn})^2 - \sum_{j=N(n)+1}^{\infty} (c'w_{jn})^2 = 1 + o(1).$$

Next let ν be as in Assumption A2. Then,

$$\begin{aligned}\sum_{j=1}^{N(n)} E[(c'w_{jn}\varepsilon_j)^2 1(|c'w_{jn}\varepsilon_j| > \delta)] &= \sum_{j=1}^{N(n)} (c'w_{jn})^2 E[\varepsilon_j^2 1(|c'w_{jn}\varepsilon_j| > \delta)] \\ \leq \sum_{j=1}^{N(n)} (c'w_{jn})^2 \left(\frac{|c'w_{jn}|}{\delta}\right)^\nu E|\varepsilon_j|^{2+\nu} &= \delta^{-\nu} \sum_{j=1}^{N(n)} |c'w_{jn}|^{2+\nu} E|\varepsilon_j|^{2+\nu} \\ \leq C\delta^{-\nu} \sum_{j=1}^{N(n)} |c'w_{jn}|^{2+\nu} &\leq C\delta^{-\nu} \max_{1\leq j\leq n} |c'w_{jn}|^\nu \sum_{j=1}^{N(n)} (c'w_{jn})^2.\end{aligned}$$

The first inequality follows from $1(|c'w_{jn}\varepsilon_j| > \delta) \leq (|c'w_{jn}\varepsilon_j|/\delta)^\nu$. With $\sum_{j=1}^{N(n)} (c'w_{jn})^2 \to 1$, (2.9.23) is verified once we show that $\max_{j\geq 1} |c'w_{jn}|^\nu \to 0$. Conditionally on $X_1, \cdots, X_n$, the following holds for any $j \geq 1$:

$$\begin{aligned}
|c'w_{jn}| &= \left| \frac{c'}{\sqrt{n}} \bar{V}_n^{-1/2} \sum_{i=1}^{n} A'P(X_i)\sigma(X_i)b_{ij} \right| \\
&\leq \|c\| \|\bar{V}_n^{-1/2}\| \frac{1}{\sqrt{n}} \max_{1 \leq j \leq n} \sum_{i=1}^{n} |b_{ij}| \|A'P(X_i)\sigma(X_i)\| \\
&= O\left(\frac{\zeta(K)^2}{\sqrt{n}} \max_{1 \leq j \leq n} \sum_{i=1}^{n} |b_{ij}| \right) = o(1), \qquad (2.9.24)
\end{aligned}$$

by Assumption B4 and the bound $\|A'P(X_i)\sigma(X_i)\| \leq C\|A\|\|P(X_i)\| \leq C\zeta^2(K)$. ■

Proof of Theorem 2.3. We will prove later that $\|\bar{V}_n - V_n\| = o_p(1)$. Then, $V_n^{-1}\bar{V}_n \rightarrow_p I$ since $\|V_n^{-1}\bar{V}_n - I\| \leq \|V_n^{-1}\|\|\bar{V}_n - V_n\| = o_p(1)$, which in turn gives $V_n\bar{V}_n^{-1} \rightarrow_p I$. It follows that

$$\|\bar{V}_n^{-1}\| \leq \|V_n^{-1}\|\|V_n\bar{V}_n^{-1}\| = O_p(1).$$

Now, to show the final statement, (2.4.6), of the Theorem 2.3, write:

$$\sqrt{n}V_n^{-1/2}(\hat{\theta} - \theta_0) = \sqrt{n}\bar{V}_n^{-1/2}(\hat{\theta} - \theta_0) + \sqrt{n}\big(V_n^{-1/2} - \bar{V}_n^{-1/2}\big)(\hat{\theta} - \theta_0).$$

The first term was shown to converge in distribution to $N(0, I_p)$ in Theorem 2.2, while the second term is negligible:

$$\|\sqrt{n}\big(V_n^{-1/2} - \bar{V}_n^{-1/2}\big)(\hat{\theta} - \theta_0)\| \leq \|\big(V_n^{-1/2}\bar{V}_n^{1/2} - I\big)\|\|\sqrt{n}\bar{V}_n^{-1/2}(\hat{\theta} - \theta_0)\| = o_p(1),$$

since $V_n^{-1/2}\bar{V}_n^{1/2} \rightarrow_p I$ from $V_n^{-1}\bar{V}_n \rightarrow_p I$, and thus $\|V_n^{-1/2}\bar{V}_n^{1/2} - I\| = o_p(1)$.

Proof of $\|\bar{V}_n - V_n\| = o_p(1)$. By definition of the spectral norm, $\|\bar{V}_n - V_n\| = o_p(1)$ follows if $|(\bar{V}_n - V_n)_{\ell p}| = o_p(1)$, for all $\ell, p = 1, \cdots, d$, where $(B)_{\ell p}$ denotes the $(\ell, p)^{th}$ element of a matrix B. Then, using notation (2.4.2),

$$\begin{aligned}
(\bar{V}_n - V_n)_{\ell p} &= \frac{1}{n} \sum_{i,j=1}^{n} \gamma_{ij} \left\{ \sigma(X_i)A'_\ell P(X_i)\sigma(X_j)P'(X_j)A_p - E(\sigma(X_i)A'_\ell P(X_i)\sigma(X_j)P'(X_j)A_p) \right\} \\
&= \frac{1}{n} \sum_{i,j=1}^{n} \gamma_{ij} \left\{ h_i^{(\ell)}h_j^{(p)} - E(h_i^{(\ell)}h_j^{(p)}) \right\}.
\end{aligned}$$

Since

$$h_i^{(\ell)}h_j^{(p)} - E(h_i^{(\ell)}h_j^{(p)}) = \left\{ \bar{h}_i^{(\ell)}\bar{h}_j^{(p)} - E(\bar{h}_i^{(\ell)}\bar{h}_j^{(p)}) \right\} + \bar{h}_j^{(p)}E(h_i^{(\ell)}) + \bar{h}_i^{(\ell)}E(h_j^{(p)}),$$

we obtain that

$$\begin{aligned}(\bar{V}_n - V_n)_{\ell p} &= \frac{1}{n}\sum_{i,j=1}^{n} \gamma_{ij}\left\{\bar{h}_i^{(\ell)}\bar{h}_j^{(p)} - E(\bar{h}_i^{(\ell)}\bar{h}_j^{(p)})\right\} \\ &+ \frac{1}{n}\sum_{i,j=1}^{n} \gamma_{ij}\bar{h}_j^{(p)}E(h_i^{(\ell)}) + \frac{1}{n}\sum_{i,j=1}^{n} \gamma_{ij}\bar{h}_i^{(\ell)}E(h_j^{(p)}) \\ &=: S_{1,n} + S_{2,n} + S_{3,n}.\end{aligned}$$

We shall show that

$$Var(S_{k,n}) = o(1), \quad k = 1, 2, 3, \tag{2.9.25}$$

which proves $\|\bar{V}_n - V_n\| = o_p(1)$.

Proof of (2.9.25), k=1. We have

$$Var(S_{1,n}) = \frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4} Cov\left(\bar{h}_{i_1}^{(\ell)}\bar{h}_{i_2}^{(p)}, \bar{h}_{i_3}^{(\ell)}\bar{h}_{i_4}^{(p)}\right).$$

Introduce the notation, $\phi_{ij}^{(\ell,p)} := Cov(\bar{h}_i^{(\ell)}, \bar{h}_j^{(p)})$ and denote by $\Phi^{(\ell,p)}$ the $n \times n$ matrix whose $(i,j)^{th}$ element is $\phi_{ij}^{(\ell,p)}$. Recall that by the Definition 2 of joint 4^{th} order cumulant,

$$Cov(Z_1Z_2, Z_3Z_4) = \kappa(Z_1, Z_2, Z_3, Z_4) + Cov(Z_1, Z_3)Cov(Z_2, Z_4) + Cov(Z_1, Z_4)Cov(Z_2, Z_3).$$

One has

$$\begin{aligned}Var(S_{1,n}) &= \frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4}\kappa(\bar{h}_{i_1}^{(\ell)}, \bar{h}_{i_2}^{(p)}, \bar{h}_{i_3}^{(\ell)}, \bar{h}_{i_4}^{(p)}) && (2.9.26) \\ &+ \frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4}\phi_{i_1 i_3}^{(\ell,\ell)}\phi_{i_2 i_4}^{(p,p)} && (2.9.27) \\ &+ \frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4}\phi_{i_1 i_4}^{(\ell,p)}\phi_{i_2 i_3}^{(p,\ell)}. && (2.9.28)\end{aligned}$$

Denote by $\Gamma = \Gamma_n$ the $n \times n$ matrix whose $(i,j)^{th}$ element is γ_{ij}. Firstly, by Assumption B7, the RHS of (2.9.26) is $o(1)$. To bound (2.9.27) and (2.9.28), write

$$\begin{aligned}\frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4}\phi_{i_1 i_3}^{(\ell,\ell)}\phi_{i_2 i_4}^{(p,p)} &= \frac{1}{n^2}tr\left(\Gamma\Phi^{(p,p)}\Gamma\Phi^{(\ell,\ell)}\right), \\ \frac{1}{n^2}\sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2}\gamma_{i_3 i_4}\phi_{i_1 i_4}^{(\ell,p)}\phi_{i_2 i_3}^{(p,\ell)} &= \frac{1}{n^2}tr\left(\Gamma\Phi^{(p,\ell)}\Gamma\Phi^{(p,\ell)}\right).\end{aligned}$$

By the properties of matrix norms given earlier, we see that

$$\left|tr\left(\Gamma\Phi^{(p,p)}\Gamma\Phi^{(\ell,\ell)}\right)\right| \leq \|\Gamma\Phi^{(p,p)}\|_E\|\Gamma\Phi^{(\ell,\ell)}\|_E \leq \|\Gamma\|^2\|\Phi^{(p,p)}\|_E\|\Phi^{(\ell,\ell)}\|_E. \quad (2.9.29)$$

Partition $\|\Phi^{(p,p)}\|_E^2 = \sum_{i,j=1}^{n}(\phi_{ij}^{(p,p)})^2 = \sum_{i=1,i=j}^{n}(\phi_{ii}^{(p,p)})^2 + \sum_{i,j=1,i\neq j}^{n}(\phi_{ij}^{(p,p)})^2$. For $i = j$, $|\phi_{ii}^{(p,p)}| = Var(\bar{h}_i^{(p)}) \leq \zeta^4(K)$. For $i \neq j$, one has $|\phi_{ij}^{(p,p)}| \leq C\zeta^4(K)\int_{\mathcal{X}^2}|f_{ij}(x,y) - f(x)f(y)|dxdy$, since $|\sigma(X_i)A_p'P(X_i)| \leq C\zeta^2(K)$. Therefore,

$$\|\Phi^{(p,p)}\|_E^2 \leq Cn\zeta^8(K) + C\zeta^8(K)\sum_{i,j=1,i\neq j}^{n}\left(\int|f_{ij}(x,y) - f(x)f(y)|dxdy\right)^2.$$

It is clear that$\int|f_{ij}(x,y) - f(x)f(y)|dxdy \leq 2$ for all i and j. Hence,

$$\sum_{i,j=1,i\neq j}^{n}\left(\int|f_{ij}(x,y) - f(x)f(y)|dxdy\right)^2 \leq 2\sum_{i,j=1,i\neq j}^{n}\int|f_{ij}(x,y)-f(x)f(y)|dxdy = 2\triangle_n.$$

Thus, for any $p = 1,\cdots,d$,

$$\|\Phi^{(p,p)}\|_E^2 = \sum_{i,j=1}^{n}(\phi_{ij}^{(p,p)})^2 \leq C\zeta^8(K)(n+\triangle_n). \quad (2.9.30)$$

Hence, by (2.9.29) and Assumption B6,

$$\frac{1}{n^2}\|\Gamma\|^2\|\Phi^{(p,p)}\|_E\|\Phi^{(\ell,\ell)}\|_E \leq \frac{1}{n^2}\left(\max_{j\geq 1}\sum_{i=1}^{n}|\gamma_{ij}|\right)^2\zeta^8(K)(n+\triangle_n) = o(1),$$

since by the property of spectral norm $\|A\|^2 \leq \|A\|_C\|A\|_R$, and by the symmetry of Γ,

$$\|\Gamma\|^2 \leq \|\Gamma\|_C^2 = \left(\max_{j\geq 1}\sum_{i=1}^{n}|\gamma_{ij}|\right)^2.$$

Similarly, it follows that $n^{-2}tr\left(\Gamma\Phi^{(p,\ell)}\Gamma\Phi^{(p,\ell)}\right) = o(1)$, which completes the proof of (2.9.25) when $k = 1$.

Proof of (2.9.25), k=2,3. Recall, $S_{n,2} = n^{-2} \sum_{i,j=1}^{n} \gamma_{ij} \bar{h}_j^{(p)} E(h_i^{(\ell)})$. Therefore,

$$\begin{aligned}
Var(S_{2,n}) &= \frac{1}{n^2} \sum_{i_1,i_2,i_3,i_4=1}^{n} \gamma_{i_1 i_2} \gamma_{i_3 i_4} E(h_{i_1}^{(\ell)}) E(h_{i_3}^{(\ell)}) E(\bar{h}_{i_2}^{(p)} \bar{h}_{i_4}^{(p)}) \\
&= \frac{1}{n^2} \sum_{i_2,i_4=1}^{n} \left(\sum_{i_1=1}^{n} \gamma_{i_1 i_2} E(h_{i_1}^{(\ell)}) \right) \left(\sum_{i_3=1}^{n} \gamma_{i_3 i_4} E(h_{i_3}^{(\ell)}) \right) \phi_{i_2 i_4}^{(p,p)} \\
&\leq \frac{1}{n^2} \left(\zeta^2(K) \left| \max_{1 \leq j \leq n} \sum_{i=1}^{n} \gamma_{ij} \right| \right)^2 \sum_{i,j=1}^{n} |\phi_{ij}^{(p,p)}| \\
&\leq \frac{1}{n^2} \left(\zeta^2(K) \max_{1 \leq j \leq n} \sum_{i=1}^{n} |\gamma_{ij}| \right)^2 \sum_{i,j=1}^{n} |\phi_{ij}^{(p,p)}|
\end{aligned}$$

using the bound $E|h_i^{(\ell)}| \leq C\zeta^2(K)$. By the same steps taken in two lines prior to (2.9.30),

$$\sum_{i,j=1}^{n} |\phi_{ij}^{(p,p)}| \leq C\zeta^4(K)(n + \triangle_n).$$

This, together with Assumption B6 yields

$$Var(S_{n,2}) \leq \frac{C\zeta^8(K)(n + \triangle_n)}{n^2} \left(\max_{j \geq 1} \sum_{i=1}^{n} |\gamma_{ij}| \right)^2 = o(1).$$

■

Proof of Theorem 2.4. We need to show $\|V_n - V\| = o(1)$, as $n \to \infty$. By the triangle inequality,

$$\|V_n - V\| \leq \|V_n - W_n\| + \|W_n - V\|,$$

where $\|W_n - V\| = o(1)$ holds by Assumption C2 (i). To bound $\|V_n - W_n\|$ note that

$$V_n - W_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_{ik} E[\sigma(X_i)\sigma(X_k)\{v_K(X_i)v_K'(X_k) - w(X_i)w'(X_k)\}].$$

We shall establish $\|V_n - W_n\| = o(1)$ by showing that elements $(V_n - W_n)_{\ell,p}$, $1 \leq \ell, p \leq d$, of $V_n - W_n$ converges to zero. We have that

$$\begin{aligned}
|(V_n - W_n)_{\ell p}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \gamma_{ik} E[\sigma(X_i)\sigma(X_k)(v_{\ell K}(X_i)v_{pK}(X_k) - w_\ell(X_i)w_p(X_k))] \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} |\gamma_{ik}| E[|\sigma(X_i)\sigma(X_k)\{v_{\ell K}(X_i)v_{pK}(X_k) - w_\ell(X_i)w_p(X_k)\}|].
\end{aligned}$$

Notice that

$$
\begin{aligned}
&E[|\sigma(X_i)\sigma(X_k)\{v_{\ell K}(X_i)v_{pK}(X_k) - w_\ell(X_i)w_p(X_k)\}|] \\
&\quad \leq CE[|v_{\ell K}(X_i)\{v_{pK}(X_k) - w_p(X_k)\}|] + CE[|\{v_{\ell K}(X_i) - w_\ell(X_i)\}w_p(X_k)|] \\
&\quad \leq C\left(E[v_{\ell K}^2(X_i)]\right)^{1/2}\left(E[\{v_{pK}(X_k) - w_p(X_k)\}^2]\right)^{1/2} \\
&\quad +C\left(E[\{v_{\ell K}(X_i) - w_\ell(X_i)\}^2]\right)^{1/2}\left(E[w_p^2(X_k)]\right)^{1/2} = o(1),
\end{aligned}
$$

because for any $p = 1, \cdots, d$, $E[w_p^2(X_i)] < \infty$ by Assumption C1 (i), $E[\{v_{pK}(X_i) - w_p(X_i)\}^2] = o(1)$ by Assumption C1 (iii) and $E[v_{pK}^2(X_i)] < \infty$. The latter follows from

$$E[v_{pK}^2(X_i)] \leq 2E[\{v_{pK}(X_i) - w_p(X_i)\}^2] + 2E[w_p^2(X_i)] < \infty. \tag{2.9.31}$$

Hence,

$$|(V_n - W_n)_{\ell p}| \leq \left[\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^n|\gamma_{ik}|\right]\cdot o(1) = o(1),$$

by Assumption C2 (ii). This completes the proof of the Theorem. ■

Proof of Theorem 2.5. Proof of Theorem 2.5 is based on Lemmas 2.1, 2.2 and 2.3 stated in Appendix B. Define the $d \times 1$ summation

$$\hat{S}_n^*(r) := \sum_{i=1}^{[rn]}\hat{A}^{*\prime}\hat{B}_K^{-1}p^K(X_i)\hat{U}_i/\sqrt{n}, \quad 0 \leq r \leq 1,$$

where $[rn]$ denotes the integer part of rn. Based on the statement of Lemma 2.2 and 2.3, one has weak convergence $\left(\hat{S}_n^*(r)\right)_{r\in[0,1]} \Rightarrow \left(V^{1/2}\{W_d(r) - rW_d(1)\}\right)_{r\in[0,1]}$ in the space $D[0,1]^d$. Observe that $\hat{C}_n = \frac{1}{n}\sum_{m=1}^n S_n^*(\frac{m}{n})S_n^*(\frac{m}{n})' \sim \int_0^1 S_n^*(r)S_n^*(r)'dr$. Therefore, continuous mapping theorem gives

$$V^{-1/2}\hat{C}_nV^{-1/2} \Rightarrow \Psi_d. \tag{2.9.32}$$

Write

$$\hat{C}_n^{-1/2}\sqrt{n}(\hat{\theta}_n - \theta_0) = (\hat{C}_n^{-1/2}V^{1/2})\left(\sqrt{n}V^{-1/2}(\hat{\theta}_n - \theta_0)\right).$$

By Lemma 2.1-2.3, $\hat{C}_n^{-1/2}V^{1/2} \Rightarrow \Psi_d^{-1/2}$, and by Theorem 2.4, $\sqrt{n}V^{-1/2}(\hat{\theta}_n - \theta_0) \to_d N(0, I_d)$, where convergence of the two terms is joint, completing the proof. ■

2.10 Appendix B. Lemmas 2.1-2.3. Propositions 2.1-2.2.

Let $X(\cdot), Y(\cdot) \in D[0,1]$, the space of all real valued functions on $[0,1]$ that are right-continuous with finite left limits. Skorohod metric $d(\cdot,\cdot)$ in $D[0,1]$ is given by:

$$d(X,Y) = \inf_{\varepsilon>0}\{\varepsilon : \|\lambda\| \leq \varepsilon,\ \sup_{r\in[0,1]} |X(r) - Y(\lambda(r))| \leq \varepsilon\}$$

where λ is any continuous mapping of $[0,1]$ onto itself with $\lambda(0) = 0$, $\lambda(1) = 1$ and

$$\|\lambda\| = \sup_{r,u\in[0,1]:r\neq u} \Big| \log \frac{\lambda(u) - \lambda(r)}{u - r} \Big|, \quad 0 \leq r < u \leq 1.$$

Denote by

$$S_n(r) := \sum_{i=1}^{[rn]} A'P(X_i)U_i/\sqrt{n}, \quad \text{and} \quad \hat{S}_n(r) := \sum_{i=1}^{[rn]} A'P(X_i)\hat{U}_i/\sqrt{n}, \quad r \in [0,1] \qquad (2.10.1)$$

the $d \times 1$ vector-valued summations.

Note that $S_n(\cdot) \in D[0,1]^d = D[0,1] \times \cdots \times D[0,1]$, where $D[0,1]^d$ is the product space. Endowing each component space $D[0,1]$ with the well-known Skorohod metric $d(\cdot,\cdot)$, stated above, we assign the following metric to the product space $D[0,1]^p$ as was done in Phillips and Durlauf (1986). For $X(\cdot) = (X_1(\cdot),\cdots,X_d(\cdot))' \in D[0,1]^d$ and $Y(\cdot) = (Y_1(\cdot),\cdots,Y_d(\cdot))' \in D[0,1]^d$, define the metric:

$$d'(X,Y) = \max_{1\leq\ell\leq d}\{d(X_\ell,Y_\ell) : X_\ell, Y_\ell \in D[0,1]\}.$$

Lemma 2.1 states functional central limit theorem (FCLT) for $S_n(r)$ in $D[0,1]^d$ equipped with the metric $d'(\cdot,\cdot)$. The notation $\Rightarrow_{D[0,1]^d}$ signifies weak convergence of the associated probability measures in $D[0,1]^d$.

Remark. The specification of $S_n(r)$, and similarly $\hat{S}_n(r)$, in (2.10.1) as a partial summation over $i = 1$ up to $[rn]$ may seem like an obvious choice, as partial summation of random variables, $S_n(r) = \sum_{i=1}^{[rn]} \eta_i$, where η_i does not depend on n, is frequently considered in FCLT literature, see e.g. Phillpis and Durlauf (1986) and Davidson and de Jong (2000). But it is worth noting that the setting here is more involved and differs somewhat from those works. This is because the summand of $S_n(r)$ takes on a *triangular array* structure. Bearing in mind $K = K(n)$ is a function of n and recalling $A = \big(D(P_{1K}), D(P_{2K}), \cdots, D(P_{KK})\big)' \in \mathbb{R}^{K\times d}$, we see the summand $A'P(X_i)U_i/\sqrt{n}$ of (2.10.1) can be written as:

$$A'P(X_i)U_i/\sqrt{n} = \sum_{l=1}^{K(n)} D(P_{lK})P_{lK}(X_i) \sum_{j=1}^{\infty} b_{ij}\varepsilon_j/\sqrt{n}.$$

Now, consider the following representation of $S_n(r)$, as a weighted summation of ε_j's over $j=1$ to ∞, with weights that are triangular arrays:

$$\begin{aligned} S_n(r) &= \sum_{i=1}^{[rn]}\Big[\sum_{l=1}^{K(n)} D(P_{lK})P_{lK}(X_i)\sum_{j=1}^{\infty} b_{ij}\varepsilon_j/\sqrt{n}\Big] \\ &= \sum_{j=1}^{\infty}\Big[\sum_{i=1}^{[rn]}\sum_{l=1}^{K(n)} D(P_{lK})P_{lK}(X_i)b_{ij}/\sqrt{n}\Big]\cdot\varepsilon_j = \sum_{j=1}^{\infty} c_j(n;r)\varepsilon_j, \quad (2.10.2) \end{aligned}$$

where we denote

$$c_j(n;r) := \Big[\sum_{i=1}^{[rn]}\sum_{l=1}^{K(n)} D(P_{lK})P_{lK}(X_i)b_{ij}/\sqrt{n}\Big], \quad r\in[0,1], \quad n\geq 1.$$

The specification $S_n(r) = \sum_{j=1}^{\infty} c_j(n;r)\varepsilon_j$ was previously considered in Kasahara and Maejima (1986) for general functional limit theorems for infinite weighted sums. It goes without saying that the alternative representations of $S_n(r)$ given by (2.10.1) and (2.10.2) are of course equivalent. For the rest of the proof, we use the form (2.10.1) instead of (2.10.2) for ease of algebra, as using U_i instead of $\sum_{j=1}^{\infty} b_{ij}\varepsilon_j$ considerably simplifies some steps, by the use of the quantity $\gamma_{ij} = Cov(U_i, U_j)$.

In the following proofs we will need some notations. For $j\geq 1$, introduce a $j\times K$ random matrix $\mathbf{P}_j = (P(X_1),\cdots,P(X_j))'$ and $j\times 1$ random vectors, $M_j = (m(X_1),\cdots,m(X_j))'$ and $\hat{M}_j = (\hat{m}(X_1),\cdots,\hat{m}(X_j))'$.

Lemma 2.1. Under Assumptions of Theorem 2.5,

$$\Big(S_n(r)\Big)_{0\leq r\leq 1} \Rightarrow_{D[0,1]^d} \big(V^{1/2}W_d(r)\big)_{0\leq r\leq 1}. \quad (2.10.3)$$

Proof of Lemma 2.1. Lemma 2.1 states weak convergence in the d-dimensional product space $D[0,1]^d$. Phillips and Durlauf (1986, pp. 487-489) had established two sufficient conditions for weak convergence of probability measures in this multi-dimensional product space. These two conditions, adapted here for (2.10.3), are; convergence of finite dimensional distributions of $S_n(\cdot)$ to those of $V^{1/2}W_d(\cdot)$, and; tightness of each component of the vector $S_n(\cdot)$.

We first establish the following two statements, which will be subsequently used to obtain the above two facts: for any $0\leq r\leq u\leq 1$,

$$ES_n(r)S_n(u)' \to r\cdot V, \quad (2.10.4)$$

$$E|S_{n\ell}(u) - S_{n\ell}(r)|^2 \leq C\Big|\frac{[un]-[rn]}{n}\Big|, \quad \ell = 1,\cdots,d \quad (2.10.5)$$

where $S_n(r) = \big(S_{n1}(r), \cdots, S_{nd}(r)\big)'$. Write

$$ES_n(r)S_n(u)' = ES_n(r)S_n(r)' + E\big(S_n(r)(S_n(u)' - S_n(r)')\big).$$

By Theorem 2.4, $E(S_nS_n') = V_n \to V$. Therefore,

$$ES_n(r)S_n(r)' = \frac{[rn]}{n}\frac{1}{[rn]}E(A'\mathbf{P}_{[rn]}'U_{[rn]}U_{[rn]}'\mathbf{P}_{[rn]}A) \to rV.$$

Hence (2.10.4) follows if we show that $E\big(S_n(r)(S_n(u)' - S_n(r)')\big) \to 0$. This is done by showing that the corresponding limit of each element of the vector is zero. For $\ell, p = 1, \cdots, d$,

$$\begin{aligned} |E\big[S_n(r)(S_n(u)' - S_n(r)')\big]_{\ell p}| &\le \frac{C}{n}\sum_{i=1}^{[rn]}\sum_{k=[rn]+1}^{[un]}|\gamma_{ik}|E|v_{\ell K}(X_i)v_{pK}(X_k)| \\ &\le \frac{C}{n}\sum_{i=1}^{[rn]}\sum_{k=[rn]+1}^{[un]}|\gamma_{ik}| = o(1), \end{aligned}$$

by Assumption C3 (i), and because

$$E|v_{\ell K}(X_i)v_{pK}(X_k)| \le \big(Ev_{\ell K}^2(X_i)Ev_{pK}^2(X_k)\big)^{1/2} < \infty$$

as shown in the proof of Theorem 2.4. This completes the proof of (2.10.4).

To prove (2.10.5), observe that

$$\begin{aligned} E|S_{n\ell}(u) - S_{n\ell}(r)|^2 &= E\big|\frac{1}{\sqrt{n}}\sum_{i=[rn]+1}^{[un]}A_\ell'P(X_i)U_i\big|^2 \\ &\le \frac{1}{n}\sum_{i,k=[rn]+1}^{[un]}|\gamma_{ik}|E|\sigma(X_i)\sigma(X_k)v_{\ell K}(X_i)v_{\ell K}'(X_k)| \\ &\le \frac{C}{n}\sum_{i,k=[rn]+1}^{[un]}|\gamma_{ik}| \le \frac{C}{n}\sum_{i=[rn]+1}^{[un]}\big[\max_{1\le i\le n}\sum_{k=1}^{n}|\gamma_{ik}|\big] \\ &\le C\big|\frac{[un]-[rn]}{n}\big|, \end{aligned}$$

by Assumption C3 (ii), which proves (2.10.5).

Next we show that finite dimensional distributions of $S_n(\cdot)$ converge to those of $V^{1/2}W_d(\cdot)$. This states that for an arbitrary integer k, and any choices of points $r_1, \cdots, r_k$ in $[0,1]$,

$$\big(S_n(r_1), \cdots, S_n(r_k)\big) \to_d \big(V^{1/2}W_d(r_1), \cdots, V^{1/2}W_d(r_k)\big).$$

Using Cramer-Wold device, it suffices to show that for any $d \times 1$ vectors $c_1', \cdots, c_k'$, the scalar random variable

$$Q_n := \sum_{l=1}^{k} c_l' S_n(r_l) \to_d \sum_{l=1}^{k} c_l' V^{1/2} W_d(r_l) =: Q. \tag{2.10.6}$$

Write $S_n(r) = \sum_{j=1}^{\infty} w_{j,[rn]}\varepsilon_j$ with $w_{j,[rn]}$ as in (2.9.21), with $\bar{V}_n$ replaced by V. Then, $Q_n = \sum_{j=1}^{\infty} w_{jn}^* \varepsilon_j$ with $w_{jn}^* = \sum_{l=1}^{k} c_l' w_{j,[r_l n]}$. Then by (2.10.4),

$$Var(Q_n) = \sum_{j=1}^{\infty} (w_{jn}^*)^2 \to Var(Q) = \sum_{l,t=1}^{k} c_l' V c_t \cdot \min\{r_l, r_t\} < \infty.$$

By (2.9.24), which holds for all $c_l' w_{j,[r_l n]}$, $l = 1, \cdots, k$, we have $\max_{j\geq 1} |w_{jn}^*| = o(1)$, and convergence (2.10.6) follows by the same argument as in the proof of asymptotic normality (2.9.14).

Finally, we establish tightness for individual component of the vector $S_n(r)$, which completes the proof of the lemma. Noting $S_{n\ell}(\cdot) \in D[0,1]$, $\ell = 1, \cdots, d$, we verify the following sufficient condition for tightness given in Billingsley (1968, Theorem 15.6, pp.128): for any $0 \leq r \leq s \leq t \leq 1$, and some $\beta \geq 0$, $\alpha > \frac{1}{2}$ and $C > 0$,

$$E[|S_{n\ell}(s) - S_{n\ell}(r)|^{2\beta} |S_{n\ell}(t) - S_{n\ell}(s)|^{2\beta}] \leq C\big|t - r\big|^{2\alpha}, \quad \ell = 1, \cdots, d. \tag{2.10.7}$$

This is in turn derived by showing that for any $0 \leq r \leq u \leq 1$,

$$E|S_{n\ell}(u) - S_{n\ell}(r)|^4 \leq C\Big|\frac{[un] - [rn]}{n}\Big|^2. \tag{2.10.8}$$

To see (2.10.8) implies (2.10.7), note that for $\beta = 1$, the LHS of (2.10.7) is

$$\begin{aligned} E[|S_{n\ell}(s) - S_{n\ell}(r)|^2 |S_{n\ell}(t) - S_{n\ell}(s)|^2] &\leq \big\{E[|S_{n\ell}(s) - S_{n\ell}(r)|^4] E[|S_{n\ell}(t) - S_{n\ell}(s)|^4]\big\}^{1/2} \\ &\leq C\Big(\Big|\frac{[sn] - [rn]}{n}\Big|^2 \Big|\frac{[tn] - [sn]}{n}\Big|^2\Big)^{1/2} \\ &= C\Big|\frac{[sn] - [rn]}{n}\Big|\Big|\frac{[tn] - [sn]}{n}\Big| \\ &\leq C\Big|\frac{[tn] - [rn]}{n}\Big|^2, \end{aligned} \tag{2.10.9}$$

where the first step uses the Cauchy-Schwarz inequality, the second inequality follows from (2.10.8) and the last inequality from $0 \leq r \leq s \leq t \leq 1$. As explained on pp.138 of Billingsley (1968), if $t - r \geq 1/n$, then (2.10.9) implies (2.10.7) with $\alpha = 1$: since $[nt_2] \leq nt_2$ and $[nt_1] \geq nt_1 - 1$,

$$\frac{[nt_2] - [nt_1]}{n} \leq \frac{nt_2 - nt_1 + 1}{n} = t_2 - t_1 + \frac{1}{n} \leq 2(t_2 - t_1).$$

On the other hand, if $t-r<1/n$, then at least one of $[sn]-[rn]=0$ or $[tn]-[sn]=0$ holds. Then the LHS's of (2.10.7) and (2.10.9) vanish, and thus (2.10.7) holds.

To verify (2.10.8), denote by e_ℓ, a d-dimensional vector, whose ℓ^{th} element is 1 and the other elements 0. Then one can write

$$S_{n\ell}(u)-S_{n\ell}(r)=\sum_{j=1}^{\infty}e_\ell'(w_{j,[un]}-w_{j,[rn]})\varepsilon_j=:\sum_{j=1}^{\infty}\lambda_{jn}\varepsilon_j.$$

Rewriting the LHS of (2.10.8) with the new notation and noting $E(\varepsilon_j^4)=\kappa<\infty$, $\forall j$ by Assumption C5, we obtain

$$\begin{aligned}E(\sum_{j=1}^{\infty}\lambda_{jn}\varepsilon_j)^4 &= \sum_{j_1,\cdots,j_4=1}^{\infty}\lambda_{j_1n}\lambda_{j_2n}\lambda_{j_3n}\lambda_{j_4n}E(\varepsilon_{j_1}\varepsilon_{j_2}\varepsilon_{j_3}\varepsilon_{j_4})\\ &= 3[\sum_{j,j'=1:j\neq j'}^{\infty}\lambda_{jn}^2\lambda_{j'n}^2]+\kappa\sum_{j=1}^{\infty}\lambda_{jn}^4\leq C[\sum_{j=1}^{\infty}\lambda_{jn}^2]^2\\ &= C\big(E|S_{n\ell}(u)-S_{n\ell}(r)|^2\big)^2\leq C|\frac{[un]-[rn]}{n}|^2,\end{aligned}$$

where the last step follows from (2.10.5). This completes the proof of the lemma. ■

Lemma 2.2. Under Assumptions of Theorem 2.5,

$$\Big(\hat{S}_n(r)\Big)_{0\leq r\leq 1}\Rightarrow_{D[0,1]^d}\big(V^{1/2}\{W_d(r)-rW_d(1)\}\big)_{0\leq r\leq 1}. \qquad (2.10.10)$$

Proof of Lemma 2.2. Since $\hat{U}_i-U_i=m(X_i)-\hat{m}(X_i)$,

$$L_n(r):=\hat{S}_n(r)-S_n(r)=\sum_{i=1}^{[rn]}A'P(X_i)\{m(X_i)-\hat{m}(X_i)\}/\sqrt{n}.$$

We can write, using $\hat{m}(X_i)=P'(X_i)\hat{\gamma}$,

$$\begin{aligned}L_n(r) &= \sum_{i=1}^{[rn]}A'P(X_i)\{m(X_i)-P'(X_i)\gamma_K\}/\sqrt{n}+\sum_{i=1}^{[rn]}A'P(X_i)P'(X_i)(\gamma_K-\hat{\gamma})/\sqrt{n}\\ &= A'\mathbf{P}_{[rn]}'(M_{[rn]}-\mathbf{P}_{[rn]}\gamma_K)/\sqrt{n}+A'\mathbf{P}_{[rn]}'\mathbf{P}_{[rn]}(\gamma_K-\hat{\gamma})/\sqrt{n},\end{aligned}$$

leading to

$$\begin{aligned}\hat{S}_n(r) &= S_n(r)+\frac{A'\mathbf{P}_{[rn]}'(M_{[rn]}-\mathbf{P}_{[rn]}\gamma_K)}{\sqrt{n}}-\frac{A'\mathbf{P}_{[rn]}'\mathbf{P}_{[rn]}(\hat{\gamma}-\gamma_K)}{\sqrt{n}}\\ &=: S_n(r)+a_n(r)-\ell_n(r).\end{aligned}$$

We shall show that

$$\sup_{r\in[0,1]} \|a_n(r)\| = o_p(1), \tag{2.10.11}$$

$$\ell_n(r) \Rightarrow_{D[0,1]^d} rV^{1/2}W_d(1), \tag{2.10.12}$$

which, together with Lemma 2.1, prove (2.10.10).

Proof of (2.10.11). One has

$$\begin{aligned} \sup_{r\in[0,1]} \|a_n(r)\| &\leq \|A'\| \sup_{r\in[0,1]} \|\mathbf{P}'_{[rn]}\| \sup_{r\in[0,1]} \|(M_{[rn]} - \mathbf{P}_{[rn]}\gamma_K)/\sqrt{n}\| && (2.10.13) \\ &= O_p(\sqrt{n}\xi^2(K)K^{-\alpha}) && (2.10.14) \end{aligned}$$

because $\|A'\| \leq \zeta(K) \leq \xi(K)$, whereas

$$\sup_{r\in[0,1]} \|\mathbf{P}'_{[rn]}\| = O_p(\sqrt{n}\xi(K)), \quad \sup_{r\in[0,1]} \|(M_{[rn]} - \mathbf{P}_{[rn]}\gamma_K)/\sqrt{n}\| = O(K^{-\alpha}),$$

by Assumption A4. Then (2.10.11) follows by Assumption C4.

Proof of (2.10.12). Recall $Y = M + U$, and one has $\hat{\gamma} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'Y = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'(M - \mathbf{P}\gamma_K) + (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'(\mathbf{P}\gamma_K + U)$. Therefore,

$$\sqrt{n}(\hat{\gamma} - \gamma_K) = \left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} \frac{\mathbf{P}'(M - \mathbf{P}\gamma_K)}{\sqrt{n}} + \left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} \frac{\mathbf{P}'U}{\sqrt{n}}.$$

Hence,

$$\begin{aligned} \ell_n(r) &= A'\left(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n}\right)\left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} \frac{\mathbf{P}'(M - \mathbf{P}\gamma_K)}{\sqrt{n}} \\ &+ A'\left(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n}\right)\left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} \frac{\mathbf{P}'U}{\sqrt{n}} =: \ell_{1,n}(r) + \ell_{2,n}(r). && (2.10.15) \end{aligned}$$

We shall show the following two results which constitute the proof of (2.10.12):

$$\sup_{r\in[0,1]} \|\ell_{1,n}(r)\| = o_p(1), \quad \ell_{2,n}(r) \Rightarrow_{D[0,1]^d} rW_d(1).$$

Noting that $(\mathbf{P}'\mathbf{P}/n)^{-1} = O_p(1)$ and $\sup_{r\in[0,1]} \left\|\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}/n\right\| = O_p(\xi^2(K))$, since

$$\|\sum_{i=1}^{[rn]} P(X_i)P'(X_i)/n\| \leq \xi^2(K),$$

we obtain

$$\begin{aligned}\|\ell_{1,n}(r)\| &\leq \|A'\| \sup_{r\in[0,1]} \left\| \frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n} \right\| \left\| (\frac{\mathbf{P}'\mathbf{P}}{n})^{-1} \right\| \left\| \frac{\mathbf{P}'(M-\mathbf{P}\gamma_K)}{\sqrt{n}} \right\| \\ &\leq \|A'\| O_p\big(\xi^2(K)\big)\|\mathbf{P}'\| \left\| \frac{(M-\mathbf{P}\gamma_K)}{\sqrt{n}} \right\| = O_p(\sqrt{n}\xi^3(K)K^{-\alpha}) = o_p(1),\end{aligned}$$

by Assumption C4(iv). Next, write

$$\ell_{2,n}(r) = rA'\mathbf{P}'U/\sqrt{n} + A'\left(\left(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n}\right)\left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} - rI\right)\frac{\mathbf{P}'U}{\sqrt{n}}.$$

Since convergence $r(A'\mathbf{P}'U/\sqrt{n}) \to_d rV^{1/2}W_d(1)$ was shown in the proofs of Theorems 2.2 and 2.4, it remains to verify that

$$\sup_{r\in[0,1]} \|A'\left(\left(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n}\right)\left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1} - rI\right)\frac{\mathbf{P}'U}{\sqrt{n}}\| = o_p(1).$$

One has $\|A\| = O(\xi(K))$ and $\|\mathbf{P}'U/\sqrt{n}\| = O(\sqrt{K})$ by Assumption C4 (ii). Next, we have

$$\begin{aligned}&\sup_{r\in[0,1]} \left\| (\frac{[rn]}{n})(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{[rn]})(\frac{\mathbf{P}'\mathbf{P}}{n})^{-1} - rI \right\| \\ \leq\ & \sup_{r\in[0,1]} \frac{[rn]}{n} \left\| \frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{[rn]} - I \right\| \left\| (\frac{\mathbf{P}'\mathbf{P}}{n})^{-1} - I \right\| \\ +\ & \sup_{r\in[0,1]} \frac{[rn]}{n} \left\| \frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{[rn]} - I \right\| + \sup_{r\in[0,1]} \frac{[rn]}{n} \left\| (\frac{\mathbf{P}'\mathbf{P}}{n})^{-1} - I \right\| + o(1/n).\end{aligned}$$

From the proof of Theorem 2.1, (2.9.12), we have

$$\|\hat{Q} - I\|^2 = \left\| \frac{\mathbf{P}'\mathbf{P}}{n} - I \right\|^2 = O_p\left(K^2\xi^4(K)(\frac{1}{n} + \frac{\triangle_n}{n^2})\right).$$

This fact, by Horn and Johnson (1990) pp 335-336, implies

$$\left\| (\frac{\mathbf{P}'\mathbf{P}}{n})^{-1} - I \right\|^2 = O_p\left(K^2\xi^4(K)(\frac{1}{n} + \frac{\triangle_n}{n^2})\right) = O_p\left(K^2\xi^4(K)/n\right),$$

with the last step following from Assumption C4(i). Similarly, one has that

$$\begin{aligned}\sup_{r\in[0,1]} (\frac{[rn]}{n})^2 \left\| \frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{[rn]} - I \right\|^2 &= \sup_{r\in[0,1]} (\frac{[rn]}{n})^2 O_p\left(K^2\xi^4(K)(\frac{1}{[rn]} + \frac{\triangle_{[rn]}}{[rn]^2})\right) \\ &= \sup_{r\in[0,1]} \frac{[rn]}{n} O_p\left(K^2\xi^4(K)(\frac{1}{n} + \frac{\triangle_{[rn]}}{n[rn]})\right) = O_p\left(K^2\xi^4(K)/n\right), \qquad (2.10.16)\end{aligned}$$

by Assumption A4 (i). Therefore,

$$\sup_{r\in[0,1]}\|A'\left(\left(\frac{\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}}{n}\right)\left(\frac{\mathbf{P}'\mathbf{P}}{n}\right)^{-1}-rI\right)\frac{\mathbf{P}'U}{\sqrt{n}}\|$$
$$= O_p(K\sqrt{K}\xi^3(K)/\sqrt{n}) = o_p(1), \tag{2.10.17}$$

with the last step following from Assumption A3 (ii). This completes the proof of Lemma 2.2. ■

Lemma 2.3. Under assumptions of Theorem 2.5, $\sup_{r\in[0,1]}\|\hat{S}_n^*(r)-\hat{S}_n(r)\| = o_p(1)$.

Proof of Lemma 2.3. Recall that $A' = A^{*\prime}B_K^{-1/2}$, $\mathbf{p}'_{[rn]} = B_K^{1/2}\mathbf{P}'_{[rn]}$. Thus,

$$\begin{aligned}\|\hat{S}_n^*(r)-\hat{S}_n(r)\| &= \|(\hat{A}^{*\prime}\hat{B}_K^{-1}\mathbf{p}'_{[rn]}\hat{U}_{[rn]} - A'\mathbf{P}'_{[rn]}\hat{U}_{[rn]})/\sqrt{n}\| \\ &= \|(\hat{A}^{*\prime}\hat{B}_K^{-1}B_K^{1/2} - A^{*\prime}B_K^{-1/2})\mathbf{P}'_{[rn]}\hat{U}_{[rn]}/\sqrt{n}\|.\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{r\in[0,1]}\|\hat{S}_n^*(r)-\hat{S}_n(r)\| &\le \|\hat{A}^{*\prime}\hat{B}_K^{-1}B_K^{1/2} - A^{*\prime}B_K^{-1/2}\| \\ &\quad\cdot \sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}\hat{U}_{[rn]}/\sqrt{n}\| =: d_{n,1}d_{n,2}. \end{aligned}\tag{2.10.18}$$

We shall show that

$$d_{n,1} = O_p(K\xi^2(K)/\sqrt{n}) + O_p\big(\xi^2(K)\big(\sqrt{\frac{tr(\Sigma)}{n}} + K^{-\alpha}\big)\big), \tag{2.10.19}$$
$$d_{n,2} = O_p(K^{1/2} + K^{-\alpha}\sqrt{n}). \tag{2.10.20}$$

Then, since $tr(\Sigma) = O_p(K)$ by Assumption C4 (ii),

$$\begin{aligned}d_{n,1}d_{n,2} &= O_p(K\xi^2(K)/\sqrt{n} + \xi^2(K)K^{-\alpha})O_p(K^{1/2} + K^{-\alpha}\sqrt{n}) \\ &= O_p(K^{3/2}\xi^2(K)/\sqrt{n} + \xi^2(K)K^{-\alpha+1/2} + \xi^2(K)K^{-\alpha}\sqrt{n}) = o_p(1),\end{aligned}$$

by Assumption B3 (iii), C4 (iv) and B3 (ii).

$$\begin{aligned}d_{n,1} &= \|\hat{A}^{*\prime}\hat{B}_K^{-1}B_K^{1/2} - A^{*\prime}B_K^{-1/2}\| \le \|\hat{A}^{*\prime} - A^{*\prime}\|\|\hat{B}_K^{-1}B_K^{1/2} - B_K^{-1/2}\| \\ &\quad+\|A^{*\prime}\|\|\hat{B}_K^{-1}B_K^{1/2} - B_K^{-1/2}\| + \|\hat{A}^{*\prime} - A^{*\prime}\|\|B_K^{-1/2}\|.\end{aligned}$$

Note that $\|A^*\| \le \xi(K)$, and by Assumption A3 (i), $\|B_K^{-1}\| = O_p(1)$. Now,

$$\|\hat{B}_K^{-1}B_K^{1/2} - B_K^{-1/2}\| \le \|\hat{B}_K^{-1} - B_K^{-1}\|\|B_K^{1/2}\| = O_p\left(K\xi^2(K)\frac{1}{\sqrt{n}}\right), \tag{2.10.21}$$

since by Assumption C4 (iii), $\|B_K\| = O(1)$, whereas $\|\hat{B}_K - B_K\|^2 = O_p\left(K^2\xi^4(K)/n\right)$ which can be shown using the same argument shown in obtaining an order of magnitude (2.9.12) for $\|\hat{Q}-I\|$ and applying Assumption C4 (i). Then, $\|\hat{B}_K^{-1}-B_K^{-1}\|^2 = O_p\left(K^2\xi^4(K)/n\right)$ follows from Horn and Johnson (1990) pp 335-336, under Assumptions C4 (iii) and A3 (i), which imply $\|B_K\| = O(1)$ and $\|B_K^{-1}\| = O(1)$, as $n \to \infty$.

To obtain (2.10.19), it remains to evaluate the term $\|\hat{A}^* - A^*\|$. Newey (1997) showed that the estimate $\hat{A}^* = (\hat{A}_1^*, \cdots, \hat{A}_d^*)$ is equal to the quantity $\big(D(p_1;\hat{m}), \cdots, D(p_K;\hat{m})\big)'$ with probability approaching one. Recalling $D(\cdot;\hat{m}) = \big(D_1(\cdot;\hat{m}), \cdots, D_d(\cdot;\hat{m})\big)$, the i^{th} column of $\hat{A}^* - A^*$ can be written as

$$\hat{A}_i^* - A_i^* = \big(D_i(p_1;\hat{m}) - D_i(p_1;m), \cdots, D_i(p_K;\hat{m}) - D_i(p_K;m)\big)', \quad i = 1, \cdots, d.$$

Using linearity of $D_i(g;\hat{m})$ in g, one writes

$$\begin{aligned}\|\hat{A}_i^* - A_i^*\|^2 &= (\hat{A}_i^* - A_i^*)'(\hat{A}_i^* - A_i^*) = |D_i\big((\hat{A}_i^* - A_i^*)'p^K;\hat{m}\big) - D_i\big((\hat{A}_i^* - A_i^*)'p^K;m\big)| \\ &\leq C|(\hat{A}_i^* - A_i^*)'p^K|_\infty|\hat{m} - m|_\infty \leq C\|\hat{A}_i^* - A_i^*\|\xi(K)|\hat{m} - m|_\infty,\end{aligned}$$

with the first inequality following from Assumption C6. Therefore, $\|\hat{A}_i^* - A_i^*\| = O_p(\xi(K)|\hat{m} - m|_\infty)$, for $i = 1, \cdots, p$. This allows us to bound

$$\begin{aligned}\|\hat{A}^* - A^*\|^2 &\leq tr\big((\hat{A}^* - A^*)'(\hat{A}^* - A^*)\big) = \sum_{i=1}^{p}(\hat{A}_i^* - A_i^*)'(\hat{A}_i^* - A_i^*) \\ &= \big(\sum_{i=1}^{p}\|\hat{A}_i^* - A_i^*\|^2\big) \leq C\xi^2(K)|\hat{m} - m|_\infty^2.\end{aligned}$$

Therefore, applying for $|\hat{m} - m|_\infty$ the bound of Theorem 2.1, we obtain

$$\|\hat{A}^* - A^*\| = O_p\left(\xi^2(K)\left[\sqrt{\frac{tr(\Sigma_n)}{n}} + K^{-\alpha}\right]\right)) = o_p(1),$$

by Assumption B3 (ii)-(iii), completing the proof of (2.10.19).

Next, decompose $d_{n,2}$ as follows.

$$\begin{aligned}d_{n,2} &\leq \sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}(\hat{U}_{[rn]} - U_{[rn]})/\sqrt{n}\| + \sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}U_{[rn]}/\sqrt{n}\| \\ &= d_{n,21} + d_{n,22}.\end{aligned}$$

As in the proof of Lemma 2.2, one can bound

$$d_{n,21} \leq \sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}(M_{[rn]} - \mathbf{P}_{[rn]}\gamma_K)/\sqrt{n}\| + \sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}/n\|\|\sqrt{n}(\hat{\gamma} - \gamma_K)\|.$$

From (2.10.14) it is seen that the first term on the RHS is $O_p(\sqrt{n}\xi(K)K^{-\alpha}) = o_p(1)$, by Assumption C4 (iv). By (2.9.11), $\|(\hat{\gamma} - \gamma_K)\sqrt{n}\| = O_p(tr(\Sigma_n)^{1/2} + K^{-\alpha}\sqrt{n}) = O_p(K^{1/2} + K^{-\alpha}\sqrt{n})$ from Assumption C4 (ii), whereas by (2.10.16), $\sup_{r\in[0,1]}\|\mathbf{P}'_{[rn]}\mathbf{P}_{[rn]}/[rn]\| = O_p(1) + O_p(K\zeta^2(K)/\sqrt{n}) = O_p(1)$ by Assumption A3 (ii). Thus, $d_{n,21} = O_p(K^{1/2} +$

$K^{-\alpha}\sqrt{n})$.

Finally, one has

$$d_{n,22} = \sup_{r\in[0,1]} (\frac{[rn]}{n})^{1/2} \|\mathbf{P}'_{[rn]} U_{[rn]}/\sqrt{[rn]}\| = O_p(\sup_{r\in[0,1]} tr^{1/2}(\Sigma_{[rn]})) = O_p(\sqrt{K}),$$

by Assumption C4(ii). Hence, $d_{n,2} = O_p(K^{1/2})$, which proves (2.10.20) and completes the proof of the lemma. ■

In the following proposition we provide the upper bound for $\triangle_n$ in (2.3.3) in the case of Gaussian random variables X_i.

Proposition 2.1. Let $X_i \sim N(0,1), i = 1, 2, \cdots$, be Gaussian variables with $\sigma_{ij}^{(X)} = Cov(X_i, X_j)$. If for some $c_0 < 1$, one has $|\sigma_{ik}^{(X)}| \le c_0, \forall i, k = 1, 2, \cdots ; i \ne k$, then,

$$\triangle_n \le C \sum_{i,k=1,i\ne k}^{n} |\sigma_{ik}^{(X)}|, \quad n \ge 1. \tag{2.10.22}$$

Proof of Proposition 2.1. Recall that the bivariate density of $X \sim N(0,1), Y \sim N(0,1), Cov(X,Y) = \rho$ is

$$\begin{aligned} \mathrm{f}_\rho(\mathrm{x},\mathrm{y}) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\big(-\mathrm{m}_\rho(\mathrm{x},\mathrm{y})\big), \\ m_\rho(x,y) &:= \frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}, \quad x, y \in \mathbb{R}. \end{aligned}$$

Then $\mathrm{f}_0(\mathrm{x},\mathrm{y}) = \mathrm{f}(\mathrm{x})\mathrm{f}(\mathrm{y})$, where $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. We shall show that for all $|\rho| \le c_0 < 1$,

$$|\mathrm{f}_\rho(\mathrm{x},\mathrm{y}) - \mathrm{f}_0(\mathrm{x},\mathrm{y})| \le \mathrm{C}\rho \exp\big(\frac{-(\mathrm{x}^2+\mathrm{y}^2)}{8}\big), \ \mathrm{x},\mathrm{y} \in \mathbb{R}, \tag{2.10.23}$$

where C does not depend on ρ. Since $f_{ik}(x,y) = \mathrm{f}_{\sigma_{\mathrm{ik}}}(\mathrm{x},\mathrm{y})$ is the bivariate density of X_i, X_k, the following holds by (2.10.23),

$$\begin{aligned} \triangle_n &= \sum_{i,k=1,i\ne k}^{n} \int |f_{ij}(x,y) - f(x)f(y)|dxdy \\ &\le C \sum_{i,k=1,i\ne k}^{n} |\sigma_{ik}^{(X)}| \int \exp\big(-(x^2+y^2)/8\big)dxdy \\ &\le C \sum_{i,k=1,i\ne k}^{n} |\sigma_{ik}^{(X)}|, \end{aligned}$$

which proves (2.10.22).

Proof of (2.10.23) By the mean value theorem, applied in $|\rho| \le c_0$,

$$|\mathrm{f}_\rho(\mathrm{x},\mathrm{y}) - \mathrm{f}_0(\mathrm{x},\mathrm{y})| \le |\rho| \sup_{|\rho|\le \mathrm{c}_0} |\mathrm{f}'_\rho(\mathrm{x},\mathrm{y})|. \tag{2.10.24}$$

Note that

$$\mathrm{f}'_\rho(\mathrm{x},\mathrm{y}) = \mathrm{f}_\rho(\mathrm{x},\mathrm{y})\left(\frac{\rho}{1-\rho^2} - \frac{\partial \mathrm{m}_\rho(\mathrm{x},\mathrm{y})}{\partial\rho}\right). \tag{2.10.25}$$

One has

$$\left|\frac{\rho}{1-\rho^2}\right| \leq \frac{c_0}{1-c_0^2}.$$

We shall show that

$$\mathrm{f}_\rho(\mathrm{x},\mathrm{y}) \leq \mathrm{c}\exp\big(-(\mathrm{x}^2+\mathrm{y}^2)/4\big) \tag{2.10.26}$$

$$\left|\frac{\partial m_\rho(x,y)}{\partial\rho}\right| \leq c(x^2+y^2),\ x,y\in\mathbb{R}, \tag{2.10.27}$$

where c does not depend on ρ and x, y, which together with (2.10.24) and (2.10.25) implies (2.10.23).

Note that

$$\begin{aligned} m_\rho(x,y) &\geq \frac{x^2+y^2-2|\rho xy|}{2(1-|\rho|^2)} \\ &= \frac{|\rho|(x^2+y^2-2|xy|)+(1-|\rho|)(x^2+y^2)}{2(1-|\rho|^2)} \\ &\geq \frac{(1-|\rho|)(x^2+y^2)}{2(1-|\rho|^2)} \geq \frac{x^2+y^2}{2(1+|\rho|)} \geq \frac{x^2+y^2}{4}, \end{aligned}$$

the second inequality following from $2|xy| \leq x^2+y^2$. This implies (2.10.26):

$$\begin{aligned} f_\rho(x,y) &= \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\big(-m_\rho(x,y)\big) \\ &\leq \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\big(-(x^2+y^2)/4\big) \\ &\leq \frac{1}{2\pi\sqrt{1-c_0^2}}\exp\big(-(x^2+y^2)/4\big). \end{aligned}$$

Next,

$$\begin{aligned} \left|\frac{\partial m_\rho(x,y)}{\partial\rho}\right| &= \left|\frac{-4(1-\rho^2)xy+4\rho(x^2+y^2-2\rho xy)}{[2(1-\rho^2)]^2}\right| \\ &\leq \frac{|xy|}{(1-\rho^2)} + \frac{|x^2+y^2-2\rho xy|}{[(1-\rho^2)]^2} \\ &\leq \frac{|xy|}{(1-c_0^2)} + \frac{x^2+y^2+2|\rho xy|}{(1-c_0^2)^2} \leq c(x^2+y^2), \end{aligned}$$

since $2|xy| \leq x^2+y^2$, which proves (2.10.27) and completes the proof of the proposition. ■

Proposition 2.2 Assume that there exists $\eta(j) \geq 0$, $j \in \mathbb{Z}$ such that $\sum_{j=-\infty}^{\infty} \eta(j) < \infty$

and $|\gamma_{ikn}| \leq \eta(i-k), \forall i, k = 1, 2, \cdots$. Then for any $r \in [0, 1]$,

$$\sum_{i=1}^{[rn]} \sum_{k=[rn]+1}^{n} |\gamma_{ikn}| = o(n).$$

Proof of Proposition 2.2. Note that $\tau_n := \sum_{|j| \geq \log n} \eta(j) \to 0$ as $n \to \infty$, and $\max_j \eta(j) \leq C < \infty$. One has

$$\begin{aligned}
&\sum_{i=1}^{[rn]} \sum_{k=[rn]+1}^{n} |\gamma_{ikn}| \leq \sum_{i=1}^{[rn]} \sum_{k=[rn]+1}^{n} \eta(i-k) \leq \sum_{i=1}^{[rn]} \sum_{k=[rn]+\log n}^{n} \eta(i-k) \\
&+ \sum_{k=[rn]+1}^{n} \sum_{i=1}^{[rn]-\log n} \eta(i-k) \quad + \quad C \sum_{i=[rn]-\log n}^{[rn]} \sum_{k=[rn]+1}^{[rn]+\log n} 1 \\
&\leq \tau_n \sum_{i=1}^{[rn]} 1 + \tau_n \sum_{k=[rn]+1}^{n} 1 + 2C \log n \leq 2\tau_n n + 2C \log n = o(n).
\end{aligned}$$

This completes the proof of the proposition. ■

3 Panel Data Model with Non-parametric Common Regression and Individual Fixed Effects

3.1 Introduction

Availability of multiple observations on a set of individuals over time, i.e. panel data, allows economists to account for unobserved individual effects, which is not possible when dealing with observations of a single cross section. There is a substantial amount of literature on estimation of linear or non-linear panel data models with individual effects, see e.g. Arellano and Honore (2001), Hahn and Kuersteiner (2002), Wooldridge (2002) and Bai (2009).

Non-parametric methods that enable consistent estimation of functions without the danger of parametric misspecification are becoming increasingly more accepted, at least in samples of moderate size. Ruckstuhl, Welsh and Carroll (2000) considered kernel estimation of regression function with panel data when the cross-sectional size, N, is fixed and there are additive individual effects that are "random", i.e. uncorrelated with regressors. Additive individual components represent the effects of unobserved time-invariant individual characteristics on the variable of interest and are often viewed as a simple yet satisfactory way of modeling individual heterogeneity in panel data. In many economic applications, it is difficult to justify the assumption of "random" effects as the unobserved individual characteristics may be correlated with the regressors. Henderson et al. (2008) consider consistent estimation of non-parametric and semi-parametric (partly linear) regression functions when additive individual "fixed" effects, that may be correlated with regressors, are present.

In this chapter, we consider a panel data model with additive individual components and time-varying "common" regressor, that is shared by all cross-sectional units, for a dataset whose time dimension, T, is large relative to its cross-sectional size, N. The type of data envisaged is when the cross-sectional units are large entities such as countries/regions or firms. It is expected that such datasets typically exhibit cross-sectional dependence in the error terms: countries or firms may be interdependent or subject to global shocks that affect everyone. Such dependence could be substantial, and it is deemed crucial in this work not to impose stringent restrictions on the strength of cross-sectional dependence.

We consider the following model for a balanced panel data set of size $N \times T$. Below Y_{it} denotes a one dimensional dependent variable, λ_i an additive individual fixed effect of individual i, Z_t is a q-dimensional vector of time-varying common regressors, common to individuals, whereas $m(\cdot)$ is the non-parametric regression function of interest, and U_{it} the error term. Y_{it} is defined as follows:

$$Y_{it} = \lambda_i + m(Z_t) + U_{it}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T. \tag{3.1.1}$$

The model can be written in N-dimensional vector form as

$$Y_{\cdot t} = \lambda + m(Z_t) 1_N + U_{\cdot t}, \quad t = 1, \cdots, T,$$

setting: $Y_{\cdot t} = (Y_{1t}, \cdots, Y_{Nt})', \lambda = (\lambda_1, \cdots, \lambda_N)', 1_N = (1, \cdots, 1)', U_{\cdot t} = (U_{1t}, \cdots, U_{Nt})'$. The cross-sectional size $N (= N_T)$ is assumed to be either fixed or increasing slowly as $T \to \infty$.

The model above was considered in Robinson (2010b) for common trend estimation, with Z_t replaced by the deterministic argument t/T . He showed how to incorporate the knowledge of cross-sectional dependence in U_{it}'s into estimating $m(t/T)$ in order to obtain an efficiency gain. In particular, a generalised least squares (GLS) type estimate under the full knowledge of cross-sectional dependence was shown to be superior in the mean square error (MSE) sense, to one that does not incorporate such information. Asymptotic equivalence between the infeasible and feasible GLS type estimates was also established.

The present chapter aims to address similar issues in the case of multivariate stochastic regressors. The random nature of the regressors, as opposed to the deterministic t/T of Robinson (2010b), gives rise to the possibility of conditional heteroscedasticity as we do not assume independence between the error term and regressors. In our setting, we allow the cross-sectional covariance matrix of the error terms to depend on the value of the concurrent regressors, which leads to the use of "local" weights in the GLS-type estimation, as opposed to the global weight used in the trend estimation in Robinson (2010b). We further relax conditions on U_{it} (and Z_t) from being independent and identically distributed (i.i.d.) across time to possible weak dependence.

Our interest in the model (3.1.1) can be more broadly motivated from a more general and applicable model where time-varying, individual-specific regressors are also present. For example, one may want to model house price indices of countries within the Eurozone, Y_{it}, in terms of the interest rate set by the European Central Bank, Z_t, and country-specific covariates X_{it} such as the country's GDP, inflation and stock market index. One could formulate for Y_{it} a partly linear regression specification:

$$Y_{it} = \lambda_i + X_{it}' \gamma + m(Z_t) + U_{it}. \tag{3.1.2}$$

The estimation of the linear parameter γ could be faciliated by the following data transformation that involves differencing across the cross section and time. Noting that $Y_{it} - Y_{i-1t} = \lambda_i - \lambda_1 + (X_{it} - X_{i-1t})'\gamma + U_{it} - U_{i-1t}, \quad i = 2, \cdots, N$, consider:

$$(Y_{it} - Y_{i-1t}) - (Y_{it-1} - Y_{i-1t-1}) = [(X_{it} - X_{i-1t}) - (X_{it-1} - X_{i-1t-1})]'\gamma \\ +(U_{it} - U_{i-1t}) - (U_{it-1} - U_{i-1t-1}), \quad t = 2, \cdots, T.$$

Based on the transformed data above, the linear parameter γ can be estimated at the parametric $\sqrt{NT}$ rate under the same conditions as those required for first difference estimation of the linear regression model with additive individual effects, see e.g. Wooldridge (2002, pp. 279-281) and Arellano et al. (2001, pp. 3233-3241). Therefore, in considering non-parametric estimation of $m(\cdot)$ in (3.1.2), one could treat γ as known and focus on the simpler model (3.1.1), instead of (3.1.2), noting that γ can be consistently estimated at a faster rate of convergence than $m(\cdot)$ under suitable conditions.

The plan of the chapter is as follows. Section 3.2 introduces the simple kernel estimate of the function $m(\cdot)$ and presents its asymptotic MSE, the consequent optimal choice of the bandwidth parameter, and establishes its asymptotic normality. Section 3.3 discusses improved estimation based on the unknown cross-sectional covariance matrix of the error terms, stating asymptotic results on the behaviour of the improved estimate. Estimates of the cross-sectional covariance matrix are considered in Section 3.4, with asymptotic justification for their use in deriving the optimal bandwidths and improved trend estimates. Section 3.5 presents a small Monte Carlo study of finite sample performance. Appendix A contains some useful lemmas, of which Lemma 3.6 constitutes an additional contribution of this work in offering a useful decomposition of U statistic of order up to 4, under serial dependence in its arguments. Proofs of theorems are provided in Appendix B.

3.2 Simple non-parametric regression estimation

In (3.1.1), it is notable that λ_i and $m(\cdot)$ are only identified up to a location shift. As noted in Robinson (2010b), an (arbitrary) identification restriction $\sum_{i=1}^{N} \lambda_i = 0$ identifies the function $m(\cdot)$ up to a vertical shift and leads to the relationship:

$$\bar{Y}_{A_t} = m(Z_t) + \bar{U}_{At}, \tag{3.2.1}$$

where we denote by $\bar{Y}_{At} := \sum_{i=1}^{N} Y_{it}/N$ and $\bar{U}_{At} := \sum_{i=1}^{N} U_{it}/N$, the cross-sectional averages. Under (3.2.1), one can non-parametrically estimate $m(\cdot)$ using the time series data $(\bar{Y}_{At}, Z_t')$. In this section, we derive the asymptotic MSE of the simple Nadaraya-Watson (N-W) estimator of $m(\cdot)$, based on (3.2.1), at a fixed point ζ, and consider optimal bandwidth choice. A temporal dependence condition on Z_t and $U_{it}, i \geq 1$ will be phrased in terms of their α-mixing coefficients. A multivariate CLT is also presented at d fixed points, $(\zeta_1, \zeta_2, \cdots, \zeta_d)$, which is nothing new in itself and included here for the sake of completeness. The main reference is Robinson (1983), with conditions amended to suit the setting under consideration.

For ease of algebra, we consider product kernels with a diagonal bandwidth matrix, which could cover the general case of different kernel function and/or bandwidth choice in each dimension of regressors. For ease of algebra and notation we will use the same kernel function and bandwidth for each element of Z_t, the relaxation of which is straight-forward.

Let $a_T = a$ be a positive scalar bandwidth parameter approaching 0 as T increases. Below, the subscript T will be suppressed for brevity. For a given real-valued bounded kernel function $k : \mathbb{R} \to \mathbb{R}$, the product kernel $K : \mathbb{R}^q \to \mathbb{R}$ is defined as

$$K(u) = \prod_{j=1}^{q} k(u_j), \quad u = (u_1, u_2, \cdots, u_q)'.$$

Definition 1. *The simple N-W non-parametric regression estimate of $m(\zeta)$ at a fixed point ζ which uses the cross sectional average $\bar{Y}_{At}$ is defined as*

$$\tilde{m}(\zeta) := \frac{\sum_{t=1}^{T} K\Big(\frac{Z_t - \zeta}{a}\Big)\bar{Y}_{At}}{\sum_{t=1}^{T} K\Big(\frac{Z_t - \zeta}{a}\Big)}.$$

Definition 2. (α-mixing) *Let $\mathcal{M}_u^v$ be the σ-field of events generated by a stationary vector process $X_t, u \le t \le v$. Then the α-mixing coefficient of X_t is defined as*

$$\alpha(\tau) := \sup_{t \in \mathbb{N}} \sup_{A \in \mathcal{M}_{-\infty}^{t}, B \in \mathcal{M}_{t+\tau}^{\infty}} |P(A \cap B) - P(A)P(B)|, \quad \tau > 0.$$

A stationary process X_t is called α-mixing if $\alpha(\tau) \to 0$ as $\tau \to \infty$.

Definition 3. *$\mathcal{K}_\ell$, $\ell \ge 1$, denotes the class of uniformly bounded even functions $k : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\int k(u)du = 1, \quad \int u^i k(u)du = 0, \quad i = 1, \cdots, \ell - 1, \quad \chi_\ell := \int_{-\infty}^{\infty} u^\ell |k(u)| du < \infty, \tag{3.2.2}$$

and such that $\sup_u (1 + |u|^{\ell+1})|k(u)| < \infty$.

Assumption 1. *For all $i \ge 1$, (Z_t, U_{it}) is a jointly stationary α-mixing processes with mixing coefficient $\alpha_i(\tau)$. Define $\alpha(\tau) := \max_i \alpha_i(\tau)$. For some $\theta > 2$,*

$$\sum_{\tau=L}^{\infty} \alpha^{1-2/\theta}(\tau) = o(L^{-1}), \quad as \quad L \to \infty.$$

Assumption 2. *The process $\{U_{it}\}$ is such that for all $i = 1, \cdots, N, \quad N \ge 1, \quad t \ge$*

1, $E(U_{it}) = 0$ and $E(U_{it}|Z_t) = 0$.

Let $(\zeta_1, \cdots, \zeta_d)$ be the set of d points in $\mathbb{R}^q$ where $m(\cdot)$ is estimated.

Assumption 3. *The process $Z_t, t \geq 1$, is stationary and has probability density function (pdf) f which is continuous and bounded. Moreover, $f(\zeta_l) > 0, \quad l = 1, \cdots, d$.*

Assumption 4. *The functions $f(\cdot)$ and $m(\cdot)$ have bounded derivatives of total order s at $z = \zeta_l$, $l = 1, \cdots, d$.*

Assumption 5. *The conditional expectation functions $\omega_{ij}(z) = E(U_{it}U_{jt}|Z_t = z), \quad i, j = 1, \cdots, N$ are bounded continuous functions of z. We denote the $N \times N$ conditional covariance matrix of $U_{\cdot t}$ by $\mathbf{\Omega}(z) = \mathbf{\Omega}(Z_t = z) = \{\omega_{ij}(z)\}$.*

The N-subscript is suppressed but it is recalled that N may be increasing with T.

Assumption 6. *$k(u)$ is a uniformly bounded and even function that belongs to $\mathcal{K}_s$.*

Assumption 6, combined with Assumption 4, leads to a bias reduction for the N-W estimation arising from the use of higher order kernels, exploiting the assumed smoothness of unknown functions m and f.

Assumption 7. *The bandwidth $a = a_T \to 0$ is such that $Ta^q \to \infty$ as $T \to \infty$.*

The randomness of the denominator in the non-parametric regression estimator $\tilde{m}(\zeta_l)$ gives rise to difficulty in obtaining the exact expression for its MSE and instead, as is conventional, we consider an "approximate" MSE. Below, we present an approximate bias expression, of which the detailed derivation for the scalar Z_t case can be found in p. 97-102 of Pagan and Ullah (1999), then combine this with the asymptotic variance expression from the CLT result in Theorem 3.3, in order to formulate the approximate MSE for $\tilde{m}(\zeta_l)$, presented in Theorem 3.1.

Denote the approximate bias expression for $\tilde{m}(\zeta)$ when using an s^{th} order kernel, by

$$Bias_s(\tilde{m}(\zeta)) := \frac{\chi_s a^s}{f(\zeta)} \Phi(\tilde{m}(\zeta)),$$

where

$$\Phi(\tilde{m}(\zeta)) = \sum_{j=1}^{q} \sum_{\ell=1}^{s} \frac{1}{\ell!} \frac{1}{s-\ell!} \frac{\partial^{(\ell)} m}{\partial z_j^{\ell}}\Big|_{z=\zeta} \frac{\partial^{(s-\ell)} f}{\partial z_j^{s-\ell}}\Big|_{z=\zeta}.$$

Theorems 3.1-3.3 are merely restatements of standard results and proofs are not given here. Let $\kappa = \int_{-\infty}^{\infty} k^2(u)du$ and χ_2 be as in (3.2.2). In the rest of the chapter, when we say "asymptotically", we mean "as T, and possibly $N = N_T$, go to ∞".

Theorem 3.1. *Under Assumptions 1-7, asymptotically, the approximate MSE is*

$$MSE\Big(\tilde{m}(\zeta)\Big) \sim \frac{\kappa^q}{Ta^q f(\zeta)} \frac{1_N' \mathbf{\Omega}(\zeta_l) 1_N}{N^2} + Bias_s^2(\tilde{m}(\zeta_l)). \tag{3.2.3}$$

The first term on the right hand side (RHS) represents the variance contribution, reflecting the variance of the simple cross-sectional average $\bar{U}_{At}$: $Var(\bar{U}_{At}) = 1_N' \mathbf{\Omega}(\zeta_l) 1_N / N^2$.

Theorem 3.2. *Under Assumptions 1-7, the bandwidth minimising the approximate*

MSE (3.2.3) is

$$a^*_{m,AMSE(\zeta_l)} = \Big(\frac{\kappa^q f(\zeta_l)}{T\chi_s^2\Phi(\tilde{m}(\zeta_l))^2}\frac{1_N'\mathbf{\Omega}(\zeta_l)1_N}{N^2}\Big)^{\frac{1}{q+2s}}.$$

Now, let $f_j(z,u)$ denote the joint pdf of (Z_t, Z_{t+j}) and $f_{j,k}(z,u,w)$ denote the joint pdf of $(Z_t, Z_{t+j}, Z_{t+j+k})$.

Assumption 8. *(i) For some $\xi > 0$, $\sup_z \|z\|^\xi f(z) < \infty$. (ii) For some $C < \infty$,*

$$\sup_{z,u} f_j(z,u) \leq C \quad \forall j \geq 1 \quad \text{and} \quad \sup_{z,u,w} f_{jk}(z,u,w) \leq C, \quad \forall j,k \geq 1.$$

Assumption 8 (i) bounds the joint densities of Z_t's and is natural given that Assumption 3 assumes boundedness of the marginal density. Assumption 8 (i) is from Hansen (2008) and is later needed to obtain a uniform rate of convergence.

Assumption 9. *For $\theta > 2$ of Assumption 1, $E|m(Z_t)|^\theta < \infty$ and $E|U_{it}|^\theta \leq C < \infty$, for $i = 1, \cdots, N$, $\quad N \geq 1, t \geq 1$.*

Assumption 10. *For some $c > \theta$ and for $z = \zeta_l$, $l = 1, \cdots, d$, $E(|U_{it}|^c|Z_t = z)$ is finite and has bounded derivatives of order s.*

Assumptions 9 and 10 are both from Robinson (1983) and are additional assumptions required for the asymptotic normality result below. For a symmetric positive definite matrix A, let $A^{1/2}$ denote its unique matrix square root. Below, we present asymptotic normality result of $\tilde{m}(\zeta_1)$ when the bandwidth parameter a is set so that the bias term is negligible compared to the variance component.

Theorem 3.3. *Let the bandwidth a be such that $Ta^{q+2s} \to 0$ as $T \to \infty$. Then under Assumptions 1-10, asymptotically,*

$$(Ta^q)^{\frac{1}{2}}\mathbf{V}_N^{-1/2}\Big(\tilde{m}(\zeta_1) - m(\zeta_1), \cdots, \tilde{m}(\zeta_d) - m(\zeta_d)\Big)' \xrightarrow{d} N_d(\mathbf{0}, \mathbf{I_d}),$$

where $\mathbf{V}_N$ is a $d \times d$ diagonal matrix whose $(l,l)^{th}$ element is $\kappa^q 1_N'\mathbf{\Omega}(\zeta_l)1_N/N^2 f(\zeta_l)$.

The quantity $1_N'\mathbf{\Omega}(\zeta_l)1_N/N^2 = \sum_{i,j}^N \omega_{i,j}(\zeta_l)/N^2$ reflects the strength of cross-sectional dependence in the error terms. In the case of increasing N, $1_N'\mathbf{\Omega}(\zeta_l)1_N/N^2 = O(N^{-1})$ is analogous to a common weak dependence assumption in time series. We are only requiring $1_N'\mathbf{\Omega}(\zeta_l)1_N/N^2 = O(1)$ in Assumption 5, therefore allowing the possibility of what is analogous to strong or long-range dependence in time series. On the other hand, since $1_N'\mathbf{\Omega}(\zeta_l)1_N/N^2$ may be of order $o(1)$, the rate of convergence of the N-W estimator is affected by the strength of cross sectional dependence if $N \to \infty$.

3.3 Improved estimation

This section considers improvement in the efficiency of the common regression estimation in a similar way to Robinson (2010b), taking into account possible conditional heteroscedasticity. Recall that the identifying condition of the non-parametric regression function m in (3.1.1) was $1_N'\lambda = 0$, leading to $\bar{Y}_{At} = m(Z_t) + \bar{U}_{At}$, where the $N \times 1$ weight vector used in $\bar{Y}_{At}$ was $1_N/N$. Replacing $1_N/N$ with an alternative weight vector gives rise to a different identification restriction. The following representation of $Y_{\cdot t}$ corresponds to a general $N \times 1$ weight vector w:

$$Y_{\cdot t} = \lambda^{(w)} + m^{(w)}(Z_t)1_N + U_{\cdot t}, \qquad w'Y_{\cdot t} = m^{(w)}(Z_t) + w'U_{\cdot t},$$

where the identifying restriction is given by $w'\lambda^{(w)} = 0$. There is a vertical shift between the functions identified under $w'\lambda^{(w)} = 0$ and $1_N'\lambda = 0$, namely, $m^{(w)}(z) - m(z) = w'\lambda$ for all z.

One may improve the efficiency of estimation of m by using an optimal weight vector. Note $Var(w'U_{\cdot t}|Z_t = z) = w'\mathbf{\Omega}(z)w$, which enters into the variance of the N-W estimator as a scale factor. Therefore, the following weight vector would give rise to the minimum variance N-W estimate out of all estimates formed this way:

$$w^*(z) = \text{argmin}_w w'\mathbf{\Omega}(z)w = (1_N'\mathbf{\Omega}(z)^{-1}1_N)^{-1}\mathbf{\Omega}(z)^{-1}1_N. \tag{3.3.1}$$

Hence, we define the optimal N-W estimator at z as:

$$\tilde{m}^*(z) := \frac{\sum_{t=1}^{T} K\Big(\frac{Z_t - z}{a}\Big) w^*(z)'Y_{\cdot t}}{\sum_{t=1}^{T} K\Big(\frac{Z_t - z}{a}\Big)}, \tag{3.3.2}$$

where $w^*(z)$ is the optimal weight vector that minimises the conditional variance of the weighted average of $w'Y_{\cdot t}$ when $Z_t = z$:

Assumption 11. *The matrix $\mathbf{\Omega}(z)$ is nonsingular at $z = \zeta_l$, $l = 1, \cdots, d$.*

Conditional heteroscedasticity implies the optimal weight vector varies across the point of estimation, leading to the additional caveat that values of the regression function $m^*(w^*)$ identified at different points have vertical differences between them. The regression function identified under $w^*(z)$ has the following vertical shift from that identified in (3.1.1):

$$m^*(z) - m(z) = w^*(z)'\lambda. \tag{3.3.3}$$

Therefore, in the improved estimation, for the sake of comparability between points of estimation, one should first carry out the optimal N-W estimation at each point of interest, then adjust back to the baseline by using an estimate of the additive fixed effect λ. One can use the following $\sqrt{T}$-consistent estimate of λ to do this in light of

(3.3.3), where $\bar{Y}_{iA} := \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ and $\bar{Y}_{AA} := \frac{1}{N}\sum_{i=1}^{N} \bar{Y}_{iA}$:

$$\begin{aligned}\hat{\lambda}_i := \bar{Y}_{iA} - \bar{Y}_{AA} &= \lambda_i + \frac{1}{T}\sum_{t=1}^{T} m(Z_t) + \frac{1}{T}\sum_{t=1}^{T} U_{it} - \left(\frac{1}{T}\sum_{t=1}^{T} m(Z_t) + \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it}\right)\\ &= \lambda_i + \frac{1}{T}\sum_{t=1}^{T} U_{it} - \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} U_{it} = \lambda_i + O_p\left(\frac{1}{\sqrt{T}}\right),\end{aligned}$$

with the last step following from Assumption 1 and 5.

Theorem 3.4. *Under Assumptions 1-7 and 11, the approximate MSE is given by*

$$MSE\Big(\tilde{m}^*(\zeta_l)\Big) \sim \frac{\kappa^q}{Ta^q f(\zeta_l)}\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1} + Bias_s^2(\tilde{m}(\zeta_l)).$$

Note that the bias is the same as in Theorem 3.1 of the previous section.

Theorem 3.5. *Under Assumptions 1-7 and 11, the bandwidth minimising the approximate MSE of $\tilde{m}^*(\zeta_l)$ is given by*

$$a^*_{m^*AMSE(\zeta_l)} = \Big(\frac{\kappa^q f(\zeta_l)}{T\chi_s^2\Phi(\tilde{m}(\zeta_l))^2}\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1}\Big)^{\frac{1}{q+2s}}.$$

Theorem 3.6. *Let a be such that $Ta^{q+2s} \to 0$ as $T \to \infty$. Then under Assumptions 1-11, asymptotically,*

$$(Ta^q)^{\frac{1}{2}}\mathbf{V}^{*-1/2}_N\Big(\tilde{m}^*(\zeta_1) - m^*(\zeta_1), \cdots, \tilde{m}^*(\zeta_d) - m^*(\zeta_d)\Big)' \to_d N_d(0, I_d),$$

where $m^*(\zeta_l) = m(\zeta_l) + w^*(\zeta_l)'\lambda$ with m and λ from (3.1.1) and $\mathbf{V}^*_N$ is a $d \times d$ diagonal matrix whose $(l,l)^{th}$ element is $\kappa^q\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1}/f(\zeta_l)$.

The result shows that the rate of convergence depends on the rate of decay (if any) of $\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1}$ as $N \to \infty$. Hence in the case of $N \to \infty$, the improved estimator may have a faster rate than that of the simple N-W estimator if the rate of decay of $\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1}$ is faster than that of $\big(1_N'\mathbf{\Omega}(\zeta_l)1_N\big)/N^2$. It was shown in Robinson (2010b) that $\big(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N\big)^{-1} < \big(1_N'\mathbf{\Omega}(\zeta_l)1_N\big)/N^2$, unless $\mathbf{\Omega}(\zeta_l)$ has an eigenvector 1_N. For a discussion of when such situation would arise in the context of the familiar factor models or spatial autoregressive models, see Section 4 of Robinson (2010b).

3.4 Feasible estimator

We need now to consider feasibility of such estimation and efficiency gain in the absence of knowledge of $\mathbf{\Omega}(\zeta_l)$. As is done in the GLS framework, it is natural to form a feasible version of $\tilde{m}^*(\zeta_l)$ by replacing $\mathbf{\Omega}(\zeta_l)$ with a consistent estimator. However, consistency may not be in itself satisfactory if the estimating error of $\mathbf{\Omega}(\zeta_l)$ does not vanish fast enough and outweigh the efficiency gain desired. Theorem 3.9 below will provide additional conditions for asymptotic negligence of the difference between the infeasible and feasible N-W estimators. We first need to establish how good an estimator of $\mathbf{\Omega}(\zeta_l)$ we have, see Theorem 3.7. The proof of that theorem involves finding the stochastic order of some quantities taking the form of U-statistics whose arguments are observations from a time series process with dependence across time. To do this, we use results of Yoshihara (1976) whose conditions involve β-mixing coefficients. Therefore some assumptions in this section will be phrased in terms of the β-mixing coefficients of Z_t and U_{it}, $i = 1, 2, \cdots$.

Definition 4. **(β-mixing)** *Let $\mathcal{M}_u^v$ be the σ-field of events generated by the vector process $X_t, u \leq t \leq v$. Then the β-mixing coefficient of X_t is defined as*

$$\beta(\tau) := \sup_{t \in \mathbb{N}} \sup_{A \in \mathcal{M}_{-\infty}^t, B \in \mathcal{M}_{t+\tau}^\infty} |P(A|\mathcal{M}_{t+\tau}^\infty) - P(A)|. \tag{3.4.1}$$

X_t is called β-mixing if $\beta(\tau) \to 0$ as $\tau \to \infty$.

In addition, to derive the uniform rate of convergence result for the non-parametric density estimator, we will use α-mixing conditions of Hansen (2008). It can be shown that $\beta(\tau) \leq \alpha(\tau)$ where $\alpha(\tau)$ is the α-mixing coefficient. Hence β-mixing condition is more restrictive than α-mixing condition, but a number of interesting processes have been shown to be β-mixing. Volkonskii and Rozanov (1961) showed if Gaussian process X_t has spectrum with j-th derivative of bounded variation, then $\beta(\tau) = O(\tau^v)$, with $v = j - 1$. Pham and Tran (1985) established that for (vector-valued) linear process $X_t = \sum_{j=0}^{\infty} b_j \varepsilon_{t-j}$, where ε_t are *i.i.d.* and b_j's are fixed weights, one has $\beta(\tau) = O\big(\sum_{k=\tau}^{\infty}(\sum_{l=k}^{\infty} \|b_j\|_E)^{\delta/(1+\delta)}\big)$, where $\delta > 0$, $\|\cdot\|_E$ denotes Euclidean norm, and ε_t satisfies a δ-th moment condition and has a pdf satisfying certain conditions. Pham (1986) showed that some random coefficient autoregressive and bilinear processes are β-mixing with $\beta(\tau) = O(\tau^v)$, for any $v > 0$.

To estimate $\mathbf{\Omega}(\zeta)$, we use the following fitted residual:

$$\begin{aligned} \hat{U}_{it} &:= Y_{it} - \bar{Y}_{iA} - \tilde{m}(Z_t) + \bar{Y}_{AA} && (3.4.2) \\ &= (\lambda_i + m(Z_t) + U_{it}) - (\lambda_i + \frac{1}{T}\sum_{t=1}^{T} m(Z_t) + \bar{U}_{iA}) - \tilde{m}(Z_t) + (\frac{1}{T}\sum_{t=1}^{T} m(Z_t) + \bar{U}_{AA}) \\ &= U_{it} - \bar{U}_{iA} - \tilde{m}(Z_t) + m(Z_t) + \bar{U}_{AA}. && (3.4.3) \end{aligned}$$

As will be made clear later, there is a need for a different bandwidth to be used in the preliminary stage regression: we will denote that bandwidth h.

No parametric structure on the conditional heteroscedasticity $\mathbf{\Omega}(\zeta)$ is pre-imposed.

We will use the kernel local smoothing estimate of $\mathbf{\Omega}(\zeta_l)$:

$$\hat{\mathbf{\Omega}}(\zeta_l) = \frac{\sum_{t=1}^{T} K_l(Z_t; h)\hat{U}_{\cdot t}\hat{U}'_{\cdot t}}{\sum_{t=1}^{T} K_l(Z_t; h)}, \tag{3.4.4}$$

where $K_l(z; h) = K^*\left((z - \zeta_l)/h\right)$, for suitable kernel function K^*, with the $*$ superscript to stress that the kernel function used in the estimation of $\mathbf{\Omega}$ need not be the same as one used in the N-W estimate of $m(\cdot)$. An additional caveat involved in the local smoothing of $\Omega(\cdot)$, namely that we need to investigate the behavior of $K_l(Z_t; h)/f_t$, where $f_t := f(Z_t)$. The function $1/f(z)$ is typically not integrable and we will get around this difficulty using a kernel K^* with a bounded support.

Assumption 12. *For all $i \geq 1$, (Z_t, U_{it}) is a jointly stationary vector β-mixing processes with mixing coefficient $\beta_i(\tau)$. Define $\beta(\tau) := \max_i \beta_i(\tau)$.*

(i) For some $0 < \gamma < 1$ and $\varepsilon > 0$, $\beta(\tau) = O(\tau^{-(2+\varepsilon)/\gamma})$ as $\tau \to \infty$.

(ii) For some $\varkappa > 1 + q + \xi^*$ and some $\xi^* > 0$, their α-mixing coefficients satisfy $\alpha(\tau) = O(\tau^{-\varkappa})$ as $\tau \to \infty$.

Assumption 12 (ii) is implied by 12 (i) if $(2 + \varepsilon)/\gamma > \varkappa$. Assumption 12 (i) is taken from Fan and Li (1999) and implies $\sum_{\tau=1}^{\infty} \tau\beta(\tau)^{\gamma} < \infty$, a fact used in the proof of Lemma 3.6. Assumption 12(ii) was required in Hansen (2008).

Assumption 9′. *For some $\theta > 4/(1 - \gamma)$, where γ is as in Assumption 12 (i), $\sup_t E|\bar{U}_{At}|^{\theta} < \infty$,. Also, θ and γ are such that $1 - 4\gamma \leq \frac{8}{\theta}$.*

Note that $\sup_t E|\bar{U}_{At}|^{\theta} < \infty$ is a stronger condition than assuming $\sup_t E|U_{it}|^{\theta} < \infty, i = 1, \cdots, N$. Assumption 9' strengthens the moment condition on the error terms from $E|U_{it}|^{\theta} < \infty, \theta > 2$ of Assumption 9 and is required in the proof of Theorem 3.7 below.

Assumption 13. *The functions $m(z)$ and $f(z)$ are s-times partially boundedly differentiable over z for some $s > 2q/\theta$.*

Assumption 13 strengthens the local differentiability conditions on the two functions m and f in Assumption 4 to the global differentiability and is needed in handling the bias of the first stage non-parametric estimates.

Assumption 14. *The kernel function $k(\cdot)$ used in the preliminary stage N-W estimation is an even and uniformly bounded, integrable function that belongs to $\mathcal{K}_s$ and satisfies $\int |k(u)|^{\frac{2(1-\gamma)}{\theta(1-\gamma)-4}} du < \infty$. Also, $|k(u)| < C|u|^{-q/\xi^*}$ for u large.*

Assumption 15. *Each element of $\mathbf{\Omega}(z)$ has bounded derivatives of total order p at $z = \zeta_l, \quad l = 1, 2, \cdots, d$.*

Assumption 16. *The kernel function $K^*(\cdot) \in \mathcal{K}_p$ used in local smoothing of $\mathbf{\Omega}(\zeta_l)$ is an even and uniformly bounded function of bounded support.*

Assumption 15 together with Assumption 16 imply that the bias of each element of the smoothing estimate (3.4.4) of $\boldsymbol{\Omega}(\zeta_l)$ is of order $O(h^p)$.

Assumption 17. *The bandwidth $h \to 0$ is such that, with γ as in Assumption 12 (i) and $\varrho = \frac{\varkappa-1-\xi^*-q}{\varkappa+3-q}$ with $\varkappa, \xi^*$ as in Assumption 12(ii), as $T \to \infty$,*

$$\text{(i)} \quad \log T/(T^{\varrho}h^q) \to 0, \quad \text{(ii)} \log T \times h^{\frac{2q}{\theta}+\frac{4\gamma q}{\theta(1-\gamma)}} \to 0.$$

Assumption 17 (i) is from Hansen (2008) and implies $Th^q \to \infty$, which is needed to make the variance component of the first stage kernel estimation of $f(\zeta_l), l = 1, \cdots, d$ to go to zero. Assumption 17 (ii) is satisfied if h takes the form of $T^{-\eta}$ for some $\eta > 0$.

Denote by $\hat{\omega}_{ij}$ the $(i, j)^{th}$ element of non-parametric conditional covariance estimator $\hat{\boldsymbol{\Omega}}(\zeta_l)$. Theorem 3.7 gives the consistency rate for each $\hat{\omega}_{ij}$. It is reminded that in this chapter, "asymptotically" means "as T, and possibly $N = N_T$, go to ∞".

Theorem 3.7. *Under Assumptions 2,3,8, 9', 12-17, asymptotically*

$$\max_{1\leq i,j\leq N} |\hat{\omega}_{ij} - \omega_{ij}| = O_p(R_{T,h}), \quad N \geq 1,$$

where

$$R_{T,h} := h^p + h^{2s-\frac{2q}{\theta}} + \frac{1}{Th^{3\gamma q+\frac{4q}{\theta}}} + \frac{1}{Th^{q+\gamma q+\frac{2q}{\theta}-1}} + \frac{1}{Th^{\frac{q}{2}+\frac{6q}{\theta(1-\gamma)}}} + \frac{1}{\sqrt{Th^{q+\frac{12q}{\theta}}}}. \tag{3.4.5}$$

The rate obtained in Theorem 3.7 will be instrumental in the proof of Theorems 3.8 and 3.9.

Recall that Theorems 3.2 and 3.5 provide optimal bandwidth choices when $\boldsymbol{\Omega}(\zeta_l)$ is known. When carrying out feasible estimation using the estimate $\hat{\boldsymbol{\Omega}}(\zeta_l)$ instead of $\boldsymbol{\Omega}(\zeta_l)$, optimal bandwidths are obtained by replacing unknown values in the expression for $a^*_{m,AMSE(\zeta_l)}$ and $a^*_{m^*,AMSE(\zeta_l)}$ by their corresponding estimates. Theorem 3.8 shows that the infeasible and feasible optimal bandwidth choices become equivalent asymptotically. To show such equivalence, we need however to impose the following additional assumptions.

Assumption 18. *The estimates $\hat{f}$ and $\hat{\Phi}$ are such that asymptotically,*

$$\begin{aligned}
\hat{f}(\zeta_l) - f(\zeta_l) &= O_p\Big(\|\boldsymbol{\Omega}(\zeta_l)\|^{-1}\|\hat{\boldsymbol{\Omega}}(\zeta_l) - \boldsymbol{\Omega}(\zeta_l)\|\Big), \\
\hat{\Phi}^2(\tilde{m}(\zeta_l)) - \Phi^2(\tilde{m}(\zeta_l)) &= O_p\Big(\|\boldsymbol{\Omega}(\zeta_l)\|^{-1}\|\hat{\boldsymbol{\Omega}}(\zeta_l) - \boldsymbol{\Omega}(\zeta_l)\|\Big), \quad l = 1, \cdots, d.
\end{aligned}$$

Assumption 18 is rather unprimitive, but is essentially required to ensure that the effect of estimating biases for quantities $f(\zeta_l)$ and $\Phi^2(\tilde{m}(\zeta_l))$, which are required to construct the optimal bandwidth choices, are negligible so as to yield asymptotic equivalence of the feasible and infeasible optimal bandwidth choices.

Assumption 19. *N and h are such that asymptotically, $NR_{T,h} = o(1)$.*

Assumption 19 requires the choice of bandwidth parameter h to be such that the rate $R_{T,h}$ obtained in Theorem 3.7 converges sufficiently fast to 0.

Assumption 20. $\mathbf{\Omega}$ *is such that*

$$\|\mathbf{\Omega}(\zeta_l)^{-1}\| + \frac{N1_N'\mathbf{\Omega}(\zeta_l)^{-2}1_N}{(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N)^2} = O(1), \quad l = 1, \cdots, d.$$

Assumption 20 was discussed in detail in Robinson (2010b), where it was noted the first term on the LHS requires the smallest eigenvalue of $\mathbf{\Omega}(\zeta_l)$ to be bounded away from zero for large N. A sufficient (but not necessary) condition for the second therm on the LHS to be bounded is that the greatest eigenvalue of $\mathbf{\Omega}(\zeta_l)$ is bounded. See Robinson (2010b) for an example where the second term on the LHS may be bounded although the greatest eigenvalue of $\mathbf{\Omega}(\zeta_l)$ may increase with N.

Theorem 3.8. *Under Assumptions 1-20, asymptotically*

$$\frac{\hat{a}^*_{m,MSE(\zeta_l)}}{a^*_{m,MSE(\zeta_l)}}, \quad \frac{\hat{a}^*_{m^*,MSE(\zeta_l)}}{a^*_{m^*,MSE(\zeta_l)}} \to_p 1, \quad l = 1, \cdots, d.$$

Next, we define a feasible optimal N-W estimate with a bandwidth a as

$$\hat{m}^*(\zeta_l) = \frac{(1_N'\hat{\mathbf{\Omega}}(\zeta_l)^{-1}1_N)^{-1}1_N'\hat{\mathbf{\Omega}}(\zeta_l)^{-1}\sum_{t=1}^{T} K\Big(\frac{Z_t - \zeta_l}{a}\Big)Y_{\cdot t}}{\sum_{t=1}^{T} K\Big(\frac{Z_t - \zeta_l}{a}\Big)}.$$

Assumption 21. *The bandwidth a is such that as $T \to \infty$, $N^3a^q \to 0$, and with $\psi = min\left\{2s - \frac{2q}{\theta}, p\right\}$, where p is as in Assumption 15, $\sqrt{N^3Ta^q}h^\psi = o(1)$.*

Assumption 21 actually requires the bandwidth h, used in the preliminary stage, to decay slower than the bandwidth a. The last condition shows that strengthening the global smoothness conditions on m and f and the local smoothness condition on $\mathbf{\Omega}$ ensures that the non-parametric estimations in the first stage yield small enough bias.

Theorem 3.9. *Under Assumptions 1-21, asymptotically,*

$$\tilde{m}^*(\zeta_l) - \hat{m}^*(\zeta_l) = o_p\left(\frac{(1_N'\mathbf{\Omega}(\zeta_l)^{-1}1_N)^{-1/2}}{(Ta^q)^{1/2}} + a^s\right), \quad l = 1, \cdots, d.$$

Based on Theorem 3.9, one could establish an asymptotic normality result for $\hat{m}^*(\zeta_l)$, with the same limiting distribution as $\tilde{m}^*(\zeta_l)$, that is presented in Theorem 3.6.

3.5 Finite sample performance

We carry out a small simulation study to compare finite sample performance of the three estimates of $m(z)$, namely the simple N-W estimate, $\tilde{m}(z)$, the infeasible optimal N-W estimate, $\tilde{m}^*(z)$, and the feasible optimal N-W estimate, $\hat{m}^*(z)$. It is of interest to see the extent to which the feasible $\hat{m}^*(z)$ matches the efficiency of the infeasible $\tilde{m}^*(z)$ and whether it is actually better than the simple $\tilde{m}(z)$, given the sampling error in estimating $\mathbf{\Omega}(z)$. The simulation design was chosen in a close resemblance to the one reported in the common trend estimation of Robinson (2010b) for the ease of comparison to the common trend estimation case.

Recall the model $Y_{it} = \lambda_i + m(Z_t) + U_{it}$. We set the regression function to be $m(z) = 1/(1+z^2)$ and fix the individual effects λ_i by first generating $\lambda_1, \cdots, \lambda_{N-1}$ independently from standard normal distribution, then taking $\lambda_N = -\lambda_1 - \cdots - \lambda_{N-1}$ and keeping these λ_i fixed across replications. Error terms were generated by the following factor model that gives rise to cross-sectional dependence, where the factor loadings were functions of Z_t, engineering the desired conditional heteroscedasticity of the covariance matrix. Factor loadings were set to be a product of a fixed $N \times 1$ vector $b = (b_1, \cdots b_N)'$, that was generated by $b \sim \mathcal{N}_N(0, 10I_N)$ and kept fixed across replications, and a function of the value of concurrent regressor Z_t: $b_{it} = b_i(1 + |Z_t|)^{(i-1)/4}$. The error terms were then defined as

$$U_{it} = b_{it}\eta_t + \sqrt{0.5}\epsilon_{it}, \qquad \Omega(z) = 0.5I + b_t b_t'.$$

The variables $\{Z_t\}, \{\eta_t\}, \{\epsilon_{it}\}, i = 1, \cdots, N$ were generated as independent Gaussian AR(1) time series, where four values of AR coefficient, $\rho = 0, 0.2, 0.5, 0.8$ are tried in order to see how our estimates perform under differing degrees of serial dependence. The choice of the points of estimation, and the second stage bandwidth parameters are in line with the choice of Robinson (2010b): the one-dimensional regressor was generated as $Z_t \sim (0.5, \frac{1}{16})$, so as to have most observations lie in the interval [0.1], making this set-up comparable to the trend estimation where the $\frac{t}{T} \in [0,1]$ and the three fixed points of estimation used were: $z = 0.25, 0.5, 0.75$. The second stage bandwidth parameters were set to be $a = 0.1, 0.5, 1$. Because of the need for oversmoothing in the first stage, required by Assumption 21, we have set the first stage bandwidth to be 1.2 times greater than that of the second stage.

Tables 3.1 and 3.2 report the Monte Carlo MSE for differing settings of the three estimates for different choices of ρ, z and a for $(N,T) = (5, 100)$, and $(N,T) = (10, 500)$, respectively. There are $2 \times 4 \times 3 \times 3 = 72$ cases in total and each case is based on 1000 replications.

Tables 3.1 and 3.2 show that the reduction in the Monte Carlo MSE by using GLS-type estimation is substantial in all the cases. It is to be stressed here that the improvement in the MSE, that of the variance to be more specific, depends crucially on the form of the cross-sectional covariance matrix. The greater the difference between $1_N'\Omega 1_N/N^2$ and $(1_N'\Omega^{-1}1_N)^{-1}$, the greater the scope for efficiency improvement via GLS type estimation. The choice of the coefficients, in particular that of b, in generating U was such that the scope for improvement in the variance was particularly large.

It is natural that the improvement in the MSE is more pronounced for cases of smaller bandwidth parameter where the variance component dominates the bias component and this is indeed seen from the results. We would expect infeasible GLS-type estimate, $\tilde{m}^*(z)$, to perform better than the feasible version $\hat{m}^*(z)$. However, there were 7 occasions out of 72 where the feasible estimate $\hat{m}^*(z)$ performed marginally better, and these were all in the case of larger bandwidth parameters (0.5 or 1).

Tables 3.3 and 3.4 report relative Monte-Carlo MSE of infeasible and feasible GLS-type estimates in relation to that of the simple N-W estimate and were designed to facilitate comparison between differing strengths of serial dependence.

Comparing results from different degrees of serial dependence in Table 3.3, it is reported that larger serial dependence often leads to (sometimes significant) improvement in the performance of infeasible GLS-type estimate in relation to the simple N-W estimate. In fact, the ratio of Monte Carlo MSE's is smaller (i.e. better relative performance of infeasible GLS-type estimate) when $\rho = 0.8$ compared to $\rho = 0$ in every case of Table 3.3. Also for the larger bandwidth cases (0.5 and 1) there is a unilateral improvement in the relative performance of the infeasible GLS-type estimate in Table 3.3 with increase in ρ. Turning to Table 3.4, similar patterns to Table 3.3 are reported: namely, there is a unilateral improvement in the relative performance of feasible GLS-type estimate with increasing ρ for the case of larger bandwidth (0.5 and 1).

Table 3.1: Monte Carlo MSE, $N = 5, T = 100$

ρ	z	a	$\widehat{MSE}_{\tilde{m}}(z)$	$\widehat{MSE}_{\tilde{m}^*}(z)$	$\widehat{MSE}_{\hat{m}^*}(z)$
0	0.25	0.1	0.4092	0.0107	0.1398
		0.5	0.1117	0.0141	0.0131
		1	0.1129	0.0246	0.0147
	0.5	0.1	0.2817	0.0062	0.0523
		0.5	0.0991	0.0022	0.0111
		1	0.095	0.0021	0.0107
	0.75	0.1	0.5918	0.011	0.1274
		0.5	0.1416	0.0157	0.0235
		1	0.123	0.0246	0.0326
0.2	0.25	0.1	0.4344	0.0115	0.1526
		0.5	0.1541	0.0151	0.0145
		1	0.1582	0.0256	0.0167
	0.5	0.1	0.3108	0.007	0.0522
		0.5	0.145	0.0031	0.0128
		1	0.1417	0.0031	0.0125
	0.75	0.1	0.6228	0.0114	0.1538
		0.5	0.1899	0.0166	0.0247
		1	0.1713	0.0256	0.0342
0.5	0.25	0.1	0.5717	0.0157	0.2047
		0.5	0.2836	0.0181	0.0223
		1	0.2953	0.0285	0.0245
	0.5	0.1	0.4658	0.01	0.0701
		0.5	0.2868	0.0061	0.0203
		1	0.2812	0.0061	0.0202
	0.75	0.1	0.8636	0.0164	0.2183
		0.5	0.3462	0.0198	0.0332
		1	0.3139	0.0286	0.0416
0.8	0.25	0.1	1.3983	0.0321	0.829
		0.5	0.8153	0.0295	0.0561
		1	0.8284	0.0398	0.056
	0.5	0.1	1.0854	0.0231	0.1601
		0.5	0.8281	0.0172	0.0523
		1	0.8193	0.0173	0.0515
	0.75	0.1	1.9009	0.0344	0.7075
		0.5	0.9368	0.0321	0.0666
		1	0.8578	0.0401	0.0727

Table 3.2: Monte Carlo MSE, $N = 10, T = 500$

ρ	z	a	$\widehat{MSE}_{\tilde{m}}(z)$	$\widehat{MSE}_{\tilde{m}^*}(z)$	$\widehat{MSE}_{\hat{m}^*}(z)$
0	0.25	0.1	0.0758	0.0014	0.0172
		0.5	0.0359	0.0126	0.0152
		1	0.0431	0.0234	0.0251
	0.5	0.1	0.0659	0.0008	0.0103
		0.5	0.0228	0.0004	0.0036
		1	0.0219	0.0004	0.0038
	0.75	0.1	0.1236	0.0014	0.0206
		0.5	0.0421	0.0134	0.0103
		1	0.0455	0.0223	0.0166
0.2	0.25	0.1	0.0851	0.0015	0.018
		0.5	0.0456	0.0128	0.0155
		1	0.0537	0.0236	0.0254
	0.5	0.1	0.0802	0.001	0.0106
		0.5	0.0336	0.0005	0.004
		1	0.0326	0.0006	0.0041
	0.75	0.1	0.1436	0.0015	0.0214
		0.5	0.0544	0.0135	0.0106
		1	0.0567	0.0225	0.0169
0.5	0.25	0.1	0.1261	0.0021	0.025
		0.5	0.0747	0.0132	0.0176
		1	0.0851	0.0241	0.0278
	0.5	0.1	0.1109	0.0014	0.013
		0.5	0.0653	0.0009	0.006
		1	0.0648	0.001	0.0061
	0.75	0.1	0.2013	0.0021	0.0276
		0.5	0.0914	0.014	0.0125
		1	0.0895	0.0229	0.0186
0.8	0.25	0.1	0.2814	0.0046	0.0664
		0.5	0.1935	0.0151	0.0285
		1	0.2097	0.0259	0.0387
	0.5	0.1	0.2623	0.0032	0.0288
		0.5	0.192	0.0026	0.0163
		1	0.1915	0.0027	0.0163
	0.75	0.1	0.4748	0.0045	0.0709
		0.5	0.2372	0.0158	0.0225
		1	0.2184	0.0247	0.0281

Table 3.3: Relative MSE: $MSE(\tilde{m}^*(z))/MSE(\tilde{m}(z))$

		$N = 5, T = 100$			
z	$a\backslash\rho$	0	0.2	0.5	0.8
0.25	0.1	0.026149	0.026473	0.027462	0.022956
	0.5	0.126231	0.097988	0.063822	0.036183
	1	0.217892	0.16182	0.096512	0.048044
0.5	0.1	0.022009	0.022523	0.021468	0.021282
	0.5	0.0222	0.021379	0.021269	0.02077
	1	0.022105	0.021877	0.021693	0.021116
0.75	0.1	0.018587	0.018304	0.01899	0.018097
	0.5	0.110876	0.087414	0.057192	0.034266
	1	0.2	0.149445	0.091112	0.046747
		$N = 10, T = 500$			
z	$a\backslash\rho$	0	0.2	0.5	0.8
0.25	0.1	0.01847	0.017626	0.016653	0.016347
	0.5	0.350975	0.280702	0.176707	0.078036
	1	0.542923	0.439479	0.283196	0.12351
0.5	0.1	0.01214	0.012469	0.012624	0.0122
	0.5	0.017544	0.014881	0.013783	0.013542
	1	0.018265	0.018405	0.015432	0.014099
0.75	0.1	0.011327	0.010446	0.010432	0.009478
	0.5	0.31829	0.248162	0.153173	0.06661
	1	0.49011	0.396825	0.255866	0.113095

Table 3.4: Relative MSE: $MSE(\hat{m}^*(z))/MSE(\tilde{m}(z))$

		$N = 5, T = 100$			
z	$a\backslash\rho$	0	0.2	0.5	0.8
0.25	0.1	0.341642	0.351289	0.358055	0.592863
	0.5	0.117278	0.094095	0.078632	0.068809
	1	0.130204	0.105563	0.082966	0.0676
0.5	0.1	0.185659	0.167954	0.150494	0.147503
	0.5	0.112008	0.088276	0.070781	0.063157
	1	0.112632	0.088215	0.071835	0.062859
0.75	0.1	0.215275	0.246949	0.252779	0.372192
	0.5	0.16596	0.130068	0.095898	0.071093
	1	0.265041	0.19965	0.132526	0.084752
		$N = 10, T = 500$			
z	$a\backslash\rho$	0	0.2	0.5	0.8
0.25	0.1	0.226913	0.211516	0.198255	0.235963
	0.5	0.423398	0.339912	0.235609	0.147287
	1	0.582367	0.472998	0.326675	0.184549
0.5	0.1	0.156297	0.13217	0.117223	0.109798
	0.5	0.157895	0.119048	0.091884	0.084896
	1	0.173516	0.125767	0.094136	0.085117
0.75	0.1	0.166667	0.149025	0.137109	0.149326
	0.5	0.244656	0.194853	0.136761	0.094857
	1	0.364835	0.29806	0.207821	0.128663

3.6 Appendix A. Proof of Theorems 3.7-3.9

Proof of Theorem 3.7. For the $(i,j)^{th}$ element $\hat{\omega}_{ij}(\zeta_l)$ of $\hat{\boldsymbol{\Omega}}(\zeta_l)$, one has

$$\hat{\omega}_{ij}(\zeta_l) - \omega_{ij}(\zeta_l) = \frac{\sum_{t=1}^{T} K_l(Z_t;h)\{\hat{U}_{it}\hat{U}_{jt} - \omega_{ij}(\zeta_l)\}}{\sum_{t=1}^{T} K_l(Z_t;h)} =: R_{ij}^{(1)} + R_{ij}^{(2)}, \tag{3.6.1}$$

with

$$R_{ij}^{(1)} = \frac{\sum_{t=1}^{T} K_l(Z_t;h)\{U_{it}U_{jt} - \omega_{ij}(\zeta_l)\}}{\sum_{t=1}^{T} K_l(Z_t;h)}, \quad R_{ij}^{(2)} = \frac{\sum_{t=1}^{T} K_l(Z_t;h)\{\hat{U}_{it}\hat{U}_{jt} - U_{it}U_{jt}\}}{\sum_{t=1}^{T} K_l(Z_t;h)}.$$

Under Assumptions 12, 15 and 16, it can be shown $|R_{ij}^{(1)}| = O_p\left(\frac{1}{\sqrt{Th^q}} + h^p\right)$ as this is the estimation error of the usual N-W estimator of the conditional expectation $E(U_{it}U_{jt}|Z_t = \zeta_l) = \omega_{ij}(\zeta_l)$, with the typical variance and bias contributions. Next, we show that $|R_{ij}^{(2)}| = O_p\left(R_{T,h}\right)$, which implies (3.4.5) of Theorem 3.7.

Denote $d_i := \bar{U}_{AA} - \bar{U}_{iA}$ and $e_t := m(Z_t) - \tilde{m}(Z_t)$. From (3.4.3),

$$\hat{U}_{it} = U_{it} - \bar{U}_{iA} + \bar{U}_{AA} + m(Z_t) - \tilde{m}(Z_t) = U_{it} + d_i + e_t.$$

Using this equality, we can decompose

$$\hat{U}_{it}\hat{U}_{jt} - U_{it}U_{jt} = (d_i + e_t)(d_j + e_t) + U_{it}(d_j + e_t) + U_{jt}(d_i + e_t). \tag{3.6.2}$$

For the rest of the proof, denote $K_t := K_l(Z_t;h)$ for brevity. We need to find the stochastic order of

$$R_{ij}^{(2)} = \tilde{f}(\zeta_l)^{-1}\frac{1}{Th^q}\sum_{t=1}^{T} K_t\{(d_i + e_t)(d_j + e_t) + U_{it}(d_j + e_t) + U_{jt}(d_i + e_t)\}, \tag{3.6.3}$$

where $\tilde{f}(\zeta_l)$ is a non-parametric kernel estimator of $f(\zeta_l)$:

$$\tilde{f}(\zeta_l) = \frac{1}{Th^q}\sum_{s=1}^{T} K\left(\frac{Z_s - \zeta_l}{h}\right),$$

which is consistent in the light of Assumptions 3, 4, 12, 14 and 17 (i). Therefore,

$$\frac{1}{\tilde{f}(\zeta_l)} = \frac{1}{f(\zeta_l) + o_p(1)} = O_p(1). \tag{3.6.4}$$

Next, we study the stochastic order of the rest of (3.6.3). Firstly, we bound sums that involve $d_i d_j, U_{it} d_j$ and $U_{jt} d_i$, namely,

$$\frac{1}{Th^q} \sum_{t=1}^{T} K_t \{d_i d_j + U_{it} d_j + U_{jt} d_i\}. \tag{3.6.5}$$

Since U_{it}'s are weakly dependent across time and Assumption 9' implies $Var(\bar{U}_{At}) \leq C$ and $Cov(\bar{U}_{At}, \bar{U}_{As}) \leq C$, one has

$$d_i = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} U_{it} - \frac{1}{T} \sum_{t=1}^{T} U_{it} = \frac{1}{T} \sum_{t=1}^{T} \bar{U}_{At} - \frac{1}{T} \sum_{t=1}^{T} U_{it} = O_p(T^{-1/2}).$$

Therefore, the upper bound of the first term in (3.6.5) is

$$\{d_i d_j\} \frac{1}{Th^q} \sum_{t=1}^{T} K_t = O_p\left(\frac{1}{T}\right) \tilde{f}(\zeta_l) = O_p\left(\frac{1}{T}\right) \times \left(f(\zeta_l) + o_p\left(1\right)\right) = O_p\left(\frac{1}{T}\right),$$

because $\tilde{f}(\zeta_l) = f(\zeta_l) + o_p\left(1\right)$. The upper bound of the second (and third) term is

$$d_j \times \frac{1}{Th^q} \sum_{t=1}^{T} K_t U_{it} = O_p\left(\frac{1}{\sqrt{T}}\right) \times O_p\left(\frac{1}{\sqrt{Th^q}}\right) = O_p\left(\frac{1}{T\sqrt{h^q}}\right),$$

because $\sum_{t=1}^{T} K_t U_{it}/Th^q$ consistently estimates $E(U_{it}|Z_t = \zeta_l) = 0$, with zero bias and the usual variance contribution. Thus, (3.6.5) satisfies the bound $|R_{ij}^{(2)}| = O_p\left(R_{T,h}\right)$. Now the terms left to analyse from the numerator of (3.6.1), are

$$\frac{1}{Th^q} \sum_{t=1}^{T} K_t \{e_t^2 + U_{it} e_t + U_{jt} e_t + d_i e_t + d_j e_t\}. \tag{3.6.6}$$

Introduce the leave-one-out counterpart of e_t, $\tilde{e}_t := (l_t - n_t)/\tilde{f}_t$, where

$$l_t := \frac{1}{Th^q} \sum_{s=1, s\neq t}^{T} K\left(\frac{Z_s - Z_t}{h}\right) \{m(Z_t) - m(Z_s)\}, \quad n_t := \frac{1}{Th^q} \sum_{s=1, s\neq t}^{T} K\left(\frac{Z_s - Z_t}{h}\right) \bar{U}_{As}.$$

The asymptotic equivalence between (3.6.6) and

$$\frac{1}{Th^q} \sum_{t=1}^{T} K_t \{\tilde{e}_t^2 + U_{it} \tilde{e}_t + U_{jt} \tilde{e}_t + d_i \tilde{e}_t + d_j \tilde{e}_t\} \tag{3.6.7}$$

will be shown below. Recall $d_i = O_p\left(T^{-1/2}\right)$. To bound (3.6.7), we need to obtain an upper bound of six quantities. Setting $K_t := K_l(Z_t;h)$, these are:

$$\begin{aligned}\left|\frac{1}{Th^q}\sum_{t=1}^T K_t\tilde{e}_t^2\right| &\leq \frac{C}{Th^q}\sum_{t=1}^T |K_t|\frac{n_t^2}{\tilde{f}_t^2}+\frac{C}{Th^q}\sum_{t=1}^T |K_t|\frac{l_t^2}{\tilde{f}_t^2} =: \mathbf{A_T}+\mathbf{B_T},\\ \left|\frac{1}{Th^q}\sum_{t=1}^T K_tU_{it}\tilde{e}_t\right| &\leq \left|\frac{1}{Th^q}\sum_{t=1}^T K_tU_{it}\frac{l_t}{\tilde{f}_t}\right|+\left|\frac{1}{Th^q}\sum_{t=1}^T K_tU_{it}\frac{n_t}{\tilde{f}_t}\right| =: \mathbf{C_T}+\mathbf{D_T}, \qquad (3.6.8)\\ \left|\frac{1}{Th^q}\sum_{t=1}^T K_t\tilde{e}_t\right| &\leq \left|\frac{1}{Th^q}\sum_{t=1}^T K_t\frac{n_t}{\tilde{f}_t}\right|+\left|\frac{1}{Th^q}\sum_{t=1}^T K_t\frac{l_t}{\tilde{f}_t}\right| =: \mathbf{E_T}+\mathbf{F_T}.\end{aligned}$$

Then, the LHS of (3.6.7) is bounded by $\mathbf{A_T}+\mathbf{B_T}+\mathbf{C_T}+\mathbf{D_T}+O(T^{-1/2})\{\mathbf{E_T}+\mathbf{F_T}\}$.

Now, we show the negligibility of the difference between (3.6.6) and (3.6.7). Notice that

$$e_t-\tilde{e}_t=\frac{1}{Th^q}K\left(\frac{Z_t-Z_t}{h}\right)\{m(Z_t)-m(Z_t)\}+\frac{1}{Th^q}K\left(\frac{Z_t-Z_t}{h}\right)\bar{U}_{At}=\frac{K(0)}{Th^q}\bar{U}_{At}.$$

We need to show negligibility of

$$\frac{1}{Th^q}\sum_{t=1}^T K_t\{(e_t^2-\tilde{e}_t^2)+U_{it}(e_t-\tilde{e}_t)+U_{jt}(e_t-\tilde{e}_t)+d_i(e_t-\tilde{e}_t)+d_j(e_t-\tilde{e}_t)\}.$$

Firstly,

$$\begin{aligned}\frac{1}{Th^q}\sum_{t=1}^T K_t\{U_{it}(e_t-\tilde{e}_t)+d_i(e_t-\tilde{e}_t)\} &= \frac{C}{(Th^q)^2}\sum_{t=1}^T K_tU_{it}\bar{U}_{At}+\frac{Cd_i}{(Th^q)^2}\sum_{t=1}^T K_t\bar{U}_{At}\\ &= O_p\left(\frac{1}{Th^q}\big(h^p+\frac{1}{\sqrt{Th^q}}\big)\right)+O_p\left(\frac{1}{Th^q}\frac{1}{\sqrt{T}}\frac{1}{\sqrt{Th^q}}\right)=o_p(R_{T,h}).\end{aligned}$$

To justify the above bound, note that $\sum_{t=1}^T K_tU_{it}\bar{U}_{At}/Th^q$ is a N-W estimator of the conditional expectation function $E(U_{it}\bar{U}_{At}|Z_t=\zeta_l)=\sum_{j=1}^N \omega_{ij}(\zeta_l)/N$ with bias of order h^p in the light of Assumption 15 and 16, and variance of order $(Th^q)^{-1}$ under Assumption 12. Similarly, in the latter term, $\frac{1}{Th^q}\sum_{t=1}^T K_t\bar{U}_{At}$ is the N-W estimator of the conditional expectation function $E(\bar{U}_{At}|Z_t=\zeta_l)=0$ with zero bias and variance of order $(Th^q)^{-1}$.

Next, recalling $e_t - \tilde{e}_t = \frac{K(0)}{Th^q}\bar{U}_{At}$,

$$\begin{aligned}\frac{1}{Th^q}\sum_{t=1}^{T} K_t(e_t^2 - \tilde{e}_t^2) &= \frac{1}{Th^q}\sum_{t=1}^{T} K_t(e_t - \tilde{e}_t)(2\tilde{e}_t + (e_t - \tilde{e}_t)) \\ &= \frac{K(0)}{Th^q}\big[2\sum_{t=1}^{T} K_t\tilde{e}_t\bar{U}_{At} - \frac{K(0)}{Th^q}\sum_{t=1}^{T} K_t\bar{U}_{At}^2\big]. \qquad (3.6.9)\end{aligned}$$

The second term can be written as

$$\frac{K(0)}{(Th^q)^2}\sum_{t=1}^{T} K_t\bar{U}_{At}^2 = \frac{K(0)}{Th^q}\cdot\frac{1}{Th^q}\sum_{t=1}^{T} K_t\bar{U}_{At}^2 = \frac{K(0)}{Th^q}O_p\left(h^p + \frac{1}{\sqrt{Th^q}}\right) = o_p(R_{T,h}),$$

noting that MSE of the N-W estimator $\frac{1}{Th^q}\sum_{t=1}^{T} K_t\bar{U}_{At}^2$ of the conditional expectation function $E(\bar{U}_{At}^2|Z_t = \zeta_l) = \sum_{i,j=1}^{N} \omega_{ij}(\zeta_l)/N^2$ is $O(h^{2p} + (Th^q)^{-1})$ in the light of Assumptions 12, 15 and 16.

The first term of (3.6.9) satisfies the same upper bound as $\mathbf{C_T} + \mathbf{D_T}$ noting the similarity of the expression $\frac{1}{Th^q}\sum_{t=1}^{T} K_t\tilde{e}_t\bar{U}_{At}$ to $\frac{1}{Th^q}\sum_{t=1}^{T} K_t\tilde{e}_t U_{it}$ in the LHS of (3.6.8). In deriving an upper bound on $\mathbf{C_T} + \mathbf{D_T}$, the condition $E|U_{it}|^\theta < \infty$, implied by Assumption 9', is repeatedly used. The same proof, and therefore the same upper bound, applies to the first term of (3.6.9) by replacing U_{it} with $\bar{U}_{At}$ and using $E|\bar{U}_{At}|^\theta < \infty$ of Assumption 9'.

To complete the proof of Theorem 3.7 we need to show that

$$\mathbf{A_T} + \mathbf{B_T} + \mathbf{C_T} + \mathbf{D_T} \le CR_{T,h}, \qquad (3.6.10)$$

$$\mathbf{E_T} + \mathbf{F_T} \le C\sqrt{T}R_{T,h}. \qquad (3.6.11)$$

The terms $\mathbf{A_T}, \cdots, \mathbf{F_T}$ can be divided into two types. Write

$$\frac{1}{\tilde{f}_t} = \frac{1}{f_t} + \frac{1}{\tilde{f}_t} - \frac{1}{f_t} = \frac{1}{f_t} + \frac{(f_t - \tilde{f}_t)}{\tilde{f}_t f_t}. \qquad (3.6.12)$$

The first type of terms contains $1/f_t$ and takes the form of a U-statistic. Finding their stochastic order of magnitude is complicated by serial dependence in Z_t and U_{it} in their arguments. These terms will be analyzed using Lemma 3.6, which provides the asymptotic order of the difference between such U-statistics and their counterparts under independence. Bounding the first type of terms, firstly, the asymptotic order of the expectation of the kernel of U-statistic under the corresponding independent process will be derived and, secondly, the remainder terms evaluated, applying Lemma 3.6.

To deal with the second type of terms that contain $(f_t - \tilde{f}_t)/\tilde{f}_t f_t$, we use the uniform rate of convergence result of Hansen (2008). Under Assumptions 8 (ii), 12 (ii), 13, 14 and 17 (i) Hansen (2008) showed that

$$\sup_{z\in\mathbb{R}^q}\left|\tilde{f}(z) - f(z)\right| = O_p\left(\left(\frac{\log T}{Th^q}\right)^{1/2} + h^s\right), \tag{3.6.13}$$

where s is the smoothness parameter on f and m appearing in Assumption 13. It is worth noting here that the term "kernel" is used also to refer to the summand of a U-statistic, as well as the kernel function of non-parametric estimation.

3.6.1 Upper bound on $\mathbf{A}_T$.

In this section, it will be shown that

$$\begin{aligned} \mathbf{A_T} &= O(r_{1T}), \qquad (3.6.14)\\ r_{1T} &= \left(\frac{1}{Th^q}\right)^3 \left[T^2 h^{2q-\frac{2q}{\theta}} + T^2 h^{3q(1-\gamma)-\frac{4q}{\theta}}\right], \end{aligned}$$

which implies (3.6.10) for $\mathbf{A}_T$.

We first divide $\mathbf{A_T}$ into two parts, using (3.6.12).

$$\begin{aligned} \mathbf{A_T} &= \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|\frac{n_t^2}{\tilde{f}_t^2} = \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|\frac{n_t^2}{f_t^2} + \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|n_t^2\left(\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right) \qquad (3.6.15)\\ &\leq \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|\frac{n_t^2}{f_t^2} + \max_{t:K_t\neq 0}\left|\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|n_t^2\\ &=: \mathbf{A'_T} + \max_{t:K_t\neq 0}\left|\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right|\mathbf{A''_T}. \end{aligned}$$

Taking *max* over $\{t : K_t \neq 0\}$ instead of over all t is facilitated by the boundedness of the support of $K_l(\cdot; h)$, since any t with corresponding $(Z_t - \zeta_l)/h$ falling outside the support of K_l is assigned a zero weight. We show that

$$\begin{aligned} E\mathbf{A'_T} &= O(r_{1T}), \qquad (3.6.16)\\ E\mathbf{A''_T} &= O(r_{1T}), \qquad (3.6.17)\\ \max_{t:K_t\neq 0}\left|\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| &= O_p\left(\left(\frac{\log T}{Th^q}\right)^{1/2} + h^s\right) = O_p(1), \qquad (3.6.18) \end{aligned}$$

which implies (3.6.14), noting non-negativity of $\mathbf{A'_T}$ and $\mathbf{A''_T}$.

Let us first find the asymptotic order of $\mathbf{A'_T}$. Denote $K\big((Z_t - Z_s)/h\big) = K_{ts}$ for brevity of presentation. Let ${\sum_{t_1,\cdots,t_k}}'$ denote a summation over non-overlapping indices

$(t_1, \cdots, t_k)$ for $k \geq 2$. Then,

$$E(\mathbf{A'_T}) = \big(\frac{1}{Th^q}\big)^3 E\Big(\sum_{t_1,t_2=1}^{T}{}' \frac{|K_{t_1}|}{f_{t_1}^2}\bar{U}_{At_2}^2 K_{t_1t_2}^2\Big) \tag{3.6.19}$$

$$+\big(\frac{1}{Th^q}\big)^3 E\Big(\sum_{t_1,t_2,t_3=1}^{T}{}' \frac{|K_{t_1}|}{f_{t_1}^2}\bar{U}_{At_2}\bar{U}_{At_3} K_{t_1t_2}K_{t_1t_3}\Big) \tag{3.6.20}$$

$$=: \big(\frac{1}{Th^q}\big)^3 (A_{1T} + A_{2T}). \tag{3.6.21}$$

To prove (3.6.16), it remains to show that for $i = 1, 2$,

$$\mathbf{A}_{iT} \leq Cr_{1T}. \tag{3.6.22}$$

Noting that A_{1T} and A_{2T} are expectations of second and third order U-statistics, we can apply Lemma 3.6 (i) and (ii) to find their upper bounds. Denote $W_t = W_{tT} = (Z_t', U_{1t}, \cdots, U_{Nt})'$, where $N = N_T$ may increase with T. Let $\{\tilde{W}_t\}$ denote an i.i.d. process with the marginal distribution function of W_t, and independent of $\{W_t\}$.

In finding the M_T quantities of Lemma 3.6, the conditions used to obtain an upper bound below are uniform over $t_1, \cdots, t_4$, meaning the maximum over indices is redundant.

Upper bound on A_{1T}. In this section we will show (3.6.22), for $i = 1$. Notice that A_{1T} is a second order U-statistic whose kernel is

$$\phi_T(W_t, W_s) := \frac{|K_t|}{f_t^2}\bar{U}_{At}^2 K_{ts}^2.$$

By Lemma 3.6 (i),

$$|A_{1T}| = |\sum_{t,s}{}' E\phi_T(W_t, W_s)| \leq T(T-1)|E\phi_T(\tilde{W}_1, \tilde{W}_2)| + CTM_{T2}^{1-\gamma}. \tag{3.6.23}$$

We will denote below expectation under independent process with a superscript *. Trivially,

$$E(\phi_T(\tilde{W}_t, \tilde{W}_s)) = E^*\left(\frac{|K_t|}{f_t^2}\bar{U}_{As}^2 K_{ts}^2\right) = E^*\left(\frac{|K_t|}{f_t^2}E^*\left(\bar{U}_{As}^2 K_{ts}^2|Z_t\right)\right).$$

Applying Holder's inequality with $p, r > 1$ and $p^{-1} + r^{-1} = 1$,

$$E^*\left(\bar{U}_{As}^2 K_{ts}^2|Z_t\right) \leq \left[E^*\left(|\bar{U}_{As}|^{2p}|Z_t\right)\right]^{\frac{1}{p}}\left[E^*\left(|K_{ts}|^{2r}|Z_t\right)\right]^{\frac{1}{r}} = \left[E\left(|\bar{U}_{As}|^{2p}\right)\right]^{\frac{1}{p}}\left[E^*\left(|K_{ts}|^{2r}|Z_t\right)\right]^{\frac{1}{r}},$$

where the last step holds because of the supposed independence between $\bar{U}_{As}$ and Z_t. The power p is selected as follows. Notice that $E|\bar{U}_{As}|^{2p} < \infty$, with $2p = \theta$, by Assumption 9. Then, $\frac{1}{r} = 1 - \frac{2}{\theta}$. Since Assumption 14 implies k is such that

$\int |k(u)|^{2r}du < \infty$, we have $E^*\left(|K_{ts}|^{2r}|Z_t = z\right) = O(h^q)$ uniformly over z by Lemma 3.1. Therefore, $E^*\left(\bar{U}_{As}^2 K_{ts}^2|Z_t = z\right) = O\left(h^{q(\theta-2)/\theta}\right)$ uniformly over z. Hence

$$\begin{aligned} E(\phi_T(\tilde{W}_t, \tilde{W}_s)) &= \int_z \frac{|K_l(z;h)|}{f(z)^2} E^*\left(\bar{U}_{As}^2 K_{ts}^2 | Z_t = z\right) f(z) dz \\ &\leq Ch^{\frac{q(\theta-2)}{\theta}} E\left(\frac{|K_t|}{f_t^2}\right) = O\left(h^{2q - \frac{2q}{\theta}}\right), \qquad (3.6.24) \end{aligned}$$

where the last step follows by Lemma 3.3.

Next, we bound the quantity

$$M_{T2} = \max_{1 \leq s < t \leq T} \left(E|\tilde{\phi}_T(W_s, W_t)|^{\frac{1}{1-\gamma}} + E|\tilde{\phi}_T(\tilde{W}_s \tilde{W}_t)|^{\frac{1}{1-\gamma}} \right),$$

where $\tilde{\phi}_T(W_s, W_t) = \phi_T(W_s, W_t) + \phi_T(W_t, W_s)$. We have

$$\begin{aligned} E|\phi_T(W_t, W_s)|^{\frac{1}{1-\gamma}} &= E\left(\left| \frac{K_t}{f_t^2} \bar{U}_{As}^2 K_{ts}^2 \right|^{\frac{1}{1-\gamma}} \right) \leq \left(E\left|\bar{U}_s\right|^{\frac{2p}{1-\gamma}} \right)^{\frac{1}{p}} \left(E \left| K_{ts}^2 \frac{K_t}{f_t^2} \right|^{\frac{r}{1-\gamma}} \right)^{\frac{1}{r}} \\ &= O\left(h^{2q(1 - \frac{2}{\theta(1-\gamma)})} \right), \end{aligned}$$

where the last step follows using Lemma 3.4 (i) and choosing p such that from setting $2p/(1-\gamma) = \theta$, by Assumption 9 we have $E|\bar{U}_{At}|^{2p/(1-\gamma)} = E|\bar{U}_{At}|^\theta < \infty$. Such choice of p gives $\frac{1}{r} = 1 - \frac{2}{\theta(1-\gamma)}$. Similarly,

$$\begin{aligned} &E|\phi_T(W_s, W_t)|^{\frac{1}{1-\gamma}} = O\left(h^{2q(1 - \frac{2}{\theta(1-\gamma)})} \right), \\ &E|\phi_T(\tilde{W}_s, \tilde{W}_t)|^{\frac{1}{1-\gamma}} = E^*(|\frac{K_t}{f_t^2} \bar{U}_{As}^2 K_{ts}^2|^{\frac{1}{1-\gamma}}) = O\left(h^{2q(1 - \frac{2}{\theta(1-\gamma)})} \right). \end{aligned}$$

This gives $M_{T2}^{1-\gamma} \leq Ch^{2q(1-\gamma) - \frac{4q}{\theta}}$ and together with (3.6.23) and (3.6.24) implies (3.6.22) for $i = 1$, because $Th^{2q(1-\gamma) - \frac{4q}{\theta}} = T^2 h^{3q(1-\gamma) - \frac{4q}{\theta}} (Th^{q(1-\gamma)})^{-1} = O(T^2 h^{3q(1-\gamma) - \frac{4q}{\theta}})$ by Assumption 17.

Upper bound on A_{2T}. We will prove (3.6.22) for A_{2T}. $A_{2T} = O\left(T^2 h^{3q(1-\gamma) - \frac{4q}{\theta}}\right)$. Recalling A_{2T} defined in (3.6.20), the kernel function is

$$\phi_T(W_t, W_s, W_r) = \frac{|K_t|}{f_t^2} \bar{U}_{As} \bar{U}_{Ar} K_{ts} K_{tr}. \qquad (3.6.25)$$

The proof follows the structure of the proof for A_{1T}. By Lemma 3.6 (ii),

$$|A_{2T}| \leq T^3 |E\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3)| + C(T^2 M_{T12}^{1-\gamma} + TM_{T3}^{1-\gamma}). \qquad (3.6.26)$$

The expectation under independence is

$$E[\phi_T(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)] = E^* \left(\frac{|K_t|}{f_t^2} E^*(\bar{U}_{As} K_{ts} | Z_t) E^*(\bar{U}_{Ar} K_{tr} | Z_t) \right) = 0,$$

because by Assumption 2, $E^*(\bar{U}_{As} K_{ts} | Z_t) = E^*[K_{ts} E^*(\bar{U}_{As} | Z_s) | Z_t] = E^*[K_{ts} \cdot 0 | Z_t] = 0$. Next, we will use Lemma 3.6 (ii) to find upper bounds on M_{T3} and M_{T12}.

As noted earlier, in obtaining the upper bounds, we use the fact that the upper bound under serial dependence between arguments dominates that obtained under independence. We will show that

$$\begin{aligned} M_{T12} &= \max_{1 \le s < t \le T} (E|\tilde{\phi}_T(\tilde{W}_t, \tilde{W}_s, W_r)|^{\frac{1}{1-\gamma}} + E|\tilde{\phi}_T(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)|^{\frac{1}{1-\gamma}}) \\ &= O(h^{3q - \frac{4}{\theta(1-\gamma)}}), \qquad (3.6.27) \\ M_{T3} &= \max_{1 \le s < t \le T} (E|\tilde{\phi}_T(\tilde{W}_t, W_s, W_r)|^{\frac{1}{1-\gamma}} + E|\tilde{\phi}_T(\tilde{W}_t, \tilde{W}_s, W_r)|^{\frac{1}{1-\gamma}}) \\ &= O(h^{2q - \frac{4}{\theta(1-\gamma)}}), \qquad (3.6.28) \end{aligned}$$

which with (3.6.26) imply $A_{2T} \le C[T^2 h^{3q(1-\gamma) - \frac{4q}{\theta}} + T h^{2q(1-\gamma) - \frac{4q}{\theta}}] \le C T^2 h^{3q(1-\gamma) - \frac{4q}{\theta}}$ because $T h^{q(1-\gamma)} \to \infty$ by Assumption 17. This proves (3.6.22) for A_{2T}.

Upper bound on M_{T12}. For M_{T12}, we need to consider when the variables that enter ϕ_T are divided into either two or three independent subsets. The methods and conditions used to derive the upper bounds apply uniformly over $1 \le r, s, t, \le T$ so the *max* over indices is redundant: we are concerned only with how the arguments W_r, W_s, W_T are divided into independent subsets. For the case of two independent subsets, the symmetry between W_s and W_r in ϕ_T means that it suffices to consider two distinct cases, namely $\{\tilde{W}_t, W_s, W_r\}$ and $\{\tilde{W}_r, W_t, W_s\}$.

For $\{\tilde{W}_t, W_s, W_r\}$, we will show that

$$E|\phi_T(\tilde{W}_t, W_s, W_r)|^{\frac{1}{1-\gamma}} = E_{t,sr} \left(\left| \frac{K_t}{f_t^2} \right|^{\frac{1}{1-\gamma}} E_{t,sr} \left(\left| \bar{U}_{As} \bar{U}_{Ar} K_{ts} K_{tr} \right|^{\frac{1}{1-\gamma}} |Z_t \right) \right) = O\left(h^{3q - \frac{4}{\theta(1-\gamma)}} \right),$$

where $E_{t,sr}$ denotes expectation taken under $\{\tilde{W}_t, W_s, W_r\}$. To show (3.6.29), note that for $p, w > 1$, $p^{-1} + w^{-1} = 1$,

$$\begin{aligned} & E_{t,sr} \left(\left| \bar{U}_{As} \bar{U}_{Ar} K_{ts} K_{tr} \right|^{\frac{1}{1-\gamma}} |Z_t = z \right) \\ & \le \left[E_{t,sr} \left(\left| \bar{U}_{As} \bar{U}_{Ar} \right|^{\frac{p}{1-\gamma}} |Z_t = z \right) \right]^{\frac{1}{p}} \left[E_{t,sr} \left(\left| K_{ts} K_{tr} \right|^{\frac{w}{1-\gamma}} |Z_t = z \right) \right]^{\frac{1}{w}} \\ & = \left[E_{t,sr} \left(\left| \bar{U}_{As} \bar{U}_{Ar} \right|^{\frac{p}{1-\gamma}} \right) \right]^{\frac{1}{p}} \left[E_{t,sr} \left(\left| K_{ts} K_{tr} \right|^{\frac{w}{1-\gamma}} |Z_t = z \right) \right]^{\frac{1}{w}} \end{aligned}$$

because of the presumed independence between $\{\bar{U}_{As}, \bar{U}_{Ar}\}$ and Z_t. By Cauchy-

Schwarz inequality and Assumption 9',

$$E_{t,sr}\left(\left|\bar{U}_{As}\bar{U}_{Ar}\right|^{\frac{p}{1-\gamma}}\right) \leq \left[E\left(\left|\bar{U}_{As}\right|^{\frac{2p}{1-\gamma}}\right)E\left(\left|\bar{U}_{Ar}\right|^{\frac{2p}{1-\gamma}}\right)\right]^{1/2} \leq C < \infty.$$

We select p setting $\frac{2p}{1-\gamma} = \theta$. Then $\frac{1}{w} = 1 - \frac{2}{\theta(1-\gamma)}$. Now,

$$\begin{aligned} E_{t,sr}\left(|K_{ts}K_{tr}|^{\frac{w}{1-\gamma}}\,|Z_t = z\right) \leq \sup_{w,y} f_{|r-s|}(v,y) \int |K(\frac{v-z}{h})|^{\frac{w}{1-\gamma}}dv \\ \int |K(\frac{y-z}{h})|^{\frac{w}{1-\gamma}}dy = O(h^{2q}) \end{aligned}$$

uniformly over z by Lemma 3.1. The above estimates together with Lemma 3.3 imply the bound (3.6.29):

$$\begin{aligned} E_{t,sr}\left(\left|\frac{K_t}{f_t^2}\right|^{\frac{1}{1-\gamma}} E_{t,sr}\left(\left|\bar{U}_{As}\bar{U}_{Ar}K_{ts}K_{tr}\right|^{\frac{1}{1-\gamma}}|Z_t\right)\right) &= E\left(\left|\frac{K_t}{f_t^2}\right|^{\frac{1}{1-\gamma}}\right)O(h^{2q-\frac{4q}{\theta(1-\gamma)}}) \\ &= O\left(h^q \times h^{2q-\frac{4q}{\theta(1-\gamma)}}\right) = O\left(h^{3q-\frac{4q}{\theta(1-\gamma)}}\right). \end{aligned}$$

The contribution of the case of $\{\tilde{W}_r, W_t, W_s\}$ in M_{T12} can be bounded by

$$\begin{aligned} E_{ts,r}|\phi_T(W_t, W_s, \tilde{W}_r)|^{\frac{1}{1-\gamma}} &= \left(E_{ts,r}\left|\frac{K_t}{f_t^2}\bar{U}_{As}K_{ts}\right|^{\frac{1}{1-\gamma}} E_{ts,r}\left(\left|\bar{U}_{Ar}K_{tr}\right|^{1-\gamma}|Z_t\right)\right) \\ &= O\left(h^{3q-\frac{3q}{\theta(1-\gamma)}}\right). \qquad (3.6.29) \end{aligned}$$

To obtain it, apply Holder's inequality on the inner conditional expectation

$$E_{ts,r}\left(\left|\bar{U}_{Ar}K_{tr}\right|^{1-\gamma}|Z_t = z\right) \leq E_{ts,r}\left(\left|\bar{U}_{Ar}\right|^{\frac{p}{1-\gamma}}\right)^{\frac{1}{p}} E_{ts,r}\left(|K_{tr}|^{\frac{w}{1-\gamma}}\,|Z_t = z\right)^{\frac{1}{w}} = O(h^{q-\frac{q}{\theta(1-\gamma)}}),$$

noting that by Lemma 3.1, $E\left(|K_{tr}|^{\frac{w}{1-\gamma}}\,|Z_t = z\right) = O(h^q)$ uniformly over z, with p defined by $\theta = \frac{p}{1-\gamma}$ and $\frac{1}{w} = 1 - \frac{1}{\theta(1-\gamma)}$. Now, since W_s and W_t are dependent,

$$E_{ts,r}\left(\left|\frac{K_t}{f_t^2}\bar{U}_{As}K_{ts}\right|^{\frac{1}{1-\gamma}}\right) \leq C\left[E_{ts,r}\left(\left|\bar{U}_{As}\right|^{\frac{p}{1-\gamma}}\right)\right]^{\frac{1}{p}}\left[E_{ts,r}\left(\left|K_{ts}\frac{K_t}{f_t^2}\right|^{\frac{w}{1-\gamma}}\right)\right]^{\frac{1}{w}} = O\left(h^{2q-\frac{2q}{\theta(1-\gamma)}}\right),$$

with p, w as above, which yields (3.6.29), and completes the proof of (3.6.27). The contribution to M_{T12} of the case of $(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)$ is not larger than that of the two cases presented above, since the steps to get to the upper bounds in the cases of $\{\tilde{W}_t, W_s, W_r\}$ and $\{\tilde{W}_r, W_t, W_s\}$ apply to the case of $(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)$.

Upper bound on M_{T3}. We obtain $M_{T3} = O\left(h^{2q-\frac{4q}{\theta(1-\gamma)}}\right)$ since under dependence between all three time periods:

$$E\left(\left|\frac{K_t}{f_t^2}\bar{U}_{As}K_{ts}\bar{U}_{Ar}K_{tr}\right|^{\frac{1}{1-\gamma}}\right) \leq C\left[E\left(\left|\bar{U}_{As}\right|^{\frac{2p}{1-\gamma}}\right)\right]^{\frac{1}{p}}\left[E\left(\left|K_{tr}\frac{K_t}{f_t^2}\right|^{\frac{w}{1-\gamma}}\right)\right]^{\frac{1}{w}} = O\left(h^{2q-\frac{4q}{\theta(1-\gamma)}}\right)$$

with p defined by $\theta = 2p/(1-\gamma)$, $\frac{1}{w} = 1 - \frac{2}{\theta(1-\gamma)}$, by Assumption 9, and Lemma 3.4 (i). This rate dominates the contributions from $(\tilde{W}_t, W_s, W_r)$ and $(W_t, W_s, \tilde{W}_r)$ presented above and proves (3.6.28).

Proof of (3.6.18). Since $f(\zeta_l) > 0, l = 1, 2, \cdots, d$, for T large enough, there exists a constant $c > 0$ such that $\min\limits_{t:K_t\neq 0} f(Z_t) \geq c$, due to the boundedness of the support of the kernel $K_l(\cdot; h)$ and continuity of $f(\cdot)$ and $h \to 0$. Now,

$$\max_{t:K_t\neq 0}\left|\frac{f_t^2 - \tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| \leq \max_{t:K_t\neq 0}\left|f_t^2 - \tilde{f}_t^2\right| \max_{t:K_t\neq 0}\left|\frac{1}{f_t^2}\right| \max_{t:K_t\neq 0}\left|\frac{1}{\tilde{f}_t^2}\right|.$$

The second term is a random variable bounded by a finite constant for T sufficiently large. As for the third term, as $T \to \infty$,

$$\max_{t:K_t\neq 0}\left|\frac{1}{\tilde{f}_t^2}\right| = \frac{1}{\min\limits_{t:K_t\neq 0}|\tilde{f}_t^2|} = O_p(1),$$

because $\min\limits_{t:K_t\neq 0}|\tilde{f}_t^2| \geq \min\limits_{t:K_t\neq 0}|f_t^2| - \max\limits_{t:K_t\neq 0}|\tilde{f}_t^2 - f_t^2| = \min\limits_{t:K_t\neq 0}|f_t| + o_p(1) \geq c + o_p(1) = c(1 + o_p(1))$. Now,

$$\begin{aligned}\max_{t:K_t\neq 0}\left|f_t^2 - \tilde{f}_t^2\right| &= \max_{t:K_t\neq 0}\left|(f_t - \tilde{f}_t)^2 + 2\tilde{f}_t(f_t - \tilde{f}_t)\right| \\ &\leq \left[\max_{t:K_t\neq 0}\left|f_t - \tilde{f}_t\right|\right]^2 + 2\max_{t:K_t\neq 0}|f_t|\max_{t:K_t\neq 0}|f_t - \tilde{f}_t| = O_p\left(\left(\frac{\log T}{Th^q}\right)^{1/2} + h^s\right),\end{aligned}$$

since by (3.6.13),

$$\max_{t:K_t\neq 0}\left|f_t - \tilde{f}_t\right| \leq \sup\left|f(z) - \tilde{f}(z)\right| = O_p\left(\left(\frac{\log T}{Th^q}\right)^{1/2} + h^s\right).$$

Therefore,

$$\max_{t:K_t\neq 0}\left|\frac{f_t^2 - \tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| = O_p\left(\left(\frac{\log T}{Th^q}\right)^{1/2} + h^s\right) = O_p(1).$$

Proof of (3.6.17) Notice the similarity of $E\mathbf{A}''_{\mathbf{T}} = \frac{1}{Th^q}\sum_{t=1}^{T}[|K_l(Z_t; h)|n_t^2]$ to $E\mathbf{A}'_{\mathbf{T}}$ in (3.6.15). Compared to equation (3.6.15), the difference is that we have $|K_l(Z_t; h)|$ instead of $|K_l(Z_t; h)|/f_t$ in the kernel of the U-statistic. The steps to get the upper bound for this terms is almost identical to the steps for $\mathbf{A}'_{\mathbf{T}}$, with results such as

$E|K_l(Z_t;h)|^a = O(h^q)$ replacing their corresponding ones, such as $E\left|\frac{K_l(Z_t;h)}{f_t^2}\right|^a = O(h^q)$, in the proof and yields the same upper bound as for $\mathbf{A'_T}$.

3.6.2 Upper bound on $\mathbf{B_T}$.

In this section, we will show that

$$\mathbf{B_T} = O_p(r_{2T}), \quad \text{where} \quad r_{2T} := \left(\frac{1}{Th^q}\right)^3 \left[T^3h^{3q+2s} + T^2h^{2q+2} + T^2h^{3q(1-\gamma)+2}\right], \tag{3.6.30}$$

which also implies (3.6.10) for $\mathbf{B_T}$. Since $\frac{1}{\tilde{f}_t^2} = \frac{1}{f_t^2} + \frac{(f_t^2-\tilde{f}_t^2)}{f_t^2\tilde{f}_t^2}$,

$$\begin{aligned}
\mathbf{B_T} &= \frac{1}{Th^q}\sum_{t=1}^T |K_t|\frac{l_t^2}{\tilde{f}_t^2} \le \frac{1}{Th^q}\sum_{t=1}^T |K_t|\frac{l_t^2}{f_t^2} + \max_{t:K_t\neq 0}\left|\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| \frac{1}{Th^q}\sum_{t=1}^T |K_t| l_t^2 \\
&=: \mathbf{B'_T} + \max_{t:K_t\neq 0}\left|\frac{f_t^2-\tilde{f}_t^2}{f_t^2\tilde{f}_t^2}\right| \mathbf{B''_T} = \mathbf{B'_T} + O_p(1)\mathbf{B''_T},
\end{aligned}$$

by (3.6.18). We will show that

$$E\mathbf{B'_T} = O(r_{2T}), \qquad E\mathbf{B''_T} = O(r_{2T}),$$

which because of non-negativity of $\mathbf{B'_T}$ and $\mathbf{B''_T}$ implies (3.6.30). Firstly, the stochastic order of $\mathbf{B'_T}$ can be found by studying

$$\begin{aligned}
E(\mathbf{B'_T}) &= \left(\frac{1}{Th^q}\right)^3 E\Big(\sum_{t_1,t_2=1}^{T}{}' \frac{|K_l(Z_{t_1};h)|}{f_{t_1}^2}\{m(Z_{t_1}) - m(Z_{t_2})\}^2 K_{12}^2\Big) \\
&+ \left(\frac{1}{Th^q}\right)^3 E\Big(\sum_{t_1,t_2,t_3=1}^{T}{}' \frac{|K_l(Z_{t_1};h)|}{f_{t_1}^2}\{m(Z_{t_1}) - m(Z_{t_2})\}K_{12}\{m(Z_{t_1}) - m(Z_{t_3})\}K_{13}\Big) \\
&=: \left(\frac{1}{Th^q}\right)^3 (B_{1T} + B_{2T}).
\end{aligned}$$

Upper bound on B_{1T}. We will show

$$B_{1T} = O\left(T^2h^{2q+2} + Th^{2q(1-\gamma)+2}\right), \tag{3.6.31}$$

$$B_{2T} = O\left(T^3h^{2q+2s} + T^2h^{3q(1-\gamma)+2}\right). \tag{3.6.32}$$

B_{1T} is expectation of a second order U-statistic with kernel

$$\phi_T(W_t, W_s) = \frac{|K_t|}{f_t^2}\{m(Z_t) - m(Z_s)\}^2 K_{ts}^2.$$

By Lemma 3.6 (i),

$$|B_{1T}| \leq CT^2|E\phi_T(\tilde{W}_t, \tilde{W}_s)| + CTM_{T2}^{1-\gamma}. \tag{3.6.33}$$

To prove (3.6.31), we show that

$$|E\phi_T(\tilde{W}_t, \tilde{W}_s)| \leq Ch^{2q+2}, \quad \text{and} \quad M_{T2} \leq Ch^{2q+\frac{2}{1-\gamma}}. \tag{3.6.34}$$

The expectation under independence is

$$\begin{aligned} E(\phi_T(\tilde{W}_t, \tilde{W}_s)) &= E^* \left(\frac{|K_t|}{f_t^2} \{m(Z_t) - m(Z_s)\}^2 K_{ts}^2 \right) \\ &= E^* \left(\frac{|K_t|}{f_t^2} E^* \left(\{m(Z_t) - m(Z_s)\}^2 K_{ts}^2 | Z_t \right) \right) = O(h^{2q+2}), \end{aligned}$$

by Lemmas 3.2 and 3.3. To bound M_{2T}, similar to A_{1T}:

$$\begin{aligned} M_{2T} \leq \quad & E\left(\left| \frac{K_t}{f_t^2} \right|^{\frac{1}{1-\gamma}} |\{m(Z_t) - m(Z_s)\}K_{ts}|^{\frac{2}{1-\gamma}} \right) \\ & + E^* \left(\left| \frac{K_t}{f_t^2} \right|^{\frac{1}{1-\gamma}} |\{m(Z_t) - m(Z_s)\}K_{ts}|^{\frac{2}{1-\gamma}} \right) = O\left(h^{2q+\frac{2}{1-\gamma}} \right), \end{aligned}$$

by Lemma 3.4 (iii), which proves (3.6.34).

Upper bound on B_{2T}. To bound B_{2T}, we will show

$$B_{2T} = O\left(T^3 h^{3q+2s} + T^2 h^{3q(1-\gamma)+2} + Th^{2q(1-\gamma)+2} \right) = O\left(T^3 h^{3q+2s} + T^2 h^{3q(1-\gamma)+2} \right), \tag{3.6.35}$$

where the last inequality follows from Assumption 17, $Th^{1-\gamma} \to \infty$, which yields (3.6.31). B_{2T} is a third order U-statistic with the kernel

$$\phi_T(W_t, W_s, W_r) = \frac{|K_t|}{f_t^2} \{m(Z_t) - m(Z_s)\}K_{ts}\{m(Z_t) - m(Z_r)\}K_{tr}. \tag{3.6.36}$$

By Lemma 3.6 (ii),

$$|B_{2T}| \leq T^3|E(\phi_T(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)| + C(T^2 M_{T12}^{1-\gamma} + TM_{T3}^{1-\gamma}). \tag{3.6.37}$$

To prove (3.6.35), we shall show that

$$|E(\phi_T(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)| \leq Ch^{2q+2s}, \tag{3.6.38}$$

$$M_{T12} \leq Ch^{3q+\frac{2}{1-\gamma}}, \tag{3.6.39}$$

$$M_{T3} \leq Ch^{2q+\frac{2}{1-\gamma}}. \tag{3.6.40}$$

Expectation under independence is

$$
\begin{aligned}
|E(\phi_T(\tilde{W}_t, \tilde{W}_s, \tilde{W}_r)| &= |E^* \left(\frac{|K_t|}{f_t^2} \{m(Z_t) - m(Z_s)\} K_{ts} \{m(Z_t) - m(Z_r)\} K_{tr} \right) | \\
&\leq E^* \left(\left| \frac{K_t}{f_t^2} \right| \left| E^*(\{m(Z_t) - m(Z_s)\} K_{ts} \middle| Z_t) \right| \left| E^*(\{m(Z_t) - m(Z_r)\} K_{tr} | \, Z_t) \right| \right) \\
&\leq C h^{2(q+s)} E^* \left(\left| \frac{K_t}{f_t^2} \right| \right) = O\left(h^{3q+2s}\right),
\end{aligned}
$$

by Lemma 3.2 (i) and Lemma 3.3. Obtaining the bound (3.6.39) for M_{T12} follows the same steps as in case of A_{2T} above. To find upper bound on M_{T12}, due to the symmetry between W_s and W_r in (3.6.36), it suffices to consider two distinct cases when there are two independent subsets.

For $(W_s, W_r, \tilde{W}_t)$,

$$
\begin{aligned}
& E_{sr,t} \left[\left| \frac{K_t}{f_t^2} \right|^{\frac{1}{1-\gamma}} E_{sr,t} \left(|\{m(Z_t) - m(Z_s)\} K_{ts} \{m(Z_t) - m(Z_r)\} K_{tr}|^{\frac{1}{1-\gamma}} \middle| Z_t \right) \right] \\
& \leq C h^{2q + \frac{2}{1-\gamma}} E \left[\left| \frac{K_t}{f_t^2} \right|^{\frac{1}{1-\gamma}} \right] = O\left(h^{3q + \frac{2}{1-\gamma}}\right),
\end{aligned}
$$

because uniformly over z, under Assumption 4, by Lemma 3.1

$$
\begin{aligned}
& E_{sr,t} \left(|\{m(Z_t) - m(Z_s)\} K_{ts} \{m(Z_t) - m(Z_r)\} K_{tr}|^{\frac{1}{1-\gamma}} \middle| Z_t \right) \\
& \leq \sup_{w,y} f_{|s-t|}(w, y) \int |\{m(z) - m(w)\} K(\frac{z-w}{h})|^{\frac{1}{1-\gamma}} dw \\
& \qquad \int |\{m(z) - m(y)\} K(\frac{z-y}{h})|^{\frac{1}{1-\gamma}} dy \\
& \leq C \big[\int \|\psi\|^{\frac{1}{1-\gamma}} K(\psi) d\psi \big]^2 = O\left(h^{2q + \frac{2}{1-\gamma}}\right).
\end{aligned}
$$

For $(W_t, W_r, \tilde{W}_s)$,

$$
\begin{aligned}
& E_{tr,s} \left[\left| \frac{K_t}{f_t^2} \{m(Z_t) - m(Z_r)\} K_{tr} \right|^{\frac{1}{1-\gamma}} E_{tr,s} \left(|\{m(Z_t) - m(Z_s)\} K_{ts}|^{\frac{1}{1-\gamma}} \middle| Z_t \right) \right] \\
& \leq C h^{q + \frac{1}{1-\gamma}} E_{tr,s} \left(\left| \frac{K_t}{f_t^2} \{m(Z_t) - m(Z_r)\} K_{tr} \right|^{\frac{1}{1-\gamma}} \right) = O\left(h^{3q + \frac{2}{1-\gamma}}\right),
\end{aligned}
$$

by Lemma 3.2 and then applying Lemma 3.4 (iii), which completes the proof of (3.6.34). The same upper bound is applicable in the case of $(\tilde{W}_r, \tilde{W}_s, \tilde{W}_t)$ as all the steps taken above apply in that case.

Proof of (3.6.40). Under dependence across all three time periods,

$$\begin{aligned} M_{T3} &= E\left[\left|\frac{K_t}{f_t^2}\{m(Z_t)-m(Z_r)\}K_{tr}\{m(Z_t)-m(Z_s)\}K_{ts}\right|^{\frac{1}{1-\gamma}}\right] \\ &\le \left[E\left|\frac{K_t}{f_t}\{m(Z_t)-m(Z_r)\}K_{tr}\right|^{\frac{2}{1-\gamma}}\right]^{1/2}\left[E\left|\frac{K_t}{f_t}\{m(Z_t)-m(Z_s)\}K_{ts}\right|^{\frac{2}{1-\gamma}}\right]^{1/2} \\ &= O\left(h^{2q+\frac{2}{1-\gamma}}\right), \end{aligned}$$

by Lemma 3.4 (iii), which yields (3.6.40) and completes the proof of (3.6.32).

Upper bound on $\mathbf{B''_T}$. Noting the similarity of $\mathbf{B''_T} = \frac{1}{Th^q}E\left(\sum_{t=1}^{T}|K_t|l_t^2\right)$ to $\mathbf{B'_T}$, all the steps of finding upper bound of $\mathbf{B''_T}$ yield the same bound as for $\mathbf{B'_T}$. This completes the proof of (3.6.30).

3.6.3 Upper bound on $\mathbf{C_T}$.

Recall that by (3.6.12),

$$\begin{aligned} \mathbf{C_T} &= \left|\frac{1}{Th^q}\sum_{t=1}^{T}K_tU_{it}\frac{l_t}{\tilde{f}_t}\right| \\ &\le \left|\frac{1}{Th^q}\sum_{t=1}^{T}K_tU_{it}\frac{l_t}{f_t}\right| + \left|\frac{1}{Th^q}\sum_{t=1}^{T}K_tU_{it}l_t\frac{f_t-\tilde{f}_t}{\tilde{f}_tf_t}\right| =: \mathbf{C'_T}+\mathbf{C''_T}. \end{aligned}$$

We shall show that

$$\begin{aligned} \mathbf{C'_T} &= O_p(r_{3T}), && (3.6.41) \\ \mathbf{C''_T} &= O_p(r_{2T}+h^{2s-\frac{2q}{\theta}}+\frac{\log T}{Th^{q+\frac{2q}{\theta}}}), && (3.6.42) \\ r_{3T} &:= (\frac{1}{Th^q})^2\left(T^3h^{2+3q(1-\frac{2}{\theta})}+T^3h^{4q(1-\gamma)-2(\frac{2q}{\theta}-1)}+T^2h^{2q(1-\gamma)-2(\frac{2q}{\theta}-1)}\right)^{1/2}, \end{aligned}$$

which implies (3.6.10) for $\mathbf{C}_T$.

Proof of (3.6.42). Using inequality $|ab| \le a^2+b^2$,

$$\begin{aligned} \mathbf{C''_T} &\le \frac{1}{Th^q}\sum_{t=1}^{T}|K_t|\{|U_{it}\frac{f_t-\tilde{f}_t}{f_t}|^2+(\frac{l_t}{\tilde{f}_t})^2\} \\ &\le \max_{t:K_t\neq 0}|\frac{f_t-\tilde{f}_t}{f_t}|^2(\frac{1}{Th^q}\sum_{t=1}^{T}|K_t|U_{it}^2)+\mathbf{B_T} =: I_T\cdot V_T+\mathbf{B_T}. \quad (3.6.43) \end{aligned}$$

By (3.6.30), $\mathbf{B_T} = O_p(r_{2T})$. Next, to bound V_T, note that by Lemma 3.3,

$$E[|K_t|U_{it}^2] \le (E|K_t|^2)^{1/2}(E|U_{it}|^4)^{1/2} \le Ch^{q-\frac{2q}{\theta}}.$$

Then $V_T \leq Ch^{-2q/\theta}$, and since $I_T = O_p(\frac{\log T}{Th^q} + h^{2s})$,

$$I_T V_T \leq O_p(\frac{\log T}{Th^{q+\frac{2q}{\theta}}} + h^{2s-\frac{2q}{\theta}}), \tag{3.6.44}$$

which proves (3.6.42).

Proof of (3.6.41). Since $\mathbf{C'_T} = O_p([E(\mathbf{C'_T})^2]^{\frac{1}{2}})$, we show that

$$E[(\mathbf{C'_T})^2] \leq C(\frac{1}{Th^q})^4(T^3h^{3q+2s-\frac{2q}{\theta}} + T^3h^{4q(1-\gamma)-2(\frac{2q}{\theta}-1)} + T^2h^{2q(1-\gamma)-2(\frac{2q}{\theta}-1)}), \tag{3.6.45}$$

which implies (3.6.41).

Write

$$\begin{aligned}
&E[(\mathbf{C'_T})^2] \\
&= (\frac{1}{Th^q})^4 \sum_{t_1,t_2=1}^{T}{}' \sum_{t_3,t_4=1}^{T}{}' E\left(\frac{K_{t_1}}{f_{t_1}}\frac{K_{t_3}}{f_{t_3}}U_{it_1}U_{it_3}K_{t_1t_2}K_{t_3t_4}\{m(Z_{t_1})-m(Z_{t_2})\}\{m(Z_{t_3})-m(Z_{t_4})\}\right) \\
&\quad = (\frac{1}{Th^q})^4 \sum_{t_1,t_2=1}^{T}{}' \sum_{t_3,t_4=1}^{T}{}' \{1_{I_1}E[\cdots] + 1_{I_2}E[\cdots] + 1_{I_3}E[\cdots]\} =: (C_{1T} + C_{2T} + C_{3T}),
\end{aligned}$$

where $I_1 \cup I_2 \cup I_3 = [1, \cdots, T]^4$,

$$\begin{aligned}
I_1 &= \{(t_1 = t_3, t_2 = t_4), (t_1 = t_4, t_2 = t_3)\}, \\
I_2 &= \{(t_1 = t_3, t_2 \neq t_4), (t_1 = t_4, t_2 \neq t_3), (t_3 = t_2, t_1 \neq t_4), (t_2 = t_4, t_1 \neq t_3)\} \\
I_3 &= \{(t_1 \neq t_3, t_2 \neq t_4)\}.
\end{aligned}$$

We will show that

$$\begin{aligned}
C_{1T} &= O\big((\frac{1}{Th^q})^4(T^2h^{2+2q(1-\frac{2}{\theta})})\big), && (3.6.46) \\
C_{2T} &= O\big((\frac{1}{Th^q})^4(T^3h^{2+3q(1-\frac{2}{\theta})})\big), && (3.6.47) \\
C_{3T} &= O\big((\frac{1}{Th^q})^4\left(T^3h^{4q(1-\gamma)-2(\frac{2q}{\theta}-1)} + T^2h^{2q(1-\gamma)-2(\frac{q}{\theta}-1)}\right)\big), && (3.6.48)
\end{aligned}$$

which proves (3.6.45).

Proof of (3.6.46) for C_{1T}. Since

$$\left|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\right| \leq \left(\frac{K_t}{f_t}U_{it}\right)^2 + \left(\frac{K_s}{f_s}U_{is}\right)^2,$$

then,

$$
\begin{aligned}
C_{1T} &\leq \sum_{t,s=1}^{T} E\big(\frac{K_t^2}{f_t^2}U_{it}^2K_{ts}^2\{m(Z_t)-m(Z_s)\}^2 \\
&\qquad + \left|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\right| K_{ts}^2\{m(Z_t)-m(Z_s)\}^2\big) \\
&\leq 3\sum_{t,s=1}^{T} E\big(\frac{K_t^2}{f_t^2}U_{it}^2K_{ts}^2\{m(Z_t)-m(Z_s)\}^2\big) \\
&\leq CT^2h^{2+2q(1-\frac{2}{\theta})},
\end{aligned}
$$

because with p such that $2p=\theta$, and $\frac{1}{r}=1-\frac{1}{p}=1-\frac{2}{\theta}$,

$$
\begin{aligned}
&\sum_{t,s=1}^{T} E\big(\frac{K_t^2}{f_t^2}U_{it}^2K_{ts}^2\{m(Z_t)-m(Z_s)\}^2\big) \\
&\leq \big(E\left|\frac{K_t}{f_t}K_{ts}\{m(Z_t)-m(Z_s)\}\right|^{2r}\big)^{1/r}\big(E\,|U_{it}|^{2p}\big)^{1/p} \\
&\leq C(h^{2q+2r})^{1/2} = Ch^{\frac{2q}{r}+2} = Ch^{2+2q(1-\frac{2}{\theta})},
\end{aligned}
$$

by Lemma 3.4 (iii) which proves (3.6.46).

Proof of (3.6.47) for C_{2T}. It suffices to show that

$$
\begin{aligned}
&E\left(1_{I_2}E\left|\frac{K_{t_1}}{f_{t_1}}\frac{K_{t_3}}{f_{t_3}}U_{it_1}U_{it_3}K_{t_1t_2}K_{t_3t_4}\{m(Z_{t_1})-m(Z_{t_2})\}\{m(Z_{t_3})-m(Z_{t_4})\}\right|\right) \\
&\quad \leq Ch^{2+3q(1-\frac{2}{\theta})}. \qquad (3.6.49)
\end{aligned}
$$

According to definition of I_2, we need to check (3.6.49) in four cases.

Case 1, $(t_1=t_3, t_2\neq t_4)$. Then, the above expectation becomes

$$
\begin{aligned}
&E\big(\frac{K_t^2}{f_t^2}U_{it}^2|K_{ts}K_{tr}\{m(Z_t)-m(Z_s)\}\{m(Z_t)-m(Z_r)\}|\big) \\
&\leq \big(E\left|\frac{K_t^2}{f_t^2}K_{ts}K_{tr}\{m(Z_t)-m(Z_s)\}\{m(Z_t)-m(Z_r)\}\right|^{w}\big)^{1/w}\big(E\,|U_{it}|^{2p}\big)^{1/p} \\
&\leq C(h^{\frac{(3q+2w)}{w}}) = Ch^{2+3q(1-\frac{2}{\theta})}, \qquad (3.6.50)
\end{aligned}
$$

selecting p such that $2p=\theta$, setting $\frac{1}{w}=1-\frac{2}{\theta}$, and using Lemma 3.4 (iv) and Assumption 9.

Case 2, $(t_1=t_4, t_2\neq t_3)$. Then the expectation on the LHS of (3.6.49) is

$$
E\big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{rt}\{m(Z_t)-m(Z_s)\}\{m(Z_r)-m(Z_t)\}\big|. \qquad (3.6.51)
$$

Since $\big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\big| \leq \big(\frac{K_t}{f_t}U_{it}\big)^2+\big(\frac{K_s}{f_s}U_{is}\big)^2$, the bound (3.6.49) follows similarly as

(3.6.50).

Case 3, $(t_3 = t_2, t_1 \neq t_4)$. Here, the expectation of the LHS of (3.6.49) is

$$E\Big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{sr}\{m(Z_t) - m(Z_s)\}\{m(Z_s) - m(Z_r)\}\Big|,$$

and (3.6.49) follows by the same argument as in Case 2.

Case 4, $(t_2 = t_4, t_1 \neq t_3)$. Here, the expectation of the LHS of (3.6.49) is

$$E\Big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{sr}\{m(Z_t) - m(Z_s)\}\{m(Z_s) - m(Z_r)\}\Big|,$$

and (3.6.49) follows the same argument as in Case 2.

Upper bound on C_{3T}. Next, we bound C_{3T}. We will show that

$$C_{3T} = O\left(T^3 h^{4q(1-\gamma)-2(\frac{2q}{\theta}-1)} + T^2 h^{2q(1-\gamma)-2(\frac{q}{\theta}-1)}\right). \tag{3.6.52}$$

C_{3T} is the expectation of a fourth order U statistic, whose kernel is

$$\phi_T(W_t, W_s, W_r, W_u) = \frac{K_t}{f_t}\frac{K_r}{f_r}U_{it}U_{ir}K_{ts}K_{ru}\{m(Z_t) - m(Z_s)\}\{m(Z_r) - m(Z_u)\}.$$

By Lemma 3.6 (iii),

$$|C_{3T}| = T^4|E\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3, \tilde{W}_4)| + C[T^3 M_{T112}^{1-\gamma} + T^2 M_{T13}^{1-\gamma} + T^2 M_{T4}^{1-\gamma}].$$

Expectation under independence is zero:

$$\begin{aligned}
E[\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3, \tilde{W}_4)] &= E^*\left(\frac{K_t}{f_t}\frac{K_r}{f_r}U_{it}U_{ir}K_{ts}K_{ru}\{m(Z_t) - m(Z_s)\}\{m(Z_r) - m(Z_u)\}\right) \\
&= E^*\left(\frac{K_t}{f_t}K_{ts}\{m(Z_t) - m(Z_s)\}E^*(U_{it}|Z_t, Z_s)\right) \\
&\quad \times E^*\left(\frac{K_r}{f_r}K_{ru}\{m(Z_r) - m(Z_u)\}E^*(U_{ir}|Z_r, Z_u)\right) = 0,
\end{aligned}$$

by Assumption 2.

We will show that

$$M_{T112} \leq Ch^{4q-\frac{2}{1-\gamma}(\frac{2q}{\theta}-1)}, \tag{3.6.53}$$

$$M_{T13}, M_{T4} \leq Ch^{2q-\frac{2}{1-\gamma}(\frac{2q}{\theta}-1)}, \tag{3.6.54}$$

which proves (3.6.52).

Proof of (3.6.53). As noted in Lemma 3.6 (iii), M_{T112} is the maximal $\frac{1}{(1-\gamma)}^{th}$ moment quantity when partitioning the four time periods into either three or four independent subsets. There are three distinct combinations of dependence to be considered in the case of three independent subsets.

For $(W_r, W_u, \tilde{W}_t, \tilde{W}_s)$, one can separate out expectations,

$$\begin{aligned}
E_{ru,t,s} \quad & \left[\left| \frac{K_t}{f_t} U_{it} \right|^{\frac{1}{1-\gamma}} E_{ru,t,s} \left(|K_{ts}\{m(Z_t) - m(Z_s)\}|^{\frac{1}{1-\gamma}} \, |Z_t \right) \right] \\
\times \quad & E_{ru,t,s} \left[\left| \frac{K_r}{f_r} U_{ir} K_{ru}\{m(Z_r) - m(Z_u)\} \right|^{\frac{1}{1-\gamma}} \right] \\
\leq \quad & h^{q+\frac{1}{1-\gamma}} \times h^{q/w} \times h^{(2q+\frac{w}{1-\gamma})/w} = O\left(h^{4q-\frac{1}{1-\gamma}\left(\frac{3q}{\theta}-2\right)} \right),
\end{aligned}$$

because by Lemma 3.2 (ii), Lemma 3.3, and Holder's inequality with Assumption 9', setting p such that $\frac{p}{1-\gamma} = \theta$ and $\frac{1}{w} = 1 - \frac{1}{p} = 1 - \frac{1}{\theta(1-\gamma)}$:

$$E_{ru,t,s}[|K_{ts}\{m(Z_t) - m(Z_s)\}|^{\frac{1}{1-\gamma}} |Z_t] = O(h^{q+\frac{1}{1-\gamma}}), \tag{3.6.55}$$

$$E|\frac{K_t}{f_t} U_{it}|^{\frac{1}{1-\gamma}} \leq (E|U_{it}|^{\frac{p}{1-\gamma}})^{1/p} (E|\frac{K_t}{f_t}|^{\frac{w}{1-\gamma}})^{1/w} = O(h^{\frac{q}{w}}), \tag{3.6.56}$$

and by Lemma 3.4 (iii),

$$\begin{aligned}
& E_{ru,t,s}|\frac{K_r}{f_r} U_{ir} K_{ru}\{m(Z_r) - m(Z_u)\}|^{\frac{1}{1-\gamma}} \leq (E|U_{ir}|^{\frac{p}{1-\gamma}})^{1/p} \\
& \quad \times (E_{ru,t,s}|\frac{K_r}{f_r} K_{ru}\{m(Z_r) - m(Z_u)\}|^{\frac{w}{1-\gamma}})^{1/w} = O(h^{(2q+\frac{w}{1-\gamma})\frac{1}{w}}) = O(h^{\frac{2q}{w}+\frac{1}{1-\gamma}}),
\end{aligned}$$

with $p = \theta(1-\gamma)$ and $\frac{1}{w} = 1 - \frac{1}{\theta(1-\gamma)}$.

For $(W_s, W_u, \tilde{W}_t, \tilde{W}_r)$, the $\frac{1}{(1-\gamma)}^{th}$ moment of the kernel is

$$\begin{aligned}
E_{su,t,r}\{ \quad & \left| \frac{K_t U_{it}}{f_t} \frac{K_r U_{ir}}{f_r} \right|^{\frac{1}{1-\gamma}} \\
\times \quad & E_{su,t,r} \left(|K_{ts}\{m(Z_t) - m(Z_s)\} K_{ru}\{m(Z_r) - m(Z_u)\}|^{\frac{1}{1-\gamma}} \, |Z_t, Z_r \right) \quad \} \\
& \leq C E_{su,t,r} \left| \frac{K_t U_{it}}{f_t} \frac{K_r U_{ir}}{f_r} \right|^{\frac{1}{1-\gamma}} \cdot h^{2q+\frac{2}{1-\gamma}} = O\left(h^{4q-\frac{2}{1-\gamma}\left(\frac{q}{\theta}-1\right)} \right).
\end{aligned} \tag{3.6.57}$$

because the inner conditional expectation evaluated at $Z_t = z$ and $Z_r = u$ is

$$\begin{aligned}
& E_{su,t,r} \left(|K_{ts}\{m(Z_t) - m(Z_s)\} K_{ru}\{m(Z_r) - m(Z_u)\}|^{\frac{1}{1-\gamma}} \, |Z_t = z, Z_s = u \right) \\
& \leq \sup_{w,y} f_{|u-s|}(w, y) \int \left| K\left(\frac{w-z}{h} \right) \{m(z) - m(w)\} \right|^{\frac{1}{1-\gamma}} dw \\
& \quad \times \int \left| K\left(\frac{y-z}{h} \right) \{m(z) - m(y)\} \right|^{\frac{1}{1-\gamma}} dy = O(h^{2q+\frac{2}{1-\gamma}})
\end{aligned}$$

uniformly over z and u due to Lemma 3.1. Noting the independence between $\tilde{W}_t$ and

$\tilde{W}_r$, by (3.6.56),

$$\begin{aligned} E_{su,t,r}\left(\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}}\right) &= E\left(\left|\frac{K_tU_{it}}{f_t}\right|^{\frac{1}{1-\gamma}}\right)E\left(\left|\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}}\right) \\ &= O(h^{\frac{2q}{w}}) = O\left(h^{2q(1-\frac{1}{\theta(1-\gamma)})}\right). \end{aligned}$$

For $(W_t, W_r, \tilde{W}_s, \tilde{W}_u)$, by (3.6.55),

$$\begin{aligned} & E_{tr,s,u}\{ \quad \left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}} E_{tr,s,u}\left(|K_{ts}\{m(Z_t)-m(Z_s)\}|^{\frac{1}{1-\gamma}}|Z_t\right) \\ & \quad \times E_{tr,s,u}\left(|K_{ru}\{m(Z_r)-m(Z_u)\}|^{\frac{1}{1-\gamma}}|Z_r\right) \quad \} \\ & \leq Ch^{2(q+\frac{1}{1-\gamma})}E_{tr,s,u}\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}} \\ & = O\left(h^{2q(1-\frac{2}{\theta(1-\gamma)})}\times h^{\frac{2q(\theta-\frac{2}{1-\gamma})}{\theta}}\right) = O\left(h^{4q-\frac{2}{1-\gamma}\left(\frac{2q}{\theta}-1\right)}\right), \end{aligned}$$

since by Lemma 3.4 (ii),

$$E\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}} \leq (E|U_{it}U_{ir}|^{\frac{p}{1-\gamma}})^{1/p}(E|\frac{K_t}{f_t}\frac{K_r}{f_r}|^{\frac{w}{1-\gamma}})^{1/w} = O(h^{\frac{2q}{w}}) = O(h^{2q-\frac{4q}{\theta(1-\gamma)}}), \quad (3.6.58)$$

setting $\frac{2p}{1-\gamma} = \theta$ and $\frac{1}{w} = 1 - \frac{2}{\theta(1-\gamma)}$. This proves (3.6.53).

Upper bound on M_{T13} and M_{T4}. For both M_{T13} and M_{T4}, one finds the upper bound that holds for all relevant combinations of dependence:

$$\begin{aligned} & E\left[\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}K_{ts}\{m(Z_t)-m(Z_s)\}K_{ru}\{m(Z_r)-m(Z_u)\}\right|^{\frac{1}{1-\gamma}}\right] \\ & \leq \left(E\left|\frac{K_tU_{it}}{f_t}K_{ts}\{m(Z_t)-m(Z_s)\}\right|^{\frac{2}{1-\gamma}} E\left|\frac{K_rU_{ir}}{f_r}K_{ru}\{m(Z_r)-m(Z_u)\}\right|^{\frac{2}{1-\gamma}}\right)^{1/2} \\ & \leq (E\left|\frac{K_t}{f_t}K_{ts}\{m(Z_t)-m(Z_s)\}\right|^{\frac{p}{1-\gamma}})^{1/2p}(E\,|U_{it}|^{\frac{2w}{1-\gamma}})^{1/w} \\ & \quad \times(E\left|\frac{K_r}{f_r}K_{ru}\{m(Z_r)-m(Z_u)\}\right|^{\frac{p}{1-\gamma}})^{1/2p}(E\,|U_r|^{\frac{2w}{1-\gamma}})^{1/w} \\ & \quad = h^{2q+\frac{2w}{1-\gamma}} = O\left(h^{2q-\frac{2}{1-\gamma}\left(\frac{2q}{\theta}-1\right)}\right), \end{aligned}$$

by setting $2p/(1-\gamma) = \theta$ and $\frac{1}{w} = 1 - \frac{2}{\theta(1-\gamma)}$ and Lemma 3.4 (iii), which proves (3.6.57).

3.6.4 Upper bound on $\mathbf{D_T}$.

By (3.6.8) and (3.6.12),

$$\begin{aligned}\mathbf{D_T} &= \left|\frac{1}{Th^q}\sum_{t=1}^{T} K_t U_{it}\frac{n_t}{\tilde{f}_t}\right| \\ &\leq \left|\frac{1}{Th^q}\sum_{t=1}^{T} K_t U_{it}\frac{n_t}{f_t}\right| + \left|\frac{1}{Th^q}\sum_{t=1}^{T} K_t U_{it} n_t \frac{f_t - \tilde{f}_t}{\tilde{f}_t f_t}\right| =: \mathbf{D_T}' + \mathbf{D_T}''.\end{aligned}$$

We will show that

$$\begin{aligned}\mathbf{D'_T} &= O_p(r_{4T}), \quad r_{4T} := \big(\frac{1}{Th^q}\big)^2 \big(T^3 h^{3q(1-\frac{4}{\theta})} + T^2 h^{3q-\frac{12q}{\theta(1-\gamma)}}\big)^{1/2}, &(3.6.59)\\ \mathbf{D''_T} &= O_p(r_{1T} + \frac{\log T}{Th^{q+\frac{2q}{\theta}}} + h^{2s-\frac{2q}{\theta}}), &(3.6.60)\end{aligned}$$

where r_{1T} is the same as in (3.6.14), which proves (3.6.22) for D_T.

Proof of (3.6.60). Similarly as in the proof of (3.6.42),

$$\mathbf{D''_T} \leq \frac{1}{Th^q}\sum_{t=1}^{T} |K_t|\{\left|U_{it}\frac{f_t - \tilde{f}_t}{\tilde{f}_t f_t}\right|^2 + \frac{n_t^2}{f_t^2}\} \leq I_T V_T + A_T.$$

By (3.6.14), $A_T = O_p(r_{1T})$ which together with (3.6.44) implies $\mathbf{D''_T} = O_p(r_{1T} + \frac{\log T}{Th^{q+\frac{2q}{\theta}}} + h^{2s-\frac{2q}{\theta}})$ proving (3.6.60).

Proof of (3.6.59). Since $\mathbf{D'_T} = O_p([E(\mathbf{D'_T})^2]^{\frac{1}{2}})$, we show that

$$E[(\mathbf{D'_T})^2] \leq C\big(\frac{1}{Th^q}\big)^4 \big(T^3 h^{3q(1-\frac{4}{\theta})} + T^2 h^{2q-\frac{8q}{\theta(1-\gamma)}}\big), \tag{3.6.61}$$

which implies (3.6.59).

Write

$$\begin{aligned}&E[(\mathbf{D'_T})^2] \\ &= \big(\frac{1}{Th^q}\big)^4 {\sum_{t_1,t_2=1}^{T}}' {\sum_{t_3,t_4=1}^{T}}' E\left(\frac{K_{t_1}}{f_{t_1}}\frac{K_{t_3}}{f_{t_3}} U_{it_1} U_{it_3} K_{t_1t_2} K_{t_3t_4} \bar{U}_{At_2} \bar{U}_{At_4}\right) \\ &\quad = \big(\frac{1}{Th^q}\big)^4 {\sum_{t_1,t_2=1}^{T}}' {\sum_{t_3,t_4=1}^{T}}' \{1_{I_1} E[\cdots] + 1_{I_2} E[\cdots] + 1_{I_3} E[\cdots]\} =: (D_{1T} + D_{2T} + D_{3T}),\end{aligned}$$

where I_1, I_2 and I_3 are as in the proof for $\mathbf{C_T}'$.

We will show that

$$\begin{aligned} D_{1T} &= O\big((\frac{1}{Th^q})^4(T^2h^{2q(1-\frac{4}{\theta})})\big), && (3.6.62)\\ D_{2T} &= O\big((\frac{1}{Th^q})^4(T^3h^{3q(1-\frac{4}{\theta})})\big), && (3.6.63)\\ D_{3T} &= O\big((\frac{1}{Th^q})^4\left(T^3h^{4q-\frac{6q}{\theta(1-\gamma)}}+T^2h^{3q-\frac{12q}{\theta(1-\gamma)}}\right)\big), && (3.6.64)\end{aligned}$$

which proves (3.6.61).

Proof of (3.6.62) for D_{1T}. Similarly as in the proof for C_{1T}, since

$$\left|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\bar{U}_{At}\bar{U}_{As}\right| \le \left(\frac{K_t}{f_t}U_{it}\bar{U}_{At}\right)^2+\left(\frac{K_s}{f_s}U_{is}\bar{U}_{As}\right)^2,$$

then,

$$\begin{aligned} D_{1T} &\le \sum_{t,s=1}^{T}E\big(\frac{K_t^2}{f_t^2}U_{it}^2K_{ts}^2\bar{U}_{As}^2+\left|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\bar{U}_{At}\bar{U}_{As}\right|K_{ts}^2\big)\\ &\le 3\sum_{t,s=1}^{T}E\big(\frac{K_t^2}{f_t^2}U_{it}^2\bar{U}_{As}^2K_{ts}^2\big)\le T^2h^{2q(1-\frac{4}{\theta})},\end{aligned}$$

because with p such that $4p=\theta$, and $\frac{1}{r}=1-\frac{1}{p}=1-\frac{4}{\theta}$,

$$\sum_{t,s=1}^{T}E\big(\frac{K_t^2}{f_t^2}U_{it}^2\bar{U}_{As}^2K_{ts}^2\big)\le \big(E\left|\frac{K_t}{f_t}K_{ts}\right|^{2r}\big)^{1/r}\big(E\left|U_{it}\bar{U}_{As}\right|^{2p}\big)^{1/p}\le Ch^{\frac{2q}{r}}=Ch^{2q(1-\frac{4}{\theta})},$$

by Lemma 3.4 (i) and Assumption 9 yielding $E\left|U_{it}\bar{U}_{As}\right|^{2p}\le\big(E\left|U_{it}\right|^{4p}E\left|\bar{U}_{As}\right|^{4p}\big)^{1/2}<\infty$, which proves (3.6.62).

Proof of (3.6.63) for D_{2T}. It suffices to show that

$$E\left(1_{I_2}E\left|\frac{K_{t_1}}{f_{t_1}}\frac{K_{t_3}}{f_{t_3}}U_{it_1}U_{it_3}\bar{U}_{At_2}\bar{U}_{At_4}K_{t_1t_2}K_{t_3t_4}\right|\right)\le Ch^{3q(1-\frac{4}{\theta})}. \qquad (3.6.65)$$

According to definition of I_2, we need to check (3.6.65) in four cases.

Case 1, $(t_1=t_3,t_2\ne t_4)$. Then, the above expectation becomes

$$\begin{aligned}&E\big(\frac{K_t^2}{f_t^2}U_{it}^2|K_{ts}K_{tr}\bar{U}_{As}\bar{U}_{Ar}|\big)\\ &\le \big(E\left|\frac{K_t^2}{f_t^2}K_{ts}K_{tr}\right|^{w}\big)^{1/w}\big(E\left|U_{it}\right|^{2p}\left|\bar{U}_{As}\bar{U}_{Ar}\right|^{p}\big)^{1/p}\\ &\le \big(E\left|\frac{K_t^2}{f_t^2}K_{ts}K_{tr}\right|^{w}\big)^{1/w}\big(E\left|U_{it}\right|^{4p}(E|\bar{U}_{As}|^{4p}E|\bar{U}_{Ar}|^{4p})^{1/2}\big)^{1/2p}\\ &\le C(h^{\frac{3q}{w}})=Ch^{3q(1-\frac{4}{\theta})}, \qquad (3.6.66)\end{aligned}$$

selecting p such that $4p = \theta$, setting $\frac{1}{w} = 1 - \frac{4}{\theta}$, and using Lemma 3.4 (v) and Assumption 9.

Case 2, $(t_1 = t_4, t_2 \neq t_3)$. Then the expectation on the LHS of (3.6.65) is

$$\begin{aligned}
&E\Big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{rt}\bar{U}_{As}\bar{U}_{At}\Big| \\
&\qquad \leq E\Big|\frac{K_t^2}{f_t^2}U_{it}^2\bar{U}_{At}^2K_{ts}K_{rt}\Big| + E\Big|\frac{K_s^2}{f_s^2}U_{is}^2\bar{U}_{As}^2K_{ts}K_{rt}\Big|,
\end{aligned}$$

since $\big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}\big| \leq \big(\frac{K_t}{f_t}U_{it}\big)^2 + \big(\frac{K_s}{f_s}U_{is}\big)^2$. The bound (3.6.65) follows similarly as in (3.6.66).

Case 3, $(t_3 = t_2, t_1 \neq t_4)$. Here, the expectation of the LHS of (3.6.65) is

$$E\Big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{sr}\bar{U}_{As}\bar{U}_{Ar}\Big|,$$

and (3.6.65) follows the same argument as in Case 2.

Case 4, $(t_2 = t_4, t_1 \neq t_3)$. Here, the expectation of the LHS of (3.6.65) is

$$E\Big|\frac{K_t}{f_t}\frac{K_s}{f_s}U_{it}U_{is}K_{ts}K_{rs}\bar{U}_{As}^2\Big|,$$

and (3.6.65) follows the same argument as in Case 2.

Upper bound on D_{3T}. We will show that

$$D_{3T} = O\left(\left(\frac{1}{Th^q}\right)^4\left[T^3h^{4q(1-\gamma)-\frac{6q}{\theta}} + T^2h^{2q(1-\gamma)-\frac{8q}{\theta}}\right]\right). \tag{3.6.67}$$

Denote

$$\phi_T(W_t, W_s, W_r, W_u) = \frac{K_t}{f_t}\frac{K_r}{f_r}U_{it}U_{ir}\bar{U}_{As}\bar{U}_{As}K_{ts}K_{ru}.$$

By Lemma 3.6 (iii),

$$|D_{3T}| = T^4|E\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3, \tilde{W}_4)| + C[T^3M_{T112}^{1-\gamma} + T^2M_{T13}^{1-\gamma} + T^2M_{T4}^{1-\gamma}].$$

The expectation under independence is zero by Assumption 2:

$$\begin{aligned}
E[\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3, \tilde{W}_4)] &= E^*\left(\frac{K_t}{f_t}\frac{K_r}{f_r}U_{it}U_{ir}K_{ts}K_{ru}\bar{U}_{As}\bar{U}_{Au}\right) \\
&= E^*\left(\frac{K_t}{f_t}K_{ts}E^*(\bar{U}_{As}|Z_t, Z_s)E^*(U_{it}|Z_t, Z_s)\right) \\
&\quad \times E^*\left(\frac{K_r}{f_r}K_{ru}E^*(\bar{U}_{Au}|Z_r, Z_u)E^*(U_{ir}|Z_r, Z_u)\right) = 0.
\end{aligned}$$

We will show that

$$M_{T112} \le Ch^{4q(1-\gamma)-\frac{6q}{\theta}}, \tag{3.6.68}$$
$$M_{T13}, M_{T4} \le Ch^{3q(1-\gamma)-\frac{12q}{\theta}}, \tag{3.6.69}$$

which proves (3.6.67).

Proof of (3.6.68). Proof is similar to that of (3.6.53). As noted in Lemma 3.6 (iii), M_{T112} is the maximal $\frac{1}{(1-\gamma)}^{th}$ moment quantity when partitioning the four time periods into either three or four independent subsets. There are three distinct combinations of dependence to be considered in the case of three independent subsets.

For $(W_r, W_u, \tilde{W}_t, \tilde{W}_s)$, one can separate out expectations,

$$\begin{aligned}
&E_{ru,t,s}\left[\left|\frac{K_tU_{it}}{f_t}K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}\right] E_{ru,t,s}\left[\left|\frac{K_rU_{ir}}{f_r}K_{ru}\bar{U}_{Au}\right|^{\frac{1}{1-\gamma}}\right]\\
&= E^*\left[\left|\frac{K_tU_{it}}{f_t}\right|^{\frac{1}{1-\gamma}} E^*(\left|K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}|\tilde{W}_t)\right] E_{ru,t,s}\left[\left|\frac{K_rU_{ir}}{f_r}K_{ru}\bar{U}_{Au}\right|^{\frac{1}{1-\gamma}}\right]\\
&= O\left(h^{2q-\frac{2q}{\theta(1-\gamma)}}\right)\times O\left(h^{2q-\frac{4q}{\theta(1-\gamma)}}\right) = O\left(h^{4q-\frac{6q}{\theta(1-\gamma)}}\right),
\end{aligned}$$

because by Lemma 3.1 and 3.3 and Assumption 9', we set $2p=\theta$ giving $\frac{1}{w}=1-\frac{2}{\theta}$

$$\begin{aligned}
E|\frac{K_t}{f_t}U_{it}|^{\frac{1}{1-\gamma}} &\le (E|U_{it}|^{\frac{p}{1-\gamma}})^{1/p}(E|\frac{K_t}{f_t}|^{\frac{w}{1-\gamma}})^{1/w} = O(h^{\frac{q}{w}}), \\
E|K_{ts}\bar{U}_{As}|^{\frac{1}{1-\gamma}} &\le (E|\bar{U}_{As}|^{\frac{p}{1-\gamma}})^{1/p}(E|K_{ts}|^{\frac{w}{1-\gamma}})^{1/w} = O(h^{\frac{q}{w}}),
\end{aligned} \tag{3.6.70}$$

and by Lemma 3.4 (i) with $2p=\theta(1-\gamma)$ and $\frac{1}{w}=1-\frac{2}{\theta(1-\gamma)}$,

$$\begin{aligned}
&E_{ru,t,s}|\frac{K_r}{f_r}K_{ru}U_{ir}\bar{U}_{Au}|^{\frac{1}{1-\gamma}} \le (E|U_{ir}|^{\frac{2p}{1-\gamma}}E|\bar{U}_{Au}|^{\frac{2p}{1-\gamma}})^{1/2p}(E_{ru,t,s}|\frac{K_r}{f_r}K_{ru}|^{\frac{w}{1-\gamma}})^{1/w}\\
&= O(h^{\frac{2q}{w}}) = O(h^{2q-\frac{4q}{\theta(1-\gamma)}}).
\end{aligned}$$

For $(W_t, W_r, \tilde{W}_s, \tilde{W}_u)$, the $\frac{1}{(1-\gamma)}^{th}$ moment of the kernel is

$$\begin{aligned}
&E_{su,t,r}\left[\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}} E_{su,t,r}\left[\left|K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}|Z_t\right] E_{su,t,r}\left[\left|K_{ru}\bar{U}_{Au}\right|^{\frac{1}{1-\gamma}}|Z_r\right]\right]\\
&= O\left(h^{2q-\frac{4q}{\theta(1-\gamma)}}\right)\times O\left(h^{q-\frac{q}{\theta(1-\gamma)}}\times h^{q-\frac{q}{\theta(1-\gamma)}}\right) = O\left(h^{4q-\frac{6q}{\theta(1-\gamma)}}\right),
\end{aligned}$$

because the inner conditional expectation is

$$E_{su,t,r}\left[\left|K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}|Z_t\right] \le (E|\bar{U}_{As}|^{\frac{p}{1-\gamma}})^{1/p}(E\left[|K_{ts}|^{\frac{w}{1-\gamma}}|Z_t\right])^{1/w} \le Ch^{q-\frac{q}{\theta(1-\gamma)}},$$

by Lemma 3.1 and Assumption 9', setting $p=\theta(1-\gamma)$ and $\frac{1}{w}=1-\frac{1}{\theta(1-\gamma)}$. Noting the independence between $\tilde{W}_t$ and $\tilde{W}_r$, by (3.6.56),

$$E_{su,t,r}\left(\left|\frac{K_tU_{it}}{f_t}\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}}\right) = E\left(\left|\frac{K_tU_{it}}{f_t}\right|^{\frac{1}{1-\gamma}}\right)E\left(\left|\frac{K_rU_{ir}}{f_r}\right|^{\frac{1}{1-\gamma}}\right) = O(h^{\frac{2q}{w}}) = O\left(h^{2q-\frac{2q}{\theta(1-\gamma)}}\right).$$

For $(W_s, W_u, \tilde{W}_r, \tilde{W}_t)$, similarly to (3.6.58) and (3.6.56),

$$E_{tr,s,u}\left[|\frac{K_rU_{ir}}{f_r}|^{\frac{1}{1-\gamma}}\left[E\left|\frac{K_tU_{it}}{f_t}K_{ru}\bar{U}_{Au}\right|^{\frac{1}{1-\gamma}}E\left[\left|K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}|Z_t\right]|Z_r\right]\right] = O\left(h^{4q-\frac{6q}{\theta(1-\gamma)}}\right).$$

since uniformly over z

$$\begin{aligned}
&E\left(\left|K_{ts}\bar{U}_{As}\right|^{\frac{1}{1-\gamma}}|Z_t = z\right) \leq [E\left(|K_{ts}|^{\frac{p}{1-\gamma}}|Z_t = z\right)]^{1/p}[E|\bar{U}_{As}|^{\frac{r}{1-\gamma}}]^{1/r} = O\left(h^{q-\frac{q}{\theta(1-\gamma)}}\right),\\
&E|\frac{K_rU_{ir}}{f_r}|^{\frac{1}{1-\gamma}} \leq [E\left|\frac{K_r}{f_r}\right|^{\frac{p}{1-\gamma}}]^{1/p}[E|U_{ir}|^{\frac{r}{1-\gamma}}]^{1/r} = O\left(h^{q-\frac{q}{\theta(1-\gamma)}}\right),
\end{aligned}$$

by Lemma 3.1 and Assumption 9', setting $\frac{2p}{1-\gamma} = \theta$ and $\frac{1}{w} = 1-\frac{2}{\theta(1-\gamma)}$ and $E\left|\frac{K_tU_{it}}{f_t}K_{ru}\bar{U}_{Au}\right| = O(h^{2q-\frac{4q}{\theta(1-\gamma)}})$ by similar argument as in the proof of (3.6.58). This proves (3.6.70).

Upper bound on M_{T13} and M_{T4}. For both M_{T13} and M_{T4}, one finds the upper bound that holds for all relevant combinations of dependence:

$$\begin{aligned}
&E\left[\left|\frac{K_tU_{it}}{f_t}\bar{U}_{Au}K_{ts}\bar{U}_{As}\frac{K_rU_{ir}}{f_r}K_{ru}\right|^{\frac{1}{1-\gamma}}\right]\\
&\leq C\left[E\left|\frac{K_t}{f_t}\frac{K_r}{f_r}K_{ts}\right|^{\frac{r}{1-\gamma}}\right]^{\frac{1}{r}}\left[E|U_{it}\bar{U}_{Au}|^{\frac{2p}{1-\gamma}}\right]^{\frac{1}{2p}}\left[E|U_{ir}\bar{U}_{As}|^{\frac{2p}{1-\gamma}}\right]^{\frac{1}{2p}}\\
&\leq C\left[E\left|\frac{K_t}{f_t}\frac{K_r}{f_r}K_{ts}\right|^{\frac{w}{1-\gamma}}\right]^{\frac{1}{w}}\left[\left(E|U_{it}|^{\bar{}\frac{4p}{1-\gamma}}\right)^{1/2}\left(E|U_{Au}|^{\frac{4p}{1-\gamma}}\right)^{1/4}\right]^{\frac{1}{p}}\\
&= O\left(h^{3q-\frac{12q}{\theta(1-\gamma)}}\right)
\end{aligned}$$

by setting $4p/(1-\gamma) = \theta$ and $\frac{1}{w} = 1 - \frac{4}{\theta(1-\gamma)}$ and Lemma 3.4 (vi) and Assumption 9', which proves (3.6.69). This completes the proof of (3.6.69).

3.6.5 Upper bounds on $\mathbf{E}_T$ and $\mathbf{F}_T$.

Notice that by Lemma 3.3, $\frac{1}{Th^q}\sum_{t=1}^{T}|K_t| = O_p(1)$. Therefore, by Holder inequality,

$$\begin{aligned}
\mathbf{E}_T + \mathbf{F}_T \quad \leq \quad & (\frac{1}{Th^q}\sum_{t=1}^{T}|K_t|)^{1/2}\{(\frac{1}{Th^q}\sum_{t=1}^{T}|K_t||\frac{n_t}{\tilde{f}_t}|^2)^{1/2}\\
& +(\frac{1}{Th^q}\sum_{t=1}^{T}|K_t||\frac{l_t}{\tilde{f}_t}|^2)^{1/2}\} = O_p(A_T^{1/2} + B_T^{1/2}).
\end{aligned}$$

Thus, by (3.6.10)

$$T^{-1/2}(\mathbf{E}_T+\mathbf{F}_T)=O_P(T^{-1/2}(\mathbf{A}_T^{1/2}+\mathbf{B}_T^{1/2}))=O_p(T^{-1}(\mathbf{A}_T+\mathbf{B}_T))=O_p(R_{Th}),$$

completing proof of (3.6.11).

We showed that

$$\begin{aligned}&\mathbf{A_T}+\mathbf{B_T}+\mathbf{C_T}+\mathbf{D_T}+T^{-1/2}(\mathbf{E_T}+\mathbf{F_T})\\ &\quad\leq C\big(\frac{\log T}{Th^{q+\frac{2q}{\theta}}}+h^{2s-\frac{2q}{\theta}}+r_{1T}+r_{2T}+r_{3T}+r_{4T}\big),\end{aligned}\tag{3.6.71}$$

where

$$\begin{aligned}r_{1T}&=&(\frac{1}{Th^q})^3\left(T^2h^{2q-\frac{2q}{\theta}}+T^2h^{3q(1-\gamma)-\frac{4q}{\theta}}\right),\\ r_{2T}&=&(\frac{1}{Th^q})^3\left(T^3h^{3q+2s}+T^2h^{2q+2}+T^2h^{3q(1-\gamma)+2}\right),\\ r_{3T}&=&(\frac{1}{Th^q})^2\left(T^3h^{2+3q(1-\frac{2}{\theta})}+T^3h^{4q(1-\gamma)-2(\frac{2q}{\theta}-1)}+T^2h^{2q(1-\gamma)-2(\frac{2q}{\theta}-1)}\right)^{1/2},\\ r_{4T}&=&(\frac{1}{Th^q})^2(T^3h^{3q(1-\frac{4}{\theta})}+T^2h^{3q-\frac{12q}{\theta(1-\gamma)}})^{1/2},\\ R_{T,h}&=&h^p+h^{2s-\frac{2q}{\theta}}+\frac{1}{Th^{3\gamma q+\frac{4q}{\theta}}}+\frac{1}{Th^{q+\gamma q+\frac{2q}{\theta}-1}}+\frac{1}{Th^{\frac{q}{2}+\frac{6q}{\theta(1-\gamma)}}}+\frac{1}{\sqrt{Th^{q+\frac{12q}{\theta}}}}.\end{aligned}$$

The proof of Theorem 3.7 is completed by showing that (3.6.71) is $O(R_{T,h})$. Firstly, by Assumption 17 (ii), and since $\frac{4q}{\theta(1-\gamma)}=\frac{4q}{\theta}+\frac{4\gamma q}{\theta(1-\gamma)}$,

$$\frac{\log T}{Th^{q+\frac{2q}{\theta}}}=\frac{(\log T)h^{\frac{2q}{\theta}+\frac{4\gamma q}{\theta(1-\gamma)}}}{Th^{q+\frac{4q}{\theta(1-\gamma)}}}=O\big(\frac{1}{Th^{q+\frac{4q}{\theta(1-\gamma)}}}\big)=O(R_{T,h}).$$

Secondly,

$$\begin{aligned}r_{2T}&\leq&h^{2s-\frac{2q}{\theta}}+r_{1T}=h^{2s-\frac{2q}{\theta}}+\frac{1}{Th^{q+\frac{2q}{\theta}}}+\frac{1}{Th^{3\gamma q+\frac{4q}{\theta}}}\\ &\leq&h^{2s-\frac{2q}{\theta}}+\frac{1}{Th^{q+\frac{4q}{\theta(1-\gamma)}}}+\frac{1}{Th^{3\gamma q+\frac{4q}{\theta}}}=O(R_{T,h}).\end{aligned}$$

Thirdly, by $1-4\gamma\leq\frac{8}{\theta}$ of Assumption 9',

$$\begin{aligned}r_{3T}&=&\frac{1}{\sqrt{Th^{q+\frac{6q}{\theta}-2}}}+\frac{1}{\sqrt{Th^{4\gamma q+\frac{4q}{\theta}-2}}}+\frac{1}{Th^{q+\gamma q+\frac{2q}{\theta}-1}}\\ &=&\frac{\sqrt{h^{\frac{6q}{\theta}+2}}}{\sqrt{Th^{q+\frac{12q}{\theta}}}}+\frac{\sqrt{h^{(\frac{8}{\theta}-(1-4\gamma))q+2}}}{\sqrt{Th^{q+\frac{12q}{\theta}}}}+\frac{1}{Th^{q+\gamma q+\frac{2q}{\theta}-1}}\\ &=&\frac{o(1)}{\sqrt{Th^{q+\frac{12q}{\theta}}}}+\frac{o(1)}{\sqrt{Th^{q+\frac{12q}{\theta}}}}+\frac{1}{Th^{q+\gamma q+\frac{2q}{\theta}-1}}=O(R_{T,h}).\end{aligned}$$

Finally,

$$r_{4T} = \frac{1}{\sqrt{Th^{q+\frac{12q}{\theta}}}} + \frac{1}{Th^{\frac{q}{2}+\frac{6q}{\theta(1-\gamma)}}} = O(R_{T,h}).$$

■

Proof of Theorem 3.8 We provide proof for the $a^*_{m,MSE(\zeta_l)}$, while the proof for $a^*_{m^*,MSE(\zeta_l)}$ follows by the same argument. Recall

$$\begin{aligned} & a^*_{m,AMSE(\zeta_l)} - \hat{a}^*_{m,AMSE(\zeta_l)} \\ & = \Big(\frac{\kappa^q}{TN^2\chi_r^2}\Big)^{\frac{1}{q+2r}} \left(\left\{\frac{f(\zeta_l)1_N'\mathbf{\Omega}(\zeta_l)1_N}{\Phi(\tilde{m}(\zeta_l))^2}\right\}^{\frac{1}{q+2r}} - \left\{\frac{\hat{f}(\zeta_l)1_N'\hat{\mathbf{\Omega}}(\zeta_l)1_N}{\hat{\Phi}(\tilde{m}(\zeta_l))^2}\right\}^{\frac{1}{q+2r}}\right). \end{aligned} \tag{3.6.72}$$

By the mean value theorem, the last term is bounded in absolute value by

$$\frac{1}{q+2r}|\tilde{r}_1||1_N'(\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l))1_N|+\frac{1}{q+2r}|\tilde{r}_2||(f(\zeta_l)-\hat{f}(\zeta_l)|+\frac{1}{q+2r}|\tilde{r}_3|\left|\Phi(\tilde{m}(\zeta_l))^2-\hat{\Phi}(\tilde{m}(\zeta_l))^2\right|, \tag{3.6.73}$$

where r_i's, $i = 1, 2, 3$ are derivatives of the expression in the curly bracket in the RHS of (3.6.72) with respect to $\mathbf{\Omega}$, f and Φ, respectively. $\tilde{r}_1$ lies in

$$\frac{1}{q+2r}\left[\left(\frac{f(\zeta_l)}{\Phi(\tilde{m}(\zeta_l))^2}\right)^{\frac{1}{(q+2r)}}(1_N'\mathbf{\Omega}(\zeta_l)1_N)^{\frac{(-q-2r+1)}{(q+2r)}},\quad \left(\frac{\hat{f}(\zeta_l)}{\hat{\Phi}(\tilde{m}(\zeta_l))^2}\right)^{\frac{1}{(q+2r)}}(1_N'\hat{\mathbf{\Omega}}(\zeta_l)1_N)^{\frac{(-q-2r+1)}{(q+2r)}}\right],$$

$\tilde{r}_2$ lies in

$$\frac{1}{q+2r}\left[\left(\frac{1_N'\mathbf{\Omega}(\zeta_l)1_N}{\Phi(\tilde{m}(\zeta_l))^2}\right)^{\frac{1}{(q+2r)}}(f(\zeta_l))^{\frac{(-q-2r+1)}{(q+2r)}},\quad \left(\frac{1_N'\hat{\mathbf{\Omega}}(\zeta_l)1_N}{\hat{\Phi}(\tilde{m}(\zeta_l))^2}\right)^{\frac{1}{(q+2r)}}(\hat{f}(\zeta_l))^{\frac{(-q-2r+1)}{(q+2r)}}\right],$$

and $\tilde{r}_3$ lies between

$$\begin{aligned} \frac{-1}{q+2r}\big[&(1_N'\mathbf{\Omega}(\zeta_l)1_Nf(\zeta_l))^{\frac{1}{(q+2r)}}(\Phi(\tilde{m}(\zeta_l))^2)^{\frac{(-q-2r-1)}{(q+2r)}}, \\ &(1_N'\hat{\mathbf{\Omega}}(\zeta_l)1_N\hat{f}(\zeta_l))^{\frac{1}{(q+2r)}}(\hat{\Phi}(\tilde{m}(\zeta_l))^2)^{\frac{(-q-2r-1)}{(q+2r)}}\big]. \end{aligned}$$

Then straightforwardly we deduce that the upper bound of (3.6.73) is

$$\begin{aligned} & O_p\big((1_N'\mathbf{\Omega}(\zeta_l)1_N)^{(-q-2r+1)/(q+2r)}\big[N\|\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l)\| + (1_N'\mathbf{\Omega}(\zeta_l)1_N)|f(\zeta_l)-\hat{f}(\zeta_l)| \\ & \qquad +(1_N'\mathbf{\Omega}(\zeta_l)1_N)\left|\Phi(\tilde{m}(\zeta_l))^2-\hat{\Phi}(\tilde{m}(\zeta_l))^2\right|\big]\big) \\ & = O_p((1_N'\mathbf{\Omega}(\zeta_l)1_N)^{(-q-2r+1)/(q+2r)}N[\|\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l)\| + \|\mathbf{\Omega}(\zeta_l)\||f(\zeta_l)-\hat{f}(\zeta_l)| \\ & \qquad +\|\mathbf{\Omega}(\zeta_l)\|\left|\Phi(\tilde{m}(\zeta_l))^2-\hat{\Phi}(\tilde{m}(\zeta_l))^2\right|]) \\ & = O_p\left((1_N'\mathbf{\Omega}(\zeta_l)1_N)^{(-q-2r+1)/(q+2r)}N\|\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l)\|\right), \end{aligned}$$

where the last step follows from Assumption 19. Thus (3.6.72) becomes

$$
\begin{aligned}
a^*_{m,MSE(\zeta_l)} - \hat{a}^*_{m,MSE(\zeta_l)} &= O_p\left(\left(\frac{1_N'\mathbf{\Omega}(\zeta_l)1_N}{TN^2}\right)^{\frac{1}{q+2r}}\left(\frac{N}{1_N'\mathbf{\Omega}(\zeta_l)1_N}\right)^{\frac{1}{q+2r}}\|\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l)\|\right)\\
&= O_p\left(\left(\frac{1_N'\mathbf{\Omega}(\zeta_l)1_N}{TN^2}\right)^{\frac{1}{q+2r}}\|\mathbf{\Omega}(\zeta_l)-\hat{\mathbf{\Omega}}(\zeta_l)\|\right) = o_p(a^*_{m,MSE(\zeta_l)}).
\end{aligned}
$$

■

Proof of Theorem 3.9

For the same reason as in Robinson(2009, pp.28-29), it is sufficient to show that

$$
NR_{T,h} = o\left(a^s + \frac{1}{\sqrt{Ta^q}}\right),
$$

which follows by Assumption 20. ■

3.7 Appendix B. Lemmas 3.1-3.6

Recall the product form of the kernel $K(u) = \prod_{j=1}^{q} k(u_j)$. We first note the multivariate version of Taylor expansion for a function $m : \mathbb{R}^q \to \mathbb{R}$. Suppose m possesses continuous partial derivatives of order r at any $z \in \mathbb{R}^q$ which are uniformly bounded. Then, for $z, w \in \mathbb{R}^q$, one may write

$$
\begin{aligned}
m(z) - m(w) &= \sum_{\ell=1}^{r-1}\frac{1}{\ell!}\sum_{i_1=1}^{q}\cdots\sum_{i_\ell=1}^{q}\frac{\partial^\ell m(t_1,\cdots,t_q)}{\partial t_{i_1}\cdots\partial t_{i_\ell}}\Big|_{t=z}\prod_{j=1}^{\ell}(z_{i_j}-w_{i_j})\\
&\quad+\frac{1}{r!}\sum_{i_1=1}^{q}\cdots\sum_{i_r=1}^{q}\frac{\partial^r m(t_1,\cdots,t_q)}{\partial t_{i_1}\cdots\partial t_{i_r}}\Big|_{t=x}\prod_{j=1}^{r}(z_{i_j}-w_{i_j}), \quad (3.7.1)
\end{aligned}
$$

where x lies on the line segment joining z and w.

Lemma 3.1. Let $\int |k(u)|(1+|u|^a)du < \infty$, for some $a \geq 0$. Then uniformly in z,

$$
\begin{aligned}
\int \|w-z\|^a \left|K\left(\frac{w-z}{h}\right)\right| dw &= h^{q+a}\int \|\psi\|^a\,|K(\psi)|d\psi\\
&\leq h^{q+a}q^a\int |u^a k(u)|du\Big(\int |k(u)|du\Big)^{q-1} = O(h^{q+a}). \quad (3.7.2)
\end{aligned}
$$

Lemma 3.2. Suppose m and f have bounded derivatives of total order up to s, $k \in \mathcal{K}_s$ and $\sup_z f(z) < \infty$.

(i) (Lemma 5 of Robinson (1988)) If Z_1 and Z_2 are independent, then, uniformly over z,

$$
\left|E\left(\{m(Z_1)-m(Z_2)\}K\left(\frac{Z_1-Z_2}{h}\right)|Z_1=z\right)\right| = O\left(h^{q+s}\right).
$$

(ii) If $\int |u^s k(u)|^a du < \infty$ for some $a > 0$ and Z_1 and Z_2 are independent, then uniformly over z,

$$E\left(\left|\{m(Z_1) - m(Z_2)\}K\left(\frac{Z_1 - Z_2}{h}\right)\right|^a |Z_1 = z\right) = O(h^{q+a}). \tag{3.7.3}$$

(iii) If Z_1 and Z_2 are dependent with joint density $f(z, y)$ satisfying $\sup_\delta \int f(z, z+\delta)dz < \infty$, then,

$$E\left(\left|\{m(Z_1) - m(Z_2)\}K\left(\frac{Z_1 - Z_2}{h}\right)\right|^a\right) = O(h^{q+a}). \tag{3.7.4}$$

Proof. (ii) Notice that (3.7.1) implies $|m(z) - m(w)|^a \leq C\|z - w\|^a$. Then by (3.7.2), the LHS of (3.7.3) is bounded by

$$C\int \|w - z\|^a |K(\frac{z - w}{h})|^a f(w)dw \leq C\int \|w - z\|^a |K(\frac{z - w}{h})|^a dw = O(h^{q+a}).$$

(iii) From (3.7.2), it follows that the LHS of (3.7.4) is bounded by

$$\begin{aligned}&\int |\{m(z) - m(w)\}K\left(\frac{z - w}{h}\right)|^a f(z, w)dzdw \\ &\qquad \leq Ch^{q+a}\int \|\psi\|^a |K(\psi)|^a \left(\int f(z, z - h\psi)dz\right)d\psi = O(h^{q+a}).\end{aligned}$$

Lemma 3.3. Let k be a kernel function with a compact support, say $[-1, 1]$, such that $\int |k(u)|^a du < \infty$ for some $a > 0$. Suppose that Z is a random variable with a continuous pdf f and $\zeta \in \mathbb{R}^q$ is such that $f(\zeta) > 0$. Then, for all $b > 1$,

$$E\left[\frac{|K((Z - \zeta)/h)|^a}{f(Z)^b}\right] = O(h^q).$$

Proof. Since f is continuous and positive at ζ, there exist $\delta > 0$ and $\varepsilon > 0$ such that $f(\zeta + w) \geq \delta$, for $|w| \leq \varepsilon$. Then $|hu| < \varepsilon$, $\forall |u| < 1$, for T large enough. Thus as $T \to \infty$,

$$\begin{aligned}E(\frac{|K((Z - \zeta)/h)|^a}{f(Z)^b}) &= \int \frac{|K((z - \zeta)/h)|^a}{f(z)^{b-1}}dz = h^q \int_{-1}^1 \frac{|K(u)|^a}{f(\zeta + hu)^{b-1}}du \\ &\leq \frac{h^q}{\delta^{b-1}}\int_{-1}^1 |K(u)|^a du = O(h^q).\end{aligned}$$

Lemma 3.4. Let Z_1, Z_2, Z_3 be random variables with joint densities $f(\cdot, \cdot, \cdot)$, $f(\cdot, \cdot)$ and marginal density $f(\cdot)$ such that $\sup_{z,u} f(z, u) < \infty$, $\sup_{z,u,w} f(z, u, w) < \infty$ and $f(\zeta) > 0$, for a fixed point ζ. Let k be a kernel function with a compact support, and ℓ be a kernel function such that $\int\{|\ell(u)|^a + |k(u)|^b\}du < \infty$ for some $a, b > 0$ and let $c \geq 0$. Then the product kernels $L(u) = \prod_{j=1}^q \ell(u_j), K(u) = \prod_{j=1}^q k(u_j)$ have the following properties:

$$\text{(i)}\quad E\left[\left|K\big(\frac{Z_1-\zeta}{h}\big)\right|^b\left|L\big(\frac{Z_1-Z_2}{h}\big)\right|^a\frac{1}{f(Z_1)^c}\right]=O(h^{2q}),$$

$$\text{(ii)}\quad E\left|\frac{K((Z_1-\zeta)/h)}{f(Z_1)}\frac{K((Z_2-\zeta)/h)}{f(Z_2)}\right|^a=O(h^{2q}),$$

$$\text{(iii)}\quad E\left[\left|K\big(\frac{Z_1-\zeta}{h}\big)\right|^b\left|\{m(Z_1)-m(Z_2)\}L\big(\frac{Z_1-Z_2}{h}\big)\right|^a\frac{1}{f(Z_1)^c}\right]=O(h^{2q+a}),$$

$$\text{(iv)}E\Big[\left|K\big(\frac{Z_1-\zeta}{h}\big)\right|^b\left|\{m(Z_1)-m(Z_2)\}\{m(Z_1)-m(Z_3)\}L\big(\frac{Z_1-Z_2}{h}\big)L\big(\frac{Z_1-Z_3}{h}\big)\right|^a\frac{1}{f(Z_1)^c}\Big]$$
$$=O(h^{3q+2a}),$$

$$\text{(v)}\quad E\left[\left|K\big(\frac{Z_1-\zeta}{h}\big)\right|^b\left|L\big(\frac{Z_1-Z_2}{h}\big)L\big(\frac{Z_1-Z_3}{h}\big)\right|^a\frac{1}{f(Z_1)^c}\right]=O(h^{3q}),$$

$$\text{(vi)}\quad E\left[\left|K\big(\frac{Z_1-\zeta}{h}\big)K\big(\frac{Z_3-\zeta}{h}\big)L\big(\frac{Z_1-Z_2}{h}\big)\right|^a\frac{1}{f(Z_1)^a}\frac{1}{f(Z_3)^a}\right]=O(h^{3q}).$$

Proof. (i) Denote $\phi=(z-\zeta)/h$ and $\psi=(w-\zeta)/h$. Since $\sup_{z,w}f(z,w)<\infty$ and $f(z)>0$ for $|z-\zeta|\le ch$ as $h\to 0$,

$$\begin{aligned}&\int\left|K\big(\frac{z-\zeta}{h}\big)\right|^b\left|L\big(\frac{z-w}{h}\big)\right|^a\frac{f(z,w)}{f(z)^c}dzdw\\&\quad\le Ch^{2q}\int|K(\phi)|^b|L(\phi-\psi)|^a d\phi d\psi\\&\qquad=h^{2q}\int|K(\phi)|^b d\phi\int|L(\psi)|^a d\psi=O(h^{2q}).\end{aligned}$$

(ii) Similarly, since $f(z)\ge c>0$ in the neighborhood of ζ,

$$\begin{aligned}&\int\left|\frac{K((z-\zeta)/h)}{f(z)}\frac{K((w-\zeta)/h)}{f(w)}\right|^a f(z,w)dzdw\\&\quad\le C\int\left|\frac{K((z-\zeta)/h)}{f(z)}\right|^a dz\int\left|\frac{K((w-\zeta)/h)}{f(w)}\right|^a dw\\&\quad\le Ch^{2q}(\int|K(\phi)|^a\,d\phi)^2=O(h^{2q}).\end{aligned}$$

(iii) As above,

$$\begin{aligned}&\int|K\big(\frac{z-\zeta}{h}\big)|^b|\{m(z)-m(w)\}L\big(\frac{z-w}{h}\big)|^a\frac{f(z,w)}{f(z)^c}dzdw\\&\quad\le Ch^{2q+a}\int|K(\psi)|^b d\psi\int\|\phi\|^a|L(\phi)|^a d\phi=O(h^{2q+a}).\end{aligned}$$

Proof of (iv) follows by the same argument as in (iii), proof of (v) is analogous to that of (i) and proof of (vi) is similar to that of (i) and (ii). ■

The next three lemmas offer convenient tools in dealing with asymptotic behaviour of U-statistics of a stationary β-mixing process.

Lemma 3.5. (Yoshihara's Inequality) Suppose $\{W_t\}$ is a strictly stationary β-mixing process with mixing coefficient $\beta(\tau)$, taking values in $\mathbb{R}^q$ with marginal distribution function F. Let $1 \leq t_1 < \cdots < t_k, k \geq 2$ be integers and $F_{t_1,\cdots,t_k}$ the joint distribution function of $(W_{t_1}, \cdots, W_{t_k})$. Denote by $\phi_T(w_1, \cdots, w_k)$ a sequence of functions on $(\mathbb{R}^q)^k$. Then for $0 < \gamma < 1$,

$$\begin{aligned}
&\left| \int \phi_T(w) dF_{t_1,\cdots,t_k} - \int \phi_T(w) dF_{t_1,\cdots,t_j} dF_{t_{j+1},\cdots,t_k} \right| \\
&\qquad \leq 4 \left(\int |\phi_T(w)|^{1/(1-\gamma)} d\{F_{t_1,\cdots,t_k} + F_{t_1,\cdots,t_j} F_{t_{j+1},\cdots,t_k}\} \right)^{1-\gamma} \times \beta(t_{j+1} - t_j)^{\gamma},
\end{aligned}$$

provided the RHS exists.

Proof can be found in Yoshihara (1976). The original lemma had ϕ, not ϕ_T and the extension is mentioned in Robinson (1991).

Before stating the next lemma, we need the following notation. By $(\pi(1), \cdots, \pi(k))$ denote a permutation of the set $(1, \cdots, k)$. For example, for $k = 3$, $(\pi(1), \cdots, \pi(3)) \in \{(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,2,1), (3,1,2)\}$. Define

$$\tilde{\phi}_T(w_1, \cdots, w_k) = \sum_{\pi(1),\cdots,\pi(k)} \phi_T(w_{\pi(1)}, \cdots, w_{\pi(k)}), \tag{3.7.5}$$

where the sum $\sum_{\pi(1),\cdots,\pi(k)}$ is taken over all permutation of the set $\{1, \cdots, k\}$. Note that $\tilde{\phi}_T$ is a symmetric function. For brevity, we write $F_{t_1,t_2,t_3} = F_{t_1,t_2,t_3}(w_1, w_2, w_3)$, $F_{t_1} F_{t_2,t_3} = F_{t_1}(w_1) F_{t_2,t_3}(w_2, w_3)$, and so on.

Define:

$$\begin{aligned}
M_{T2} &:= \max_{1 \leq t_1 < t_2 \leq T} \int_{\mathbb{R}^{2q}} |\tilde{\phi}_T(w_1, w_2)|^{1/(1-\gamma)} d\{F_{t_1,t_2} + F_{t_1} F_{t_2}\}, \\
M_{T3} &:= \max_{1 \leq t_1 < t_2 < t_3 \leq T} \int_{\mathbb{R}^{3q}} |\tilde{\phi}_T(w_1, w_2, w_3)|^{1/(1-\gamma)} d\{F_{t_1,t_2,t_3} + F_{t_1} F_{t_2,t_3} + F_{t_1,t_2} F_{t_3}\}, \\
M_{T12} &:= \max_{1 \leq t_1 < t_2 < t_3 \leq T} \int_{\mathbb{R}^{3q}} |\tilde{\phi}_T(w_1, w_2, w_3)|^{1/(1-\gamma)} d\{F_{t_1} F_{t_2,t_3} + F_{t_1,t_2} F_{t_3} + F_{t_1} F_{t_2} F_{t_3}\},
\end{aligned}$$

$$\begin{aligned}
M_{T4} &:= \max_{1\le t_1<t_2<t_3<t_4\le T}\int_{\mathbb{R}^{4q}}|\tilde{\phi}_T(w_1,w_2,w_3,w_4)|^{1/(1-\gamma)}d\{F_{t_1,t_2,t_3,t_4}+F_{t_1}F_{t_2,t_3,t_4}\\
&\qquad +F_{t_1,t_2}F_{t_3,t_4}+F_{t_1,t_2,t_3}F_{t_4}\},\\
M_{T13} &:= \max_{1\le t_1<t_2<t_3<t_4\le T}\int_{\mathbb{R}^{4q}}|\tilde{\phi}_T(w_1,w_2,w_3,w_4)|^{1/(1-\gamma)}d\{F_{t_1}F_{t_2,t_3,t_4}+F_{t_1,t_2}F_{t_3,t_4}\\
&\qquad +F_{t_1,t_2,t_3}F_{t_4}+F_{t_1,t_2}F_{t_3}F_{t_4}+F_{t_1}F_{t_2,t_3}F_{t_4}+F_{t_1}F_{t_2}F_{t_3,t_4}\},\\
M_{T112} &:= \max_{1\le t_1<t_2<t_3<t_4\le T}\int_{\mathbb{R}^{4q}}|\tilde{\phi}_T(w_1,w_2,w_3,w_4)|^{1/(1-\gamma)}d\{F_{t_1,t_2}F_{t_3}F_{t_4}+F_{t_1}F_{t_2,t_3}F_{t_4}\\
&\qquad +F_{t_1}F_{t_2}F_{t_3,t_4}+F_{t_1}F_{t_2}F_{t_3}F_{t_4}\}.
\end{aligned}$$

Let $\{\tilde{W}_t\}$ denote an i.i.d. process with the marginal distribution function F, and $\sum_{t_1,\cdots,t_k}{}'$ is a summation over non-overlapping indices $(t_1,\cdots,t_k)$.

Lemma 3.6. In addition to assumptions of Lemma 3.5, assume that for some $0<\gamma<1$ and $\epsilon>0$, the β-mixing coefficient of W_t satisfies $\beta(\tau)=O(\tau^{-(2+\epsilon)/\gamma})$ as $\tau\to\infty$. Then, for some $0<C<\infty$,

$$\text{(i)}\qquad \left|\sum_{t_1,t_2}{}'E\left(\phi_T(W_{t_1},W_{t_2})\right)-T(T-1)E\left(\phi_T(\tilde{W}_1,\tilde{W}_2)\right)\right|\le CTM_{T2}^{1-\gamma}.$$

$$\begin{aligned}
\text{(ii)}\qquad &\left|\sum_{t_1,t_2,t_3}{}'E\phi_T(W_{t_1},W_{t_2},W_{t_3})-T(T-1)(T-2)E\left(\phi_T(\tilde{W}_1,\tilde{W}_2,\tilde{W}_3)\right)\right|\\
&\qquad\le CT^2M_{T12}^{1-\gamma}+CTM_{T3}^{1-\gamma}.
\end{aligned}$$

$$\begin{aligned}
\text{(iii)}\qquad &\left|\sum_{t_1,t_2,t_3,t_4}{}'E\left(\phi_T(W_{t_1},W_{t_2},W_{t_3},W_{t_4})\right)-T(T-1)(T-2)(T-3)E\left(\phi_T(\tilde{W}_1,\tilde{W}_2,\tilde{W}_3,\tilde{W}_4)\right)\right|\\
&\qquad\le CT^3M_{T112}^{1-\gamma}+CT^2M_{T13}^{1-\gamma}+CT^2M_{T4}^{1-\gamma}.
\end{aligned}$$

Proof. (i) One can write

$$\sum_{1\le t_1,t_2\le T}{}'E\left(\phi_T(W_{t_1},W_{t_2})\right)=\sum_{1\le t_1<t_2\le T}E\left(\phi_T(W_{t_1},W_{t_2})+\phi_T(W_{t_2},W_{t_1})\right).$$

For all $1\le t_1<t_2\le T$, Yoshihara's inequality yields:

$$\begin{aligned}
\left|E[\phi_T(W_{t_1},W_{t_2})-\phi_T(\tilde{W}_1,\tilde{W}_2)]\right|&\le CM_{T2}^{1-\gamma}\beta^\gamma(t_2-t_1),\\
\left|E[\phi_T(W_{t_2},W_{t_1})-\phi_T(\tilde{W}_1,\tilde{W}_2)]\right|&\le CM_{T2}^{1-\gamma}\beta^\gamma(t_2-t_1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\left|\sum_{1\le t_1,t_2\le T}{}'E[(\phi_T(W_{t_1},W_{t_2}))-\phi_T(\tilde{W}_1,\tilde{W}_2)]\right|&\le CM_{T2}^{1-\gamma}\sum_{1\le t_1<t_2\le T}\beta^\gamma(t_2-t_1)\\
&\le CTM_{T2}^{1-\gamma}\sum_{\tau=1}^{T-1}\beta^\gamma(\tau)\le CTM_{T2}^{1-\gamma},
\end{aligned}$$

because of the assumption $\beta(\tau) = O(\tau^{-(2+\epsilon)/\gamma})$ and $E\phi_T(\tilde{W}_s, \tilde{W}_t) = E\phi_T(\tilde{W}_1, \tilde{W}_2)$ for $t \neq s$.

(ii) One has

$$\begin{aligned}
&\sum_{t_1,t_2,t_3}{}' E[\phi_T(W_{t_1}, W_{t_2}, W_{t_3})] \\
&\quad = \sum_{1\le t_1<t_2<t_3\le T} E\left[\phi_T(W_{t_1}, W_{t_2}, W_{t_3}) + \cdots + \phi_T(W_{t_3}, W_{t_2}, W_{t_1})\right] \\
&\qquad\qquad = \sum_{1\le t_1<t_2<t_3\le T} E\tilde{\phi}(W_{t_1}, W_{t_2}, W_{t_3}),
\end{aligned}$$

where $\tilde{\phi}_T$ is as in (3.7.5). For any $1 \le t_1 < t_2 < t_3 \le T$, define $t^* := \max\{t_3 - t_2, t_2 - t_1\}$ and $t_* := \min\{t_3 - t_2, t_2 - t_1\}$. Then by stationarity and Yoshihara's inequality,

$$\begin{aligned}
&\left|E[\tilde{\phi}_T(W_{t_1}, W_{t_2}, W_{t_3})] - d_T(t_1, t_2, t_3)\right| \le C M_{T3}^{1-\gamma} \beta^\gamma(t^*), \\
&d_T(t_1, t_2, t_3) := \int\int \tilde{\phi}_T(w_1, w_2, w_3) dF_{0,t^*}(w_1, w_2) F(w_3), \\
&|d_T(t_1, t_2, t_3) - \int \tilde{\phi}_T(w_1, w_2, w_3) dF(w_1)F(w_2)F(w_3)| \le 4M_{T12}^{1-\gamma}\beta^\gamma(t_*).
\end{aligned}$$

Therefore ,

$$\begin{aligned}
&\left|E\tilde{\phi}_T(W_{t_i}, W_{t_j}, W_{t_k}) - \int \phi_T(w_1, w_2, w_3) dF(w_1)dF(w_2)dF(w_3)\right| \\
&\qquad \le C M_{T3}^{1-\gamma}\beta^\gamma(t^*) + C M_{T12}^{1-\gamma}\beta^\gamma(t_*).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\left|\sum_{t_1,t_2,t_3}{}' E\left(\phi_T(W_{t_1}, W_{t_2}, W_{t_3})\right) - T(T-1)(T-2)E(\phi_T(\tilde{W}_1, \tilde{W}_2, \tilde{W}_3))\right| \\
&\qquad \le C M_{T3}^{1-\gamma} \sum_{1\le t_1<t_2<t_3\le T} \beta^\gamma(t^*) + C M_{T12}^{1-\gamma} \sum_{1\le t_1<t_2<t_3\le T} \beta^\gamma(t_*) \\
&\qquad \le C[T M_{T3}^{1-\gamma} + T^2 M_{T12}^{1-\gamma}]. \qquad (3.7.6)
\end{aligned}$$

To verify (4.7.4), note that from definition of t^* and t_*, and $\beta(\tau)^\gamma \le c\tau^{-(2+\varepsilon)}$,

$$\begin{aligned}
\beta(t^*) &\le c|t_3 - t_2|^{-(1+\varepsilon/2)}|t_2 - t_1|^{-(1+\varepsilon/2)}, \\
\beta(t_*) &\le c(|t_3 - t_2|^{-(2+\varepsilon)} + |t_2 - t_1|^{-(2+\varepsilon)}).
\end{aligned}$$

Thus,

$$\sum_{1\le t_1<t_2<t_3\le T} \beta^\gamma(t^*) \le C\left(\sum_{t_1=1}^T 1\right)\left(\sum_{s=1}^T s^{-(1+\varepsilon/2)}\right)^2 \le CT,$$
$$\sum_{1\le t_1<t_2<t_3\le T} \beta^\gamma(t_*) \le C\sum_{1\le t_1<t_2\le T} |t_2-t_1|^{-(2+\varepsilon)}\left(\sum_{t_3=1}^T 1\right) \le C\left(\sum_{s=1}^T s^{-(2+\varepsilon)}\right)T^2 \le CT^2.$$

(iii) For any $1\le t_1<t_2<t_3<t_4\le T$, define $t^* := \max\{t_4-t_3, t_3-t_2, t_2-t_1\}$, $t_* := \min\{t_4-t_3, t_3-t_2, t_2-t_1\}$ and $t_m := \{t_4-t_3, t_3-t_2, t_2-t_1\}\backslash\{t^*, t_*\}$. By similar steps to (ii), one has

$$\begin{aligned}\left|\sum_{t_1,t_2,t_3,t_4}{}' E\left(\phi_T(W_{t_1}, W_{t_2}, W_{t_3}, W_{t_4})\right) - T(T-1)(T-2)(T-3)E\left(\phi_T(\tilde W_1, \tilde W_2, \tilde W_3, \tilde W_4)\right)\right|& \\ \le CM_{T112}^{1-\gamma}\sum_{1\le t_1<t_2<t_3<t_4\le T}\beta^\gamma(t_*) + CM_{T13}^{1-\gamma}\sum_{1\le t_1<t_2<t_3<t_4\le T}\beta^\gamma(t_m)& \\ +CM_{T4}^{1-\gamma}\sum_{1\le t_1<t_2<t_3<t_4\le T}\beta^\gamma(t^*)& \\ \le C\left[M_{T112}^{1-\gamma}T^3 + M_{T13}^{1-\gamma}T^2 + M_{T4}^{1-\gamma}T^2\right].& \qquad (3.7.7)\end{aligned}$$

The last bounds in (3.7.7) follows noting that $\beta(\tau)^\gamma \le c\tau^{-(2+\varepsilon)}$, and therefore

$$\begin{aligned}&\beta^\gamma(t^*) \le c|t_3-t_2|^{-(1+\varepsilon/2)}|t_2-t_1|^{-(1+\varepsilon/2)},\\ &\beta^\gamma(t_m) \le c|t_3-t_2|^{-(1+\varepsilon/2)}|t_2-t_1|^{-(1+\varepsilon/2)},\\ &\beta^\gamma(t_*) \le c(|t_4-t_3|^{-(2+\varepsilon)} + |t_3-t_2|^{-(2+\varepsilon)} + |t_2-t_1|^{-(2+\varepsilon)}).\end{aligned}$$

Hence

$$\sum_{1\le t_1<t_2<t_3<t_4\le T} |\beta^\gamma(t^*) + \beta^\gamma(t_m)| \le C\left(\sum_{t_1,t_4=1}^T 1\right)\left(\sum_{s=1}^T s^{-(1+\varepsilon/2)}\right)^2 \le CT^2,$$
$$\begin{aligned}\sum_{1\le t_1<t_2<t_3<t_4\le T} \beta^\gamma(t_*) &\le C\sum_{1\le t_1<t_2\le T}|t_2-t_1|^{-(2+\varepsilon)}\left(\sum_{t_1,t_4=1}^T 1\right)\\ &\le CT^3\left(\sum_{s=1}^T s^{-(2+\varepsilon)}\right) \le CT^3,\end{aligned}$$

which proves (3.7.7) and completes the proof of (iii). ■

4 Efficiency Improvement in Estimation of Semi-parametric Pure Spatial Autoregressive Model

4.1 Introduction

Spatial econometric data typically feature irregular spacing, for example, when observations are recorded across cities, regions or countries. In numerous applications of interest in Economics, correlation across observations may be characterized by some general notion of economic distance (e.g. differences in household income or product characteristics) that does not necessarily have a geographical interpretation, see, e.g. Conley and Dupor (2003). These two features render much of the spatial statistics inapplicable to economic data. As a result, Spatial Autoregressive (SAR) models of Cliff and Ord (1968), that can cater for the two afore-mentioned features, have gained popularity in applications (see, e.g. Arbia (2006)), and received much attention in the theoretical literature, see, e.g. Kelejian and Prucha (1998), Lee (2002), Lee (2004) and Rossi (2010).

In this chapter, we consider the so-called pure SAR model, which describes spatial dependence in the absence of any regressors, modeled parametrically by a linear transformation of underlying shocks. Let $y = (y_1, \cdots, y_n)^T$ be a vector of observations having the same (unknown) mean, $E(y_i) = \mu_0$, and with y^T denoting transposition. The model is given by

$$(I - \lambda_0 W)(y - \mu_0 1_n) = \sigma_0 \varepsilon, \tag{4.1.1}$$

where 1_n is a $n \times 1$ vector of 1's, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is a vector of independent identically distributed random variables with zero mean and unit variance, and σ_0 and λ_0 are unknown scalar parameters. The $n \times n$ weight matrix, $W = W_n$, is fixed and assumed to be known *a priori*, having real-valued (i, j)-th element $w_{ij} = w_{ijn}$ such that

$$w_{ii} = 0, \qquad \sum_{j=1}^{n} w_{ij} = 1,\ i = 1, \cdots, n, \quad \text{i.e.} \quad W 1_n = 1_n. \tag{4.1.2}$$

It is noted that the elements w_{ijn} of the weight matrix may change with n but the n subscript is suppressed below for brevity. The w_{ij} are typically interpreted as inverse economic distances (see, e.g. Arbia (2006)).

The meaning of the row-normalisation restriction of (4.1.2) becomes more tangible once we write the model in a scalar form:

$$y_i - \mu_0 = \lambda_0 \big[\sum_{j=1}^{n} w_{ij} (y_j - \mu_0) \big] + \sigma_0 \varepsilon_i.$$

The summation inside the square bracket on the right hand side (RHS) is called the "spatial lag" of unit i and the row normalisation naturally requires this to be a weighted average.

When ε, and thus y, is Gaussian, the model (4.1.1) can be thought of as primarily describing the covariance matrix of y, since this, and μ_0, describe the distribution of y completely. The parameter vector $\theta_0 = (\lambda_0, \mu_0, \sigma_0)^T$ can be asymptotically efficiently estimated by the maximum likelihood estimate (MLE) $\tilde{\theta} = \left(\tilde{\lambda}, \tilde{\mu}, \tilde{\sigma}\right)^T$. It has been explicitly established in Lee (2004) that, under some regularity conditions, $\tilde{\theta}$ is consistent and asymptotically normal. In fact, these latter properties hold over a much more general class of distributions of the ε_i, in which case the estimate $\tilde{\theta}$ is termed a (Gaussian) pseudo MLE (PMLE).

However $\tilde{\theta}$ is not asymptotically efficient when it is only a PMLE. Given a (non-Gaussian) parametric specification of the distribution of ε_1, we can construct a (non-Gaussian) MLE as follows. Let $f(x; \zeta_0) = \mathbb{R}^{1+q} \to \mathbb{R}^1$ be the probability density function of ε_1, a given function of all its arguments, with ζ_0 being an unknown $q \times 1$ parameter vector. Write $\theta_0 = \left(\lambda_0, \mu_0, \sigma_0, \zeta_0^T\right)^T$, and denote by $\theta = \left(\lambda, \mu, \sigma, \zeta^T\right)^T$ any admissible value of θ_0. Introducing the notation $S(\lambda) := I - \lambda W$ allows us to write the log likelihood as

$$L(\theta) = \sum_{i=1}^{n} \log f\left(\frac{S_i^T(\lambda)\,(y - \mu 1_n)}{\sigma}; \zeta \right) + \log\det\{S(\lambda)\} - \frac{n}{2} \log \sigma^2, \tag{4.1.3}$$

where $S_i^T(\lambda)$ denotes the i-th row of $S(\lambda)$. The MLE $\bar{\tau} = \left(\bar{\lambda}, \bar{\mu}, \bar{\sigma}, \bar{\zeta}^T\right)^T$ of τ_0 maximizes (4.1.3) over a suitable compact set, and can be expected to be asymptotically efficient. Unfortunately there are rarely strong prior grounds for specifying f, and misspecification of a non-Gaussian probability density f can lead to inconsistent estimation.

In practice, λ_0 is often the main feature of interest, with μ_0 and σ_0 being nuisance parameters (and our results on estimation of λ_0 are unaffected if $\mu_0 = 0$ is known *a priori*). In this chapter we establish an estimate $\hat{\lambda}$ of λ_0 that achieves the same asymptotic distribution as the MLE $\bar{\lambda}$, in the presence of only non-parametric assumptions on the distribution of ε_1. Specifically, $\hat{\lambda}$ takes a Newton step from the Gaussian PMLE $\tilde{\lambda}$, using non-parametric (series) estimation of the score function.

This kind of "adaptive" property was previously established in a spatial context by Robinson (2010a), for the model

$$(I - \lambda_0 W)\, y = \mu_0 + X\beta_0 + \sigma_0 \varepsilon, \tag{4.1.4}$$

where X is a $n \times k$ matrix of observed regressors and β_0 is a vector of unknown parameters. Although it may seem that pure SAR is a special case of the mixed SAR with $\beta = 0$, it has been shown in literature, see Lee (2004), that the asymptotic behaviour of the parameter estimates of λ_0 under the two models are radically different, with different rates of convergence. Consequently, the feasibility and implementation of such adaptive estimation in the pure SAR model need to be established separately.

The method of estimation we employ is very similar to that of Robinson (2010a), but the asymptotic variance matrix of his estimate of $(\lambda_0, \beta_0^T)^T$ corresponds to that found in the classical adaptive estimation literature, whereas the asymptotic variance matrix of our estimate of λ_0 differs from the classical one. In particular, the gain in efficiency of $\hat{\lambda}$ over $\tilde{\lambda}$ can be either less or more (typically less) than in the classical outcome.

Section 4.2 presents the information matrix corresponding to estimation based on (4.1.3), its form suggesting both the potential for adapting to unknown distributional form of ε_1 in the estimation of λ_0, and the scope for efficiency gains described in the previous paragraph. Sections 4.3 and 4.4 describe, respectively, our estimate $\hat{\lambda}$ and its asymptotic distribution. Section 4.5 reports a Monte Carlo study of finite-sample behaviour of this estimator.

4.2 Block-diagonality of the information matrix

The feasibility of adaptive estimation of λ_0 w.r.t. unknown error distribution in the pure SAR model is shown via establishing the block-diagonality of the information matrix. Firstly, we introduce restrictions on the weight matrix W. Define $S(\lambda) := S_n(\lambda) = I - \lambda W$.

Assumption 1. (i) *$W = (w_{ij})_{i,j=1,\cdots,n}$ is row-normalized, i.e. $W 1_n = 1_n$ and is uniformly bounded in both row and column sums, i.e.*

$$\max_{1 \le i \le n} \sum_{j=1}^{n} |w_{ij}| = O(1) \quad \text{and} \quad \max_{1 \le j \le n} \sum_{i=1}^{n} |w_{ij}| = O(1).$$

(ii) *For some $h = h_n \to \infty$ and $h = o(n)$ as $n \to \infty$, $\max\limits_{1 \le i,j \le n} |w_{ij}| = O\big(\frac{1}{h}\big)$.*

(iii) *$S := S(\lambda_0)$ is non-singular and S^{-1} is uniformly bounded in both row and column sums.*

The sequence h is important in the asymptotic analysis, defining the rate of convergence of estimates of the parameter λ_0.

The row and column absolute summability of W are used routinely in the SAR literature to control the degree of dependence, e.g. in Kelejian and Prucha (1998, 2001), Lee (2002, 2004). In fact, all those works also assume Assumption 1 (iii), which in turn leads to row and column absolute summability of the $n \times n$ covariance matrix $E(yy') = \sigma_0^2 S(S')^{-1}$. This implies $\sum_{i,j=1}^{n} |Cov(y_i, y_j)| = O(n)$, which is our definition of weak dependence as mentioned in Chapter 1. So the existing literature on pure SAR model only covers weak spatial dependence.

We shall use notation $G(\lambda) := W(I - \lambda W)^{-1}$ and set $G := G(\lambda_0) = (g_{ij})$. Lee (2002, pp. 258) has shown that under Assumption 1, the matrix G has the property,

$$\max_{1 \le i,j \le n} |g_{ij}| = O\big(\frac{1}{h}\big). \tag{4.2.1}$$

Assumption 1 also implies that G is uniformly bounded in both row and column sums:

$$\max_{1 \le i \le n} \sum_{j=1}^{n} |g_{ij}| = O(1) \quad \text{and} \quad \max_{1 \le j \le n} \sum_{i=1}^{n} |g_{ij}| = O(1). \tag{4.2.2}$$

We assume the following limits exist and are non-zero,

$$\omega_1 := \lim_{n \to \infty} \frac{h}{n} tr(GG^T), \quad \omega_2 := \lim_{n \to \infty} \frac{h}{n} tr(G^2).$$

Assumption 2. *The h and W are such that there exist finite limits $\omega_1 \neq 0$, $\omega_2 \neq 0$.*

To show feasibility of adaptive estimation of λ_0 w.r.t. unknown error distribution, we need to establish the block-diagonality of the information matrix between the parameter of interest λ_0 and the other (nuisance) parameters of the model. Let f denote the probability density function (pdf) of ε_i. Suppose f is parametric, i.e. $f(x) = f(x; \zeta_0)$, where $f : \mathbb{R} \times \mathbb{R}^d \Rightarrow \mathbb{R}$ is a known function of its arguments and ζ_0 is $d \times 1$ vector of unknown parameter. Recall the log likelihood of θ is given by

$$L(\theta) = \sum_{i=1}^{n} \log f\Big(\frac{S_i^T(\lambda)(y - \mu 1_n)}{\sigma}; \zeta\Big) + \log\det\{\mathrm{S}(\lambda)\} - \frac{\mathrm{n}}{2} \log \sigma^2, \tag{4.2.3}$$

writing $\theta = \big(\lambda, \mu, \sigma^2, \zeta^T\big)^T$ and denoting by $S_i^T(\lambda)$ the i-th row of $S(\lambda)$,.

To derive the information matrix of the model, we need the following quantities:

$$\begin{aligned} \psi_i &= -\frac{\partial}{\partial \varepsilon_i} \log f(\varepsilon_i; \zeta_0), \quad \chi_i = -\frac{\partial}{\partial \zeta} \log f(\varepsilon_i; \zeta_0), \quad i = 1, 2, \cdots, n, \\ \mathcal{J} &= E(\psi_i^2), \quad D = diag\big\{(n/h)^{\frac{1}{2}}, \quad n^{\frac{1}{2}} I_{d+2}\big\}. \end{aligned}$$

Define $\Xi := \lim_{n \to \infty} D^{-1} E\big(-\frac{d^2 L(\theta_0)}{d\theta d\theta^T}\big) D^{-1}$.

Lemma 4.1. Under Assumptions 1-7,

$$\Xi = \begin{pmatrix} \mathcal{J}\omega_1 + \omega_2 & & & \\ 0 & \left(\frac{1-\lambda_0}{\sigma_0}\right)^2 \mathcal{J} & & \\ 0 & \left(\frac{1-\lambda_0}{2\sigma_0^3}\right) E(\varepsilon_i \psi_i^2) & \frac{1}{4\sigma_0^4} E(\varepsilon_i^2 \psi_i^2 - 1) & \\ 0 & 0 & -\frac{1}{2\sigma_0^2} E(\varepsilon_i \psi_i \chi_i) & E(\chi_i \chi_i^T) \end{pmatrix}.$$

Noting the zero non-diagonal elements of the first column, the feasibility of adaptive estimation of λ_0 with respect to unknown error distribution is established. The proof of Lemma 4.1 is given in the Appendix.

4.3 Adaptive estimation

Our objective is to construct adaptive estimate $\hat{\lambda}$ based on a preliminary estimator $\tilde{\lambda}$ of λ_0. Recall the score function

$$\psi(s) = -\frac{f'(s)}{f(s)}, \quad s \in \mathbb{R},$$

where prime denotes differentiation. To form our adaptive estimator, we will use series estimation of the score function, of which the advantages over kernel estimation are discussed in Robinson (2010a). To formulate the adaptive estimator, we will need some additional notations. Let $\phi_\ell(s), \quad \ell = 1, 2, \cdots$ be a sequence of smooth functions, which will be used in the series estimation of $\psi(\cdot)$. For an integer $L \geq 1$, where $L = L_n$ will be regarded as increasing with n, define the $L \times 1$ vectors

$$\begin{aligned} &\phi^{(L)}(s) = (\phi_1(s), \cdots, \phi_L(s))^T, \quad \bar{\phi}^{(L)}(s) = \phi^{(L)}(s) - E\left\{\phi^{(L)}(\varepsilon_i)\right\}, \\ &\phi^{'(L)}(s) = (\phi_1^{'}(s), \cdots, \phi_L^{'}(s))^T. \end{aligned} \tag{4.3.1}$$

$L = L_n$ is the number of approximating functions that are used in the series estimation of $\psi(\cdot)$ for a sample size n. Allowing $L \to \infty$ as $n \to \infty$ facilitates non-parametric estimation $\psi(\cdot)$. Consider first the case when $\psi(s)$ has a parametric form

$$\psi(s, a^{(L)}) = \bar{\phi}^{(L)}(s)^T a^{(L)}, \tag{4.3.2}$$

where $a^{(L)} = (a_1, \cdots, a_L)^T$ is an unknown vector, and $\bar{\phi}^{(L)}(\varepsilon_i)$ has zero mean. As mentioned in Robinson (2010a), under some mild conditions on f, integration-by-parts allows $a^{(L)}$ to be identified by

$$a^{(L)} = \left[E\left\{\bar{\phi}^{(L)}(\varepsilon_i)\bar{\phi}^{(L)}(\varepsilon_i)^T\right\}\right]^{-1} E\left\{\phi^{'(L)}(\varepsilon_i)\right\}. \tag{4.3.3}$$

Given a vector of observable proxies $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \cdots, \tilde{\varepsilon}_n)^T$, we shall approximate parametric $a^{(L)}$ by $\tilde{a}^{(L)}$, a sample analogue of (4.3.3) constructed as follows. For a generic vector $x = (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, define

$$\tilde{a}^{(L)}(x) = W^{(L)}(x)^{-1} w^{(L)}(x)$$

where

$$W^{(L)}(x) = \frac{1}{n}\sum_{i=1}^{n} \Phi^{(L)}(x_i)\Phi^{(L)}(x_i)^T, \quad \Phi^{(L)}(x_i) = \phi^{(L)}(x_i) - \frac{1}{n}\sum_{j=1}^{n} \phi^{(L)}(x_j),$$

and $w^{(L)}(x) := \frac{1}{n}\sum_{i=1}^{n} \phi'^{(L)}(x_i)$. Next, for given $x = (x_1, \cdots, x_n)^T$ and x_i, $i = 1, \cdots, n$, define the function

$$\psi^{(L)}\left(x_i; \tilde{a}^{(L)}(x)\right) := \Phi^{(L)}(x_i)^T \tilde{a}^{(L)}(x), \quad i = 1, \cdots, n.$$

The estimator $\tilde{\psi}_{iL} := \psi^{(L)}\left(\tilde{\varepsilon}_i; \tilde{a}^{(L)}(\tilde{\varepsilon}_i)\right)$ of $\psi(\varepsilon_i)$ for a given vector $\tilde{\varepsilon}$, which will be later used to construct the Newton step term of our adaptive estimator (2.9.8).

The above discussion is based on a given vector of proxy $\tilde{\varepsilon}$ for ε. We now construct the specific proxy $\tilde{\varepsilon}$ that will be used in the adaptive estimation of λ_0. Consider the $n \times 1$ vector,

$$e(\lambda) := (e_1(\lambda), \cdots, e_n(\lambda))^T = (I - \lambda W)y = S(\lambda)y, \quad \lambda \in [0, 1].$$

Since (4.1.1) gives a mean-adjusted expression of y, namely

$$\sigma_0 \varepsilon = S(\lambda_0)y - \mu_0 S(\lambda_0)1_n = S(\lambda_0)y - E\left\{S(\lambda_0)y\right\},$$

we denote,

$$\epsilon_i(\lambda) := e_i(\lambda) - \frac{1}{n}\sum_{j=1}^{n} e_j(\lambda), \quad i = 1, \cdots, n.$$

Using the $n \times n$ matrix $H := I - \frac{1}{n}1_n 1_n^T$, we can write

$$\epsilon(\lambda) = (\epsilon_1(\lambda), \cdots, \epsilon_n(\lambda))^T = HS(\lambda)y, \quad i = 1, \cdots, n. \tag{4.3.4}$$

For a given estimate $\tilde{\lambda}$ of λ_0, we shall estimate σ_0^2 by

$$\tilde{\sigma}^2(\tilde{\lambda}) := \frac{1}{n}\epsilon(\tilde{\lambda})^T \epsilon(\tilde{\lambda}).$$

This leads to the definition of our proxy $\tilde{\varepsilon}$ for errors ε based on $\tilde{\lambda}$:

$$\tilde{\varepsilon} := \frac{\epsilon(\tilde{\lambda})}{\tilde{\sigma}}.$$

For convenience, set $\tilde{\psi}_{iL} := \tilde{\psi}_{iL}(\tilde{\lambda}, \tilde{\sigma})$, where $\tilde{\psi}_{iL}(\lambda, \sigma) := \Phi^L(\epsilon_i(\lambda)/\sigma)^T \tilde{a}^L(\epsilon(\lambda)/\sigma)$.

Introduce the estimate of the information measure $\mathcal{J} := E\left(\psi^2(\varepsilon_i)\right)$, denoted by $\tilde{\mathcal{J}}_L := \tilde{\mathcal{J}}_L(\tilde{\lambda}, \tilde{\sigma})$, where

$$\tilde{\mathcal{J}}_L(\lambda, \sigma) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}_{iL}^2(\lambda, \sigma). \tag{4.3.5}$$

We are now ready to define our adaptive estimator of λ_0, based on a preliminary estimate $\tilde{\lambda}$, as follows:

$$\begin{aligned} \hat{\lambda} &= \tilde{\lambda} + \left(\tilde{\mathcal{J}}_L \cdot \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})\mathrm{G}(\tilde{\lambda})^{\mathrm{T}}\right\} + \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})^2\right\}\right)^{-1} \left(\sum_{i=1}^{n} \tilde{\psi}_{iL}\frac{E_i'}{\tilde{\sigma}} - \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})\right\}\right) \qquad (4.3.6) \\ &= \tilde{\lambda} + \left(\tilde{\mathcal{J}}_L \cdot \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})\mathrm{G}(\tilde{\lambda})^{\mathrm{T}}\right\} + \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})^2\right\}\right)^{-1} \left(\frac{1}{\tilde{\sigma}}(\tilde{\psi}_{1L}, \cdots, \tilde{\psi}_{nL})HWy - \mathrm{tr}\left\{\mathrm{G}(\tilde{\lambda})\right\}\right). \end{aligned}$$

The second term of (4.3.6) represents the approximate Newton step, based on the non-parametric estimate of the score function $\psi(\cdot)$. The estimator $\hat{\lambda}$ can be written alternatively as follows. Introduce the $n \times 1$ vector of derivatives $e' := (e_1', \cdots, e_n')^T = \frac{\partial e(\lambda)}{\partial \lambda} = -Wy$, which do not depend on λ. Denote by $\epsilon_i' = e_i' - \frac{1}{n}\sum_{j=1}^{n} e_j'$, $\quad i = 1, 2, \cdots, n$, the sample-mean-adjusted form of $e_i^{'}$, which can be written as

$$\epsilon' = -HWy.$$

Write,

$$\begin{aligned} r_L(\lambda, \sigma) &:= \sum_{i=1}^{n} \tilde{\psi}_{iL}(\lambda, \sigma)\frac{\epsilon_i'}{\sigma} - \mathrm{tr}\left\{\mathrm{G}(\lambda)\right\} \qquad (4.3.7) \\ &= \frac{1}{\sigma}\left(\tilde{\psi}_{1L}(\lambda, \sigma), \cdots, \tilde{\psi}_{nL}(\lambda, \sigma)\right) HWy - \mathrm{tr}\left\{\mathrm{G}(\lambda)\right\} \\ &= \frac{\sigma_0}{\sigma}\left(\tilde{\psi}_{1L}(\lambda, \sigma), \cdots, \tilde{\psi}_{nL}(\lambda, \sigma)\right) HG\varepsilon - \mathrm{tr}\left\{\mathrm{G}(\lambda)\right\}. \end{aligned}$$

Note that $HWy = HW(S^{-1}\sigma_0\varepsilon - \mu_0 1_n) = \sigma_0 HG\varepsilon$ due to $HW1_n = H1_n = 0$. Hence $\hat{\lambda}$ of (4.3.6) can be written as

$$\hat{\lambda} - \lambda_0 = (\tilde{\lambda} - \lambda_0) + \left(\tilde{\mathcal{J}}_L \cdot \mathrm{tr}\left\{G(\tilde{\lambda})G(\tilde{\lambda})^T\right\} + \mathrm{tr}\left\{G(\tilde{\lambda})^2\right\}\right)^{-1} r_L(\tilde{\lambda}, \tilde{\sigma}). \tag{4.3.8}$$

4.4 Asymptotic normality and efficiency

Assumption 3. *$\{\varepsilon_i\}$ is a sequence of i.i.d. random variables with zero mean, unit variance and twice differentiable probability density function $f(\cdot)$ such that $sf'(s) \to 0$ and $s^2 f''(s) \to 0$ as $|s| \to \infty$ and satisfy the following moment conditions:*

$$E|\varepsilon_1|^4 < \infty, \quad E|\psi(\varepsilon_1)|^4 < \infty, \quad E|\varepsilon_1\psi(\varepsilon_1)|^{2+\delta} < \infty.$$

Assumption 4. *In (4.3.1) and (4.3.2), $\phi_\ell(s) = \phi^\ell(s)$, $\ell = 1, \cdots, L$, where $\phi(s)$ is* *strictly increasing and thrice differentiable function such that for some* $\kappa \geq 0$*,* $K > 0$*,*

$$|\phi(s)| \leq 1 + |s|^{\kappa}, \quad |\phi'(s)| + |\phi''(s)| + |\phi'''(s)| \leq C(1 + |\phi(s)|^{K}), \quad s \in \mathbb{R}. \tag{4.4.1}$$

Define $\eta := 1 + \sqrt{2}$ and $\varphi := (1 + |\phi(s_1)|)/\{\phi(s_2) - \phi(s_1)\}$, with $[s_1, s_2]$ being an interval on which $f(s)$ is bounded away from zero.

Assumption 5. *The sequences* h *and* L *of (4.3.1) satisfy one of the following conditions with* κ *as in (4.4.1).*

(i) $\kappa = 0$, $E(\varepsilon_i^4) < \infty$, and for some $A > \eta \max(\varphi, 1)$,

$$L \log L \leq \frac{\log h}{8 \log A}, \quad n \to \infty. \tag{4.4.2}$$

(ii) $\kappa > 0$, for some $\omega > 0$ and $t > 0$, $E\left(e^{t|\varepsilon_i|^{\omega}}\right) < \infty$, and for some $B > 8\kappa \max(1, \frac{1}{\omega})$,

$$L \log L \leq \frac{\log h}{B}, \quad n \to \infty. \tag{4.4.3}$$

(iii) $\kappa > 0$, the random variables ε_i's are almost surely bounded, and for some $C > 4\kappa$,

$$L \log L \leq \frac{\log h}{C}, \quad n \to \infty. \tag{4.4.4}$$

Assumption 6. *As* $n \to \infty$*,*

$$E\left\{\bar{\phi}^{(L)}(\varepsilon_i)^T a^{(L)} - \psi(\varepsilon_i)\right\}^2 = o\left(h/n\right), \quad E\left\{\bar{\phi}'(\varepsilon_i)^T a^{(L)} - \psi'(\varepsilon_i)\right\}^2 = o(1).$$

Assumption 7. *As* $n \to \infty$*,*

$$|\tilde{\lambda} - \lambda_0| = O_p((h/n)^{1/2}), \quad |\tilde{\sigma} - \sigma_0| = O_p(n^{-1/2}).$$

Recalling that the object being estimated by series estimation is score function $\psi(\cdot) = -f'(\cdot)/f(\cdot)$, it is of interest to allow for the possibility that $\psi(\cdot)$ may be unbounded. Assumption 4 imposes a restriction on the rate at which the tail of $\phi(\cdot)$ and its derivatives may diverge by the choice of κ. If we restrict the series functions to be bounded by setting $\kappa = 0$, the relatively mild fourth order moment condition suffices in Assumption 5 (i). For unbounded $\phi(\cdot)$, we have a choice between moment generating function (ii) and boundedness (iii) requirements on ε_i of Assumption 5. Part (ii) of Assumption 5 holds with $\omega = 1$ for Laplace ε_i and with $\omega = 2$ for Gaussian ε_i.

Implication of Assumption 5 on the rate of increase in L as $n \to \infty$ is the same across all three cases considered, namely $L \log L = O(\log h)$. The different constants in the upper bound of $L \log L$ are stated here for the sake of precision. The condition $L \log L = O(\log h)$ was also imposed in Assumption 5 (ii) and (iii) of Robinson (2010a) and is difficult to verify in practice, as it is rare that the sequence $h = h_n$ is known in terms of more tangible quantities such as n. An exception to this is the following block-diagonal weight matrix of Case (1991), which was introduced for m number of districts with equal number of farmers r, hence $n = mr$:

$$W = \frac{1}{r-1}\begin{pmatrix} 1_r 1_r' - I_r & 0 & 0 & \cdots \\ 0 & 1_r 1_r' - I_r & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_r 1_r' - I_r \end{pmatrix}. \qquad (4.4.5)$$

With the above weight matrix, $h = r - 1$ and Assumption 1 requires both r and m to increase with n. Assumption 5 requires $L \log L = O(\log r)$, meaning the faster the rate of increase in r as $n \to \infty$, the less restrictive is Assumption 5.

Assumption 6 requires the choice of series functions to yield a series approximation error of the estimator of the unknown score function $\psi(\cdot)$ that decreases at a suitably fast rate as n increases, which is a typical condition imposed in series estimation literature. Assumption 6 is stronger than Assumption 7 of Robinson (2010a), necessitated by the slower rate of convergence of the estimate of λ_0 in the pure SAR model.

It may be of interest to relax Assumption 5, which together with Assumption 6 requires that series functions approximate the score function at a sufficiently fast rate. In Robinson (2010a), the rate restriction on L of Assumption 5 in part (i) was in fact milder at $\log L = O(\log h)$. In this work, this milder restriction was sufficient for the part of technical interest in the proof of Theorem 4.1, but would have resulted in an untenable length of proof for less interesting results that are required for the theorem to hold.

Assumption 7 requires availability of preliminary estimates $\tilde{\lambda}$ of λ_0 and $\tilde{\sigma}$ of σ_0 that have the above rates of convergence. The quasi-maximum likelihood estimators (QMLE) $\tilde{\lambda}^{QMLE}, \tilde{\sigma}^{QMLE}$ of Lee (2004) satisfy Assumption 7 and will be used in the first stage of our adaptive estimation.

The following theorem states asymptotic normality of the adaptive estimator $\hat{\lambda}$ of (4.3.6).

Theorem 4.1. *Let y follow the model (4.3.1) with $\lambda_0 \in (-1, 1)$ and Assumptions 1 - 7 be satisfied. Then, as $n \to \infty$,*

$$\sqrt{\frac{n}{h}}\left(\hat{\lambda} - \lambda_0\right) \to_d N(0, \{\mathcal{J}\omega_1 + \omega_2\}^{-1}).$$

4.5 Efficiency comparison of adaptive estimate and Gaussian PMLE

In Lee (2004) it was shown that

$$\sqrt{\frac{n}{h}}(\tilde{\lambda}^{QMLE} - \lambda_0) \to_d N\left(0, \{\omega_1 + \omega_2\}^{-1}\right).$$

It is of interest to compare the asymptotic variance of $\tilde{\lambda}^{QMLE}$, to that of $\hat{\lambda}$ given in Theorem 4.1 and see how the efficiency improvement attained via adaptive estimation in the spatial setting contrasts to that in time series setting.

Under general enough conditions on W, it may be possible that $\mathrm{tr}(G^2) < 0$, $\omega_2 < 0$. However, if all elements of G are non-negative, which is implied if $w_{ij} \geq 0$ and $\lambda_0 \geq 0$, or if W is symmetric, then $\omega_2 > 0$. In any case, it is possible to show that $\mathrm{tr}\big(G(G+G^T)\big) > 0$, so since $\mathrm{tr}(GG^T) \geq 0$ also, we have $\omega_1 > 0$ and $\omega_1 + \omega_2 > 0$, implying

$$\mathcal{J}\omega_1 + \omega_2 \geq \omega_1 + \omega_2 > 0, \quad \text{because} \quad \mathcal{J} \geq 1.$$

This shows that $\hat{\lambda}$ is better than $\tilde{\lambda}^{QMLE}$. The relative efficiency of $\hat{\lambda}$ to $\tilde{\lambda}^{QMLE}$ is given by

$$\frac{\omega_1 + \omega_2}{\mathcal{J}\omega_1 + \omega_2} = \frac{1 + \omega_2/\omega_1}{\mathcal{J} + \omega_2/\omega_1}.$$

In the time series setting, where W is lower triangular matrix, $\omega_2 = 0$ and therefore the efficiency improvement is $1/\mathcal{J}$. If $\omega_2 > 0$, then the efficiency improvement of adaptive estimator is smaller than in the time series situation. For example if W is symmetric, the relative efficiency is $2/(\mathcal{J}+1)$. On the contrary, $\omega_2 < 0$ yields greater efficiency improvement than under time series setting. The latter case is not ruled out by any conditions of this chapter.

4.6 Monte Carlo study of finite sample performance

In this section, we report results from a small Monte Carlo study of the finite sample performance of the adaptive estimator $\hat{\lambda}$. We study the efficiency improvement achieved by the adaptive $\hat{\lambda}$ relative to the preliminary estimate $\tilde{\lambda}^{QMLE}$ under differing error distributions, sample sizes, and the magnitude of spatial dependence. 1000 replications were carried out in each setting considered.

We use the block diagonal weight matrix of Case (1991) introduced in (4.4.5). The sample size is $n = mr$ and we have $h = r - 1$. We take values of (m, r) same as in the Monte Carlo study of Robinson (2010a): $(m, r) = (12, 8), (18, 11)$ and $(28, 14)$ with the corresponding sample sizes $n = 96, 198$ and 392. To investigate effects of differing strength of spatial dependence, we consider three different values of $\lambda_0 = 0.2, 0.4, 0, 8$. As was done in the Monte Carlo study of Robinson (2010a), the following four different distributions of ε_i are used with the asymptotic relative efficiency (ARE) $(= 2/(\mathcal{J}+1))$ of $\hat{\lambda}$ to $\tilde{\lambda}^{QMLE}$ as reported below.

(a) Bimodal mixture normal, $\varepsilon_i = u/\sqrt{10}$, where the pdf of ε is

$$f(u) = \frac{0.5}{\sqrt{2\pi}} exp\Big(-\frac{(u-3)^2}{2}\Big) + \frac{0.5}{\sqrt{2\pi}} exp\Big(-\frac{(u+3)^2}{2}\Big), \quad u \in \mathbb{R} \quad ARE = 0.188.$$

(b)Unimodal mixture normal, $\varepsilon_i = u/\sqrt{2.2}$ where

$$f(u) = \frac{0.05}{\sqrt{50\pi}} exp\big(-\frac{u^2}{50}\big) + \frac{0.95}{\sqrt{2\pi}} exp\big(-\frac{u^2}{2}\big), \quad u \in \mathbb{R} \quad ARE = 0.679.$$

(c) Laplace, $f(u) = \exp(-|s|\sqrt{2})\sqrt{2}$, $ARE = 0.666$.

(d) Student t_5, $\varepsilon_i = u\sqrt{3/5}$, where $u \sim t_5$, $ARE = 0.685$.

The ARE was calculated from the reported values of $1/\mathcal{J}$ from Robinson (2010a).

Three choices of the number of series functions in series estimation were tried, $L = 1, 2, 4$. It was set that $\phi_\ell(s) = \phi^\ell(s), \ell = 1, \cdots, L$ and two choices of $\phi(s)$ were used:

$$(i) \quad \phi(s) = s, \qquad (ii) \quad \phi(s) = \frac{s}{(1+s^2)^{1/2}}.$$

Based on the 1000 replications, the Monte Carlo variance and MSE of the two estimates of λ_0 were computed in each setting considered, and their ratios are presented in Table 4.1 and 4.2. The ratio taking a value smaller than 1 indicates an efficiency improvement.

Across all the cases, it appears that the choice $L = 1$ led to poor approximation to the score function, resulting in disappointing performance of the adaptive estimator, especially for the choice (i) of $\psi(\cdot)$. The relative performance of the adaptive estimator is best for $L = 4$ in all cases and the improvements are substantial in the cases of (a) and (b), which were also observed in Robinson (2010a). Table 4.2 reports the relative MSE to ascertain whether the bias has been adversely affected by the adaptive estimation. In fact, the relative MSE reported often greater improvement than the relative variance, suggesting the bias has been also reduced. A distinctive contrast to the results reported in the mixed SAR case of Robinson (2010a) is that the efficiency improvement is greater under larger values of λ_0. It is possible to take more than a single Newton-Raphson step, subsequently iterating the adaptive estimation and it is expected that this would yield some further improvement.

Table 4.1: Relative Monte Carlo Variance, $Var(\hat{\lambda})/Var(\tilde{\lambda}^{QMLE})$

λ_0	0.2			0.4			0.8		
ϕL	(12,8)	(18,11)	(28,14)	(12,8)	(18,11)	(28,14)	(12,8)	(18,11)	(28,14)
(a) (i) 1	2.1912	1.858	1.7041	1.3752	1.1709	1.1491	1	1	1.0001
2	2.1467	1.8419	1.693	1.3726	1.1515	1.1449	0.9783	0.9975	0.9929
4	0.623	0.5256	0.4811	0.3994	0.3118	0.3026	0.2315	0.2326	0.2144
(ii) 1	1.9314	1.6553	1.5123	1.3698	1.2129	1.215	1.5921	1.6036	1.7006
2	1.7766	1.612	1.4534	1.2862	1.1232	1.1676	1.3956	1.4979	1.5807
4	0.5331	0.4387	0.388	0.3115	0.243	0.2573	0.1582	0.1747	0.1692
(b) (i) 1	2.1525	1.8964	1.7921	1.3475	1.1279	1.074	1	1	1.0001
2	2.0793	1.8681	1.78	1.3259	1.1107	1.079	0.9778	0.9805	0.9896
4	1.2754	1.2546	1.286	0.7531	0.6707	0.7388	0.493	0.5742	0.5968
(ii) 1	1.502	1.2897	1.1915	0.8544	0.6116	0.5971	0.3871	0.3545	0.3335
2	1.3285	1.1955	1.1465	0.7838	0.5705	0.5907	0.3618	0.3391	0.325
4	0.3033	0.2476	0.2274	0.2011	0.1225	0.1143	0.1117	0.0963	0.0942
(c) (i) 1	2.1835	1.9367	1.803	1.3102	1.2397	1.0924	1	1	1
2	2.1701	1.9235	1.7901	1.2915	1.2007	1.0898	0.9795	0.9848	0.993
4	2.0268	1.7568	1.6724	1.1888	1.065	0.966	0.8725	0.8517	0.8663
(ii) 1	2.0655	1.7866	1.6708	1.1707	1.081	0.9327	0.8184	0.7775	0.7744
2	2.0558	1.7879	1.6644	1.1628	1.0717	0.9395	0.8091	0.78	0.7734
4	1.8495	1.5566	1.4152	1.0579	0.9311	0.8062	0.7904	0.75	0.739
(d) (i) 1	2.1609	1.7507	1.6618	1.3784	1.1419	1.0921	1	1	1.0001
2	2.1383	1.7384	1.6374	1.3713	1.1235	1.0614	0.959	0.9771	0.9884
4	2.0261	1.6514	1.5371	1.3212	1.0477	1.0221	0.9103	0.8992	0.9186
(ii) 1	2.0034	1.6464	1.5368	1.3185	1.0383	1.0067	0.8883	0.8891	0.9031
2	2.012	1.6332	1.5335	1.3017	1.05	1.0082	0.894	0.8979	0.9033
4	1.9794	1.623	1.5315	1.3266	1.0534	1.0166	0.9115	0.8811	0.9111

Table 4.2: Relative Monte Carlo MSE, $MSE(\hat{\lambda})/MSE(\tilde{\lambda}^{QMLE})$

λ_0	0.2			0.4			0.8		
ϕL	(12,8)	(18,11)	(28,14)	(12,8)	(18,11)	(28,14)	(12,8)	(18,11)	(28,14)
(a) (i) 1	2.416	1.9759	1.795	1.3647	1.162	1.1409	1	1	1.0001
2	2.3677	1.9568	1.7847	1.3476	1.139	1.138	0.972	0.9983	0.9897
4	0.6339	0.5154	0.4696	0.3564	0.2655	0.2697	0.2092	0.2093	0.1915
(ii) 1	2.4891	2.1387	1.9654	1.8218	1.7635	1.8302	2.4763	2.7575	3.3888
2	2.1592	2.0142	1.8636	1.5547	1.5418	1.7266	1.9975	2.4224	3.0505
4	0.5393	0.4248	0.3771	0.2744	0.2019	0.2242	0.1379	0.1538	0.1481
(b) (i) 1	2.353	2.0171	1.8689	1.335	1.125	1.0718	1	1	1.0001
2	2.2473	1.9833	1.8614	1.2925	1.1065	1.072	0.964	0.9773	0.9903
4	1.3087	1.2746	1.2818	0.6782	0.6346	0.7045	0.4588	0.547	0.5618
(ii) 1	1.4966	1.2269	1.0941	0.7056	0.5168	0.5075	0.3169	0.3427	0.3774
2	1.3109	1.1418	1.0534	0.6437	0.4822	0.5021	0.2948	0.3255	0.3636
4	0.3033	0.2413	0.2124	0.169	0.1107	0.1039	0.1006	0.0914	0.0913
(c) (i) 1	2.3959	2.0794	1.8871	1.298	1.226	1.09	1	1	1.0001
2	2.3464	2.0589	1.8747	1.2709	1.1895	1.0878	0.9731	0.9811	0.9924
4	2.1503	1.8614	1.7165	1.1455	1.0438	0.962	0.8478	0.8331	0.8571
(ii) 1	2.1842	1.8482	1.6713	1.0989	1.0106	0.8798	0.7419	0.7006	0.6936
2	2.1496	1.8532	1.6679	1.0862	1.0034	0.8879	0.7362	0.7056	0.6945
4	1.9216	1.6038	1.4314	1.0143	0.9098	0.7955	0.759	0.7396	0.7424
(d) (i) 1	2.3796	1.8576	1.737	1.36	1.1347	1.0874	1	1	1.0001
2	2.3352	1.8327	1.6972	1.3529	1.1218	1.0565	0.9572	0.9716	0.9875
4	2.1958	1.7309	1.586	1.2881	1.0355	0.9997	0.9004	0.8849	0.9055
(ii) 1	2.1805	1.7256	1.5744	1.2771	1.0096	0.9673	0.8684	0.8529	0.8674
2	2.1745	1.7145	1.5711	1.2619	1.0216	0.9698	0.877	0.8622	0.8705
4	2.1169	1.6951	1.5724	1.2788	1.0259	0.9779	0.898	0.8588	0.889

4.7 Proofs

Proof of Lemma 4.1. Recall $\Xi := \lim_{n\to\infty} D^{-1}E\big(-\frac{d^2}{d\theta d\theta^T}L(\theta_0)\big)D^{-1}$, $S = S(_0) = I - \lambda_0 W$, $G = WS^{-1}$ and that the log likelihood $L(\theta)$ is given by

$$L(\theta) = \sum_{i=1}^{n} \log f\Big(\frac{S_i^T(\lambda)(y-\mu 1_n)}{\sigma};\zeta\Big) + \log\det\{\mathrm{S}(\lambda)\} - \frac{\mathrm{n}}{2}\log\sigma^2,$$

where $S_i^T(\lambda)$ denotes the i-th row of $S(\lambda)$. Firstly, notice from (4.1.1),

$$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T = \frac{S(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}, \quad \varepsilon_i = \frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}, \quad i = 1, \cdots, n.$$

The first derivatives of $L(\theta)$ w.r.t $\theta = (\lambda, \mu, \sigma^2, \zeta)^T$ at $\theta_0 = (\lambda_0, \mu_0, \sigma_0^2, \zeta_0)^T$ is given by

$$\begin{aligned}
\frac{\partial L(\theta_0)}{\partial \lambda} &= \sum_{i=1}^{n} \frac{W_i^T(y-\mu_0 1_n)}{\sigma_0}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big) - \mathrm{tr(G)},\\
\frac{\partial L(\theta_0)}{\partial \mu} &= \sum_{i=1}^{n} \frac{S_i^T 1_n}{\sigma_0}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big),\\
\frac{\partial L(\theta_0)}{\partial \sigma^2} &= \sum_{i=1}^{n} \frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{2\sigma_0^3}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big) - \frac{n}{2\sigma_0^2},\\
\frac{\partial L(\theta_0)}{\partial \zeta} &= -\sum_{i=1}^{n} \chi_i,
\end{aligned}$$

taking into account that

$$\frac{d\log\{\det(S(\lambda_0))\}}{d\lambda} = \mathrm{tr}\Big(\mathrm{S^T}(\lambda_0)^{-1}\frac{\mathrm{dS^T}(\lambda_0)}{\lambda}\Big) = \mathrm{tr}\Big(\mathrm{S}^{-1}(\lambda_0)^{\mathrm{T}}(-\mathrm{W^T})\Big) = -\mathrm{tr}(\mathrm{G^T}) = -\mathrm{tr}(\mathrm{G}).$$

The following facts are repeatedly used in deriving the second order derivative matrix:

$$\begin{aligned}
\frac{\partial \psi(s)}{\partial s} &= \frac{(f'(s))^2 - f''(s)f(s)}{f^2(s)} = \psi^2(s) - \frac{f''(s)}{f(s)},\\
\frac{\partial}{\partial \lambda}\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0} &= \frac{-W_i^T(y-\mu_0 1_n)}{\sigma_0} = -W_i^T S(\lambda_0)\varepsilon = -G_i^T\varepsilon,\\
\frac{\partial}{\partial \mu}\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0} &= \frac{-S_i^T(\lambda_0)1_n}{\sigma_0},\\
\frac{\partial}{\partial \sigma^2}\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0} &= \frac{-S_i^T(\lambda_0)(y-\mu_0 1_n)}{2\sigma_0^3} = \frac{-\varepsilon_i}{2\sigma_0^2},
\end{aligned}$$

where G_i^T denotes the ith row of G. Next, we derive the elements of Ξ. For brevity, we denote $\psi_i = \psi(\varepsilon_i)$.

$(1,1)^{th}$ *element of* Ξ. We note first that

$$\frac{\partial \mathrm{tr(G)}}{\partial\lambda} = \mathrm{tr}\Big(\mathrm{W}\cdot\frac{\partial(\mathrm{I}-\lambda_0\mathrm{W})^{-1}}{\partial\lambda}\Big) = \mathrm{tr}\Big(\mathrm{W}(\mathrm{I}-\lambda_0\mathrm{W})^{-1}\mathrm{W}(\mathrm{I}-\lambda_0\mathrm{W})^{-1}\Big) = \mathrm{tr}(\mathrm{G}^2).$$

$$\begin{aligned}
&\frac{\partial}{\partial\lambda_0}\Big\{\sum_{i=1}^n \frac{W_i^T(y-\mu_0 1_n)}{\sigma_0}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big)\Big\}\\
&= \sum_{i=1}^n G_i^T\varepsilon\Big(\psi^2(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0) - \frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)\frac{\partial}{\partial\lambda}\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\\
&= -\sum_{i=1}^n G_i^T\varepsilon\Big(\psi^2(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0) - \frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)G_i^T\varepsilon\\
&= -\sum_{i=1}^n (G_i^T\varepsilon)^2\cdot\Big(\psi_i^2 - \frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big). \qquad (4.7.1)
\end{aligned}$$

Then, in the last line of (4.7.1), expectation of the first term is

$$\begin{aligned}
-\sum_{i=1}^n E[(G_i^T\varepsilon)^2\psi_i^2] &= -\sum_{i=1}^n\sum_{j=1}^n g_{ij}^2 E[\varepsilon_j^2\psi^2(\varepsilon_i)]\\
&= -E(\psi_1^2)E(\varepsilon_1^2)\cdot\sum_{i=1}^n\sum_{j=1}^n g_{ij}^2 + \big(E(\psi_1^2)E(\varepsilon_1^2) - E(\varepsilon_1^2\psi_1^2)\big)\cdot\sum_{i=1}^n g_{ii}^2\\
&= -\mathcal{J}\cdot\mathrm{tr}(GG^T) + O\Big(\frac{n}{h^2}\Big),
\end{aligned}$$

since $g_{ii} = O(1/h)$ uniformly in i, see (4.2.1). Next, taking the expectation of the second product of (4.7.1) and noting $E\big(f''(\varepsilon_i)/f(\varepsilon_i)\big) = 0$,

$$\sum_{i=1}^n E\Big((G_i^T\varepsilon)^2\frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big) = E(\varepsilon_1^2 f''(\varepsilon_1)/f(\varepsilon_1))\cdot\sum_{i=1}^n g_{ii}^2 = 2\mathrm{tr}(G^2) = O\big(\frac{n}{h^2}\big),$$

since under Assumption 3, $E(\varepsilon_1^2 f''(\varepsilon_1)/f(\varepsilon_1)) = 2$. Therefore, the $(1,1)^{th}$ element of Ξ is given by

$$\lim_{n\to\infty}\frac{h}{n}E\big(-\frac{d^2L(\theta_0)}{d\lambda^2}\big) = \lim_{n\to\infty}\frac{h}{n}\Big(\mathcal{J}\mathrm{tr}(\mathrm{GG^T}) + \mathrm{tr}(\mathrm{G}^2)\Big) = \mathcal{J}\omega_1 + \omega_2.$$

$(2,2)^{th}$ *element.* We have

$$E\Big(-\frac{\partial^2}{\partial\mu^2}L(\theta_0)\Big) = \sum_{i=1}^n\Big(\frac{S_i^T 1_n}{\sigma_0}\Big)^2 E\Big(\psi^2(\varepsilon_i) - \frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big) = \frac{n(1-\lambda_0)^2}{\sigma_0^2}\mathcal{J},$$

since $S_i^T 1_n = (\ell_i^T - \lambda_0 W_i^T)1_n = 1-\lambda_0$, due to $W_i^T 1_n = 1 \quad \forall i$. Therefore, the $(2,2)^{th}$

element of Ξ is

$$\lim_{n\to\infty}\frac{n\mathcal{J}(1-\lambda_0)^2}{n}=\mathcal{J}(1-\lambda_0)^2.$$

$(3,3)^{th}$ *element.* The second order derivative w.r.t. σ^2 is given by

$$\begin{aligned}\frac{-\partial^2}{\partial(\sigma^2)^2}L(\theta_0) &= \sum_{i=1}^{n}\frac{3S_i^T(\lambda_0)(y-\mu_0 1_n)}{4\sigma_0^5}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big)\\ &+ \sum_{i=1}^{n}\frac{\big(S_i^T(\lambda_0)(y-\mu_0 1_n)\big)^2}{4\sigma_0^6}\Big(\psi^2\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big)\\ &\qquad -\frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)-\frac{n}{2\sigma_0^4}\\ &= \sum_{i=1}^{n}\Big[\frac{3\varepsilon_i}{4\sigma_0^4}\psi(\varepsilon_i)+\frac{\varepsilon_i^2}{4\sigma_0^4}\big(\psi^2(\varepsilon_i)-\frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\big)\Big]-\frac{n}{2\sigma_0^4}.\end{aligned}$$

Taking expectation, noting $E(\varepsilon_i\psi_i)=1$ and $E(\varepsilon_i^2 f''(\varepsilon_i)/f(\varepsilon_i))=2$, yields

$$\begin{aligned}E\Big(\frac{-\partial^2}{\partial(\sigma^2)^2}L(\theta_0)\Big) &= \frac{1}{4\sigma_0^4}\sum_{i=1}^{n}\Big(E\big(\varepsilon_i^2\psi_i^2\big)-E(\varepsilon_i^2 f''(\varepsilon_i)/f(\varepsilon_i))+3E(\varepsilon_i\psi_i)\Big)-\frac{n}{2\sigma_0^4}\\ &= \frac{n}{4\sigma_0^4}\Big(E\big(\varepsilon_i^2\psi_i^2\big)-1\Big).\end{aligned}$$

Therefore, the $(3,3)^{th}$ element of Ξ is given by $E\Big(\varepsilon_i^2\psi_i^2-1\Big)/4\sigma_0^4$.
$(1,2)^{th}$ *element.* One has

$$\begin{aligned}\frac{-\partial^2}{\partial\mu\partial\lambda}L(\theta_0) &= \sum_{i=1}^{n}\frac{W_i^T 1_n}{\sigma_0}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big)\\ &\quad+\sum_{i=1}^{n}\frac{W_i^T(y-\mu_0 1_n)}{\sigma_0}\Big(\psi^2(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)\\ &\quad+\frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)\frac{S_i^T(\lambda_0)1_n}{\sigma_0}\\ &= \sum_{i=1}^{n}\frac{W_i^T 1_n}{\sigma_0}\psi_i+\sum_{i=1}^{n}G_i^T\varepsilon\Big(\psi_i^2+\frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big)\frac{S_i^T(\lambda_0)1_n}{\sigma_0}.\end{aligned}$$

Taking expectation, and noting (4.2.1),

$$E\Big(\frac{-\partial^2}{\partial\mu\partial\lambda}L(\theta_0)\Big)=\frac{(1-\lambda_0)}{\sigma_0}\sum_{i=1}^{n}g_{ii}\Big(E(\varepsilon_i\psi_i^2)+E\big(\varepsilon_i\frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\big)\Big)=O\Big(\sum_{i=1}^{n}|g_{ii}|\Big)=O\Big(\frac{n}{h}\Big).$$

Therefore, the $(1,2)^{th}$ element of Ξ is of order $O\Big(\frac{n}{h}\Big)\times\frac{\sqrt{h}}{n}=O\Big(\frac{1}{\sqrt{h}}\Big)=o(1)$.

$(1,3)^{th}$ *element.* One has

$$\begin{aligned}
\frac{-\partial^2}{\partial\sigma^2\partial\lambda}L(\theta_0) &= \sum_{i=1}^n \frac{W_i^T(y-\mu_0 1_n)}{2\sigma_0^3}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big) \\
&\quad + \sum_{i=1}^n \frac{S_i^T(y-\mu_0 1_n)}{2\sigma_0^3}\Big(\psi^2(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0) \\
&\quad + \frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)\frac{W_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0} \\
&= \sum_{i=1}^n \frac{G_i^T\varepsilon}{2\sigma_0^2}\psi(\varepsilon_i) + \sum_{i=1}^n \frac{\varepsilon_i}{2\sigma_0^2}\Big(\psi^2(\varepsilon_i) + \frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big)G_i^T\varepsilon.
\end{aligned}$$

Taking expectation yields

$$E\Big(\frac{-\partial^2}{\partial\sigma^2\partial\lambda}L(\theta_0)\Big) = \frac{1}{2\sigma_0^2}\sum_{i=1}^n g_{ii}\Big(E\big(\varepsilon_i\psi_i\big) + E\big(\varepsilon_i^2\psi_i^2\big) + 2\Big) = O\Big(\sum_{i=1}^n |g_{ii}|\Big) = O\Big(\frac{n}{h}\Big)$$

Therefore, the $(1,3)^{th}$ element of Ξ is of order $O\Big(\frac{n}{h}\Big) \times \frac{\sqrt{h}}{n} = O\Big(\frac{1}{\sqrt{h}}\Big) = o(1)$.
$(2,3)^{th}$ *element.* One has

$$\begin{aligned}
\frac{-\partial^2}{\partial\sigma^2\partial\mu}L(\theta_0) &= \sum_{i=1}^n \frac{S_i^T(\lambda_0)1_n}{2\sigma_0^3}\psi\Big(\frac{S_i^T(\lambda_0)(y-\mu_0 1_n)}{\sigma_0}\Big) \\
&\quad - \sum_{i=1}^n \frac{S_i^T(y-\mu_0 1_n)}{2\sigma_0^3}\Big(\psi^2(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0) \\
&\quad + \frac{f''(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}{f(S_i^T(\lambda_0)(y-\mu_0 1_n)/\sigma_0)}\Big)\frac{S_i^T(\lambda_0)1_n}{\sigma_0} \\
&= \sum_{i=1}^n \frac{S_i^T(\lambda_0)1_n}{2\sigma_0^3}\psi(\varepsilon_i) - \sum_{i=1}^n \frac{\varepsilon_i}{2\sigma_0^2}\Big(\psi_i^2 + \frac{f''(\varepsilon_i)}{f(\varepsilon_i)}\Big)\frac{S_i^T(\lambda_0)1_n}{\sigma_0}.
\end{aligned}$$

Taking expectation, noting $E\psi_i = 0$ and $E(\varepsilon_i f''(\varepsilon_i)/f(\varepsilon_i)) = 0$ yields

$$E\Big(\frac{-\partial^2}{\partial\sigma^2\partial\mu}L(\theta_0)\Big) = \frac{(1-\lambda_0)}{2\sigma_0^3}\sum_{i=1}^n E\big(\varepsilon_i\psi^2(\varepsilon_i)\big).$$

Therefore, the $(2,3)^{th}$ element of Ξ is $\frac{(1-\lambda_0)}{2\sigma_0^3}E\big(\varepsilon_1\psi_1^2\big)$.
$(4,4)^{th}$ *element.* Under mild regularity conditions on f,

$$E\Big(\frac{-\partial^2}{\partial\zeta\partial\zeta^T}L(\theta_0)\Big) = E\Big(\frac{\partial L(\theta_0)}{\partial\zeta}\frac{\partial L(\theta_0)}{\partial\zeta^T}\Big) = nE\Big(\chi_i\chi_i^T\Big).$$

$(1,4)^{th}$ *element.* In deriving the $(1,4)^{th}$, $(2,4)^{th}$ and $(3,4)^{th}$ elements of Ξ, the following result is used repeatedly.

$$\begin{aligned}\frac{\partial\psi(\varepsilon_i;\zeta_0)}{\partial\zeta} &= -\frac{f(\varepsilon_i;\zeta_0)\frac{\partial^2}{\partial\varepsilon_i d\zeta}f(\varepsilon_i;\zeta_0)-\frac{\partial}{\partial\varepsilon_i}f(\varepsilon_i;\zeta_0)\frac{\partial}{\partial\zeta}f(\varepsilon_i;\zeta_0)}{f^2(\varepsilon_i;\zeta_0)}\\ &= -\big[\frac{d^2f(\varepsilon_i;\zeta_0)}{d\varepsilon_i d\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)+\chi_i\psi_i.\end{aligned}$$

The cross-second order derivative of $L(\theta_0)$ w.r.t. λ and ζ is

$$-\frac{\partial^2}{\partial\lambda\partial\zeta}L(\theta_0) = -\sum_{i=1}^n G_i^T\varepsilon\frac{\partial\psi(\varepsilon_i;\zeta_0)}{\partial\zeta} = \sum_{i=1}^n G_i^T\varepsilon\big[\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)-G_i^T\varepsilon\chi_i\psi_i.$$

Taking expectation yields

$$\begin{aligned}E\Big(-\frac{\partial^2}{\partial\mu\partial\zeta}L(\theta_0)\Big) &= \sum_{i=1}^n g_{ii}E\Big(\varepsilon_i\big[\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)\Big)-\sum_{i=1}^n g_{ii}E\Big(\varepsilon_i\chi_i\psi_i\Big)\\ &= O(1)\sum_{i=1}^n g_{ii} = O\Big(\frac{n}{h}\Big).\end{aligned}$$

Therefore, the (1,4)th element of Ξ is of order $O\Big(\frac{n}{h}\Big)\times\frac{\sqrt{h}}{n} = O\Big(\frac{1}{\sqrt{h}}\Big) = o(1)$.
$(2,4)^{th}$ *element.* The cross-second order derivative of $L(\theta_0)$ w.r.t. μ and ζ is

$$\begin{aligned}-\frac{\partial^2 L(\theta_0)}{\partial\mu\partial\zeta} &= -\sum_{i=1}^n\frac{S_i^T(\lambda_0)1_n}{\sigma_0}\frac{\partial\psi(\varepsilon_i;\zeta_0)}{\partial\zeta}\\ &= -\sum_{i=1}^n\frac{(1-\lambda_0)}{\sigma_0}\big[\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)+\sum_{i=1}^n\frac{(1-\lambda_0)}{\sigma_0}\chi_i\psi_i.\end{aligned}$$

Taking expectation yields

$$\frac{n(1-\lambda_0)}{\sigma_0}\Big[E\Big(\big[\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)\Big)E(\chi_i\psi_i)\Big] = 0,$$

because $E(\chi_i\psi_i)=0$ and

$$E\Big(\big[\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}\big]f^{-1}(\varepsilon_i;\zeta_0)\Big) = 0.$$

$(3,4)^{th}$ *element.* The cross-second order derivative of $L(\theta_0)$ w.r.t. σ^2 and ζ is

$$\begin{aligned}-\frac{\partial^2}{\partial\sigma^2\partial\zeta}L(\theta_0) &= -\sum_{i=1}^n\frac{\varepsilon_i}{2\sigma_0^2}\frac{\partial\psi(\varepsilon_i;\zeta_0)}{\partial\zeta}\\ &= \sum_{i=1}^n\frac{\varepsilon_i}{2\sigma_0^2}\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}f^{-1}(\varepsilon_i;\zeta_0)-\sum_{i=1}^n\frac{\varepsilon_i}{2\sigma_0^2}\chi_i\psi_i.\end{aligned}$$

The $(3,4)^{th}$ element of Ξ is given by

$$\begin{aligned}\frac{1}{n}E\Big(-\frac{\partial^2}{\partial\sigma^2\partial\zeta}L(\theta_0)\Big) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\sigma_0^2}E\Big(\varepsilon_i\frac{\partial^2 f(\varepsilon_i;\zeta_0)}{\partial\varepsilon_i\partial\zeta}f^{-1}(\varepsilon_i;\zeta_0)\Big)-\frac{1}{n}\sum_{i=1}^{n}E\Big(\frac{\varepsilon_i}{2\sigma_0^2}\chi_i\psi_i\Big)\\ &= \frac{-1}{2\sigma_0^2}E\Big(\varepsilon_i\chi_i\psi_i\Big).\end{aligned}$$

Proof of Lemma is completed. ■

Proof of Theorem 4.1. Let $\hat{\lambda}$ be as in (4.3.6) and recall $G(\lambda) := W(I-\lambda W)^{-1}$. Set $\tilde{\omega}_1 = (h/n)\text{tr}\left(G(\tilde{\lambda})G(\tilde{\lambda})^T\right)$, $\tilde{\omega}_2 = (h/n)\text{tr}\left(G(\tilde{\lambda})^2\right)$.

By the mean value theorem applied to $r_L(\tilde{\lambda},\tilde{\sigma})$ in (4.3.7),

$$r_L(\hat{\lambda},\hat{\sigma}) = r_L(\lambda_0,\sigma_0)+\bar{s}_{1L}(\hat{\sigma}-\sigma_0)+\bar{s}_{2L}(\hat{\lambda}-\lambda_0),$$

where $\bar{s}_{1L} = (\partial/\partial\lambda)r_L(\bar{\lambda},\bar{\sigma})$ and $\bar{s}_{2L} = (\partial/\partial\sigma)r_L(\bar{\lambda},\bar{\sigma})$ are the first derivatives of r_L at some $(\bar{\lambda},\bar{\sigma})$ such that $|\bar{\lambda}-\lambda_0|\le|\tilde{\lambda}-\lambda_0|$ and $|\bar{\sigma}-\sigma_0|\le|\tilde{\sigma}-\sigma_0|$. Thus,

$$\begin{aligned}\hat{\lambda}-\lambda_0 &= (\tilde{\lambda}-\lambda_0)\left[1+\left\{\tilde{\mathcal{J}}\tilde{\omega}_1+\tilde{\omega}_2\right\}^{-1}\frac{h}{n}\cdot\bar{s}_{1L}\right] \\ &\quad +\left\{\tilde{\mathcal{J}}\tilde{\omega}_1+\tilde{\omega}_2\right\}^{-1}\frac{h}{n}\left[\bar{s}_{2L}(\tilde{\sigma}-\sigma_0)+r_L(\lambda_0,\sigma_0)\right].\end{aligned}\tag{4.7.2}$$

Let $\mathcal{N} = \left(\lambda,\sigma : |\lambda-\lambda_0|\le\sqrt{h/n}, |\sigma-\sigma_0|\le\sqrt{1/n}\right)$ be a small neighborhood of (λ_0,σ_0), which takes into account the different rates of convergence of MLE for the two parameters λ and σ in pure SAR model.

As in Robinson (2010a), the proof of consistency and asymptotic normality of the adaptive estimators $(\hat{\lambda},\hat{\sigma})$ consist of showing

$$\sqrt{\frac{h}{n}}r_L(\lambda_0,\sigma_0)\to_d N(0,\mathcal{J}\omega_1+\omega_2),\tag{4.7.3}$$

in addition to

$$\begin{aligned}&\tilde{\omega}_1\to_p\omega_1,\quad \tilde{\omega}_2\to_p\omega_2, && (4.7.4)\\ &\frac{h}{n}\cdot s_{1L}(\lambda_0,\sigma_0)\to_p-(\mathcal{J}\omega_1+\omega_2), && (4.7.5)\\ &\frac{h}{n}s_{2L}(\lambda_0,\sigma_0)\to_p 0, && (4.7.6)\\ &\tilde{\mathcal{J}}_L(\lambda_0,\sigma_0)\to_p\mathcal{J}, && (4.7.7)\\ &\sup_N|s_{iL}(\lambda,\sigma)-s_{iL}(\lambda_0,\sigma_0)| = o_p\left(\frac{n}{h}\right),\quad i=1,2, && (4.7.8)\\ &\sup_N|\tilde{\mathcal{J}}_L(\lambda,\sigma)-\tilde{\mathcal{J}}_L(\lambda_0,\sigma_0)| = o_p(1). && (4.7.9)\end{aligned}$$

Proof of (4.7.3).

Recall the sample log likelihood $L(\theta_0)$ from (4.2.3). We verify (4.7.3), by establishing

$$\sqrt{\frac{h}{n}}\frac{\partial L(\theta_0)}{\partial\lambda} \to_d N(0, \mathcal{J}\omega_1 + \omega_2), \tag{4.7.10}$$

$$r_L(\lambda_0, \sigma_0) - \frac{\partial L(\theta_0)}{\partial\lambda} = o_p(1). \tag{4.7.11}$$

To prove (4.7.10), write

$$\begin{aligned}\frac{\partial L(\theta_0)}{\partial\lambda} &= \sum_{i=1}^{n} \frac{W_i^T(y - \mu_0\ell)}{\sigma_0}\psi(\varepsilon_i) - tr(G) &(4.7.12)\\ &= (\psi(\varepsilon_1), \cdots, \psi(\varepsilon_n))\, G\varepsilon - tr(G) = \sum_{i=1}^{n}\eta_i, &(4.7.13)\end{aligned}$$

as the sum of martingale differences $\eta_i := (\varepsilon_i\psi(\varepsilon_i) - 1)g_{ii} + \varepsilon_i\sum_{j<i}\psi(\varepsilon_j)g_{ij} + \psi(\varepsilon_i)\sum_{j<i}\varepsilon_j g_{ji}$, which satisfy $E(\eta_i|\mathcal{F}_{i-1}) = 0$, $\mathcal{F}_i = \sigma(\varepsilon_j, j \le i)$. Therefore, we establish (4.7.10) by verifying the following sufficient conditions of central limit theorem for martingale differences, see Hall and Heyde (1980):

$$\frac{h}{n}\sum_{i=1} E(\eta_i^2|\mathcal{F}_{i-1}) \to_p \mathcal{J}\omega_1 + \omega_2, \tag{4.7.14}$$

$$\left|\frac{h}{n}\right|^{2+\delta}\sum_{i=1}^{n} E|\eta_i|^{2+\delta} \to 0. \tag{4.7.15}$$

Proof of (4.7.15). Firstly, noting $E(\varepsilon_i\psi_i) = 1, E(\varepsilon_i^2) = 1$ and $E(\psi_i^2) = \mathcal{J}$ and using *i.i.d.* property of $\{\varepsilon_i\}$,

$$\begin{aligned} E(\eta_i^2) &= g_{ii}^2[E(\varepsilon_i^2\psi_i^2) - 1] + \mathcal{J}\sum_{1\le j<i} g_{ij}^2 + \mathcal{J}\sum_{1\le j<i} g_{ji}^2 + 2\sum_{1\le j<i} g_{ij}g_{ji},\\ \sum_{i=1}^{n} E(\eta_i^2) &= \sum_{i=1}^{n} g_{ii}^2[E(\varepsilon_i^2\psi_i^2) - 2 - \mathcal{J}] + \mathcal{J}\sum_{i,j=1}^{n} g_{ij}^2 + \sum_{i,j=1}^{n} g_{ij}g_{ji}\\ &= O(1)\sum_{i=1}^{n} g_{ii}^2 + \mathcal{J}\mathrm{tr}(GG^T) + \mathrm{tr}(G^2).\end{aligned}$$

Therefore, by Assumption 2,

$$\frac{h}{n}\sum_{i=1}^{n} E(\eta_i^2) \to \mathcal{J}\omega_1 + \omega_2, \tag{4.7.16}$$

since $\frac{h}{n}\sum_{i=1}^{n} g_{ii}^2 = O\big(\frac{h}{n} \times \frac{n}{h^2}\big) = O\big(\frac{1}{h}\big) = o(1)$. Now, direct calculation, noting that

$E\varepsilon_i^2\psi_i = 0$ under Assumption 3, gives

$$\begin{aligned}
E(\eta_i^2|\mathcal{F}_{i-1}) - E(\eta_i^2) = &\quad \sum_{j,j'<i:j\neq j'} \psi_j\psi_{j'}g_{ij}g_{ij'} + \sum_{j<i} g_{ij}^2(\psi_j^2 - \mathcal{J}) \\
&+ \quad \mathcal{J} \sum_{j,j'<i:j\neq j'} \varepsilon_j\varepsilon_{j'}g_{ji}g_{j'i} + \mathcal{J}\sum_{j<i} g_{ji}^2(\varepsilon_j^2 - 1) \\
&+ \quad 2g_{ii}E(\psi_i^2\varepsilon_i)\sum_{j<i} g_{ji}\varepsilon_j \\
&+ \quad 2\sum_{j<i}\sum_{j'<i:j\neq j'} \psi_j\varepsilon_{j'}g_{ij}g_{j'i} + 2\sum_{j<i}(\psi_j\varepsilon_j - 1)g_{ij}g_{ji} \\
&=: \quad m_{1i} + \cdots + m_{7i}.
\end{aligned}$$

In view of (4.7.16), to prove (4.7.15), it suffices to show that

$$\begin{aligned}
&\frac{h}{n}\sum_{i=1}^n \big[E(\eta_i^2|\mathcal{F}_{i-1}) - E(\eta_i^2)\big] \\
&\quad = \frac{h}{n}\sum_{i=1}^n m_{1i} + \cdots + \frac{h}{n}\sum_{i=1}^n m_{7i} = o_p(1),
\end{aligned}$$

which is verified once we establish

$$E\big[\big(\frac{n}{h}\sum_{i=1}^n m_{di}\big)^2\big] = o\big(1\big), \quad \text{for} \quad d = 1, \cdots 7. \tag{4.7.17}$$

We first verify (4.7.17) for d=1.

$$\begin{aligned}
E\big[\big(\sum_{i=1}^n m_{1i}\big)^2\big] &= E\big[\big(\sum_{i=1}^n \sum_{j,j'<i:j\neq j'} \psi_j\psi_{j'}g_{ij}g_{ij'}\big)^2\big] \\
&\leq 2\sum_{i,i'=1}^n \sum_{j<i}\sum_{k<i'} |g_{ij}g_{i'j}g_{ik}g_{i'k}|E(\psi_j^2)E(\psi_k^2) \\
&\leq C\sum_{i,i',j,k=1}^n |g_{ij}g_{i'j}g_{ik}g_{i'k}|.
\end{aligned}$$

Thus recalling (4.2.2),

$$\begin{aligned}
E\big[\big(\sum_{i=1}^n m_{1i}\big)^2\big] &\leq Ch^{-1}\sum_{i,i',j,k=1}^n |g_{ij}g_{i'j}g_{i'k}| \\
&\leq Ch^{-1}\big(\sum_{i=1}^n 1\big)\cdot \max_{i'}\sum_{k=1}^n |g_{i'k}| \max_j \sum_{i'=1}^n |g_{i'j}| \max_i \sum_{j=1}^n |g_{ij}| \\
&\leq C\Big(\frac{n}{h}\Big) = o\Big(\frac{n^2}{h^2}\Big).
\end{aligned}$$

Verification of (4.7.17) for $d = 3, 6$ follows similar steps as in the proof for $d = 1$.

To establish (4.7.17) for $d = 2$, recall that $\psi_i^2 - \mathcal{J} = \psi_i^2 - E\psi_i^2$ is an *i.i.d.* sequence. Thus,

$$\begin{aligned}
E\big[\big(\sum_{i=1}^n m_{2i}\big)^2\big] &= E\big[\big(\sum_{i=1}^n \sum_{j<i} g_{ij}^2(\psi_j^2 - \mathcal{J})\big)^2\big] \le E\big((\psi_1^2 - E\psi_1^2)^2\big) \sum_{i,i',j=1}^n g_{ij}^2 g_{i'j}^2 \\
&= C\sum_{j=1}^n \Big(\sum_{i=1}^n g_{ij}^2\Big)^2 \le C\sum_{j=1}^n \Big(\max_i |g_{ij}| \sum_{i=1}^n |g_{ij}|\Big)^2 \\
&\le \big(\max_{i,j} |g_{ij}|\big)^2 \sum_{j=1}^n \max_j \sum_{i=1}^n |g_{ij}| \sum_{i'=1}^n |g_{i'j}| \le Ch^{-2}nO(1)O(1) = o\Big(\frac{n^2}{h^2}\Big).
\end{aligned}$$

Verifications of (4.7.17) for $d = 4, 5, 7$ follows similar steps.

Proof of (4.7.15). It holds $|a+b|^{2+\delta} \le C(|a|^{2+\delta} + |b|^{2+\delta})$. Therefore,

$$\begin{aligned}
\sum_{i=1}^n E|\eta_i|^{2+\delta} \quad &\le \quad C\Big(\sum_{i=1}^n |g_{ii}|^{2+\delta} E|\varepsilon_i\psi_i|^{2+\delta} + \sum_{i=1}^n E|\varepsilon_i|^{2+\delta} E|\sum_{j<i} g_{ij}\psi_j|^{2+\delta} \\
&\qquad + \sum_{i=1}^n E|\psi_i|^{2+\delta} E|\sum_{j<i} g_{ji}\varepsilon_j|^{2+\delta}\Big) \\
&\le \quad C\Big(\sum_{i=1}^n |g_{ii}|^{2+\delta} + \sum_{j<i} E|g_{ij}\psi_j|^{2+\delta} + \sum_{j<i} E|g_{ji}\varepsilon_j|^{2+\delta}\Big) =: C(p_{1n} + p_{2n} + p_{3n}).
\end{aligned}$$

To prove (4.7.15), we need to verify that $p_{dn} = o\big((n/h)^{2+\delta}\big)$ for $d = 1, 2, 3$. Firstly, for $d = 1$, using $|g_{ii}| = O(1/h)$,

$$p_{1n} = O\big(\frac{n}{h^{2+\delta}}\big) = o\big(\frac{n^{2+\delta}}{h^{2+\delta}}\big).$$

For $d = 2$, by Burkholder inequality (see Burkholder (1973)),

$$\sum_{i=1}^n E|\sum_{j<i} g_{ij}\psi_j|^{2+\delta} \le C\sum_{i=1}^n \big[\sum_{j=1}^n E(g_{ij}^2\psi_j^2)\big]^{(2+\delta)/2},$$

where for any $i = 1, \cdots, n$, by Assumption 3,

$$\Big(\sum_{j=1}^n E(g_{ij}^2\psi_j^2)\Big)^{(2+\delta)/2} \quad = \quad C\Big(\sum_{j=1}^n g_{ij}^2\Big)^{(2+\delta)/2} \le C\Big(\max_j |g_{ij}| \sum_{j=1}^n |g_{ij}|\Big)^{(2+\delta)/2} = O\big(\frac{1}{h^{1+\delta/2}}\big).$$

Therefore, $p_{2n} = O\big(\frac{n}{h^{1+\delta/2}}\big) = o\big((n/h)^{2+\delta}\big)$. Proof of $p_{3n} = o\big((n/h)^{2+\delta}\big)$ follows similar steps.

Proof of (4.7.11). Let, for the brevity, r_L, G and $\tilde{\psi}_{iL}$ denote quantities evaluated at the true parameter values (θ_0, σ_0) and ψ_i abbreviates $\psi(\varepsilon_i)$. Then we can write

$$\begin{aligned} r_L - \frac{\partial L(\theta_0)}{\partial \lambda} &= \left(\tilde{\psi}_{1L}, \cdots, \tilde{\psi}_{nL}\right) HG\varepsilon - tr\left\{G\right\} - \left(\psi_1, ..., \psi_n\right) G\varepsilon + tr(G) \\ &= \left(\tilde{\psi}_{1L} - \psi_1, \cdots, \tilde{\psi}_{nL} - \psi_n\right) HG\varepsilon + \left(\psi_1, \cdots, \psi_n\right)(H - I)G\varepsilon \\ &= q_{n,1} + q_{n,2}. \end{aligned}$$

It remains to show $q_{n,i} = o_p(\sqrt{n/h})$, $i = 1, 2$. First consider $i = 2$. Denote $G = (g_{ij})$. Then,

$$\begin{aligned} q_{n,2} &= \left(\psi_1, \cdots, \psi_n\right)(H - I)G\varepsilon = -\frac{1}{n}\left(\psi_1, \cdots, \psi_n\right)(1_n 1_n^T)G\varepsilon \\ &= -n^{-1}\Big[\sum_{j=1}^n \varepsilon_j \big(\sum_{i=1}^n g_{ij}\big)\Big] \cdot \sum_{m=1}^n \psi_m, \end{aligned}$$

where $\sum_{m=1}^n \psi_m = \sum_{m=1}^n \psi(\varepsilon_m) = O_p(\sqrt{n})$ due to ε_j's being *i.i.d.* By Assumption 1, $\max_{1\le j\le n} \big(\sum_{i=1}^n |g_{ij}|\big) < C$ uniformly over j. Then, $n^{-1}E\big(\sum_{j=1}^n \varepsilon_j\big(\sum_{i=1}^n g_{ij}\big)\big) = 0$, and

$$\mathrm{Var}\big(n^{-1}\sum_{j=1}^n \varepsilon_j\big(\sum_{i=1}^n g_{ij}\big)\big) = n^{-2}\big(\sum_{j=1}^n\big(\sum_{i=1}^n g_{ij}\big)^2\big) = O\left(n^{-1}\right).$$

Hence, $q_{n,2} = O_p(n^{-1/2})O_p(n^{1/2}) = O_p(1) = o_p(\sqrt{n/h})$, because $n/h \to \infty$.

Next, we show $q_{n,1} = o_p(\sqrt{n/h})$. In the following quantities introduced below, the triangular array structure is present but the n-subscript is suppressed. Let

$$t_{ij} := \ell_j^T G^T \ell_i - \sum_{m=1}^n \ell_j^T G^T 1_n/n = g_{ij} - \frac{1}{n}\sum_{m=1}^n g_{mj}, \quad \chi_i := \varepsilon^T G^T(\ell_i - n^{-1}1_n) = \sum_{j=1}^n \varepsilon_j t_{ij},$$

where ℓ_i stands for the i^{th} column of I and the equality $\sum_{i=1}^n \chi_i = 0$ holds, arising from $\sum_{i=1}^n t_{ij} = 0$ for $j = 1, \cdots, n$. As pointed out in Robinson (2010, pp. 18), Assumption 1 implies $|t_{ij}| = O(1/h)$ uniformly over i and j, following from $\max_{1\le i,j\le n} |g_{ij}| = O\big(\frac{1}{h}\big)$. Let

$$a_l := \sum_{i=1}^n b_{li}\chi_i, \quad l = 2, 3, 4.$$

Write

$$\begin{aligned}\tilde{\psi}^{(L)}(\lambda_0,\sigma_0)-\psi(\varepsilon_i) &= \left[\bar{\psi}^{(L)}\left(\varepsilon_i;a^{(L)}\right)-\psi(\varepsilon_i)\right]+\left[\psi^{(L)}\left(\varepsilon_i;\tilde{a}^{(L)}(\varepsilon)\right)-\bar{\psi}^{(L)}\left(\varepsilon_i;a^{(L)}\right)\right]\\ &+ \left[\tilde{\psi}^{(L)}(\lambda_0,\sigma_0)-\psi^{(L)}\left(\varepsilon_i;\tilde{a}^{(L)}(\varepsilon)\right)\right] =: c_{2i}+c_{3i}+c_{4i}.\end{aligned}$$

We can rewrite

$$q_{n,1}=\left(\tilde{\psi}_{1L}-\psi_1,\cdots,\tilde{\psi}_{nL}-\psi_n\right)HG\varepsilon=\sum_{i=1}^n c_{2i}\chi_i+\sum_{i=1}^n c_{3i}\chi_i+\sum_{i=1}^n c_{4i}\chi_i := a_2+a_3+a_4.$$

To prove $q_{n,1}=o_p(\sqrt{n/h})$, we show that

$$a_\ell=o_p(\sqrt{n/h}),\quad \ell=2,3,4. \tag{4.7.18}$$

Proof of (4.7.18) for $i=2$. It requires the projection error, arising from projecting the score function onto the space spanned by the functionals of our series estimation, to be of small enough order, as required in Assumption 6.

Write down a_2 as in (A.27) of Robinson (2010a):

$$a_2=\sum_{i=1}^n c_{2i}\varepsilon_i t_{ii}+\sum_{i,j=1:j\neq i}^n c_{2i}\varepsilon_i t_{ij}, \tag{4.7.19}$$

recalling $c_{2i}=\bar{\psi}^{(L)}\left(\varepsilon_i;a^{(L)}\right)-\psi(\varepsilon_i)$. Then,

$$E\Big|\sum_{i=1}^n c_{2i}\varepsilon_i t_{ii}\Big|\le\left\{E\left(c_{2i}^2\right)\right\}^{\frac{1}{2}}\sum_{i=1}^n|t_{ii}|=o\Big(\sqrt{\frac{h}{n}}\Big)\cdot O\Big(\frac{n}{h}\Big)=o\left(\sqrt{\frac{n}{h}}\right), \tag{4.7.20}$$

by Assumptions 1 and 6. The second term of (4.7.19) has zero mean and

$$\begin{aligned}\mathrm{Var}\Big(\sum_{i,j=1:j\neq i}^n c_{2i}\varepsilon_i t_{ij}\Big) &= E\Big[\big(\sum_{i<j}c_{2i}\varepsilon_i t_{ij}\big)+\big(\sum_{j\le i}c_{2i}\varepsilon_i t_{ij}\big)\Big]^2\\ &\le 2E\Big[\big(\sum_{i<j}c_{2i}\varepsilon_i t_{ij}\big)^2\Big]+2E\Big[\big(\sum_{j\le i}c_{2i}\varepsilon_i t_{ij}\big)^2\Big].\end{aligned} \tag{4.7.21}$$

The first expectation can be bounded by

$$\begin{aligned}&2\sum_{i<j}\sum_{i'<j'}\big|E[c_{2i}\varepsilon_j c_{2i'}\varepsilon_{j'}]t_{ij}t_{i'j'}\big|\\ &\le 2\sum_{i<j}E(c_{2i}^2)E(\varepsilon_i^2)|t_{ij}t_{i'j'}|=2\sum_{i<j}O_p\big(\frac{h}{n}\big)O_p\big(\frac{1}{h}\big)O_p\big(\frac{1}{h}\big)=o_p\big(\frac{n}{h}\big),\end{aligned}$$

using independence of ε_j's, the bound $Ec_{i2}^2=o_p(h/n)$ from Assumption 6 and $t_{ij}=O(1/h)$.

The same bound holds for the second term in (4.7.21) which yields $a_2=o_p(\sqrt{n/h})$.

To prove (4.7.18) for $i = 3, 4$, we shall use the following notation. Let

$$\begin{aligned} \pi_L &:= (\log L)\eta^{2L}1(\varphi < 1) + (L\log L)\eta^{2L}1(\varphi = 1) + (\log L)(\eta\varphi)^{2L}1(\varphi > 1) \\ &\leq L(\log L)A^{2L}, \end{aligned} \tag{4.7.22}$$

with $A = \eta\max(\varphi, 1)$. Note that $A > 1$.

Set

$$\begin{aligned} \rho_{uL} &= CL, \quad \text{if } u = 0, \\ &= (CL)^{uL/\omega}, \quad \text{if } u > 0 \text{ and Assumption 5(ii) holds}, \\ &= C^L, \quad \text{if } u > 0 \text{ and Assumption 5(iii) holds}. \end{aligned}$$

Proof of (4.7.18) for $i = 3$. Proof is based on an extensive use of Assumption 5. Equations (A.31)-(A.39) of pages 19-20 of Robinson (2010a) yield the upper bound on the stochastic order of a_3:

$$\begin{aligned} a_3 &= O_p\left(\frac{\sqrt{n}}{h}L^{3/2}\rho_{2\kappa L}\rho_{4\kappa L}^{\frac{1}{2}}\pi_L^2\right) \\ &= O_p\left(\sqrt{\frac{n}{h}}\frac{H_3}{\sqrt{h}}\right), \quad H_3 := L^{3/2}\rho_{2\kappa L}\rho_{4\kappa L}^{\frac{1}{2}}\pi_L^2. \end{aligned} \tag{4.7.23}$$

To prove (4.7.18) for $i = 3$, it remains to show

$$H_3 = o(\sqrt{h}). \tag{4.7.24}$$

Case 1. Let Assumption 5 (i) hold. Then, $\rho_{2\kappa L} = \rho_{4\kappa L} = CL$ and $H_3 = C^{3/2}L^3\pi_L^2$. Notice that for any $p > 0$ and $\varepsilon > 0$,

$$L^p = o\big((1+\varepsilon)^L\big). \tag{4.7.25}$$

Hence, as $L \to \infty$,

$$\pi_L^2 = o\big((1+\varepsilon)^L A^{4L}\big), \quad \forall\varepsilon > 0. \tag{4.7.26}$$

Combining (4.7.25) and (4.7.26), we obtain $H_3 = o\big([(1+\varepsilon)A]^{4L}\big), \quad \forall\varepsilon > 0.$

Thus, to prove that $H_3 = o(\sqrt{h})$, it suffices to show that

$$\begin{aligned} &[(1+\varepsilon)A]^{4L} \leq \sqrt{h}, \quad \text{i.e.} \\ &4L\log[(1+\varepsilon)A] \leq (1/2)\log h, \quad \text{or} \quad L \leq \frac{\log h}{8\log[(1+\varepsilon)A]}, \end{aligned} \tag{4.7.27}$$

which is valid for small $\varepsilon \geq 0$ by Assumption 5 (i).

Case 2. Let Assumption 5 (ii) hold. Then, $\rho_{aL} = (CL)^{\frac{aL}{\omega}}$ and

$$H_3 = L^{\frac{3}{2}}\rho_{2\kappa L}\rho_{4\kappa L}^{\frac{1}{2}}\pi_L^2 = L^{\frac{3}{2}}C^{\frac{4\kappa L}{\omega}}L^{\frac{4\kappa L}{\omega}}\pi_L^2.$$

Observe that for any $C > 0$, $p > 0$, $a > 0$ and $\varepsilon > 0$,

$$L^p = o(L^{\varepsilon L}), \quad C^{aL} = o(L^{\varepsilon L}). \tag{4.7.28}$$

Hence by (4.7.22),

$$\pi_L^2 = o(L^{\varepsilon L}), \quad \forall\varepsilon > 0, \tag{4.7.29}$$

and $H_3 = o\big(L^{L(\frac{4\kappa}{\omega}+\varepsilon)}\big)$, $\forall\varepsilon > 0$. Thus, $H_3 = o(\sqrt{h})$ holds if

$$\begin{aligned} &L^{L(\frac{4\kappa}{\omega}+\varepsilon)} \le \sqrt{h}, \quad \text{i.e.} \\ &\big(\frac{4\kappa}{\omega}+\varepsilon\big)L\log L \le \frac{1}{2}\log h, \quad \text{or} \quad L\log L \le \frac{\log h}{2\big(\frac{4\kappa}{\omega}+\varepsilon\big)}, \end{aligned} \tag{4.7.30}$$

which is valid for small $\varepsilon > 0$ by Assumption 5 (ii).

Case 3. Let Assumption 5(iii) hold. Then, $\rho_{aL} = C^L$, and $H_3 = L^{\frac{3}{2}}\rho_{2\kappa L}\rho_{4\kappa L}^{\frac{1}{2}}\pi_L^2 = L^{\frac{3}{2}}C^{\frac{3L}{2}}\pi_L^2$. Then by (4.7.28) and (4.7.29), $H_3 = o(L^{\varepsilon L}), \quad \forall\varepsilon > 0$. Thus, $H_3 = o(\sqrt{h})$, if

$$\begin{aligned} &L^{\varepsilon L} \le \sqrt{h}, \quad \text{i.e.} \\ &\varepsilon L\log L \le \frac{1}{2}\log h, \quad \text{or} \quad L\log L \le \frac{1}{2\varepsilon}\log h, \end{aligned} \tag{4.7.31}$$

which is valid for sufficiently small $\varepsilon > 0$ by Assumption 5 (ii).

Now, we prove (4.7.18) for $i = 4$.

Following (A.45)-(A.56) of Robinson (2010a), we obtain the following upper bound

$$\begin{aligned} &a_4 = O_p\big(\frac{\sqrt{n}}{h}H_4\big), \\ &H_4 := \rho_{2\kappa L}\pi_L \times \Big\{C^{\kappa L}L^{\frac{7}{2}} + \rho_{2\kappa L}\pi_L L^2 + \rho_{2\kappa L}\pi_L(CL)^{4\kappa L+3}n^{-\frac{1}{2}}\log n\Big\}. \end{aligned}$$

It remains to show that

$$H_4 = o_p(\sqrt{h}). \tag{4.7.32}$$

Case 1. Under Assumption 5 (i), $\rho_{2\kappa L} = CL$, and

$$H_4 = \pi_L L^{\frac{9}{2}} + \pi_L^2 L^4 + \pi_L^2 L^5 n^{-\frac{1}{2}}\log n.$$

By (4.7.25) and (4.7.26),

$$H_4 = o\big([(1+\varepsilon)A]^{2L} + [(1+\varepsilon)A]^{4L}(1+n^{-1/2}\log n)\big) = o\big([(1+\varepsilon)A]^{4L}\big).$$

Hence $H_4 = o(\sqrt{h})$, if $[(1+\varepsilon)A]^{4L} \leq \sqrt{h}$, which is true for small $\varepsilon > 0$ as shown in (4.7.27).

Case 2. Let Assumption 5 (ii) hold. Then $\rho_{aL} = (CL)^{\frac{aL}{\omega}}$ and

$$H_4 = C^{\kappa L}(CL)^{2\kappa L/\omega}L^{7/2}\pi_L + (CL)^{4\kappa L/\omega}L^2\pi_L^2 + \frac{(CL)^{4\kappa L(1+1/\omega)}L^3\pi_L^2}{\sqrt{n}/\log n}.$$

By (4.7.28) and (4.7.29),

$$\begin{aligned} H_4 &= o\big(L^{(\frac{2\kappa}{\omega}+\varepsilon)L} + L^{(\frac{4\kappa}{\omega}+\varepsilon)L} + \frac{L^{(4\kappa(1+1/\omega)+\varepsilon)L}}{\sqrt{n}/\log n}\big) \\ &= o\Big(L^{(\frac{4\kappa}{\omega}+\varepsilon)L}(1+\frac{L^{4\kappa L}}{\sqrt{n}/\log n})\Big). \end{aligned}$$

By (4.7.30), $L^{(\frac{4\kappa}{\omega}+\varepsilon)L} \leq \sqrt{h}$, if $\varepsilon > 0$ is small. Next, for any $\delta > 0$, $\sqrt{n}/\log n \geq n^{\frac{1}{2}-\delta} \geq h^{\frac{1}{2}-\delta}$. Hence by the same arguments as in proving (4.7.27), we obtain that

$$\frac{L^{4\kappa L}}{n^{1/2}/\log n} \leq \frac{L^{4\kappa L}}{n^{\frac{1}{2}-\delta}} \leq 1, \tag{4.7.33}$$

if $L\log L \leq (\frac{1}{2}-\delta)\log h/4\kappa$ which holds for small δ. Hence $H = o(\sqrt{h})$, and (4.7.32) holds.

Case 3. Under Assumption 5(iii), $\rho_{aL} = C^L$ and

$$H_4 = C^{(\kappa+1)L}L^{7/2}\pi_L + C^{2L}L^2\pi_L^2 + C^{(4\kappa+2)L+3}L^{4\kappa L+3}\pi_L^2 n^{-\frac{1}{2}}\log n.$$

By (4.7.28) and (4.7.29), $H_4 = o\big(L^{\varepsilon L} + \frac{L^{(4\kappa+\varepsilon)L}}{\sqrt{n}/\log n}\big)$. By (4.7.31), $L^{\varepsilon L} \leq \sqrt{h}$. Hence, to prove that $H_4 = o(\sqrt{h})$, it remains to show that

$$\frac{L^{(4\kappa+\varepsilon)L}}{\sqrt{n}/\log n} \leq \frac{L^{(4\kappa+\varepsilon)L}}{\sqrt{h}/\log h} \leq \sqrt{h},$$

where the first inequality holds because $h \leq n$. For that we shall verify that for small $\delta > 0$, $L^{(4\kappa+\varepsilon)L} \leq h^{1-\delta}$, i.e.

$$(4\kappa+\varepsilon)L\log L \leq (1-\delta)\log h, \quad \text{or} \quad L\log L \leq \frac{(1-\delta)}{(4\kappa+\delta)}\log h,$$

which follows from Assumption 5 (iii) when δ and ε are small enough. This completes the proof of (4.7.3), which is by far the most difficult and distinctive part of the Theorem proofs. ■

4.7.1 Proofs of (4.7.4)-(4.7.9)

In the rest of this appendix, we will give a detailed proof of (4.7.4) and (4.7.5) and comments on the proofs of (4.7.6) to (4.7.9).

Some preliminaries

In the proofs below, the vector norm used is Euclidean norm, denoted $\|\cdot\|$, and four matrix norms are used: the spectral norm $\|\cdot\|$, Euclidean norm $\|\cdot\|_E$, the maximum column sum norm $\|\cdot\|_C$, and the maximum row sum norm $\|\cdot\|_R$. For a $n\times 1$ vector $a=(a_1,\cdots,a_n)^T$ and $n\times n$ matrix A, one has

$$\|a\|=\sum_{i=1}^n a_i^2, \qquad \|A\|^2:=\bar{\lambda}(A'A), \qquad \|A\|_E^2:=\big(\sum_{i,j=1}^n a_{ij}^2\big),$$

$$\|A\|_C:=\max_{1\le j\le n}\big(\sum_{i=1}^n|a_{ij}|\big),\quad \|A\|_R:=\max_{1\le i\le n}\big(\sum_{j=1}^n|a_{ij}|\big),$$

where $\bar{\lambda}(A'A)$ is the largest eigenvalue of the matrix $A'A$. The afore-mentioned matrix norms are all *submultiplicative*, i.e. for conformable square matrices A and B, it holds $\|AB\|\le\|A\|\|B\|$. For square matrices A and B, the following inequalities will be useful later:

$$\|A\|\le\|A\|_E,\quad \|A\|^2\le\|A\|_R\|A\|_C,\quad |tr(AB)|\le\|A\|_E\|B\|_E,$$
$$\|AB\|_E\le\|A\|_E\|B\|,\quad \|AB\|_E\le\|A\|_E\|B\|_E. \tag{4.7.34}$$

For a square matrix of functions of a scalar parameter λ, $A=A(\lambda)$, the following three results will be used in the proofs:

$$\frac{d}{d\lambda}A^{-1}=-A^{-1}\big(\frac{d}{d\lambda}A\big)A^{-1}, \tag{4.7.35}$$
$$\frac{d}{d\lambda}\log|A|=\mathrm{tr}\big(A^{-1}\frac{d}{d\lambda}A\big),\quad \text{where } |\cdot| \text{ denotes the determinant,}$$
$$\|A(\lambda_1)-A(\lambda_2)\|\le|\lambda_1-\lambda_2|\big\|\frac{d}{d\lambda}A(\lambda)\big\| \quad \text{for some } \theta,\quad |\lambda-\lambda_2|\le|\lambda_1-\lambda_2|.$$

The above facts can be found in Searle (1982), Horn and Johnson (1990) and the Appendix of Davies (1973).

Next, we establish some properties for the matrices that appear frequently in the proofs. Assumption 1 stated that the weight matrix W and the matrix $S^{-1}=(I-\lambda_0W)^{-1}$ are both uniformly bounded in row and column sums, i.e. $\|W\|_R$, $\|W\|_C$, $\|S^{-1}\|_R$, $\|S^{-1}\|_C=O(1)$. Assumption 1 also requires $\max_{1\le i,j,\le n}|w_{ij}|=O(1/h)$. Hence, by submultiplicative property of the norms $\|\cdot\|_C$ and $\|\cdot\|_R$, the matrix $G=WS^{-1}$ is also uniformly bounded in both row and column sums. Furthermore, the elements of G are uniformly bounded by $O(1/h)$. The (i,j)-th element is $g_{ij}=W_iS^{-1}\ell_j$, where W_i is the i-th row of W, and ℓ_j is the j-th column of I. Denote by $(S^{-1})_{kj}$, the

(k,j)-th element of S^{-1}. Then, uniformly in i and j,

$$|g_{ij}| = |\sum_{k=1}^{n} w_{ik} \cdot (S^{-1})_{kj}| \le \max_{k} |w_{ik}| \sum_{k=1}^{n} |(S^{-1})_{kj}| = O(\frac{1}{h}). \tag{4.7.36}$$

We introduced earlier in Section 4.3 the $n \times n$ matrix $H = I - \frac{1}{n} 1_n 1_n^T$, which has bounded row and column sums. We denote $T := HG = \{t_{ij}\}$, which is used in defining $(\chi_1, \cdots, \chi_n)^T = HG\varepsilon = T\varepsilon$, where $\chi_i = \sum_{j=1}^{n} t_{ij}\varepsilon_j, i = 1, \cdots, n$. One can verify $\|T\|_R, \|T\|_C = O(1)$ and $\max_{1\le i,j,\le n} |t_{ij}| = O(1/h)$ using the same argument as in the case of G. These properties also hold for products G^TG, GG^T, T^TT, and TT^T, which can be verified by the same reasoning.

Recall that Assumption 4 set out the following form for the series function used in the estimation of the score function: $\phi_\ell(s) = \phi^\ell(s)$, $\ell = 1, \cdots, n$, where $\phi(s)$ is strictly increasing and thrice differentiable function such that for some $\kappa \ge 0$ and $K > 0$,

$$|\phi(s)| \le 1 + |s|^\kappa, \quad |\phi'(s)| + |\phi''(s)| + |\phi'''(s)| \le C(1 + |\phi(s)|^K), \quad s \in \mathbb{R},$$

where C denotes a generic constant throughout this proof. For $\kappa = 0$, $|\phi(s)|, |\phi'(s)|, |\phi''(s)|$ and $|\phi'''(s)|$ are bounded. For $\kappa > 0$, Assumption 4 allows tails of series functions $\phi_\ell(\cdot)$ and their derivatives to diverge, at a rate increasing with ℓ. We introduce the quantity $\mu_c = 1 + E|\varepsilon_i|^c, c > 0$, which is useful in bounding the moments of above functions.

Recall $\phi_\ell(s) = \phi^\ell(s)$. We have

$$\begin{aligned}
&|\phi_\ell(s)| \le C^\ell(1 + |s|^{\kappa\ell}),\\
&|\phi_\ell'(s)| = \ell|\phi'(s)\phi^{\ell-1}(s)| \le C^\ell \ell\big(1 + |s|^{\kappa(\ell-1+K)}\big),\\
&|\phi_\ell''(s)| = |\ell(\ell-1)\phi^{\ell-2}(s)(\phi'(s))^2 + \ell\phi^{\ell-1}(s)\phi''(s)| \le C^\ell \ell^2\big(1 + |s|^{\kappa(\ell-1+2K)}\big),\\
&|\phi_\ell'''(s)| \le C^\ell \ell^3\big(1 + |s|^{\kappa(\ell-1+3K)}\big).
\end{aligned}$$

Therefore, for $r > 0$,

$$\begin{aligned}
&E|\phi_\ell(\varepsilon_1)|^r \le C^{\ell r}\mu_{\kappa r\ell},\\
&E|\phi_\ell'(\varepsilon_1)|^r \le C^{\ell r}\ell^r\mu_{\kappa r(\ell-1+K)} \le C^{\ell r}\ell^r\mu_{\kappa r(\ell+K)},\\
&E|\phi_\ell''(\varepsilon_1)|^r \le C^{\ell r}\ell^{2r}\mu_{\kappa r(\ell-1+2K)} \le C^{\ell r}\ell^{2r}\mu_{\kappa r(\ell+2K)},\\
&E|\phi_\ell'''(\varepsilon_1)|^r \le C^{\ell r}\ell^{3r}\mu_{\kappa r(\ell-1+3K)} \le C^{\ell r}\ell^{3r}\mu_{\kappa r(\ell+3K)}.
\end{aligned} \tag{4.7.37}$$

Lemma 9 of Robinson (2005) established that $\sum_{\ell=1}^{L} \mu_{a\ell+b} \le \rho_{aL}$ for any $a, b \ge 0$. Trivially, $|\mu_a|^r \le \mu_{ar}$ for $a, r \ge 0$.

Proof of (4.7.4). We will prove (4.7.4) for $i = 2$. Recall $G(\lambda) = W(I - \lambda W)^{-1}$, for ease of notation, denote $\tilde{G} = G(\tilde{\lambda})$ and $G = G(\lambda_0)$. We assumed $\omega_2 = \lim_{n\to\infty} \frac{h}{n}\mathrm{tr}(G^2)$ is finite and nonzero, and denote $\tilde{\omega}_2 = \frac{h}{n}\mathrm{tr}(\tilde{G}^2)$. Therefore, (4.7.4) for $i = 2$ is established if we show $\big|\frac{h}{n}\big(\mathrm{tr}(\tilde{\mathrm{G}}^2) - \mathrm{tr}(\mathrm{G}^2)\big)\big| = \mathrm{o_p}(1)$. By linearity of the trace operator and (4.7.34),

$$\begin{aligned} \big|\mathrm{tr}(\tilde{G}^2) - \mathrm{tr}(G^2)\big| &= \big|\mathrm{tr}(\tilde{G}^2 - G^2)\big| = \big|\mathrm{tr}\big[\tilde{G}(\tilde{G} - G)\big] + \mathrm{tr}\big[(\tilde{G} - G)G\big]\big| \\ &\leq \|\tilde{G}\|_E\|\tilde{G} - G\|_E + \|G\|_E\|\tilde{G} - G\|_E. \end{aligned} \quad (4.7.38)$$

Lee (2004, pp.1918) established that if $S^{-1}(\lambda_0)$ exhibits row and column summability, as assumed in Assumption 1 (iii), then the same holds uniformly in λ for $S^{-1}(\lambda)$, for λ's in some neighbourhood of λ_0. Another consequence of row and column summability of $S^{-1}(\tilde{\lambda})$ is that every element of $G(\tilde{\lambda})$ is uniformly of order $O_p(1/h)$, by the same argument as in (4.7.36). Hence, it follows that

$$\|\tilde{G}\|_E = \left(\sum_{i,j=1}^{n} \tilde{g}_{ij}^2\right)^{1/2} \leq \left(\sum_{i=1}^{n} \max_j |\tilde{g}_{ij}| \sum_{j=1}^{n} |\tilde{g}_{ij}|\right)^{1/2} = O_p\big(\sqrt{\frac{n}{h}}\big),$$

and $\|G\|_E = O\big(\sqrt{n/h}\big)$ applying the same steps as above.

To find the upper bound on the RHS of (4.7.38), we need to find that on $\|\tilde{G} - G\|_E$. We have that

$$\|\tilde{G} - G\|_E \leq |\tilde{\lambda} - \lambda_0|\|\frac{dG(\lambda)}{d\lambda}\| \quad \text{for some } \lambda, \quad |\lambda - \lambda_0| \leq |\tilde{\lambda} - \lambda_0|,$$

where

$$\begin{aligned} \frac{dG(\lambda)}{d\lambda} = W\frac{d(I - \lambda W)^{-1}}{d\lambda} &= -W(I - \lambda W)^{-1}\frac{d(I - \lambda W)}{d\lambda}(I - \lambda W)^{-1} \\ &= W(I - \lambda W)^{-1}W(I - \lambda W)^{-1} = G^2(\lambda). \end{aligned}$$

By (4.7.34), $\|G^2\|_E \leq \|G\|_E\|G\| \leq \|G\|_E\sqrt{\|G\|_R\|G\|_C} = O_p\big(\sqrt{n/h}\big)$, and therefore,

$$\|\tilde{G} - G\|_E = O_p(\sqrt{\frac{h}{n}})O\big(\sqrt{\frac{n}{h}}\big) = O_p(1).$$

Hence,

$$\begin{aligned} &\big|\mathrm{tr}(\tilde{G}^2) - \mathrm{tr}(G^2)\big| \leq \|\tilde{G}\|_E\|\tilde{G} - G\|_E + \|G\|_E\|\tilde{G} - G\|_E = O_p\big(\sqrt{\frac{n}{h}}\big), \\ &\frac{h}{n}\big|\mathrm{tr}(\tilde{G}^2) - \mathrm{tr}(G^2)\big| = O_p\Big(\sqrt{\frac{h}{n}}\Big). \end{aligned}$$

Proof of (4.7.5). Recall $s_{1L} = (\partial/\partial\lambda)r_L(\lambda_0, \sigma_0)$ is the first derivatives of r_L w.r.t. λ at the true value of parameters (λ_0, σ_0). Recall that the fitted residuals at (λ, σ) are,

$$\frac{\epsilon(\lambda)}{\sigma} = \frac{HS(\lambda)y}{\sigma}, \qquad \frac{\epsilon(\lambda_0)}{\sigma_0} = \frac{HS(\lambda_0)y}{\sigma_0} = H\varepsilon.$$

From here on, for brevity we denote $\epsilon = \epsilon(\lambda_0)$ and $\epsilon_i = \epsilon_i(\lambda_0)$. Below, we get $\epsilon_i/\sigma_0 = \varepsilon_i - \bar{\varepsilon}$ where $\bar{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i/n$. The following derivative is used repeatedly throughout the proof,

$$\frac{d\big(\epsilon(\lambda)/\sigma\big)}{d\lambda}\Big|_{\sigma_0,\lambda_0} = \frac{1}{\sigma_0}\frac{dHS(\lambda)y}{d\lambda}\Big|_{\sigma_0} = -\frac{1}{\sigma_0}HW(S^{-1}\sigma_0\varepsilon + \mu_0 1_n) = -HG\varepsilon,$$

because $HW1_n = H1_n = 0$ and $G = WS^{-1}$. Therefore, for $i = 1, \cdots, n$,

$$\frac{d\big(\epsilon_i(\lambda)/\sigma\big)}{d\lambda}\Big|_{\sigma_0,\lambda_0} = -H_iG\varepsilon = -\chi_i = -\sum_{j=1}^{n} t_{ij}\varepsilon_j.$$

Recall that $r_L(\lambda_0, \sigma_0) = r_L = \sum_{i=1}^{n} \tilde{\psi}_{iL}\chi_i - \mathrm{tr}\left\{\mathrm{G}(\lambda_0)\right\}$. Since $\tilde{\psi}_{iL} = \Phi^{(L)T}(\epsilon_i/\sigma_0)\tilde{a}^{(L)}$, we have

$$\begin{aligned} \frac{h}{n}\cdot s_{1L} &= \frac{h}{n}\left(\sum_{i=1}^{n}\frac{\partial\tilde{\psi}_{iL}}{\partial\lambda}\chi_i - \mathrm{tr}\left\{G^2(\lambda_0)\right\}\right) \\ &= \frac{h}{n}\sum_{i=1}^{n}\left(\frac{\partial\Phi^{(L)T}(\epsilon_i/\sigma_0)}{\partial\lambda}\tilde{a}^{(L)}\chi_i + \Phi^{(L)T}(\epsilon_i/\sigma_0)\frac{\partial\tilde{a}^{(L)}}{\partial\lambda}\chi_i\right) - \frac{h}{n}\mathrm{tr}\left\{G^2(\lambda_0)\right\}. \end{aligned}$$

By Assumption 2, the limit of the latter term $\lim_{\to\infty} h\mathrm{tr}\left\{G^2(\lambda)\right\}/n = -\omega_2$. We will show for

$$A := \frac{h}{n}\sum_{i=1}^{n}\frac{\partial\Phi^{(L)T}(\frac{\epsilon_i}{\sigma_0})}{\partial\lambda}\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0})\chi_i, \quad B := \frac{h}{n}\sum_{i=1}^{n}\Phi^{(L)T}(\frac{\epsilon_i}{\sigma_0})\frac{\partial\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0})}{\partial\lambda}\chi_i,$$

that

$$A \to_p -\mathcal{J}\omega_1, \qquad B \to_p 0, \tag{4.7.39}$$

which completes the proof of (4.7.5), $hs_{1L}/n \to_p -\mathcal{J}\omega_1 - \omega_2$.

Recall the notations $\phi(s)^{(L)} = (\phi_1(s), \cdots \phi_L(s))^T$, $\bar{\phi}(s)^{(L)} = (\bar{\phi}_1(s), \cdots, \bar{\phi}_L(s))^T$ where $\bar{\phi}_\ell(\varepsilon_i) = \phi_\ell(\varepsilon_i) - E(\phi_\ell(\varepsilon_i))$, $\bar{\phi}'(s)^{(L)} = (\bar{\phi}_1'(s), \cdots, \bar{\phi}_L'(s))^T$ where $\bar{\phi}_\ell'(\varepsilon_i) = \phi_\ell'(\varepsilon_i) - E(\phi_\ell'(\varepsilon_i))$ and $\bar{\phi}''(s)^{(L)} = (\bar{\phi}_1''(s), \cdots, \bar{\phi}_L''(s))^T$ where $\bar{\phi}_\ell''(\varepsilon_i) = \phi_\ell''(\varepsilon_i) - E(\phi_\ell''(\varepsilon_i))$.

Proof of $A \to_p -\mathcal{J}\omega_1$.

Since $\frac{\partial(\epsilon_i(\lambda)/\sigma)}{\partial\lambda}|_{\lambda_0,\sigma_0} = -H_iG\varepsilon = -\chi_i$, we can write

$$\begin{aligned} A &= -\frac{h}{n}\sum_{i=1}^{n}[\bar{\phi}'(\varepsilon_i)^{(L)} + \phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \sum_{j=1}^{n}\phi'(\frac{\epsilon_j}{\sigma_0})^{(L)} - \bar{\phi}'(\varepsilon_i)^{(L)}]^T[a^{(L)} + \tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - a^{(L)}]\chi_i^2 \\ &= -\frac{h}{n}\sum_{i=1}^{n}[\bar{\phi}'(\varepsilon_i)^{(L)} + r_i]^T[a^{(L)} + l^{(L)}]\chi_i^2 \\ &= -\frac{h}{n}\sum_{i=1}^{n}\bar{\phi}'(\varepsilon_i)^{(L)T}a^{(L)}\chi_i^2 - \frac{h}{n}\sum_{i=1}^{n}\bar{\phi}'(\varepsilon_i)^{(L)T}l^{(L)}\chi_i^2 - \frac{h}{n}\sum_{i=1}^{n}r_i^Ta^{(L)}\chi_i^2 - \frac{h}{n}\sum_{i=1}^{n}r_i^Tl^{(L)}\chi_i^2 \\ &=: A_1 + A_2 + A_3 + A_4, \end{aligned}$$

using the $L \times 1$ vectors,

$$l^{(L)} := \tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - a^{(L)}, \quad \text{and} \quad r_i := \big(\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \sum_{j=1}^{n}\phi'(\frac{\epsilon_j}{\sigma_0})\big) - \bar{\phi}'(\varepsilon_i)^{(L)}.$$

We will now show that $A_1 \to_p -\mathcal{J}\omega_1$ and $A_i = o_p(1)$, for $i = 2, 3, 4$.

Showing $A_1 \to_p -\mathcal{J}\omega_1$. Denote,

$$A_1 = -\frac{h}{n}\sum_{i=1}^{n}\bar{\phi}'(\varepsilon_i)^Ta^{(L)}\chi_i^2 = -\frac{h}{n}\sum_{i=1}^{n}[\psi'(\varepsilon_i) + (\bar{\phi}'(\varepsilon_i)^Ta^{(L)} - \psi'(\varepsilon_i))]\chi_i^2 =: A_{11} + A_{12}.$$

We establish $A_{11} \to_p -\mathcal{J}\omega_1$ and $A_{12} = o_p(1)$.

We start with A_{11}. It will be shown that $E(A_11) \to -\mathcal{J}\omega_1$ and $E[A_{11} - E(A_11)]^2 = o(1)$. Taking expectation of A_{11}, we obtain,

$$\begin{aligned} E(A_{11}) &= -\frac{h}{n}\sum_{i,j,k=1}^{n}t_{ij}t_{ik}E(\psi'(\varepsilon_i)\varepsilon_j\varepsilon_k) \\ &= -\frac{h}{n}\sum_{i,j=1}^{n}t_{ij}^2E(\psi'(\varepsilon_i))E(\varepsilon_j^2) - \frac{h}{n}\sum_{i=1}^{n}t_{ii}^2[E(\psi'(\varepsilon_i)\varepsilon_i^2) - E(\psi'(\varepsilon_i))E(\varepsilon_j^2)] \\ &= -\mathcal{J}\frac{h}{n}\sum_{i,j=1}^{n}t_{ij}^2 + O(\frac{1}{h}), \end{aligned}$$

because $E(\psi'(\varepsilon_i)) = E(\psi^2(\varepsilon_i)) = \mathcal{J}$ and $\sum_{i=1}^{n}t_{ii}^2 = O(n/h^2)$. Now, denoting $\bar{g}_j :=$

$\frac{1}{n}\sum_{m=1}^{n} g_{mj}$ and recalling that $t_{ij} = g_{ij} - \bar{g}_j$ from $T = HG$, allow us to write

$$\begin{aligned}\frac{h}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}^2 &= \frac{h}{n}\sum_{i,j=1}^{n}(g_{ij} - \bar{g}_j)^2 = \frac{h}{n}\sum_{i,j=1}^{n}(g_{ij}^2 - 2\bar{g}_j g_{ij} + \bar{g}_j^2)\\ &= \frac{h}{n}\sum_{i,j=1}^{n} g_{ij}^2 - 2h\sum_{j=1}^{n}\bar{g}_j(\frac{1}{n}\sum_{i=1}^{n} g_{ij}) + \frac{h}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{g}_j^2\\ &= \frac{h}{n}\sum_{i,j=1}^{n} g_{ij}^2 + h\sum_{j=1}^{n}\bar{g}_j^2 = \frac{h}{n}\sum_{i,j=1}^{n} g_{ij}^2 - O(\frac{h}{n}) \to \omega_1,\end{aligned}$$

as $\bar{g}_j^2 \leq [\frac{1}{n}\sum_{m=1}^{n}|g_{mj}|]^2 = O(\frac{1}{n^2})$ due to absolute column summability of the matrix G and $\mathrm{tr}(GG^T) = \sum_{i,j=1}^{n} g_{ij}^2$.

Next, we show $E[(A_{11} - E(A_{11}))^2] = o(1)$, which together with above completes the proof of $A_1 \to_p -\mathcal{J}\omega_1$. Recall

$$E(A_{11}^2) = (\frac{h}{n})^2\sum_{i,j,k,=1}^{n}\sum_{i',j',k'=1}^{n} t_{ij}t_{ik}t_{i'j'}t_{i'k'}E(\psi'(\varepsilon_i)\varepsilon_j\varepsilon_k\psi'(\varepsilon_{i'})\varepsilon_{j'}\varepsilon_{k'}),$$

so that,

$$\begin{aligned}E(A_{11}^2) - [E(A_{11})]^2 = (\frac{h}{n})^2\sum_{i,j,k=1}^{n}\sum_{i',j',k'=1}^{n} & t_{ij}t_{ik}t_{i'j'}t_{i'k'}E(\psi'(\varepsilon_i)\varepsilon_j\varepsilon_k\psi'(\varepsilon_{i'})\varepsilon_{j'}\varepsilon_{k'})\\ & -\mathcal{J}^2(\sum_{i,j=1}^{n} t_{ij}^2)^2\\ = C_1(\frac{h}{n})^2\big[C_1\sum_{j,j'=1}^{n}\big(\sum_{i=1}^{n} t_{ij}t_{ij'}\sum_{i'=1}^{n} t_{i'j}t_{i'j'}\big) &+ C_2\sum_{i,i',j=1}^{n} t_{ij}^2 t_{i'j}^2\\ +C_3\sum_{i,i',j=1}^{n} t_{ii}t_{ij}t_{i'j}^2 + C_4\sum_{i,i',j=1}^{n} t_{ii}t_{i'i'}t_{ij}t_{i'j} &+ C_5\sum_{i,j=1}^{n} t_{ij}^2 t_{ii}^2\\ +C_6\sum_{i,j=1}^{n} t_{ij}t_{jj}^3 + C_7\sum_{i,j=1}^{n} t_{ii}^2 t_{jj}^2 + C_8\sum_{i,j=1}^{n} t_{ij}^4 &+ C_9\sum_{i,j=1}^{n} t_{ii}t_{ij}^3\big], \quad (4.7.40)\end{aligned}$$

where $C_m, m = 1, \cdots, 9$ denotes some constants. First we bound the summations that are multiplied to $C_1 - C_4$ above. We have,

$$\begin{aligned}
&\sum_{j,j'=1}^{n}\big(\sum_{i=1}^{n} t_{ij}t_{ij'}\sum_{i'=1}^{n} t_{i'j}t_{i'j'}\big) = \sum_{j,j'=1}^{n}\big(\sum_{i=1}^{n} t_{ij}t_{ij'}\big)^2 = \sum_{j,j'=1}^{n}(T^TT)^2_{jj'} \\
&\qquad\qquad \le \sum_{j=1}^{n} \max_{j'}|(T^TT)_{jj'}|\sum_{j'=1}^{n}|(T^TT)_{jj'}| = O\big(\frac{n}{h}\big) = o\left(\frac{n^2}{h^2}\right), \\
&\sum_{i,i',j=1}^{n} t_{ij}^2 t_{i'j}^2 = \big(\frac{h}{n}\big)^2\sum_{j=1}^{n}\big(\sum_{i=1}^{n} t_{ij}^2\big)^2 = \sum_{j=1}^{n}(T^TT)^2_{j,j} = O\big(\frac{n}{h^2}\big) = o\left(\frac{n^2}{h^2}\right), \\
&\Big|\sum_{i,i',j=1}^{n} t_{ii}t_{ij}t_{i'j}^2\Big| = \Big|\sum_{j=1}^{n}\sum_{i=1}^{n} t_{ii}t_{ij}\sum_{i'=1}^{n} t_{i'j}^2\Big| \le \sum_{j=1}^{n}\max_i|t_{ii}|\sum_{i=1}^{n}|t_{ij}||(T^TT)_{j,j}| \\
&\qquad\qquad = O\big(\frac{n}{h}\big)O\big(\frac{1}{h}\big) = o\left(\frac{n^2}{h^2}\right), \\
&\Big|\sum_{i=1}^{n} t_{ii}\sum_{i'=1}^{n} t_{i'i'}\sum_{j=1}^{n} t_{ij}t_{i'j}\Big| \le \sum_{i=1}^{n}|t_{ii}|\sum_{i'=1}^{n}|t_{i'i'}||(TT^T)_{i,i'}| = O\big(\frac{n^2}{h^3}\big) = o\left(\frac{n^2}{h^2}\right).
\end{aligned}$$

We can bound the summations corresponding to $C_5 - C_9$, since $\max_{1\le i,j\le n}|t_{ij}| = O(1/h)$, as explained in lines following (4.7.36),

$$\big(\frac{h}{n}\big)^2\sum_{i,j=1}^{n}\left[t_{ij}^2t_{ii}^2 + |t_{ij}t_{jj}^3| + t_{ii}^2t_{jj}^2 + t_{ij}^4|t_{ii}t_{ij}^3|\right] = O\big(\frac{h^2}{n^2}\times\frac{n^2}{h^4}\big) = O\big(\frac{1}{h^2}\big) = o(1).$$

Applying these bounds in (4.7.40) implies $E[(A_{11} - E(A_{11}))^2] = o(1)$.

Showing $A_{12} = o_p(1)$. Recall that by Assumption 6, $E[\big(\bar{\phi}'(\varepsilon_i)^Ta^{(L)} - \psi'(\varepsilon_i)\big)^2] = o(1)$. Therefore

$$\begin{aligned}
E|A_{12}| &\le \frac{h}{n}\sum_{i=1}^{n}E|(\bar{\phi}'(\varepsilon_i)^{(L)T}a^{(L)} - \psi'(\varepsilon_i))\chi_i^2| \\
&\le \frac{h}{n}\Big(E\big(\bar{\phi}'(\varepsilon_1)^{(L)T}a^{(L)} - \psi'(\varepsilon_1)\big)^2\Big)^{1/2}\sum_{i=1}^{n}\big\{E(\chi_i^4)\big\}^{1/2} \\
&= \frac{h}{n}o(1)O\big(\frac{n}{h}\big) = o(1),
\end{aligned}$$

because

$$\sum_{i=1}^{n}\big(E(\chi_i^4)\big)^{1/2} = O\big(\frac{n}{h}\big), \tag{4.7.41}$$

which follows from

$$\max_i E(\chi_i^4) = 3\max_i\sum_{j=1}^{n}\sum_{k=1}^{n}t_{ij}^2t_{ik}^2 = 3\max_i\big(\sum_{j=1}^{n}t_{ij}^2\big)^2 \le 3\big(\max_{i,j}|t_{ij}|\sum_{j=1}^{n}|t_{ij}|\big)^2 = O\big(\frac{1}{h^2}\big). \tag{4.7.42}$$

Proof of $A_i = o_p(1)$, $i = 2, 3, 4$. Recall

$$A_2 + A_3 + A_4 = -\frac{h}{n}\sum_{i=1}^{n} \bar{\phi}'(\varepsilon_i)^{(L)T} l^{(L)} \chi_i^2 - \frac{h}{n}\sum_{i=1}^{n} r_i^T a^{(L)} \chi_i^2 - \frac{h}{n}\sum_{i=1}^{n} r_i^T l^{(L)} \chi_i^2.$$

Using notations $\phi'(\varepsilon_i)^{(L)} = (\phi_1'(\varepsilon_i), \cdots, \phi_L'(\varepsilon_i))^T$ and $\phi''(\varepsilon_i)^{(L)} = (\phi_1''(\varepsilon_i), \cdots, \phi_L''(\varepsilon_i))^T$, decompose r_i into two parts:

$$\begin{aligned} r_i &= \big[\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \frac{1}{n}\sum_{j=1}^{n} \phi'(\frac{\epsilon_j}{\sigma_0})^{(L)}\big] - \bar{\phi}'(\varepsilon_i)^{(L)} \\ &= \Big\{\big[\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \frac{1}{n}\sum_{j=1}^{n} \phi'(\frac{\epsilon_j}{\sigma_0})^{(L)}\big] - \big[\phi'(\varepsilon_i)^{(L)} - \frac{1}{n}\sum_{i=1}^{n} \phi'(\varepsilon_i)^{(L)}\big]\Big\} \\ &\quad + \Big\{\big[\phi'(\varepsilon_i)^{(L)} - \frac{1}{n}\sum_{i=1}^{n} \phi'(\varepsilon_i)^{(L)}\big] - \bar{\phi}'(\varepsilon_i)^{(L)}\Big\} =: r_{1i} + r_2. \end{aligned}$$

Since $\bar{\phi}'(\varepsilon_i)^{(L)} = \phi'(\varepsilon_i)^{(L)} - E\big(\phi'(\varepsilon_i)^{(L)}\big)$, we can express r_2 as follows, explaining the lack of i subscript in the $L \times 1$ vector r_2:

$$\begin{aligned} r_2 &= \big[\phi'(\varepsilon_i)^{(L)} - \frac{1}{n}\sum_{j=1}^{n} \phi'(\varepsilon_j)^{(L)}\big] - \big[\phi'(\varepsilon_i)^{(L)} - E(\phi'(\varepsilon_i)^{(L)})\big] \\ &= E\big(\phi'(\varepsilon_i)^{(L)}\big) - \frac{1}{n}\sum_{j=1}^{n} \phi'(\varepsilon_j)^{(L)}. \end{aligned} \tag{4.7.43}$$

For each element of the $L \times 1$ vector $r_{1i} = (r_{1i1}, \cdots, r_{1iL})^T$, the mean value theorem (MVT) yields

$$r_{1i\ell} = \big\{\phi_\ell''(\varepsilon_i^*) - \frac{1}{n}\sum_{j=1}^{n} \phi_\ell''(\varepsilon_j^*)\big\}\bar{\varepsilon}, \quad \ell = 1, \cdots, L, \tag{4.7.44}$$

letting ε_i^* denote some point that lies between ϵ_i/σ_0 and ε_i, such that, $\phi_\ell'(\epsilon_i/\sigma_0) - \phi_\ell'(\varepsilon_i) = \bar{\varepsilon}\phi_\ell''(\epsilon_i^*)$, and recalling $\epsilon_i/\sigma_0 = \varepsilon_i - \bar{\varepsilon}$. The ε_i^*'s may differ across ℓ's but we suppress the reference to ℓ for brevity.

Since the $L \times 1$ vector $l^{(L)}$ is the estimation error in estimating the unknown coefficients $a^{(L)}$, an upper bound on $\|l^{(L)}\|$ can be established by combining Lemma 10 and 19 of Robinson (2005):

$$\begin{aligned} &\|l^{(L)}\| = \|\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - a^{(L)}\| \le \|\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - \tilde{a}^{(L)}(\varepsilon)\| + \|\tilde{a}^{(L)}(\varepsilon) - a^{(L)}\| = O_p(R_l), \\ &R_l := \frac{L}{n^{1/2}}\rho_{2\kappa L}^{1/2}\pi_L(1 + L^{1/2}\rho_{4\kappa L}^{1/2}\pi_L) + \rho_{2\kappa L}^{3/2}\pi_L^2\Big(\frac{L^2}{n^{1/2}} + \frac{(CL)^{4\kappa L+3}}{n}\log n\Big). \end{aligned} \tag{4.7.45}$$

To complete the proof of (4.7.39), we will show negligibility of the terms, A_2, A_3, A_4, written using the above decomposition of r_i :

$$\begin{aligned}|A_2|+|A_3|+|A_4| &= \frac{h}{n}\Big|\sum_{i=1}^{n}\bar{\phi}'(\varepsilon_i)^{(L)T}l^{(L)}\chi_i^2\Big|+\frac{h}{n}\Big|\sum_{i=1}^{n}(r_{1i}+r_2)^T a^{(L)}\chi_i^2\Big| \\ &\quad +\frac{h}{n}\Big|\sum_{i=1}^{n}(r_{1i}+r_2)^T l^{(L)}\chi_i^2\Big|.\end{aligned}$$

Showing $|A_2| = o_p(1)$. It suffices to show $E|A_2| \to 0$. Note that $h\sum_{i=1}^{n}(E\chi_i^4)^{1/2}/n = O(1)$ by (4.7.41) and from (4.7.37) by Lemma 9 of Robinson (2005),

$$E\|\bar{\phi}'(\varepsilon_1)^{(L)}\|^2 = \sum_{\ell=1}^{L}E(\bar{\phi}'_\ell(\varepsilon_1)^2) \le C^{2L}\sum_{\ell=1}^{L}\ell^2\mu_{2\kappa(\ell+K)} \le C^{2L}L^2\rho_{2\kappa L}.$$

Therefore,

$$\begin{aligned}&E|A_2| \le \frac{h}{n}\sum_{i=1}^{n}E|\bar{\phi}'(\varepsilon_i)^{(L)T}l^{(L)}\chi_i^2| \le \|l^{(L)}\|\left[E\|\bar{\phi}'(\varepsilon_1)^{(L)}\|^2\right]^{1/2}\frac{h}{n}\sum_{i=1}^{n}[E\chi_i^4]^{1/2} \\ &= \|l^{(L)}\|\left[E\|\bar{\phi}'(\varepsilon_1)^{(L)}\|^2\right]^{1/2}O(1) = O_p(q_n), \\ &q_n := R_l \times C^L L\rho_{2\kappa L}^{1/2}.\end{aligned}$$

It remains to show $q_n = o(1)$ in order to establish $|A_2| = o_p(1)$. Trivially, we have for any $p, \varepsilon > 0$ and $C < \infty$, $L^p \le L^{\varepsilon L}$ and $C^L \le L^{\varepsilon L}$. We also have $\pi_L \le L^{(1+\varepsilon)L}$ for any $\varepsilon > 0$ and recall $\rho_{uL} \le \max\{CL, (CL)^{uL/\omega}\}$ from Assumption 5. Hence, there exist $\alpha_1, \alpha_2 > 0$ large enough, so that we can write

$$q_n = o\Big(\frac{L^{\alpha_1 L}}{\sqrt{n}}+\frac{L^{\alpha_2 L}}{n}\log n\Big) = o\Big(\frac{L^{\alpha_1 L}}{\sqrt{n}}+\frac{L^{\alpha_2 L}}{\sqrt{n}}\frac{\log n}{\sqrt{n}}\Big) = o\Big(\frac{L^{(\alpha_1+\alpha_2)L}}{\sqrt{n}}\Big) = o(1), \qquad (4.7.46)$$

since Assumption 5 and 1 (ii) imply that $L\log L = o(\log n)$, hence $L^{(\alpha_1+\alpha_2)L}/\sqrt{n} = O(1)$.

Showing $|A_3| = o_p(1)$. We bound $|A_3|$ using $r_i = r_{1i} + r_2$,

$$|A_3| \le \Big|\frac{h}{n}\sum_{i=1}^{n}r_{1i}^T a^{(L)}\chi_i^2\Big| + \Big|\frac{h}{n}\sum_{i=1}^{n}r_2^T a^{(L)}\chi_i^2\Big| =: |A_{31}| + |A_{32}|.$$

We have by Cauchy-Schwarz inequality,

$$E|A_{31}| \le \frac{h}{n}\sum_{i=1}^{n}E|r_{1i}^T a^{(L)}\chi_i^2| \le \|a^{(L)}\|\frac{h}{n}\sum_{i=1}^{n}E\|r_{1i}\chi_i^2\| \le \|a^{(L)}\|\frac{h}{n}\sum_{i=1}^{n}\big(E\|r_{1i}\|^2E\chi_i^4\big)^{1/2}.$$

Denoting $r_{1i\ell} = \big\{\phi''_\ell(\varepsilon_i^*) - \frac{1}{n}\sum_{j=1}^{n} \phi''_\ell \varepsilon_j^*)\big\}\bar{\varepsilon} =: p''_{\ell i}\bar{\varepsilon}$ allows us to write

$$\begin{aligned}
&\frac{h}{n}\sum_{i=1}^{n}\big[E\|r_{1i}\|^2 E(\chi_i^4)\big]^{1/2} \le \frac{h}{n}\max_i\big(E(\chi_i^4)\big)^{1/2}\sum_{i=1}^{n}\big[\sum_{\ell=1}^{L}E(r_{1i\ell}^2)\big]^{1/2} \\
&= O_p\big(\frac{1}{n}\big)\sum_{i=1}^{n}\big[\sum_{\ell=1}^{L}E(\bar{\varepsilon}^2(p''_{\ell i})^2)\big]^{1/2} \le O_p\big(\frac{1}{n}\big)\sum_{i=1}^{n}\big[\sum_{\ell=1}^{L}\big([E\bar{\varepsilon}^4][E(p''_{\ell i})^4]\big)^{1/2}\big]^{1/2} \\
&\le O_p\big(\frac{1}{n}\big)\sum_{i=1}^{n}(E\bar{\varepsilon}^4)^{1/4}\big[\sum_{\ell=1}^{L}\big(E(p''_{\ell i})^4\big)^{1/2}\big]^{1/2} \\
&\le O_p\big(\frac{1}{n}\big)\sum_{i=1}^{n}(E\bar{\varepsilon}^4)^{1/4}\big[\sum_{\ell=1}^{L}C^{2\ell}\ell^4\mu_{4\kappa(\ell+2K)}^{1/2}\big]^{1/2} \le O_p\big(\frac{C^L}{\sqrt{n}}\big)L^2\rho_{4\kappa L}^{1/2}, \qquad (4.7.47)
\end{aligned}$$

using Cauchy-Schwarz inequality, and since $\max_i\big(E\chi_i^4\big)^{1/2} = O(1/h)$ by (4.7.42), $E(\bar{\varepsilon}^4) = O(1/n^2)$ and Lemma 9 of Robinson (2005):

$$\sum_{\ell=1}^{L}C^{2\ell}\ell^4\mu_{4\kappa(\ell+2K)}^{1/2} \le C^{2L}\sum_{\ell=1}^{L}\ell^4\mu_{4\kappa(\ell+2K)} \le C^{2L}L^4\rho_{4\kappa L}.$$

We have $E((p''_{\ell i})^4) \le C^{4\ell}\ell^8\mu_{4\kappa(\ell+2K)}$ since

$$|\phi''_\ell(\epsilon_1^*)| \le C^\ell\ell^2\big(1 + |\varepsilon_1|^{\kappa(\ell-1+2K)} + |\bar{\varepsilon}|^{\kappa(\ell-1+2K)}\big).$$

Therefore, using Lemma 10 of Robinson (2005) which states $\|a^{(L)}\| = O(L\rho_{2\kappa L}^{1/2}\pi_L)$, we conclude:

$$E|A_{31}| \le O\left(\frac{C^L L^3\rho_{2\kappa L}^{1/2}\rho_{4\kappa L}^{1/2}\pi_L}{\sqrt{n}}\right) = o(1),$$

where the last bound can be established by the same argument as in the proof of $q_n = o(1)$ in (4.7.46).

Our next task is finding an upper bound on $E|A_{32}|$.

$$\begin{aligned}
E|A_{32}| \quad &\le \quad \|a^{(L)}\|\frac{h}{n}\sum_{i=1}^{n}E\|r_2\chi_i^2\| \le \|a^{(L)}\|[E\|r_2\|^2]^{1/2}\frac{h}{n}\sum_{i=1}^{n}(E\chi_i^4)^{1/2} \\
&= \|a^{(L)}\|O_p(\big[E\|r_2\|^2\big]^{1/2}),
\end{aligned}$$

since by (4.7.42), $\frac{h}{n}\sum_{i=1}^{n}\big(E\chi_i^4\big)^{1/2} = O(1)$. Now, introducing the $L \times 1$ vector $r_2 = (r_{21}, \cdots, r_{2L})^T$, we will find an upper bound on $E\|r_2\|^2 = \sum_{\ell=1}^{L}E(r_{2\ell}^2)$. For each ℓ, recalling the notation $\bar{\phi}'_\ell(\varepsilon) = \phi'_\ell(\varepsilon) - E[\phi'_\ell(\varepsilon)]$ and independence of ε_j's,

$$\begin{aligned} E(r_{2\ell}^2) &= \frac{1}{n^2}E[\sum_{j=1}^n \phi'_\ell(\varepsilon_j) - E(\phi'_\ell(\varepsilon_1))]^2 = \frac{1}{n^2}E[\sum_{j=1}^n \bar{\phi}'_\ell(\varepsilon_j)]^2 \\ &= \frac{1}{n^2}\sum_{j=1}^n E[\bar{\phi}'^2_\ell(\varepsilon_j)] = \frac{C^{2\ell}\ell^2\mu_{2\kappa(\ell+K)}}{n}. \end{aligned}$$

Therefore,

$$E\|r_2\|^2 = \sum_{\ell=1}^L E(r_{2\ell}^2) \le \sum_{\ell=1}^L \frac{C^{2\ell}\ell^2\mu_{2\kappa(\ell+K)}}{n} = O_p(\frac{C^{2L}L^2\rho_{2\kappa L}}{n}), \tag{4.7.48}$$

by Lemma 9 of Robinson (2005). Putting together terms:

$$E|A_{32}| \le \|a^{(L)}\| \left[E\|r_2\|^2\frac{h}{n}\sum_{i=1}^n E\chi_i^4\right]^{1/2} = O(L\rho_{2\kappa L}^{1/2}\pi_L)O(\frac{C^L L\rho_{2\kappa L}^{1/2}}{\sqrt{n}}) = o(1),$$

the last equality follows by the same reasoning as in the proof of $q_n = o(1)$ in (4.7.46).

Showing $|A_4| = o_p(1)$. Now, the remaining task to complete the proof of $A \to_p -\mathcal{J}\omega_1$ is to show

$$|A_4| \le |\frac{h}{n}\sum_{i=1}^n r_{1i}^T l^{(L)}\chi_i^2| + |\frac{h}{n}\sum_{i=1}^n r_2^T l^{(L)}\chi_i^2| =: |A_{41}| + |A_{42}| = o_p(1).$$

Firstly, using the previous results (4.7.45) and (4.7.47), it follows

$$|A_{41}| \le \|l^{(L)}\|\frac{h}{n}\sum_{i=1}^n \|r_{1i}\chi_i^2\| = O_p(R_l)O_p(\frac{C^L L^2\rho_{4\kappa L}^{1/2}}{\sqrt{n}}) = o_p(1),$$

where the last equality can be established by the same argument as in (4.7.46).

Secondly, from (4.7.48) and (4.7.45),

$$|A_{42}| \le \|r_2\|\|l^{(L)}\|\frac{h}{n}\sum_{i=1}^n \chi_i^2 = O_p(\frac{C^L L\rho_{2\kappa L}^{1/2}}{\sqrt{n}}) \times O_p(R_l) = o_p(1).$$

since $\frac{h}{n}\sum_{i=1}^n E\chi_i^2 \le \frac{h}{n}O(\frac{n}{h}) = O(1)$, as $E\chi_i^2 \le \max_j |t_{ij}|\sum_{j=1}^n |t_{ij}| = O(1/h)$ uniformly over i while the bound $o_p(1)$ follows by the same reasoning as in (4.7.46).

Proof of $B \to_p 0$.

Recall,

$$\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) = \left(\tilde{W}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0})\right)^{-1}\tilde{w}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}).$$

For ease of notation, let

$$\tilde{W}^{(L)} = \tilde{W}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) \qquad \text{and} \qquad \tilde{w}^{(L)} = \tilde{w}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}).$$

We decompose B as follows, using chain rule:

$$\begin{aligned} B &= \left[\frac{h}{n}\sum_{i=1}^{n}\chi_i\Phi^{(L)T}(\frac{\epsilon_i}{\sigma_0})\right]\cdot\left[\frac{\partial\tilde{a}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0})}{\partial\lambda}\right] \\ &= \left[\frac{h}{n}\sum_{i=1}^{n}\chi_i\Phi^{(L)T}(\frac{\epsilon_i}{\sigma_0})\right]\cdot\left[\frac{\partial(\tilde{W}^{(L)})^{-1}}{\partial\lambda}\tilde{w}^{(L)} + (\tilde{W}^{(L)})^{-1}\frac{\partial(\tilde{w}^{(L)})^{-1}}{\partial\lambda}\right] \\ &=: D[F\tilde{w}^{(L)} + (\tilde{W}^{(L)})^{-1}J]. \end{aligned}$$

Next, we find upper bounds on $\|D\|$, $\|(\tilde{W}^{(L)})^{-1}\|$, $\|F\|$, $\|\tilde{w}^{(L)}\|$, and $\|J\|$.

Upper bound of $\|D\|$. We decompose the $L \times 1$ vector D as follows:

$$\begin{aligned} D &= \frac{h}{n}\sum_{i=1}^{n}\chi_i\Phi^{(L)}(\frac{\epsilon_i}{\sigma_0}) = \frac{h}{n}\sum_{i=1}^{n}\chi_i\Big(\Phi^{(L)}(\frac{\epsilon_i}{\sigma_0}) - \Phi^{(L)}(\varepsilon_i)\Big) + \frac{h}{n}\sum_{i=1}^{n}\chi_i\Big(\Phi^{(L)}(\varepsilon_i) - \bar{\phi}^{(L)}(\varepsilon_i)\Big) \\ &\qquad -\frac{h}{n}\sum_{i=1}^{n}\chi_i\bar{\phi}^{(L)}(\varepsilon_i) =: D_1 + D_2 + D_3. \end{aligned}$$

We verify below that

$$\|D\| \le \|D_1\| + \|D_2\| + \|D_3\| = O_p(C^L L\rho_{2\kappa L}^{1/2}).$$

Upper bound of $\|D_1\|$. By the MVT, we have $\phi(\epsilon_i/\sigma_0) - \phi(\varepsilon_i) = \bar{\varepsilon}\phi'(\varepsilon_i^*)$, implying that the ℓ-th element of the vector $\Phi^{(L)}(\epsilon_i/\sigma_0) - \Phi^{(L)}(\varepsilon_i)$ is

$$\bar{\varepsilon}\phi'_\ell(\varepsilon_i^*) - \frac{1}{n}\sum_{j=1}^{n}\bar{\varepsilon}\phi'_\ell(\varepsilon_j^*) = \bar{\varepsilon}\Big[\phi'_\ell(\varepsilon_i^*) - \frac{1}{n}\sum_{j=1}^{n}\phi'_\ell(\varepsilon_j^*)\Big] =: \bar{\varepsilon}p'_{i\ell}.$$

Hence, triangular inequality gives

$$\|D_1\| \le \frac{h}{n}\sum_{i=1}^{n}|\chi_i|\left[\sum_{\ell=1}^{L}\bar{\varepsilon}^2\Big(p'_{i\ell}\Big)^2\right]^{1/2} = \bar{\varepsilon}\frac{h}{n}\sum_{i=1}^{n}|\chi_i|\left[\sum_{\ell=1}^{L}\Big(p'_{i\ell}\Big)^2\right]^{1/2}.$$

Next, we find an upper bound on the latter term of above expression, using triangular and Cauchy-Schwarz inequalities,

$$\frac{h}{n}\sum_{i=1}^{n}E\Big(|\chi_i|\big[\sum_{\ell=1}^{L}(p'_{i\ell})^2\big]^{1/2}\Big) \le \frac{h}{n}\sum_{i=1}^{n}(E\chi_i^2)^{1/2}\Big[E\sum_{\ell=1}^{L}\big(p'_{i\ell}\big)^2\Big]^{1/2} = O(\sqrt{h})O(C^L L\rho_{2\kappa L}^{1/2}),$$

because for each $1 \leq \ell \leq L$, in view of $|\phi_\ell'(s)| \leq C^\ell \ell(1 + |\varepsilon_i|^{\kappa(\ell+K)} + |\bar{\varepsilon}|^{\kappa(\ell+K)})$,

$$\sum_{\ell=1}^{L} E\big(p_{i\ell}'\big)^2 \leq \sum_{\ell=1}^{L} C^{2\ell}\ell^2 \mu_{2\kappa(\ell+K)} \leq C^{2L}L^2\rho_{2\kappa L},$$

with the last step by Lemma 9 of Robinson (2005). On the other hand, $\frac{h}{n}\sum_{i=1}^{n}(E\chi_i^2)^{1/2} = O(\sqrt{h})$, as $E\chi_i^2 \leq \max_j |t_{ij}| \sum_{j=1}^{n} |t_{ij}| = O(1/h)$. Therefore, $\|D_1\| = O(C^L L \rho_{2\kappa L}^{1/2})$.

Upper bound of $\|D_2\|$.

$$\begin{aligned} D_2 &= \frac{h}{n}\sum_{i=1}^{n}\chi_i\Big(\Phi^{(L)}(\varepsilon_i) - \bar{\phi}^{(L)}(\varepsilon_i)\Big) = \frac{h}{n}\sum_{i=1}^{n}\chi_i\Big(E\big(\phi^{(L)}(\varepsilon_i)\big) - \frac{1}{n}\sum_{j=1}^{n}\phi^{(L)}(\varepsilon_j)\Big) \\ &= \Big[-\frac{h}{n}\sum_{j=1}^{n}\bar{\phi}^{(L)}(\varepsilon_j)\Big]\Big[\frac{1}{n}\sum_{i=1}^{n}\chi_i\Big], \\ \|D_2\| &\leq \Big\|\frac{h}{n}\sum_{j=1}^{n}\bar{\phi}^{(L)}(\varepsilon_j)\Big\|\Big|\frac{1}{n}\sum_{i=1}^{n}\chi_i\Big| = O_p\big(\frac{hC^L\rho_{2\kappa L}^{1/2}}{\sqrt{n}}\big)O_p\big(\frac{1}{\sqrt{n}}\big). \end{aligned}$$

We have, since $\|TT^T\|_C = O(1)$,

$$E\Big(\sum_{i=1}^{n}\chi_i\Big)^2 = \sum_{i,j,k=1}^{n} t_{ik}t_{jk} \leq \sum_{i,j=1}^{n}|(TT^T)_{ij}| = \sum_{i=1}^{n}O(1) = O(n).$$

Since, $E\big(\bar{\phi}_\ell^2(\varepsilon_j)\big) \leq C^{2\ell}\mu_{2\kappa\ell}$,

$$E\Big\|\frac{h}{n}\sum_{j=1}^{n}\bar{\phi}^{(L)}(\varepsilon_j)\Big\|^2 = \big(\frac{h}{n}\big)^2\sum_{\ell=1}^{L}\sum_{j=1}^{n}E\big(\bar{\phi}_\ell^2(\varepsilon_j)\big) \leq \frac{2h^2}{n}\sum_{\ell=1}^{L}C^{2\ell}\mu_{2\kappa\ell} = O\big(\frac{h^2}{n}C^{2L}\rho_{2\kappa L}\big).$$

This proves $\|D_2\| = O_p(C^L L \rho_{2\kappa L}^{1/2})$.

Upper bound of $\|D_3\|$.

$$\begin{aligned} E\|D_3\|^2 &= E\Big\|\frac{h}{n}\sum_{i=1}^n \chi_i\bar{\phi}^{(L)}(\varepsilon_i)\Big\|^2 = \big(\frac{h}{n}\big)^2\sum_{\ell=1}^L E\Big(\sum_{i=1}^n \chi_i\bar{\phi}_\ell(\varepsilon_i)\Big)^2 \\ &= \big(\frac{h}{n}\big)^2\sum_{\ell=1}^L\sum_{i,j=1}^n E\big[\chi_i\chi_j\bar{\phi}_\ell(\varepsilon_i)\bar{\phi}_\ell(\varepsilon_j)\big] \\ &= \big(\frac{h}{n}\big)^2\sum_{\ell=1}^L\sum_{i,jk,m=1}^n E\big[t_{ik}\varepsilon_k t_{jm}\varepsilon_m\bar{\phi}_\ell(\varepsilon_i)\bar{\phi}_\ell(\varepsilon_j)\big] \\ &= \big(\frac{h}{n}\big)^2\sum_{\ell=1}^L\sum_{i,j=1}^n \Big[t_{ij}^2 E(\bar{\phi}_\ell(\varepsilon_i)^2) + (t_{ii}t_{jj}+t_{ij}t_{ji})\Big(E\big(\varepsilon_i\bar{\phi}_\ell(\varepsilon_i)\big)\Big)^2\Big] \\ &\le \sum_{\ell=1}^L \frac{h}{n}C^{2\ell}\mu_{2\kappa\ell} + \sum_{\ell=1}^L C^{2\ell}\mu_{\kappa\ell+1}^2 = O\Big(C^{2L}\rho_{2\kappa L}\Big), \end{aligned} \tag{4.7.49}$$

since we have $\mu_{\kappa(\ell+K)+1}^2 \le \mu_{2\kappa(\ell+K)+2}$ and as explained in lines following (4.7.36),

$$\sum_{i,j=1}^n t_{ij}^2 \le \sum_{i=1}^n \max_j |t_{ij}|\sum_{j=1}^n |t_{ij}| = O\Big(\frac{n}{h}\Big), \qquad \sum_{i=1}^n |t_{ii}|\sum_{j=1}^n |t_{jj}| = O\Big(\frac{n^2}{h^2}\Big),$$
$$\Big|\sum_{i,j=1}^n t_{ij}t_{ji}\Big| \le \sum_{i=1}^n |(TT^T)_{ii}| = O\Big(\frac{n}{h}\Big).$$

Therefore, $\|D_3\| = O_p(C^L L\rho_{2\kappa L}^{1/2})$.

Upper bound of $\|\tilde{w}^{(L)}\|$.

We decompose the $L\times 1$ vector $\tilde{w}^{(L)}$ into three parts and establish the following upper bound on their norms:

$$\begin{aligned} \|\tilde{w}^{(L)}\| &\le \|E(\phi'(\varepsilon_1)^{(L)})\| + \Big\|\frac{1}{n}\sum_{i=1}^n\Big(\phi'(\varepsilon_i)^{(L)} - E(\phi'(\varepsilon_1)^{(L)})\Big)\Big\| \\ &\quad + \Big\|\frac{1}{n}\sum_{i=1}^n\Big(\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \phi'(\varepsilon_i)^{(L)}\Big)\Big\| = O_p(C^L L\rho_{2\kappa L}^{1/2}), \end{aligned} \tag{4.7.50}$$

with the last step by Lemma 9 of Robinson (2005). Similarly, the first term of the RHS of (4.7.50) has the following upper bound:

$$\begin{aligned} \|E(\phi'(\varepsilon_1)^{(L)})\|^2 &= \sum_{\ell=1}^L\big[E(\phi'_\ell(\varepsilon_1))\big]^2 \le \sum_{\ell=1}^L C^{2\ell}\ell^2\mu_{\kappa(\ell+K)}^2 \\ &\le \sum_{\ell=1}^L C^{2\ell}\ell^2\mu_{2\kappa(\ell+K)} \le C^{2L}L^2\rho_{2\kappa L} = O(C^{2L}L^2\rho_{2\kappa L}). \end{aligned}$$

The second term of the RHS of (4.7.50) has the following upper bound, recalling the notation $\bar{\phi}'(\varepsilon_i)^{(L)} = \phi'(\varepsilon_i)^{(L)} - E(\phi'(\varepsilon_1)^{(L)})$:

$$E\Big\|\frac{1}{n}\sum_{i=1}^{n}\bar{\phi}'(\varepsilon_i)^{(L)}\Big\|^2 = \frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i=1}^{n}E(\bar{\phi}'(\varepsilon_i)^2) \leq C^{2L}L^2\rho_{2\kappa L} = O(C^{2L}L^2\rho_{2\kappa L}).$$

The third term of the RHS of (4.7.50) has the following upper bound, by the MVT,

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\Big(\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \phi'(\varepsilon_i)^{(L)}\Big)\Big\|^2 = \bar{\varepsilon}^2\cdot\frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i,j=1}^{n}\phi''_\ell(\varepsilon_i^*)\phi''_\ell(\varepsilon_j^*) = O_p\big(\frac{C^{2L}L^4\rho_{2\kappa L}}{n}\big),$$

because

$$\begin{aligned}\frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i,j=1}^{n}E|\phi''_\ell(\varepsilon_i^*)\phi''_\ell(\varepsilon_j^*)| &\leq \frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}\big[E(\phi''_\ell(\varepsilon_i^*)^2\big] \\ &\leq \sum_{\ell=1}^{L}C^{2\ell}\ell^4\mu_{2\kappa(\ell+2K)} \leq C^{2L}L^4\rho_{2\kappa L}.\end{aligned}$$

Upper bound of $\|(\tilde{W}^{(L)})^{-1}\|$.

We use the following matrix result, see e.g. Davies (1973, pp.496). If $W^{(L)}$ is non-singular and $\|\tilde{W}^{(L)} - W^{(L)}\|\|W^{(L)-1}\| < 1$, then

$$\|(\tilde{W}^{(L)})^{-1}\| \leq \frac{\|W^{(L)-1}\|}{1 - \|W^{(L)-1}\|\|\tilde{W}^{(L)} - W^{(L)}\|}.$$

Lemma 8 of Robinson (2005) states that $\|(W^{(L)})^{-1}\| = O(\pi_L)$. Lemma 10 and 19 of Robinson (2005) state, respectively:

$$\begin{aligned}&\|\tilde{W}^{(L)}(\varepsilon) - W^{(L)}\| = O_p\big((L\rho_{4\kappa L}/n)^{1/2}\big),\\ &\|\tilde{W}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - \tilde{W}^{(L)}(\varepsilon)\| = O_p\big(\frac{\rho_{2\kappa L}(Ln^{1/2} + L^{2\kappa L+2}(\log n))}{n} + \frac{L^{2\kappa L+1}\rho_{2\kappa L}^{1/2}(\log n)^{1/2}}{n}\big).\end{aligned}$$

We hence obtain

$$\begin{aligned}&\|\tilde{W}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - W^{(L)}\|\|(W^{(L)})^{-1}\|\\ &\leq \big(\|\tilde{W}^{(L)}(\varepsilon) - W^{(L)}\| + \|\tilde{W}^{(L)}(\frac{\epsilon(\lambda_0)}{\sigma_0}) - \tilde{W}^{(L)}(\varepsilon)\|\big)\|(W^{(L)})^{-1}\| \leq 1\end{aligned}$$

because by the same reasoning used in showing $q_n = o(1)$ in (4.7.46), we have

$$\Big(\frac{\sqrt{L\rho_{4\kappa L}}}{\sqrt{n}} + \frac{\rho_{2\kappa L}(L\sqrt{n} + L^{2\kappa L+2}(\log n))}{n} + \frac{L^{2\kappa L+1}\rho_{2\kappa L}^{1/2}(\log n)^{1/2}}{n}\Big)L(\log L)A^{2L} = o(1).$$

Therefore we conclude $\|\tilde{W}^{(L)}\| = O_p(\pi_L)$.

Upper bound of $\|F\|$. We obtain an upper bound on $\|F\|$ as, $\|F\| \leq \|\tilde{W}^{(L)}\|^2\|\partial\tilde{W}^{(L)}/\partial\lambda\|$, following from (4.7.35).

Introduce a $L \times 1$ vector,

$$n_i = \Phi^{(L)}(\frac{\epsilon_i}{\sigma_0}) - \bar{\phi}^{(L)}(\varepsilon_i) = \big[\Phi^{(L)}(\frac{\epsilon_i}{\sigma_0}) - \Phi^{(L)}(\varepsilon_i)\big] + \big[\Phi^{(L)}(\varepsilon_i) - \bar{\phi}^{(L)}(\varepsilon_i)\big] =: n_{1i} + n_2,$$

where similar as in r_i,

$$n_{1i\ell} = \big(\phi'_\ell(\epsilon^*_i) - \frac{1}{n}\sum_{j=1}^n \phi'_\ell(\epsilon^*_j)^{(L)}\big)\bar{\varepsilon},$$
$$n_2 = E(\phi^{(L)}(\varepsilon_i)) - \frac{1}{n}\sum_{j=1}^n \phi^{(L)}(\varepsilon_j).$$

We have, with some abuse of notation,

$$\begin{aligned} \frac{\partial \tilde{W}^{(L)}}{\partial \lambda} &= \frac{2}{n}\sum_{i=1}^n \chi_i \Phi^{(L)}(\frac{\epsilon_i}{\sigma_0}) \left(\phi'(\frac{\epsilon_i}{\sigma_0})^{(L)} - \frac{1}{n}\sum_{j=1}^n \phi'(\frac{\epsilon_j}{\sigma_0})^{(L)}\right)^T \\ &= \frac{2}{n}\sum_{i=1}^n \chi_i \left(\bar{\phi}^{(L)}(\varepsilon_i) + n_i\right)\left(\bar{\phi}'(\varepsilon_i)^{(L)} + r_i\right)^T \\ &= \frac{2}{n}\sum_{i=1}^n \chi_i \bar{\phi}^{(L)}(\varepsilon_i)\bar{\phi}'(\varepsilon_i)^{(L)T} + \frac{2}{n}\sum_{i=1}^n \chi_i \bar{\phi}^{(L)}(\varepsilon_i) r_i^T + \frac{2}{n}\sum_{i=1}^n \chi_i n_i \bar{\phi}'\varepsilon_i)^{(L)T} + \\ &\qquad + \frac{2}{n}\sum_{i=1}^n \chi_i n_i r_i^T =: F_1 + F_2 + F_3 + F_4. \end{aligned}$$

Below we will find upper bounds on $\|F_1\|_E, \|F_2\|_E, \|F_3\|_E$ and $\|F_4\|_E$ then conclude

$$\|\frac{\partial \tilde{W}^{(L)}}{\partial \lambda}\| = O_p((\frac{1}{\sqrt{n}} + \frac{1}{h})C^{2L}L^{3/2}\rho_{2\kappa L}^{1/2}),$$
$$\|F\| \leq \|(\tilde{W}^{(L)})^{-1}\|^2 \|\frac{\partial \tilde{W}^{(L)}}{\partial \lambda}\| = O_p(\pi_L^2)O_p((\frac{1}{\sqrt{n}} + \frac{1}{h})C^{2L}L^{3/2}\rho_{2\kappa L}^{1/2}).$$

Upper bound of $\|F_1\|$. Firstly, the (m,ℓ)-th element of F_1 is

$$(F_1)_{m\ell} = \frac{2}{n}\sum_{i=1}^n \chi_i \bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i).$$

Therefore, for $C_1 - C_5$ denoting constants,

$$\begin{aligned}
E\|F_1\|_E^2 &= \frac{4}{n^2}\sum_{m,\ell=1}^{L}\sum_{i,j=1}^{n}\sum_{i',j'=1}^{n} t_{ij}t_{i'j'}E(\varepsilon_j\varepsilon_{j'}\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i)\bar{\phi}_m(\varepsilon_{i'})\bar{\phi}'_\ell(\varepsilon_{i'})) \\
&= \frac{1}{n^2}\sum_{m,\ell=1}^{L}\Big[C_1\sum_{i,i',j=1}^{n} t_{ij}t_{i'j}E(\varepsilon_j^2)[E(\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))]^2 \\
&\quad +C_2\sum_{i,i'=1}^{n} t_{ii}t_{i'i'}[E(\varepsilon_i\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))]^2 + C_3\sum_{i,i'=1}^{n} t_{ii'}t_{i'i}[E(\varepsilon_i\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))]^2 \\
&\quad +C_4\sum_{i,i'=1}^{n} t_{ii}t_{i'i}E(\varepsilon_i^2\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))E(\bar{\phi}_m(\varepsilon_{i'})\bar{\phi}'_\ell(\varepsilon_{i'})) \\
&\quad +C_5\sum_{i=1}^{n} t_{ii}^2E(\varepsilon_i^2\bar{\phi}_m^2(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i)^2)\Big].
\end{aligned}$$

We have, based on the lines following (4.7.36),

$$\begin{aligned}
&\sum_{i,i',j=1}^{n}|t_{ij}t_{i'j}| = \sum_{j=1}^{n}\sum_{i=1}^{n}|t_{ij}|\sum_{i'=1}^{n}|t_{i'j}| = O(n), \quad \sum_{i,i'=1}^{n}|t_{ii}t_{i'i'}| = O(\frac{n^2}{h^2}), \\
&\sum_{i,i'=1}^{n}|t_{ii'}t_{i'i}| = O(\frac{n^2}{h^2}), \quad \sum_{i,i'=1}^{n}|t_{ii}t_{i'i}| = O(\frac{n^2}{h^2}), \quad \sum_{i=1}^{n}|t_{ii}^2| = O(\frac{n}{h^2}).
\end{aligned}$$

We also have that

$$\begin{aligned}
&\Big|E(\varepsilon_j^2)[E(\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))]^2\Big|, |E(\varepsilon_i^2\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))E(\bar{\phi}_m(\varepsilon_{i'})\bar{\phi}'_\ell(\varepsilon_{i'}))|, \\
&[E(\varepsilon_i\bar{\phi}_m(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i))]^2, E(\varepsilon_i^2\bar{\phi}_m^2(\varepsilon_i)\bar{\phi}'_\ell(\varepsilon_i)^2) \leq C^{2(m+\ell)}\ell^2\mu_{2(1+\kappa(m+\ell+K))},
\end{aligned}$$

since $|s^2\bar{\phi}_m(s)^2\bar{\phi}'_\ell(s)^2| \leq C^{2m+2\ell}\ell^2\big(1+|s|^{2(1+m\kappa+\kappa(\ell+K))}\big)$. Therefore,

$$\begin{aligned}
E\|F_1\|_E^2 &= \frac{4}{n^2}\sum_{m,\ell=1}^{L} O\Big(C^{2(m+\ell)}\ell^2\mu_{2(1+\kappa(m+\ell+K))}\big(n+\frac{n^2}{h^2}\big)\Big) \\
&\leq C^{4L}\big(\frac{1}{n}+\frac{1}{h^2}\big)\sum_{m=1}^{L}\Big[\sum_{\ell=1}^{L}\ell^2\mu_{2(1+\kappa(m+\ell+K))}\Big] \\
&= C^{4L}\big(\frac{1}{n}+\frac{1}{h^2}\big)\sum_{m=1}^{L}L^2\rho_{2\kappa L} = O\big((\frac{1}{n}+\frac{1}{h^2})C^{4L}L^3\rho_{2\kappa L}\big).
\end{aligned}$$

Upper bound for $\|F_2\|$. Using $r_1 = r_{1i} + r_2$, we decompose

$$F_2 = \frac{2}{n}\sum_{i=1}^{n}\chi_i\bar{\phi}^{(L)}(\varepsilon_i)r_{1i}^T + \frac{2}{n}\sum_{i=1}^{n}\chi_i\bar{\phi}^{(L)}(\varepsilon_i)r_2^T =: F_{21} + F_{22}.$$

Recalling $r_{1i\ell} = \{\phi''_\ell(\varepsilon^*_i) - \frac{1}{n}\sum_{j=1}^n \phi''_\ell(\varepsilon^*_j)\}\bar{\varepsilon}$, the (m,ℓ)-th argument of F_{21} is given by

$$(F_{21})_{m,\ell} = \frac{2}{n}\sum_{i=1}^n \chi_i \bar{\phi}_m(\varepsilon_i)\bar{\varepsilon}\{\phi''_\ell(\varepsilon^*_i) - \frac{1}{n}\sum_{j=1}^n \phi''_\ell(\varepsilon^*_j)\}.$$

Introduce notation $p''_{i\ell} := \phi''_\ell(\varepsilon^*_i) - \frac{1}{n}\sum_{j=1}^n \phi''_\ell(\varepsilon^*_j)$. We will show that

$$\|F_{21}\|_E^2 = \bar{\varepsilon}^2 \tfrac{4}{n^2}\sum_{m,\ell=1}^L \sum_{i,j=1}^n \Big[\chi_i\chi_j\bar{\phi}_m(\varepsilon_i)p''_{i\ell}\bar{\phi}_m(\varepsilon_j)p''_{j\ell}\Big] = O_p(\frac{1}{n})O_p(\frac{C^{4L}L^3\rho_{4\kappa L}}{nh}).$$

Indeed, by Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{4}{n^2}\sum_{m,\ell=1}^L\sum_{i,j=1}^n E\Big|\chi_i\chi_j\bar{\phi}_m(\varepsilon_i)p''_{i\ell}\bar{\phi}_m(\varepsilon_j)p''_{j\ell}\Big| &\le C\frac{4}{n^2}\sum_{m,\ell=1}^L\sum_{i,j=1}^n \big[E\chi_i^2\chi_j^2\big]^{1/2}\big[E(\bar{\phi}_m(\varepsilon_i)^4 p''^{\,4}_{i\ell})\big]^{1/2}\\ &\le O(\frac{1}{h})\frac{1}{n^2}\sum_{i,j=1}^n\sum_{m,\ell=1}^L C^{2(m+\ell)}\ell^2\mu^{1/2}_{4\kappa(m+\ell+2K)}\\ &\le O(\frac{n}{h})\sum_{m=1}^L\sum_{\ell=1}^L C^{2(m+\ell)}\ell^2\mu_{4\kappa(m+\ell+2K)}\\ &\le \frac{1}{nh}\sum_{m=1}^L C^{4L}L^2\rho_{4\kappa L} = O\big(\frac{C^{4L}L^3\rho_{4\kappa L}}{nh}\big).\end{aligned}$$

because $\max_{1\le i,j\le n}\big[E\chi_i^2\chi_j^2\big]^{1/2} \le \max_{1\le i,j\le n}\big[E\chi_i^4 E\chi_j^4\big]^{1/4} = O(1/h)$ by (4.7.42), with the last inequality following from Lemma 9 of Robinson (2005).

As for F_{22}, we already have $\|r_2\| = O(C^L L\rho^{1/2}_{2\kappa L}/\sqrt{n})$ by (4.7.48). From evaluation of D_3, (4.7.49), we also have

$$\|\frac{2}{n}\sum_{i=1}^n \chi_i\bar{\phi}^{(L)}(\varepsilon_i)\| = O\Big(\frac{C^L\rho^{1/2}_{2\kappa L}}{h}\Big),$$

yielding

$$\|F_{22}\|_E \le O_p(\frac{C^{2L}L\rho_{2\kappa L}}{\sqrt{n}h}).$$

Therefore, $\|F_2\| = O_p(\frac{C^{2L}L\rho_{2\kappa L}}{\sqrt{n}h})$.
Upper bound of $\|F_3\|$. We decompose

$$\frac{2}{n}\sum_{i=1}^n \chi_i(n_{1i}+n_2)\bar{\phi}'(\varepsilon_i)^{(L)T} = \frac{2}{n}\sum_{i=1}^n \chi_i n_{1i}\bar{\phi}'(\varepsilon_i)^{(L)T} + \frac{2}{n}\sum_{i=1}^n \chi_i n_2\bar{\phi}'(\varepsilon_i)^{(L)T} =: F_{31} + F_{32}.$$

By the MVT, the ℓ-th element of n_{1i} is

$$n_{1i\ell} = \big(\phi'_\ell(\epsilon_i^*) - \frac{1}{n}\sum_{j=1}^{n}\phi'_\ell(\epsilon_j^*)\big)\bar{\varepsilon} =: p'_{i\ell}\bar{\varepsilon}.$$

This means that the (m,ℓ)-th element of F_{31} is

$$(F_{31})_{m,\ell} = \frac{2}{n}\sum_{i=1}^{n}\bar{\varepsilon}\chi_i p'_{im}\bar{\phi}'_\ell(\varepsilon_i).$$

Therefore,

$$\begin{aligned}\|F_{31}\|_E^2 &= \frac{4}{n^2}\sum_{m,\ell=1}^{L}\sum_{i,j=1}^{n}\big[\chi_i\chi_j\bar{\varepsilon}^2 p'_{im}\bar{\phi}'_\ell(\varepsilon_i)p'_{jm}\bar{\phi}'_\ell(\varepsilon_j)\big]\\ &= \bar{\varepsilon}^2 \times \frac{4}{n^2}\sum_{m,\ell=1}^{L}\sum_{i,j=1}^{n}\big[\chi_i\chi_j\bar{\varepsilon}^2 p'_{im}\bar{\phi}'_\ell(\varepsilon_i)p'_{jm}\bar{\phi}'_\ell(\varepsilon_j)\big].\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{4}{n^2}\sum_{m,\ell=1}^{L}\sum_{i,j=1}^{n}E\big|\chi_i\chi_j\bar{\varepsilon}^2 p'_{im}\bar{\phi}'_\ell(\varepsilon_i)p'_{jm}\bar{\phi}'_\ell(\varepsilon_j)\big| &\leq \frac{C}{n^2}\sum_{m,\ell=1}^{n}\sum_{i,j=1}^{n}[E\chi_i^2\chi_j^2]^{1/2}\big[E{p'_{im}}^4\bar{\phi}'_\ell(\varepsilon_i)^4\big]^{1/2}\\ = O(\frac{1}{h})\sum_{m,\ell=1}^{L}C^{2(m\ell)}\ell^2 m^2\mu_{4\kappa(m+\ell+2K)}^{1/2} &\leq O(\frac{1}{h})C^{4L}L^5\rho_{4\kappa L}.\end{aligned}$$

Therefore,

$$\|F_{31}\|_E^2 = O(\frac{1}{n}\times\frac{C^{4L}L^5\rho_{4\kappa L}}{h}) = O(\frac{C^{4L}L^5\rho_{4\kappa L}}{nh}).$$

Now, usingthe $L\times 1$ vector $n_2 = (n_{21},\cdots,n_{2L})^T$, we will find an upper bound on $E\|n_2\|^2 = \sum_{\ell=1}^{L}E(n_{2\ell}^2)$. For each ℓ, recalling the notation $\bar{\phi}_\ell(\varepsilon) = \phi_\ell(\varepsilon) - E[\phi_\ell(\varepsilon)]$,

$$\begin{aligned}E(n_{2\ell}^2) = \frac{1}{n^2}E[\sum_{j=1}^{n}\phi_\ell(\varepsilon_j) - E(\phi_\ell(\varepsilon_1))]^2 &= \frac{1}{n^2}E[\sum_{j=1}^{n}\bar{\phi}_\ell(\varepsilon_j)]^2\\ = \frac{1}{n^2}\sum_{j=1}^{n}E[\bar{\phi}_\ell^2(\varepsilon_j)] &= \frac{C^{2\ell}\mu_{2\kappa\ell}}{n}.\end{aligned}$$

Therefore, by Lemma 9 of Robinson (2005),

$$E\|n_2\|^2 = \sum_{\ell=1}^{L}E(n_{2\ell}^2) \leq \sum_{\ell=1}^{L}\frac{C^{2\ell}\mu_{2\kappa\ell}}{n} = O(\frac{C^{2L}\rho_{2\kappa L}}{n}). \tag{4.7.51}$$

Hence,

$$\|F_{32}\| \le \|n_2\| \Big\| \frac{2}{n} \sum_{i=1}^{n} \chi_i \bar{\phi}'(\varepsilon_i) \Big\| = O_p\Big(\frac{C^{2L} L \rho_{2\kappa L}}{\sqrt{n} h} \Big),$$

because by the same argument as in the proof of (4.7.49) with $\bar{\phi}'^{(L)}$ and $\ell^2 \mu_{2\kappa(\ell+K)}$ replacing $\bar{\phi}^{(L)}$ and $\mu_{2\kappa\ell}$, respectively, we obtain

$$\| \frac{2}{n} \sum_{i=1}^{n} \chi_i \bar{\phi}'^{(L)}(\varepsilon_i) \| = O\Big(\frac{C^L L \rho_{2\kappa L}^{1/2}}{h} \Big).$$

Therefore, $\|F_3\| = O_p(C^{2L} L^{5/2} \rho_{4\kappa L}^{1/2} / \sqrt{n} h)$.

Upper bound for $\|F_4\|$. Write

$$\begin{aligned} F_4 = \frac{2}{n} \sum_{i=1}^{n} \chi_i (n_{1i} + n_2)(r_{1i} + r_2)^T &= \frac{2}{n} \sum_{i=1}^{n} \chi_i \big(n_{1i} r_{1i}^T + n_{1i} r_2^T + n_2 r_{1i}^T + n_2 r_2^T \big) \\ &=: F_{41} + F_{42} + F_{43} + F_{44}. \end{aligned}$$

We have $\|r_2\| = O_p(C^L L \rho_{2\kappa L}^{1/2} / \sqrt{n})$ by (4.7.48) and $\|n_2\| = O_p\big(C^L \rho_{2\kappa L}^{1/2} / \sqrt{n}\big)$ from (4.7.51). Recall notations,

$$\begin{aligned} n_{1i\ell} &= \big(\phi'_\ell(\epsilon_i^*) - \frac{1}{n} \sum_{j=1}^{n} \phi'_\ell(\epsilon_j^*) \big) \bar{\varepsilon} =: p'_{i\ell} \bar{\varepsilon}, \\ r_{1i\ell} &= \big(\phi''_\ell(\epsilon_i^*) - \frac{1}{n} \sum_{j=1}^{n} \phi''_\ell(\epsilon_j^*) \big) \bar{\varepsilon} =: p''_{i\ell} \bar{\varepsilon}. \end{aligned}$$

Firstly, we show that

$$\|F_{41}\|_E^2 \le \Big\| \frac{1}{n} \sum_{i=1}^{n} \chi_i n_{1i} r_{1i}^T \Big\|_E^2 = \bar{\varepsilon}^4 \times \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{m,\ell=1}^{L} \chi_i \chi_j p'_{i\ell} p''_{i\ell} p'_{jm} p''_{jm} = O_p\Big(\frac{C^{4L} L^6 \rho_{8\kappa L}^2}{n^2 h} \Big). \quad (4.7.52)$$

By Cauchy-Schwarz inequality, using the argument we used repeatedly above,

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{m,\ell=1}^{L} E|\chi_i \chi_j p'_{i\ell} p''_{i\ell} p'_{jm} p''_{jm}| \\ &\qquad \le \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{m,\ell=1}^{L} (E\chi_i^2 \chi_i^2)^{1/2} \big\{ E({p'_{i\ell}}^4 {p''_{i\ell}}^4) E({p'_{jm}}^4 {p''_{jm}}^4) \big\}^{1/4} \\ &\qquad \le \frac{1}{h} \sum_{m,\ell=1}^{L} C^{2(\ell+m)} \ell^3 m^3 \mu_{4\kappa(\ell+3K)}^{1/4} \mu_{4\kappa(m+3K)}^{1/4} \\ &\qquad \le \frac{C^{4L} L^6}{h} \Big(\sum_{\ell=1}^{L} \mu_{4\kappa(\ell+3K)}^{1/4} \Big)^2 \le \frac{C^{4L} L^6}{h} \rho_{8\kappa L}^2. \end{aligned}$$

Since $\bar{\varepsilon}^4 = O(1/n^2)$, this proves (4.7.52). To bound $\|F_{42}\|$, we use that

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\chi_i n_{1i}\Big\|^2 = \frac{\bar{\varepsilon}^2}{n^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}\chi_i\chi_j p'_{i\ell}p'_{j\ell} = O_p\Big(\frac{1}{n}\Big)O_p(\frac{C^{4L}L^4\rho_{4\kappa L}}{h}),$$

which follows noting that by Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}E|\chi_i\chi_j p'_{i\ell}p'_{j\ell}| &\le \frac{1}{n}\sum_{i,j=1}^{n}(E\chi_i^2\chi_j^2)^{1/2}\sum_{\ell=1}^{L}\big(E(p'_{i\ell})^4\big)^{1/2}\\ &\le \frac{C}{h}\sum_{\ell=1}^{L}C^{4\ell}\ell^4\mu_{4\kappa(\ell+K)} = O(\frac{C^{4L}L^4\rho_{4\kappa L}}{h}).\end{aligned}$$

To bound $\|F_{43}\|$, we use that

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\chi_i r_{1i}\Big\|^2 = \frac{\bar{\varepsilon}^2}{n^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}\chi_i\chi_j p''_{i\ell}p''_{j\ell} = O_p\Big(\frac{1}{n}\Big)O_p(\frac{C^{4L}L^8\rho_{4\kappa L}}{h}),$$

which follows noting that by Cauchy-Schwarz inequality,

$$\begin{aligned}\frac{1}{n^2}\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}E|\chi_i\chi_j p''_{i\ell}p''_{j\ell}| &\le \frac{1}{n}\sum_{i,j=1}^{n}(E\chi_i^2\chi_j^2)^{1/2}\sum_{\ell=1}^{L}\big(E(p''_{i\ell})^4\big)^{1/2}\\ &\le \frac{C}{h}\sum_{\ell=1}^{L}C^{4\ell}\ell^8\mu_{4\kappa(\ell+2K)} = O(\frac{C^{4L}L^8\rho_{4\kappa L}}{h}).\end{aligned}$$

To bound $\|F_{44}\|$, we use that

$$\frac{1}{n}\sum_{i=1}^{n}E|\chi_i| \le \frac{1}{n}\sum_{i,j=1}^{n}t_{ij}^2 \le \frac{1}{n}\sum_{i=1}^{n}\max_j|t_{ij}|\sum_{j=1}^{n}|t_{ij}| = O(\frac{1}{h}).$$

Combining results above gives

$$\begin{aligned}\|F_{41}\| &= O_p\Big(\frac{C^{2L}L^3\rho_{8\kappa L}}{n\sqrt{h}}\Big),\\ \|F_{42}\| &= O_p(\frac{C^{2L}L^2\rho_{4\kappa L}^{1/2}}{\sqrt{nh}})O_p(C^L L\rho_{2\kappa L}^{1/2}/\sqrt{n}) = O_p(\frac{C^{3L}L^3\rho_{4\kappa L}}{n\sqrt{h}}),\\ \|F_{43}\| &= O_p\Big(\frac{C^{2L}L^4\rho_{4\kappa L}^{1/2}}{\sqrt{nh}}\Big)O_p(C^L\rho_{2\kappa L}^{1/2}/\sqrt{=}n) = O_p\Big(\frac{C^{3L}L^4\rho_{4\kappa L}}{n\sqrt{h}}\Big).\\ \|F_{44}\| &= O_p\Big(C^L L\rho_{2\kappa L}^{1/2}/\sqrt{n}\Big)O_p(C^L\rho_{2\kappa L}^{1/2}/\sqrt{n})O_p(\frac{1}{h}) = O_p\Big(\frac{C^{2L}L\rho_{2\kappa L}}{nh}\Big).\end{aligned}$$

Recall that for any fixed $a, b, c \ge 0$, we may find $\alpha > 0$ large enough so that

$C^{aL}L^b\rho_{cL} = O(L^{\alpha L}) = o(\sqrt{h})$ from Assumption 5. Therefore, we conclude

$$\|F_4\| = O_p\Big(\frac{C^{2L}L^3(\rho_{8\kappa L} + L\rho_{4\kappa L})}{n\sqrt{h}}\Big).$$

We have found

$$\begin{aligned}
\|F_1\| &= O_P\Big((\frac{1}{\sqrt{n}} + \frac{1}{h})C^{2L}L^{3/2}\rho_{2\kappa L}^{1/2}\Big),\\
\|F_2\| &= O_P\Big(\frac{C^{2L}L^{1/2}\rho_{2\kappa L}^{1/2}}{n\sqrt{h}}\Big),\\
\|F_3\| &= O_P\Big(\frac{C^{2L}L^{5/2}\rho_{4\kappa L}^{1/2}}{\sqrt{nh}}\Big),\\
\|F_4\| &= O_p\Big(\frac{C^{2L}L^3(\rho_{8\kappa L} + L\rho_{4\kappa L})}{n\sqrt{h}}\Big).
\end{aligned}$$

By the same reasoning as above, we find that the rate of $\|F_1\|$ dominates. Therefore,

$$\|F\| = O_P\Big((\frac{1}{\sqrt{n}} + \frac{1}{h})C^{2L}L^{3/2}\pi_L^2\rho_{2\kappa L}^{1/2}\Big).$$

Upper bound of $\|J\|$. Write

$$\begin{aligned}
J = \frac{\partial(\tilde{w}^{(L)})^{-1}}{\partial\lambda} &= \frac{1}{n}\sum_{i=1}^{n}\phi''(\frac{\epsilon_i}{\sigma_0})^{(L)}\chi_i\\
&= \frac{1}{n}\sum_{i=1}^{n}\phi''(\varepsilon_i)^{(L)}\chi_i + \frac{1}{n}\sum_{i=1}^{n}\big[\phi''(\frac{\epsilon_i}{\sigma_0})^{(L)} - \phi''(\varepsilon_i)^{(L)}\big]\chi_i =: J_1 + J_2.
\end{aligned}$$

We find upper bounds on $\|J_1\|$ and $\|J_2\|$. Firstly,

$$\begin{aligned}
E\big\|J_1\big\|^2 &= \frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i,j=1}^{n}E(\chi_i\chi_j\phi_\ell''(\varepsilon_i)\phi_\ell''(\varepsilon_j)^{(L)})\\
&\leq \frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i,j=1}^{n}[E(\chi_i^2\chi_j^2)]^{1/2}[E\phi_\ell''(\varepsilon_i)^4]^{1/2}\\
&= O(\frac{1}{h})\sum_{\ell=1}^{L}C^{2\ell}\ell^4\mu_{4\kappa(\ell+m+2K)}^{1/2} = O(\frac{C^{2L}L^4\rho_{4\kappa L}}{h}).
\end{aligned}$$

because $\max_{1\leq i,j\leq n}[E(\chi_i^2\chi_j^2)]^{1/2} = O(1/h)$ by (4.7.42) and using Lemma 9 of Robinson (2005).

Now, note that the MVT implies that the ℓ-th element of $\phi''(\frac{\epsilon_i}{\sigma_0})^{(L)} - \phi''(\varepsilon_i)^{(L)} =$

$\bar{\varepsilon}\phi_\ell'''(\varepsilon_i^*)$. Hence,

$$E\|J_2\|^2 = \frac{1}{n^2}\sum_{\ell=1}^{L}\sum_{i,j=1}^{n} E(\chi_i\chi_j\bar{\varepsilon}^2\phi_\ell'''(\varepsilon_i^*)\phi_\ell'''(\varepsilon_j^*))$$
$$\leq \frac{1}{n^2}\bar{\varepsilon}^2\sum_{i,j=1}^{n}\sum_{\ell=1}^{L}[E(\chi_i^2\chi_j^2)]^{1/2}[E(\phi_\ell'''(\varepsilon_i^*)\phi_\ell'''(\varepsilon_j^*))^2]^{1/2} \leq O(\frac{1}{nh})\sum_{\ell=1}^{L}[E(\phi_\ell'''(\varepsilon_i^*)^4)]^{1/2}$$
$$= O(\frac{1}{nh})\sum_{\ell=1}^{L}C^{2\ell}\ell^6\mu_{4\kappa(\ell+3K)} = O\Big(\frac{C^{2L}L^6\rho_{4\kappa L}}{nh}\Big),$$

since $\bar{\varepsilon}^2 = O(1/n)$ and $[E(\chi_i^2\chi_j^2)]^{1/2} = O(1/h)$ uniformly in i and j, and using Lemma 9 of Robinson (2005). Therefore, $\|J\| = O_p(C^L L^2\rho_{4\kappa L}^{1/2}/\sqrt{h})$.

To complete the proof of $\|B\| = o_p(1)$, we list our findings:

$$\|D\| = O_p(C^L L\rho_{2\kappa L}^{1/2}),$$
$$\|F\| = O_p\Big(\pi_L^2(\frac{1}{\sqrt{n}}+\frac{1}{h})C^{2L}L^{3/2}\rho_{2\kappa L}^{1/2}\Big).$$
$$\|\tilde{w}^{(L)}\| = O_p(C^L L\rho_{2\kappa L}^{1/2}),$$
$$\|(\tilde{W}^{(L)})^{-1}\| = O_p(\pi_L),$$
$$\|J\| = O_p(\frac{C^L L^2\rho_{4\kappa L}^{1/2}}{\sqrt{h}}).$$

Therefore,

$$\begin{aligned}\|B\| &\leq \|D\|[\|F\|\|\tilde{w}^{(L)}\| + \|(\tilde{W}^{(L)})^{-1}\|\|J\|]\\ &= O_p\Big(\frac{\pi_L C^{2L}L^3\rho_{4\kappa L}}{\sqrt{h}}\big(1+\pi_L(\frac{h}{n}+\frac{1}{\sqrt{h}})C^{2L}L^{1/2}\rho_{2\kappa L}^{1/2}\big)\Big) = o_p(1),\end{aligned}$$

where the last step follows by again noting for any $a, b, c, d > 0$, we may find $\alpha > 0$ large enough such that $C^{aL}L^b\rho_{cL}\pi_L^d = O(L^{\alpha L}) = o(\sqrt{h})$ by Assumption 5. This completes the proof of (4.7.5).

Comments on the proofs of (4.7.6) to (4.7.9).

Proof of (4.7.6) follows similar steps to those used in the proof of (4.7.5), with derivatives of r_L and $\epsilon(\lambda)/\sigma$ w.r.t. λ replaced by those w.r.t σ.

For the proof of (4.7.7), one has

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_{iL}^2(\lambda_0,\sigma_0) - \mathcal{J} = \frac{1}{n}\sum_{i=1}^{n}[\tilde{\psi}_{iL}^2(\lambda_0,\sigma_0) - \psi^2(\varepsilon_i)] + \Big[\frac{1}{n}\sum_{i=1}^{n}\psi^2(\varepsilon_i) - \mathcal{J}\Big]. \quad (4.7.53)$$

The second term in (4.7.53) is of order $o_p(1)$ since ε_i's are *i.i.d.* random variables and $\mathcal{J} = E(\psi^2(\varepsilon_1))$. To show that the first term in (4.7.53) is of order $o_p(1)$, we need to establish that uniformly in $i = 1, \cdots, n$,

$$\begin{aligned}|\tilde{\psi}_{iL}^2(\lambda_0, \sigma_0) - \psi^2(\varepsilon_i)| &\leq |\tilde{\psi}_{iL}(\lambda_0, \sigma_0) - \psi(\varepsilon_i)||\tilde{\psi}_{iL}(\lambda_0, \sigma_0) + \psi(\varepsilon_i)| \\ &\leq |\tilde{\psi}_{iL}(\lambda_0, \sigma_0) - \psi(\varepsilon_i)|\Big(|2\psi(\varepsilon_i)| + |\tilde{\psi}_{iL}(\lambda_0, \sigma_0) - \psi(\varepsilon_i)|\Big) = o_p(1),\end{aligned}$$

which is in turn implied by

$$|\tilde{\psi}_{iL}(\lambda_0, \sigma_0) - \psi(\varepsilon_i)| = o_p(1).$$

Recalling $\tilde{\psi}_{iL}(\lambda_0, \sigma_0) = \Phi(\epsilon_i(\lambda_0)/\sigma_0)^{(L)T}\tilde{a}^{(L)}(\epsilon(\lambda)/\sigma_0)$, the above statement can be verified using the following decomposition of $\Phi(\epsilon_i(\lambda_0)/\sigma_0)^T$ and $\tilde{a}^{(L)}(\epsilon(\lambda)/\sigma_0)$ then following steps similar to those in the proof of (4.7.5):

$$\begin{aligned}&\Phi(\frac{\epsilon_i(\lambda_0)}{\sigma_0})^{(L)} = \Phi(\frac{\epsilon_i(\lambda_0)}{\sigma_0})^{(L)} - \Phi(\varepsilon_i)^{(L)} + \Phi(\varepsilon_i)^{(L)} - \bar{\phi}^{(L)}(\varepsilon_i) + \bar{\phi}^{(L)}(\varepsilon_i), \\ &\tilde{a}^{(L)}(\frac{\epsilon(\lambda)}{\sigma_0}) - a^{(L)} = \tilde{a}^{(L)}(\frac{\epsilon(\lambda)}{\sigma_0}) - \tilde{a}^{(L)}(\varepsilon) + \tilde{a}^{(L)}(\varepsilon) - a^{(L)}.\end{aligned}$$

The proofs of (4.7.8) and (4.7.9) are broadly similar to the proofs of (4.7.5), (4.7.6) and (4.7.9). They rely on the mean value theorem to find upper bounds on the difference between quantities evaluated at $(\lambda, \sigma) \in \mathcal{N}$ and at (λ_0, σ_0), where $\mathcal{N} = \Big(\lambda, \sigma : |\lambda - \lambda_0| \leq \sqrt{h/n}, |\sigma - \sigma_0| \leq \sqrt{1/n}\Big)$. In the proof of (4.7.5), the difference between the fitted residual at (λ_0, σ_0), ϵ_i/σ_0, and the true error term, ε_i, was $\bar{\varepsilon} = O_p(1/\sqrt{n})$. Below, we will show that the difference between fitted residuals at (λ_0, σ_0) and any $(\lambda, \sigma) \in \mathcal{N}$, is also of order $O_p(1/\sqrt{n})$. Recall that $\epsilon(\lambda) = HS(\lambda)y$, where $S(\lambda) = I - \lambda W$. We have

$$\begin{aligned}\frac{\epsilon(\lambda)}{\sigma} - \frac{\epsilon(\lambda_0)}{\sigma_0} &= \frac{\epsilon(\lambda)}{\sigma} - \frac{\epsilon(\lambda_0)}{\sigma} + \frac{\epsilon(\lambda_0)}{\sigma} - \frac{\epsilon(\lambda_0)}{\sigma_0} \\ &= \frac{(\lambda_0 - \lambda)\sigma_0 HG\varepsilon}{\sigma} + \frac{\sigma_0 - \sigma}{\sigma}H\varepsilon,\end{aligned}$$

with the second inequality following from

$$\begin{aligned}\epsilon(\lambda) - \epsilon(\lambda_0) &= H\big(S(\lambda) - S(\lambda_0)\big)y = (\lambda_0 - \lambda)HWy \\ &= (\lambda_0 - \lambda)HW(S^{-1}\sigma_0\varepsilon + \mu_0 1_n) = (\lambda_0 - \lambda)\sigma_0 HG\varepsilon, \\ \epsilon(\lambda_0) &= HSy = HS(S^{-1}\sigma_0\varepsilon + \mu_0 1_n) = \sigma_0 H\varepsilon,\end{aligned}$$

since $HS1_n = H(I - \lambda_0 W)1_n = H(1 - \lambda_0)1_n = 0$ as $W1_n = 1_n$. This means that for $i = 1, \cdots, n$, and $(\lambda, \sigma) \in \mathcal{N}$,

$$\begin{aligned}|\frac{\epsilon_i(\lambda)}{\sigma} - \frac{\epsilon_i(\lambda_0)}{\sigma_0}| &\leq |\lambda_0 - \lambda||\frac{\sigma_0}{\sigma}||\chi_i| + |\frac{\sigma_0 - \sigma}{\sigma}||\varepsilon_i - \bar{\varepsilon}| \\ &= O(\frac{\sqrt{h}}{\sqrt{n}})O(1)O_p(\frac{1}{\sqrt{h}}) + O(\frac{1}{\sqrt{n}})O_p(1) = O_p(\frac{1}{\sqrt{n}}),\end{aligned}$$

because $\chi_i = O_P\big((E\chi_i^2)^{1/2}\big)$ where $E(\chi_i^2) = \sum_{j=1}^{n} t_{ij}^2 = O(1/h)$. ■

References

Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric models. *Econometrica*, **59**, 307-345.

Andrews, D.W.K. (2003). Cross-section regression with common shocks. *Discussion Paper 1428*, Cowles Foundation, Yale University.

Andrews, D.W.K. (2005). Cross-section regression with common shocks. *Econometrica*, **73**, 1551-1585.

Andrews, D.W.K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, **60**, 953-966.

Arbia, G. (2006). *Spatial Econometrics: Statistical Foundation and Applications to Regional Analysis.* Springer-Verlag, Berlin.

Arellano, M. and Honore, B. (2001). Panel data models: some recent developments. In: *The Handbook of Econometrics*, ed. J. J. Heckman and E. E. Leamer. Vol. 6, North-Holland.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, **77**, 1229-1279.

Beran, R. (1976). Adaptive estimates for autoregressive processes. *Annals of the Institute of Statistical Mathematics*, **26**, 77-89.

Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley and Sons, Inc., New York.

Brillinger, D.R. (1981). *Time Series: Data Analysis and Theory.* Holden Day, Inc., San Francisco.

Burkholder, D.L. (1973). Distribution function inequalities for martingales. *The Annals of Probability*, **1**, 19-42.

Case, A.C. (1991). Spatial patterns in household demand. *Econometrica*, **59**, 953-965.

Chamberlain, G. (1986). Notes on semiparametric regression. *unpublished manuscript*, Department of Economics, Harvard University.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. eds. James J. Heckman and Edward E. Leamer, In: *Handbook of Econometrics.* Vol. 6B, Chapter 76, North-Holland.

Chen, X., Liao, Z. and Sun, Y. (2010). On inference of sieve M-estimation of functionals of semi-nonparametric time series models. *Preprint.*

Chen, X. and Shen, X. (1998). Sieve extreme estimates for weakly dependent data. *Econometrica*, **66**, 289-314.

Cliff, A. and J.K. Ord, (1968). The problem of spatial autocorrelation. *Joint Discussion Paper, University of Bristol: Department of Economics*, **26**, *Department of Geography, Series A*, **15**.

Cliff, A. and Ord, J. (1981). *Spatial Processes, Models and Applications.* Pion, London.

Conley, T.G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, **92**, 1-45.

Davies, R.B. (1973). Asymptotic inference in stationary Gaussian time-series. *Advances in Applied Probability*, **5**, 469-497.

Davidson, J. and de Jong, R.M. (2000). The functional central limit theorem and weak convergence of stochastic integrals II. *Econometric Theory*, **16**, 643-666.

Dehling, H. (2006). Limit theorems for dependent U-statistics . *Lecture Notes in Statistics*, **187**, 65-86.

de Jong, R.M. (2002). A note on Convergence rates and asymptotic normality for series estimators: uniform convergence rates. *Journal of Econometrics*, **111**, 1-9.

Den Haan, W. J. and Levin, A. (1997). A practitioner's guide to robust covariance matrix estimation. eds. by G. S. Maddala and C. R. Rao, In: *Handbook of Statistics: Robust Inference.* Vol. 15, Chapter 12, New York: Elsevier.

Fan, Y. and Li, Q. (1999). Root-n-consistent estimation of partially linear time series models. *Journal of Nonparametric Statistics*, **11**, 251-269.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both "n" and "T" are large. *Econometrica*, **70**, 1639-1657.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

Hannan, E. J. (1957). The variance of the mean of a stationary process. *Journal of the Royal Statistical Society Series B*, **19**, 282-285.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24**, 726-748.

Henderson, D. J., Carroll, R. J. and Li, Q. (2008). Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics*, **144**, 257 - 275.

Hidalgo, J. (1997). Nonparametric estimation with strongly dependent multivariate time series. *Journal of Time Series Analysis*, **18**, 97-122.

Horn, R.A. and Johnson, C.R. (1990). *Matrix Analysis.* Cambridge University Press.

Kasahara, Y. and Maejima, M. (1986). Functional limit theorems for weighted sums of I.I.D. random variables. *Probability Theory and Related Fields.* **72**, 161-183.

Kelejian, H.H. and Prucha, I.R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, **17**, 99-121.

Kelejian, H.H. and Prucha, I.R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, **40**, 509-533.

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics*, **104**, 219-257.

Kelejian, H.H. and Prucha, I.R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, **140**, 131-154.

Kiefer, M., Vogelsang T.J. and Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, **68**, 695-714.

Lee, L.F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, **18**, 252-277.

Lee, L.F. (2004). Asymptotic distribution of quasi-maximum likelihood estimates for spatial autoregressive models. *Econometrica*, **72**, 1899-1925.

LeSage, J. P. and Pace, R. K. (2004). A matrix exponential spatial specification. *Journal of Econometrics*, **140**, 190-214.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, **17**, 571-599.

Newey, W. K. (1988). Adaptive Estimation of regression models via moment restrictions. *Journal of Econometrics*, **38**, 301-339.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, **79**, 147-168.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, 703-708.

Pagan, A. and Ullah, A (1999). *Nonparametric Econometrics.* Cambridge University Press, New York.

Penrose, R. (1955). A generalized inverse for matrices. *Proc. Cambridge Phil. Soc.*, **51**, 406-413.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, **74**, 967-1012.

Pham, D.T. (1986). The mixing property of bilinear and generalised random coefficient autoregressive models. *Stochastic Processes and their Applications*, **23**, 291-300.

Pham, D. T. and Tran, L. T. (1985). Some mixing properties of time series models. *Stochastic Processes and their Applications*, **19**, 297-303.

Phillips, P.C.B. and Durlauf, S. (1986). Multiple time series regression with integrated processes. *Review of Economic Studies*, **53**, 473-496.

Pinkse, J., Shen, L. and Slade, M. E. (2007). Central limit theorem for endogenous locations and complex spatial interactions. *Journal of Econometrics*, **140**, 215-225.

Pinkse, J., Slade, M. E. and Brett, B. (2002). Spatial price competition: a semiparametric approach. *Econometrica*, **70**, 1111-1153.

Powell, J. L., Stock, J. H. and Stocker, T. M. (1989). Semiparametric estimation of the index coefficients. *Econometrica*, **57**, 1043-1430.

Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185-207.

Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

Robinson, P. M. (1991). Hypothesis testing in semiparametric and nonparametric models for econometric time series. *Review of Economic Studies*, **56**, 511-534.

Robinson, P.M. (1997). Large-sample inference for nonparametric regression with dependent errors. *Annals of Statistics*, **28**, 2054-2083.

Robinson, P.M. (2005). Efficiency improvements in inference on stationary and nonstationary fractional time series. *Annals of Statistics*, **33**, 1800-1842.

Robinson, P.M. (2010a). Efficient estimation of the semiparametric spatial autoregressive model. *Journal of Econometrics*, **157**, 6-17.

Robinson, P. M. (2010b). Nonparametric trending regression with cross-sectional dependence. *Journal of Econometrics*, forthcoming.

Robinson, P.M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, **165**, 5-19.

Robinson, P. M., and Hidalgo, F. J. (1997). Time series regression with long-range dependence. *Annals of Statistics*, **25**, 77-104.

Robinson, P.M. and Thawornkaiwong, S. (2010). Statistical inference on regression with spatial dependence. *Journal of Econometrics*, forthcoming.

Rosenblatt, M. (1971). Curve estimates. *Annals of Mathematical Statistics*, **42**, 1815-1842.

Rossi, F. (2010). Improved test statistics for pure spatial autoregressive models. *Preprint.*

Roussas, G. G. (1969). Nonparametric estimation in Markov processes. *Annals of Institute of Statistical Mathematics*, **21**, 73-88.

Ruckstuhl, A.F., Welsh, A. H. and Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica*, **10**, 51-71.

Scott, D.J. (1973). Central limit theorems for martingales using a Skorokhod representation approach. *Advances in Applied Probability*, **5**, 119-137.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley and Sons, Inc., New York.

Volkonskii, V. A. and Rozanov, Y. A. (1961). Some limit theorems for random functions II. *Theory of Probability and its Applications*, **6**, 186-198.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press., Cambridge, Massachusetts.

Yatchew, Y. (2003). *Semiparametric Regression for the Applied Econometrician.* Cambridge University Press.

Yatchew, Y. and No, J. A. (2003). Household gasoline demand in Canada. *Econometrica*, **69**, 1697-1710.

Yoshihara, K. (1976). Limiting behavior of U-statistics for stationary, absolutely regular processes. *Probability Theory and Related Fields.* **35**, 237-252.