Essays on Nonparametric Econometrics

Young Jun Lee

A DISSERTATION

Submitted to the Department of Economics

of

University College London

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

August 2019

I, Young Jun Lee, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This dissertation consists of three chapters that focus on the nonparametric method on time-varying parameter models and optimal transport problem.

The first chapter, which is jointly authored with Dennis Kristensen, develops a novel asymptotic theory for local polynomial (quasi-) maximum-likelihood estimators of time-varying parameters in a broad class of nonlinear time series models. Under weak regularity conditions, we show the proposed estimators are consistent and follow normal distributions in large samples. We demonstrate the usefulness of our general results by applying our theory to local (quasi-) maximum-likelihood estimators of a time-varying VAR's, ARCH and GARCH, and Poisson autogressions.

The second chapter proposes a sieve M-estimation of the solution to the optimal transport problem. Many problems in economics, including matching models and quantile methods, have the structure of an optimal transport problem. The sieve M-estimator is consistent under very little structure on the underlying optimal transport problem being solved. I then derive convergence rates for the estimator and its derivative when the surplus function $\Phi(X, Y) = X'Y$. The derived convergence rates are the same as the optimal rate in the context of regression and density estimations. The results can be extended to the conditional optimal transport problem having the conditional vector quantiles as an application.

In the third chapter, I consider the multidimensional matching as one of primary applications of the optimal transport problem. We employ the sieve simultaneous minimum distance estimation method to estimate the parameters in the equilibrium wage and assignment functions. Our estimation results show that worker-job complementarities in manual skills strongly decreased, where as complementarities in cognitive skills increased. This phenomenon is consistent to the one of Lindenlaub (2017).

Impact statement

This thesis provides a novel asymptotic theory for nonparametric estimators in two econometric models, time-varying parameter model and optimal transport problem.

For the time-varying parameter estimation, the asymptotic theory imposes very little structure on the chosen objective function used for estimation and on the underlying model being estimated. In particular, in contrast to the existing literature on kernel-based estimation of time-varying parameters, we impose substantially weaker smoothness and moment conditions on the likelihood and its derivatives. Our theory also contributes to the literature on asymptotic analysis of local polynomial estimators of varying-coefficient models by extending existing results to cover situations where the objective functions is non-concave. This is an important extension since the quasi-likelihoods of most non-linear models are non-concave, and the analysis of this case requires some new technical tools.

The optimal transport problem has been a very active research area in mathematics and more recently also in economics. Many problems in economics have the structure of an optimal transport problem. Despite of its many applications, the asymptotic theory for the optimal transport problem has not been fully established. In this thesis, the asymptotic theory for the sieve M-estimator impose very little structure on the chosen objective function used for estimation and on the underlying optimal transport problem being solved. In particular, in contrast to the existing literature on nonparametric estimation of the solution to the optimal transport problem, the conditions under which we derive our results are more easily verified. Moreover, the derived convergence rates for estimators are the same as the optimal rate in the context of regression and density estimations.

I have shared this research to academic audiences in the U.K. and Europe. I will further disseminate this research through scholarly publications.

Acknowledgements

I am indebted to my supervisor, Dennis Kristensen. His generosity with his time and his advice have made this thesis possible. I am also grateful to my second advisor, Daniel Wilhelm. His guidance and advice have greatly helped me during my PhD studies.

I also greatly benefited from interacting with many faculty members and fellow PhD students at the University College London. I am grateful for their insightful advice, comments, and suggestions.

I have also benefited from the great research environment at the University College London. As well as this, my PhD studies have been funded by scholarships from ESRC. I am grateful to both institutions for these opportunity.

My parents and sister have been an enormous source of encouragement. I owe them much of my achievements in this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Local Polynomial Estimation of Time-Varing Parameters in Nonlinear Models

1.1 Introduction

We provide a novel asymptotic theory for local polynomial estimators of time-varying parameters in a broad class of non-linear time series models. The theory imposes very little structure on the chosen objective function used for estimation and on the underlying model being estimated. In particular, in contrast to the existing literature on kernel-based estimation of time-varying parameters, we impose substantially weaker smoothness and moment conditions on the likelihood and its derivatives. For example, in the case of local linear estimators we do not require the existence of so-called derivative processes. And for the local constant version we only need the first-order derivative process to exist while the existing literature require higher-order derivatives to be well-defined.[1] Finally, again compared to existing theories, our results hold under much weaker restrictions on the bandwidth sequence used in the estimation thereby allowing for standard bandwidth selection procedures to be used. These features of our theory in turn imply that our asymptotic results take a simpler form and more closely resemble those found in the literature on local maximum

[1] For example, Theorem 3 inDahlhaus and Subba Rao (2006) requires the third order derivative process to obtain the asymptotic bias, which involves the second order derivative process.

likelihood estimation in a cross-sectional setting. Our theory also applies to GARCH-type models and for this class we show that additional biases appear due to the local polynomial approximation being less precise.

We demonstrate the aforementioned attractive features of our theory in two ways: First, we re-visit some specific models that have been analyzed elsewhere in the literature and show that our theory allows us to substantially weaken existing regularity conditions for the estimators to be well-behaved. Second, we apply our theory to models that fall outside the framework of existing theories. A simulation study investigates the finite-sample performance of the estimators and an empirical application shows the usefulness of the proposed methodology in practice.

To motivate and further discuss our results, consider the following class of models,

$$Y_{n,t} = G(X_{n,t}, \varepsilon_t; \theta_{n,t}), \quad \theta_{n,t} = \theta\left(t/n\right), \tag{1.1}$$

for $t = 1, 2, ..., n$ where $Y_{n,t}$ and $X_{n,t}$ are observed, ε_t is an unobserved error, and $\theta_{n,t} \in \mathbb{R}^{d_\theta}$ is sequence of a possibly time-varying parameters generated by an underlying function $\theta : [0,1] \mapsto \mathbb{R}^{d_\theta}$. Here, $X_{n,t}$ may contain lags of $Y_{n,t}$ and so the above class of models includes m-order Markov models. However, our theory goes beyond the above and also covers many other models such as generalized autoregressive models that include, for example, GARCH as special case. Assuming that $\theta(\cdot)$ is a smooth deterministic function, we develop and analyze nonparametric estimators of $\theta(u)$ for any given $u \in [0,1]$. Our proposed estimation method is based on the local maximum likelihood principle (see Tibshirani and Hastie, 1987; Fan et al., 1995): It takes as input a given (quasi-)likelihood function of the model in the stable case where $\theta_t = \theta$ is assumed constant. We then develop a kernel-weighted version of this objective function where $\theta(t/n)$ is approximated by a polynomial in t/n. Maximizing this w.r.t. the coefficients of the polynomial, we arrive at a local polynomial estimator of $\theta(u)$ and its derivatives.

We develop a novel asymptotic theory showing that the local polynomial estimators are pointwise (in time) consistent and asymptotically normally distributed. The proof strategy pursued here is different from the standard one found in the existing literature in that we rely on a alternative expansion of the score function in order to obtain expressions of the

leading bias and variance components. This allows us to obtain a simpler expression of the leading bias and variance terms under weaker regularity conditions compared to, e.g., Dahlhaus et al. (2017) and the references therein.

Our estimation method includes as special cases the local constant estimator and the local linear estimator. The local constant estimator suffers from asymptotic biases involving the so-called derivative process of the stationary approximation to data, but the local linear estimator does not. Moreover, the local linear estimator enjoys the well-known automatic boundary adjustment property meaning that at the beginning and end of the sample, this estimator will perform better than the local constant one.

Our general theory encompasses most existing results for nonparametric estimators of with time-varying parameters which are mainly for local constant estimators; see, e.g., Kristensen (2012), Robinson (1989), Dahlhaus and Subba Rao (2006) and Fryzlewicz et al. (2008), and in many cases lead to weaker conditions for existing results to hold. We demonstrate this feature by revisiting specific models analyzed in these papers and showing that their asymptotic results carry through under substantially weaker moment and parameter restrictions. Moreover, it allows us to analyze estimators of models that, as far as we can tell, cannot be handled by the existing theory, such as Poisson autoregressions with time-varying parameters. Our theory also contributes to the literature on asymptotic analysis of local polynomial estimators of varying-coefficient models by extending existing results (as in Fan et al., 1995; Loader, 2006) to cover situations where the objective functions is non-concave. This is an important extension since the quasi-likelihoods of most non-linear models are non-concave, and the analysis of this case requires some new technical tools.

The remainder of the chapter is organized as follows: Framework and estimators are introduced in Section 1.2. Section 1.3 presents the asymptotic theory of the estimators. In Section 1.4, we extend the theory to cover GARCH-type models. We then apply our general theory to particular models in Section 1.5. We present the results of two simulation studies and an empirical application in Sections 1.6 and 1.7, respectively. All lemmas and proofs have been relegated to the Appendix.

1.2 Framework

We are given n observations, $Z_{n,t}$, $t = 1, \ldots, n$, from a nonlinear time-series model with associated (quasi-) log-likelihood $\ell_{n,t}(\theta) = \ell_t(Z_{n,t}, Z_{n,t-1}, ...Z_{n,0}; \theta) \in \mathbb{R}$ where $\theta \in \Theta$. The quasi-likelihood is assumed to identify the data-generating parameters when these are in fact constant. That is, when $\theta_{n,t} = \theta_0$ is constant, the data-generating parameter value is the maximizer of $\theta \mapsto \lim_{n\to\infty} \sum_{t=1}^{n} \mathbb{E}[\ell_{n,t}(\theta)]/n$. A natural estimator in the time-invariant case would then be the M-estimator maximizing the sample analogue, $\sum_{t=1}^{n} \ell_{n,t}(\theta)$. The choice of $\ell_{n,t}(\theta)$ is, of course, model specific. For example, in a regression setting, we could choose $\ell_{n,t}(\theta)$ as the least squares criteria, while in (G)ARCH models it could be the Gaussian (quasi-)log-likelihood.

Now, returning to the case where $\theta_{n,t}$ is potentially varying over time, we then wish to estimate $\theta(u)$ for some given value $u \in [0,1]$. We propose to do this using local polynomial estimators where $\theta(t/n)$ is approximated by the following polynomial of order $m \geq 0$ for $t/n \approx u$,

$$\theta_u^*(t/n) := \beta_1 + \beta_2(t/n - u) + \cdots + \beta_{m+1}(t/n - u)^m/m! = D(t/n - u)\beta_0, \tag{1.2}$$

where $\beta_0 = \left(\beta_{0,1}', ..., \beta_{0,m+1}'\right)' \in \mathbb{R}^{(m+1)d_\theta}$ with $\beta_{0,i+1} = \theta^{(i)}(u) = \partial^i\theta(u)/\partial u^i \in \mathbb{R}^{d_\theta}$ and

$$D(u) = \left[1, u, u^2/2 \ldots, u^m/m!\right] \otimes I_{d_\theta} \in \mathbb{R}^{d_\theta \times (m+1)d_\theta}.$$

Next, to control the approximation error, $\theta(t/n) - \theta_u^*(t/n)$, we introduce a kernel weighted version of the global quasi-log-likelihood and substitute in the polynomial approximation,

$$L_n(\beta|u) = \frac{1}{n}\sum_{t=1}^{n} K_b(t/n - u)\,\ell_{n,t}(D(t/n - u)\beta),$$

where $K_b(\cdot) = K(\cdot/b)/b$, $K : \mathbb{R} \mapsto \mathbb{R}$ is a kernel function, and $b = b_n > 0$ a bandwidth. We then estimate the polynomial coefficients by

$$\hat{\beta} = \arg\max_{\beta\in\mathcal{B}} L_n(\beta|u),$$

where $\mathcal{B} \subseteq \mathbb{R}^{(m+1)d_\theta}$ will be specified below, so that $\hat{\theta}(u) = \hat{\beta}_0$ and $\hat{\theta}^{(i)}(u) = \hat{\beta}_{i+1}$, $i = 0, ..., m$. When $m = 0$, we recover the standard local-constant estimator.

Special care has to be taken with the implementation of local polynomial estimators when the chosen objective function is not well-defined for all value of θ and/or Θ is compact. A simple example is ARCH models where parameters have to remain positive for the volatility process to be well-defined. In such cases, we have to ensure that $D(t/n - u)\beta$ satisfies these constraints for $t = 1, ..., n$. To this end, it proves useful to introduce rescaled versions of $\hat{\beta}$ using the following weighting matrix,

$$U_n = diag\{1, b, ..., b^m\} \otimes I_{d_\theta} \in \mathbb{R}^{(m+1)d_\theta \times (m+1)d_\theta}.$$

We then define $\hat{\alpha} = U_n\hat{\beta} = \left(\hat{\theta}(u)', b\hat{\theta}^{(1)}(u)', ..., b^m\hat{\theta}^{(m)}(u)'\right)'$ which satisfies

$$\hat{\alpha} = \arg\max_{\alpha \in \mathcal{A}} Q_n(\alpha|u), \quad Q_n(\alpha|u) = \frac{1}{n}\sum_{t=1}^{n} K_b(t/n - u)\,\ell_{n,t}(D_b(t/n - u)\,\alpha),$$

where $D_b(u) = D(u/b)$. The asymptotic analysis proves to be much simpler to carry out in terms of $\hat{\alpha}$ since each component of $\hat{\alpha}$ has the same convergence rate as we shall see in the following section. U_n contains the relative rates of convergence of $\hat{\beta}$. Importantly, $D_b(t/n - u)$ and $K_b(t/n - u)$ depend on the same argument which facilitates the derivation of precise restrictions on the parameter space $\mathcal{A}$ so that $Q_n(\alpha|u)$ is well-defined for all $\alpha \in \mathcal{A}$. The corresponding parameter space for β then takes the form $\mathcal{B}_n = \{\beta = U_n^{-1}\alpha : \alpha \in \mathcal{A}\}$ which expands as $b \to 0$.

1.3 Asymptotic theory

To establish an asymptotic theory for the proposed class of local polynomial estimators, we will rely on the concept of local stationarity as introduced by Dahlhaus (1997); see also Dahlhaus and Subba Rao (2006) and Dahlhaus et al. (2017). We first generalize this concept to sequences of random functions:

Definition 1.1. A triangular family of random sequences $W_{n,t}(\theta)$, $\theta \in \Theta$, $t = 1, 2, ..., n$ and $n \geq 1$, is uniformly locally stationary on Θ (ULS(p, q, Θ)) for some $p, q > 0$ if there exists a

family of processes $W_t^*(\theta|u)$, $u \in [0,1]$, such that: (i) The process $\{W_t^*(\theta|u)\}$ is stationary and ergodic for all $(\theta, u) \in \Theta \times [0,1]$; (ii) for some $C < \infty$ and $\rho < 1$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta} \|W_{n,t}(\theta) - W_t^*(\theta|u)\|^p\right]^{1/p} \leq C\left(\left|\frac{t}{n} - u\right|^q + \frac{1}{n^q} + \rho^t\right).$$

Compared to existing definitions of local stationarity, we allow for an additional term ρ^t to appear in the approximation error. This is needed in order to allow for the initial value of the (non-stationary) data-generating process to be arbitrary. In contrast, most of the existing literature implicitly assumes that the data-generating process has been initialized at $Z_{n,0} = Z_0^*(u)$ where $Z_0^*(u)$ is its stationary approximation. This has as consequence that the data-generating process changes as the researcher varies u in the local log-likelihood which is a rather peculiar assumption. Moreover, in the estimation of GARCH-type models, the conditional variance process entering the likelihood is normally initialized at a fixed value and so again an additional error term will appear when comparing this with its stationary version. The above definition again allows for this feature. To see how the additional error is generated in Markov models, we refer the reader to Theorem 1.7 in Appendix 1.8.4 which allow for an arbitrary initialization of the data-generating process. The additional error term due to different initialization is here assumed to decay geometrically and so our definition rules out long-memory type processes. This is mostly for simplicity and we expect that most of our results can be generalized to allow for slower decay rates. Appendix 1.8.1 contain a number of novel results for kernel weighted averages of parameter-dependent locally stationary processes which will be used in the following analysis of our polynomial estimators.

We will then require that $\ell_{n,t}(\theta)$ is ULS(p, q, Θ) with stationary approximation $\ell_t^*(\theta|u)$. To illustrate, consider (1.1): The stationary approximation will here take the form $\ell_t^*(\theta|u) = \ell(Z_t^*(u), \theta)$ where $Z_t^*(u) = (Y_t^*(u), X_t^*(u))$ is the stationary solution to the model when $\theta_t = \theta(u)$ is constant,

$$Y_t^*(u) = G(X_t^*(u), \varepsilon_t; \theta(u)), \quad t = 1, 2, \ldots$$

If the data-generating process is locally stationary, it follows under great generality that the

likelihood and its derivatives are also locally stationary as shown in the following theorem:

Theorem 1.1. *Suppose that $Z_{n,t}(\theta)$ is ULS(p, q, Θ) with stationary approximation satisfying $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_t^*(\theta|u)\right\|^p\right] < \infty$; (ii) ε_t is i.i.d. and independent of $(Z_{n,t}(\theta), Z_t^*(\theta|u))$; and (iii) for some $r > 0$, $\mathbb{E}\left[\left\|f(z, \varepsilon_t; \theta) - f(z', \varepsilon_t; \theta)\right\|\right] \leq C\left(1 + \|z\|^r + \|z'\|^r\right)\|z - z'\|$ for all $\theta \in \Theta$ and $z, z' \in \mathcal{Z}$. Then $f(Z_{n,t}, \varepsilon_t; \theta)$ is ULS$(p/(r+1), q, \Theta)$.*

This result generalizes Proposition 2.5 in Dahlhaus et al. (2017) in two directions: First, it allows for $Z_{n,t}(\theta)$ to be parameter dependent and second it allows for an i.i.d. component, ε_t, to enter the transformation. Allowing for parameter dependence means we can apply the above result to GARCH-type models, among others. The reason why we allow for the presence of the additional component ε_t is best illustrated by again considering (1.1): In this model, we can rewrite $Z_{n,t}$ and thereby the likelihood $\ell(Z_{n,t}; \theta)$ as a function of $X_{n,t}$ and the error term ε_t. Doing so allows for easier verification of local stationarity of the likelihood and its derivatives; see Section 1.5 for examples of this.

Under ULS, the nonstationary local likelihood function and its derivatives are well-approximated by their stationary versions. For example, $\sup_{\alpha\in A}\left|Q_n^*(\alpha|u) - Q_n^*(\alpha|u)\right| = o_p(1)$ where $Q_n^*(\alpha|u) = \frac{1}{n}\sum_{t=1}^n K_b(t/n - u)\,\ell_t^*(D_b(t/n - u)\,\alpha|u)$. The next step is then to develop a uniform Law of Large Numbers (ULLN) for $Q_n^*(\alpha|u)$. Furthermore, in order to analyze the bias properties of the local constant version, we need to be able to expand the stationary version of the score function $s_t^*(\theta|u) = \partial\ell_t^*(\theta|u)/(\partial\theta)$ w.r.t. u. To this end, we introduce the following additional concepts:

Definition 1.2. A stationary process $W_t^*(\theta|u)$ is said to be *L_p-continuous* w.r.t. θ if the following holds for all $\theta \in \Theta$: $\mathbb{E}\left[\|W_t^*(\theta|u)\|^p\right] < \infty$ and

$$\forall\epsilon > 0 \exists\delta > 0 : \mathbb{E}\left[\sup_{\theta':\|\theta-\theta'\|<\delta}\left\|W_t^*\left(\theta'|u\right) - W_t^*(\theta|u)\right\|^p\right]^{1/p} < \epsilon.$$

The process is said to be *L_p-differentiable* w.r.t. u if there exists a stationary and ergodic process $\partial_u W_t^*(\theta|u)$ with $\mathbb{E}\left[\|\partial_u W_t^*(\theta|u)\|^p\right] < \infty$ such that

$$\mathbb{E}\left[\|W_t^*(\theta|u+\Delta) - W_t^*(\theta|u) - \partial_u W_t^*(\theta|u)\,\Delta\|^p\right]^{1/p} = o(\Delta),\ \Delta \to 0.$$

Our definition of time differentiability is slightly different from the one found in Dahlhaus et al. (2017) and other papers where differentiability w.r.t. u has to hold almost surely; our version is slightly weaker since we only require it to hold in the L_p-norm. The definition of L_p-continuity w.r.t. θ is also weaker than almost sure continuity: If $\theta \mapsto W_t^*(\theta|u)$ is almost surely continuous with $\mathbb{E}\left[\sup_{\theta\in\Theta} \|W_t^*(\theta|u)\|^p\right] < \infty$ the process is also L_p-continuous since $D_t(\delta) = \sup_{\|\theta-\theta'\|\leq\delta} \|W_t^*(\theta|u) - W_t^*(\theta'|u)\|^p$, $\delta > 0$, will then satisfy $\lim_{\delta\to 0} D_t(\delta) = 0$ almost surely and so, by dominated convergence, $\lim_{\delta\to 0} \mathbb{E}[D_t(\delta)] = 0$. It is easily verified that L_p-continuity w.r.t. θ implies stochastic equicontinuity of $Q_n^*(\alpha|u)$ and so a ULLN holds, c.f. Lemma 1.1(i) in Appendix 1.8.1.

We are now ready to state the regularity conditions under which our estimators are consistent:

Assumption 1.1. *(i) $K(\cdot) \geq 0$ has compact support $\mathcal{K}$ and $\int K(v)\,dv = 1$ (ii) K is symmetric around 0; (iii) for some $\Lambda < \infty$, $|K(v) - K(v')| \leq \Lambda |v - v'|$, $v, v' \in \mathbb{R}$.*

Assumption 1.2. *The parameter space $\mathcal{A} = \left\{\alpha \in \mathbb{R}^{(m+1)d_\theta} : D(v)\,\alpha \in \Theta, \forall v \in \mathcal{K}\right\}$ where Θ is compact. The true value $\theta(u) \in \Theta$.*

Assumption 1.3. *(i) $\ell_{n,t}(\theta)$ is ULS(p, q, Θ) for some $p \geq 1$ and $q > 0$ with stationary approximation $\ell_t^*(\theta|u)$; (ii) $\theta \mapsto \ell_t^*(\theta|u)$ is L_1-continuous; (iii) $\theta \mapsto \mathbb{E}[\ell_t^*(\theta|u)]$ has a unique maximum at $\theta(u) \in \Theta$.*

Assumption 1.1(i) imposes stronger than usual assumptions on K and excludes, among others, the Gaussian kernel and higher-order kernels. It includes, on the other hand the Epanechnikov and the triangular kernel. The restriction that $K(\cdot) \geq 0$ is used to ensure identification of the parameters when $m > 0$; without this, identification is not necessarily guaranteed; see below for further discussion. The compact support assumption appears to be quite important for the analysis of local polynomial estimation of non-concave models: In order to establish uniform convergence of the likelihood we need Θ to be compact as is standard in the literature. But under this restriction, it is easily checked that $D_b(v)\,\alpha \notin \Theta$ as $b \to 0$ for any given $\alpha = (\alpha_1, ..., \alpha_{m+1})$ with $\alpha_i \neq 0$ for some $i \geq 1$ and any $v \neq 0$. Thus, to allow for kernels with unbounded support, we would generally need the parameter space $\mathcal{A}$ to collapse at $\{(\alpha_1, 0, ..., 0) : \alpha_1 \in \Theta\}$ as $b \to 0$. Such shrinking behaviour in turn

means that a Taylor expansion of $\ell_{n,t}(D_b(v)\alpha)$ w.r.t. α is not possible and so standard arguments to establish asymptotic normality of $\hat{\alpha}$ cannot be applied. On the other hand, by restricting the support $\mathcal{K}$ to be compact, it is easily checked that with $\mathcal{A}$ defined in Assumption 1.2, $K_b(v)\ell_{n,t}(D_b(v)\alpha)$ is well-defined for all $\alpha \in \mathcal{A}$ and $v \in \mathbb{R}$ (where we set $K_b(v)\ell_{n,t}(D_b(v)\alpha) = 0$ for $v/b \notin \mathcal{K}$). Moreover, $(\alpha_1, 0, ..., 0)$ is an interior point of $\mathcal{A}$ and so in our analysis of $\hat{\alpha}$ we can employ standard arguments involving a Taylor expansion of the score function around this point. Thus, it appears as if the compact support assumption is needed for standard asymptotic arguments to apply. One could replace the definition of $\mathcal{A}$ with

$$\mathcal{A}_n(u) = \left\{\alpha \in \mathbb{R}^{(m+1)d_\theta} : D_b(v-u)\alpha \in \Theta, \forall v \in \{v \in [0,1] : K_b(v-u) > 0\}\right\}.$$

This allows for a larger parameter space in finite samples. However, $\mathcal{A}_n(u) \to \mathcal{A}$ as $b \to 0$, and so we maintain the above definition of $\mathcal{A}$ for simplicity.

Assumption 1.3(ii)-(iii) are standard in the analysis of "global" extremum estimators of stationary models on the form $\tilde{\theta}(u) = \arg\max_{\theta\in\Theta}\sum_{t=1}^{n}\ell_t^*(\theta|u)$. In particular, for a given time series model, we can import existing results for verification of Assumption 1.3(ii)-(iii); see Section 1.5 for more details. 1.3(iii) in conjunction with the assumption that $K(\cdot) \geq 0$ ensures that the local polynomial estimator identifies $\theta(u)$. If we allow for kernels that take negative values, we have to replace 1.3(iii) with the following more abstract identification condition: The function $Q^*(\alpha|u) = \int K(v)\mathbb{E}[\ell_t^*(D(v)\alpha|u)]dv$ satisfies $Q^*(\alpha|u) < Q^*((\theta(u), 0, ..., 0)|u)$ for any $\alpha \neq (\theta(u), 0, ..., 0)$. We have not been able to provide primitive conditions for this to hold when K can take negative values and so instead impose the positivity constraint on K.

If the objective function $\theta \mapsto \ell_{n,t}(\theta)$ is concave and Θ is concave, we can replace Assumption 1.3(i)-(ii) with the following pointwise versions: For any $\theta \in \Theta$, $\ell_{n,t}(\theta)$ is locally stationary and $\mathbb{E}[|\ell_t^*(\theta)|] < \infty$; see Theorem 2.7 in Newey and McFadden (1994). Under the above assumptions, the following consistency result holds:

Theorem 1.2. *Let Assumptions 1.1-1.3 hold. Then, as $b \to 0$ and $nb \to \infty$, $\hat{\alpha} \to^p (\theta(u), 0, ..., 0)'$. In particular, $\hat{\theta}(u) \to^p \theta(u)$.*

Note that the above theorem only shows consistency of $\hat{\theta}(u)$ and so at this stage we cannot make any statements regarding $\hat{\theta}^{(i)}(u)$, $i = 1, ..., m$. This is similar to other results for nonlinear extremum estimators where parameters associated with components appearing in the objective function that grow (shrink) with a slower (faster) rate than the leading one will not be identified; see, e.g., Theorem 9 in Han and Kristensen (2014) where a global consistency result is only provided for the component with the fastest rate.

However, with some further regularity conditions on the quasi-likelihood function, we can provide a more precise analysis of the estimators. With $s_{n,t}(\theta) = \partial \ell_{n,t}(\theta) / (\partial \theta) \in \mathbb{R}^{d_\theta}$ and $h_{n,t}(\theta) = \partial^2 \ell_{n,t}(\theta) / \left(\partial \theta \partial \theta'\right) \in \mathbb{R}^{d_\theta \times d_\theta}$, $D_{n,t}(u) = D_b(t/n - u)$ and $K_{n,t}(u) = K_b(t/n - u)$, we introduce the score and hessian,

$$S_n(\alpha|u) = \frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' s_{n,t}(D_{n,t}(u)\alpha) \in \mathbb{R}^{(m+1)d_\theta},$$
$$H_n(\alpha|u) = \frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' h_{n,t}(D_{n,t}(u)\alpha) D_{n,t}(u) \in \mathbb{R}^{(m+1)d_\theta \times (m+1)d_\theta}.$$

It is easily checked that $\alpha_0 := U_n \beta_0$ belongs to the interior of $\mathcal{A}$ for all n large enough due to Assumption 1.4(ii) in conjunction with Assumption 1.2 and, due to the consistency result, so will $\hat{\alpha}$ w.p.a.1. Thus, $\hat{\alpha}$ will satisfy the first-order condition which combined with the mean-value theorem yield

$$0 = S_n(\hat{\alpha}|u) = S_n(\alpha_0|u) + H_n(\bar{\alpha}|u)(\hat{\alpha} - \alpha_0), \tag{1.3}$$

where $\bar{\alpha}$ is situated on the line segment connecting $\hat{\alpha}$ and α_0. We then decompose the score function into the bias and variance component, $S_n(\alpha_0|u) = B_n(u) + S_n(u)$, where

$$B_n(u) = \frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' b_{n,t},\ S_n(u) = \frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' s_{n,t}(\theta(t/n)), \tag{1.4}$$

and $b_{n,t} = s_{n,t}(\theta_u^*(t/n)) - s_{n,t}(\theta(t/n))$ with $\theta_u^*(t/n)$ defined in eq. (1.2). This decomposition is different from the one usual employed in the analysis of kernel estimators of time-varying coefficients where $s_{n,t}(\theta(t/n))$ is replaced by the stationary version of the score function evaluated at $\theta(u)$, $s_t^*(\theta(u)|u)$; see, e.g., Dahlhaus et al. (2017) and Dahlhaus and Subba Rao (2006). This choice has as consequence that the corresponding bias term in

their case generally involves the time derivative process of the score function and so their analysis tend to impose stronger regularity conditions. By instead centering the analysis around $s_{n,t}(\theta(t/n))$, our version of the first-order bias component can be obtained through a standard Taylor expansion w.r.t. θ,

$$b_{n,t} \cong h_{n,t}(\theta_u^*(t/n))\{\theta_u^*(t/n) - \theta(t/n)\} \cong -h_{n,t}(\theta(u))\frac{\theta^{(m+1)}(u)}{(m+1)!}\{t/n-u\}^{m+1}. \quad (1.5)$$

Thus, our approach allows for a simpler derivation of the leading bias and variance terms under the following weak regularity conditions:

Assumption 1.4. *(i) $\ell_{n,t}(\theta)$ is twice continuously differentiable; and (ii) $\theta(u)$ lies in the interior of Θ and is $m+1$ times continuously differentiable.*

Assumption 1.5. *(i) $s_{n,t}(\theta(t/n))$ is a martingale difference (MGD) array w.r.t. $\mathcal{F}_{n,t} = \mathcal{F}\{Z_{n,t}, Z_{n,t-1}, \ldots\}$; (ii) $\omega_{n,t}(\theta) = s_{n,t}(\theta)s_{n,t}(\theta)' \in \mathbb{R}^{d_\theta \times d_\theta}$ is ULS$(p, q, \{\theta : \|\theta - \theta(u)\| < \epsilon\})$ for some $p \geq 1$ and $q, \epsilon > 0$ with $\omega_t^*(\theta|u)$ being L_1-continuous at $\theta = \theta(u)$.*

Assumption 1.6. *$h_{n,t}(\theta)$ is ULS$(p, q, \{\theta : \|\theta - \theta(u)\| < \epsilon\})$ for some $p \geq 1$ and $q, \epsilon > 0$ with L_1-continuous stationary approximation $h_t^*(\theta|u)$ and $H(u) \equiv \mathbb{E}[h_t^*(\theta(u)|u)]$ is non-singular.*

Assumption 1.5 is non-standard compared to the existing literature (as discussed above) and allows us to apply a martingale central limit theorem for locally stationary sequences (see Lemma 1.1(iii) in Appendix 1.8.1) to $S_n(u)$. The MGD assumption amounts to assuming that the time-varying model is correctly specified and has to be verified on a case-by-case basis. Finally, Assumption 1.6 together with the expansion in eq. (1.5) is used to derive the limits of $B_n(u)$ and $H_n(\bar{\alpha}|u)$,

$$\sqrt{nb}S_n(u) \to^d N(0, \mathbb{K}_2 \otimes \Omega(u)), \ \Omega(u) = \mathbb{E}[\omega_t^*(\theta(u)|u)], \quad (1.6)$$

$$(i)\, H_n(\bar{\alpha}|u) \to^p \mathbb{K}_1 \otimes H(u), \ (ii)\, B_n(u) = b^{m+1}\left(\mu_1 \otimes H(u)\frac{\theta^{(m+1)}(u)}{(m+1)!} + o_P(1)\right), \quad (1.7)$$

where $\mu_i = \int K(v)v^{m+i}D(v)dv$ and $\mathbb{K}_i = \int K^i(v)D(v)D(v)'dv$, $i \geq 1$. Combining these limit results, we obtain:

Theorem 1.3. *Suppose that Assumptions 1.1-1.6 hold. Then, as* $b \to 0$ *and* $nb \to \infty$,

$$\sqrt{nb}U_n \left\{ \hat{\beta} - \beta_0 - R_n \left(Bias\left(u\right) + o_P\left(1\right) \right) \right\} \to^d N\left(0, \mathbb{K}_1^{-1}\mathbb{K}_2\mathbb{K}_1^{-1} \otimes H\left(u\right)^{-1} \Omega\left(u\right) H\left(u\right)^{-1}\right),$$

where $R_n = diag\left\{b^{m+1}, b^m, ..., b\right\} \otimes I_{d_\theta}$ *and* $Bias\left(u\right) = \mathbb{K}_1^{-1}\mu_1 \otimes \frac{\theta^{(m+1)}(u)}{(m+1)!}$. *In particular, for* $i = 0, 1, ..., m$,

$$\begin{aligned} &\sqrt{nb^{2i+1}} \left\{ \hat{\theta}^{(i)}\left(u\right) - \theta^{(i)}\left(u\right) - b^{m+1-i}\left(Bias_i\left(u\right) + o_P\left(1\right) \right) \right\} \\ &\to^d N\left(0, \kappa_{2,i} H\left(u\right)^{-1} \Omega\left(u\right) H\left(u\right)^{-1}\right), \end{aligned} \tag{1.8}$$

where $Bias_i\left(u\right) = \kappa_{1,i}\frac{\theta^{(m+1)}(u)}{(m+1)!} + o_P\left(1\right)$ while $\kappa_{1,i}$ and $\kappa_{2,i}$ denotes the ith element of $\mathbb{K}_1^{-1}\mu_1$ and (i,i)th element of $\mathbb{K}_1^{-1}\mathbb{K}_2\mathbb{K}_1^{-1}$, respectively.

Similar to existing results for local polynomial estimators in a cross-sectional setting, the bias component only depends on $\theta^{(m+1)}\left(u\right)$ and so the estimators adapt to the curvature of $\theta\left(u\right)$. The asymptotic variance in Theorem 1.3 can be estimated using plug-in methods: It follows from the proof of Theorem 1.3 that

$$\hat{V}\left(u\right) = \frac{1}{n}\sum_{t=1}^{n} K_{n,t}^2\left(u\right) D_{n,t}\left(u\right)' s_{n,t}\left(D_{n,t}\left(u\right)\hat{\alpha}\right) s_{n,t}\left(D_{n,t}\left(u\right)\hat{\alpha}\right)' D_{n,t}\left(u\right),$$

satisfies $\hat{V}\left(u\right) \to^p \mathbb{K}_2 \otimes \Omega\left(u\right)$ while $H_n\left(\hat{\alpha}|u\right) \to^p \mathbb{K}_1 \otimes H\left(u\right)$.

Comparing the above limit results and the conditions under which they are derived with the corresponding ones found in Dahlhaus et al. (2017) and the references therein, we note that our bandwidth restrictions are much weaker than theirs. In particular, standard bandwidth selection rules can be employed here but not in their set-up. Moreover, the existing literature requires time derivatives of the stationary score function to exist and be well-behaved with these entering the bias expressions. We on the other hand are able to obtain results that are analogous to the ones found in the literature on local polynomial likelihood estimators; see, e.g., Theorem 1b of Fan et al. (1995).

Equation (1.8) holds for any value of $m \geq 0$ and $i = 0, ..., m$. However, when $m - i$ is even, $\kappa_{1,i} = 0$ since all odd moments of K are zero due to the symmetry assumption. For example, for the local constant estimator ($m = i = 0$), Theorem 1.3 only informs us that

the bias component of $\hat{\theta}(u)$ is $o_p(b)$. To obtain the leading bias term in this case, a higher-order expansion in eq. (1.4) is necessary. This expansion requires additional assumptions involving time derivatives and standard derivatives w.r.t. θ of $h_t^*(\theta(u)|u)$:

Assumption 1.7. *$h_t^*(\theta|u)$ is time-differentiable in the L_1-sense at $(\theta(u), u)$ with time-derivative $\partial_u h_t^*(\theta(u)|u) \in \mathbb{R}^{d_\theta \times d_\theta}$.*

Assumption 1.8. *For $i = 1, ..., d_\theta$, $\partial h_{n,t}(\theta)/\partial\theta_i$ exists and is ULS$(p, q, \{\theta : \|\theta - \theta(u)\| < \epsilon\})$ for some $p \geq 1$ and $q, \epsilon > 0$ with L_1-continuous stationary approximation $\partial h_t^*(\theta|u)/\partial\theta_i$.*

Assumption 1.9. $\sum_{s=1}^{\infty} \left| Cov\left(h_{ij,t}^*(\theta(u)|u), h_{ij,t+s}^*(\theta(u)|u) \right) \right| < \infty$, $i, j = 1, ..., d_\theta$.

The time-derivative $\partial_u h_t^*(\theta(u)|u)$ will generally involve time-derivatives of the underlying stationary approximation of data. For example, if $h_{n,t}(\theta) = h(Z_{n,t}(\theta); \theta)$ where the right-hand side is differentiable w.r.t. $Z_{n,t}(\theta) \in \mathbb{R}^{d_Z}$, then it takes the form

$$\partial_u h_t^*(\theta|u) = \sum_{i=1}^{d_Z} \frac{\partial h(Z_t^*(\theta|u); \theta)}{\partial z_i} \partial_u Z_{i,t}^*(\theta|u),$$

where $\partial_u Z_{i,t}^*(\theta|u)$ is the time derivative of $Z_t^*(\theta|u)$. Assuming in addition that $\theta(u)$ is $m+2$ times continuously differentiable, the following asymptotic expansion of $b_{n,t}$ under Assumptions 1.7-1.8 holds:

$$\begin{aligned} b_{n,t} \cong & -h_t^*(\theta(u)|u) \left\{ \frac{\theta^{(m+1)}(u)}{(m+2)!}(t/n-u)^{m+1} + \frac{\theta^{(m+2)}(u)}{(m+2)!}(t/n-u)^{m+2} \right\} \\ & - \partial_u h_t^*(\theta(u)|u) \frac{\theta^{(m+1)}(u)}{(m+1)!}(t/n-u)^{m+2} \\ & - \sum_{i=1}^{d_\theta} \theta_i^{(1)}(u) \partial_{\theta_i} h_t^*(\theta(u)|u) \frac{\theta^{(m+1)}(u)}{(m+1)!}(t/n-u)^{m+2} \\ & + \frac{\{t/n-u\}^{2m+2}}{2\{(m+1)!\}^2} \sum_{i=1}^{d_\theta} \theta_i^{(m+1)}(u) \partial_{\theta_i} h_t^*(\theta(u)|u) \theta^{(m+1)}(u) \end{aligned} \tag{1.9}$$

The short memory condition imposed in Assumption 1.9 is used to control the variance component of the first-order bias term derived in Theorem 1.3. A sufficient condition for this assumption to hold is that $h_t^*(\theta(u)|u)$ is a geometric moment contraction, c.f. Proposition 2 in Wu and Shao (2004). We then obtain the following higher-order expansion of the bias component to be used when $m - i$ is even:

Theorem 1.4. *Suppose Assumptions 1.1-1.9 hold and $\theta(\cdot)$ is $m+2$ times continuously differentiable. Then, as $b \to 0$ and $nb \to \infty$,*

$$\begin{aligned} B_n(u) =& b^{m+2}\left[Bias_1(u) + o_p(1)\right] + b^{2m+2}\left[Bias_2(u) + o_p(1)\right] \\ &+ b^{m+1}\left[\mu_1 H(u)\frac{\theta^{(m+1)}(u)}{(m+1)!} + O_P\left(1/n^q\right) + O_p\left(\frac{1}{\sqrt{nb}}\right)\right], \end{aligned} \tag{1.10}$$

where, with $\partial_u H(u) = \mathbb{E}\left[\partial_u h_t^*\left(\theta(u)|u\right)\right]$ and $\partial_{\theta_i} H(u) = \mathbb{E}\left[\partial h_t^*\left(\theta(u)|u\right)/\partial\theta_i\right]$,

$$Bias_1(u) = \mu_2 H(u)\frac{\theta^{(m+2)}(u)}{(m+2)!} + \mu_2 \partial_u H(u)\frac{\theta^{(m+1)}(u)}{(m+1)!} + \mu_2 \sum_{i=1}^{d_\theta} \theta_i^{(1)}(u)\,\partial_{\theta_i} H(u)\frac{\theta^{(m+1)}(u)}{(m+1)!},$$

$$Bias_2(u) = -\frac{\mu_{m+2}}{2\left\{(m+1)!\right\}^2}\sum_{i=1}^{d_\theta}\theta_i^{(m+1)}(u)\,\partial_{\theta_i} H(u)\,\theta^{(m+1)}(u).$$

Corollary 1.1. *The local constant estimator ($m = 0$) satisfies, as $b \to 0$, $nb^3 \to \infty$ and $n^q b \to \infty$,*

$$\sqrt{nb}\left\{\hat{\theta}(u) - \theta(u) - b^2\left\{H^{-1}(u)\,Bias_0(u) + o_p(1)\right\}\right\} \to^d N\left(0, \kappa_{2,0} H^{-1}(u)\,\Omega(u)\,H^{-1}(u)\right), \tag{1.11}$$

where $\kappa_{2,0} = \int K^2(v)\,dv$ and, with $\kappa_{1,0} = \int K(v)\,v^2 dv$,

$$Bias_0(u) = \kappa_{1,0}\left\{H(u)\frac{\theta^{(2)}(u)}{2} + \partial_u H(u)\,\theta^{(1)}(u) + \frac{1}{2}\sum_{i=1}^{d_\theta}\theta_i^{(1)}(u)\,\partial_{\theta_i} H(u)\,\theta^{(1)}(u)\right\}.$$

To our knowledge this is the first complete characterization of the bias components of local constant estimators in general time-varying parameter models. Compared to existing results for specific models (see, e.g., Dahlhaus and Subba Rao, 2006) , we see that our bias expression takes a different form. In particular, ours only involves the first-order time derivative process, $\partial_u h_t^*\left(\theta(u)|u\right)$, while existing results involve higher-order derivatives. This is due to the aforementioned different proof techniques. One can show that our and theirs bias expressions are equivalent under their stronger regularity conditions. Comparing Theorems 1.3 and 1.4, we see that the local linear and local constant estimators share the same convergence rate and asymptotic variance, but that the local constant estimator suffers from additional biases. This is consistent with the theory found for local constant

and local linear estimators in a cross-sectional setting. However, compared with the theory in a cross-sectional setting (as in Fan et al., 1995), our bias takes a slightly different form. This is due to the fact that the data-generating process in our setting is non-stationary with the stationary approximation generating additional biases. Similar to the results found in a cross-sectional regression context, c.f. Fan (1993), we expect the additional biases of the local constant estimator to translate into reduced precision and efficiency compared to the local linear one.

Moreover, as is well-known, local polynomial estimators have the advantage of exhibiting automatic boundary carpentering. This property also holds in our setting near the end points of the sample ($u = 0$ and $u = 1$). Formally, we analyze the properties of the estimators at $u = cb$ and $u = 1 - cb$, respectively, for some $c > 0$. The following corollary reports the properties for the first case, a similar result holds for the latter one. We leave out the proof since it follows along the same arguments as Theorems 1.3 and 1.4, except that the asymptotic bias and variance terms take a slightly different form.

Corollary 1.2. *Let $\hat{\theta}_m(u)$ be the local polynomial estimator of order $m \in \{0,1\}$. Under the same conditions as in Theorem 1.4, with $\kappa_{i,j}^c = \int_{-c} K^i(v) v^j dv$,*

$$\begin{aligned}&\sqrt{nb}\left\{\hat{\theta}_m(cb) - \theta(cb) - b^{1+m}\left(\kappa_{1,m} Bias_m\left(0^+\right) + o_p(1)\right)\right\}\\ &\to^d N\left(0, a_m H^{-1}\left(0^+\right)\Omega\left(0^+\right)H^{-1}\left(0^+\right)\right),\end{aligned}$$

where $\Omega(0^+) = \lim_{u\downarrow 0}\Omega(u)$, $H(0^+) = \lim_{u\downarrow 0} H(u)$, and $Bias_m(0^+) = B_m\theta^{(m+1)}(0^+)$ with

$$\begin{aligned}&B_0 = \kappa_{1,1}^c/\kappa_{1,0}^c,\ B_1 = \frac{1}{2}\left[\left(\kappa_{1,2}^c\right)^2 - \kappa_{1,1}^c\kappa_{1,3}^c\right]/\left[\kappa_{1,0}^c\kappa_{1,2}^c - \left(\kappa_{1,1}^c\right)^2\right],\\ &a_0 = \kappa_{2,0}^c/\left(\kappa_{1,0}^c\right)^2,\ a_1 = \left[\left(\kappa_{1,2}^c\right)^2\kappa_{2,0}^c - 2\kappa_{1,1}^c\kappa_{1,2}^c\kappa_{2,1}^c + \left(\kappa_{1,1}^c\right)^2\kappa_{2,2}^c\right]/\left[\kappa_{1,0}^c\kappa_{1,2}^c - \left(\kappa_{1,1}^c\right)^2\right]^2.\end{aligned}$$

This corollary shows that the asymptotic biases and variances for the local constant and linear estimators at the boundaries are different. While the difference between two asymptotic variances is only a scale, the bias of the local constant estimators vanishes at a slower rate than the local linear one.

1.4 Extension to time-varying generalized autoregressive models

The theory developed in Section 1.3 requires $s_{n,t}(\theta(t/n))$ to be a Martingale difference. This assumption is violated in time-varying GARCH-type models as we shall see. We here demonstrate how our proof strategy can be generalized to cover the following class of generalized autoregressive models (GAR's),

$$Y_{n,t} = G_Y(\lambda_{n,t}, \varepsilon_t), \quad \lambda_{n,t} = G_\lambda(Y_{n,t-1}, \lambda_{n,t-1}, \theta(t/n)).$$

This class includes GARCH and Poisson Autogressions, amongst others. Since $\lambda_{n,t}$ is not directly observed, the likelihood takes the form

$$\ell_{n,t}(\theta) = \ell(Y_{n,t}, \lambda_{n,t}(\theta)), \quad \lambda_{n,t}(\theta) = G_\lambda(Y_{n,t-1}, \lambda_{n,t-1}(\theta), \theta),$$

where $\lambda_{n,t}(\theta)$ is initialized at $\lambda_{n,0}(\theta) = \lambda_0$ for some fixed λ_0 and $\ell(\cdot)$ depends on the functional form of G_Y and the assumed distribution of ε_t.

We will here only provide a theory for local constant estimators since the analysis of local polynomial estimators requires a completely different proof strategy compared to the one pursued in this chapter. To see the complications that arise when analyzing local polynomial estimators of GAR's, first recall that we need to replace $\theta(t/n)$ in the model by its local polynomial approximation, $\theta_u^*(t/n)$. But this implies that instead of using $\lambda_{n,t}(\theta)$ in the computation of the likelihood, we should use

$$\lambda_{n,t}(\theta_u^*(\cdot)) = G_\lambda(Y_{n,t-1}, \lambda_{n,t-1}(\theta_u^*(\cdot)), \theta_u^*(t/n)).$$

This in turn implies that the likelihood becomes a functional of $\theta_u^*(\cdot)$ and so the analysis of local polynomial estimators for this class of models will require a completely different proof strategy involving, amongst other things, the use of functional derivatives.

In the case of the local constant estimator, on the other hand, $\theta_u^*(t/n) = \beta_0$ is constant and most of the assumptions and arguments used in Section 1.3 carry over to GAR's assuming we can show that $\lambda_{n,t}(\theta)$ and its derivatives are ULS. However, Assumption 1.5 will

no longer hold in general. To see this, observe that

$$s_{n,t}(\theta) = \frac{\partial \ell(Y_{n,t}, \lambda_{n,t}(\theta))}{\partial \lambda} \partial_\theta \lambda_{n,t}(\theta), \ \partial_\theta \lambda_{n,t}(\theta) = \nabla_\theta G_\lambda(\partial_\theta \lambda_{n,t-1}(\theta), Y_{n,t-1}, \lambda_{n,t-1}(\theta), \theta)$$

with initial conditions $\partial_\theta \lambda_{n,t}(\theta) = 0$ and $\nabla_\theta G_\lambda(\partial_\theta \lambda, Y, \lambda, \theta) := \frac{\partial G_\lambda(Y,\lambda,\theta)}{\partial \theta} + \frac{\partial G_\lambda(Y,\lambda,\theta)}{\partial \lambda} \partial_\theta \lambda$. Here, $\partial \ell(Y_{n,t}, \lambda_{n,t}) / (\partial \lambda)$ is a MGD under great generality while $\partial \ell(Y_{n,t}, \lambda_{n,t}(\theta(t/n))) / (\partial \lambda)$ will not enjoy this property since $\lambda_{n,t-1}(\theta(t/n)) \neq \lambda_{n,t-1}$. This in turn implies that $s_{n,t}(\theta(t/n))$ will generally not be a MGD. Instead, for the arguments in Section 1.3 to apply to estimators of GAR models, we here propose to replace $s_{n,t}(\theta)$ in the definition of $S_n(u)$ by the following alternative version,

$$\bar{s}_{n,t}(\theta) = \frac{\partial \ell(Y_{n,t}, \lambda_{n,t})}{\partial \lambda} v_{n,t}(\theta)$$

for some process $v_{n,t}(\theta) \in \mathcal{F}_{n,t-1}$ as chosen by us. A natural choice is $v_{n,t}(\theta) = \partial_\theta \lambda_{n,t}(\theta)$ but we here allow for added flexibility since in some applications other choices facilitate the verification of the following high-level assumption (see the proof of Theorem 1.5 for one such example):

Assumption 1.10. *(i) $\bar{s}_{n,t}(\theta(t/n))$ is a MGD w.r.t. $\mathcal{F}_{n,t}$; (ii) $\bar{\omega}_{n,t}(\theta) = \bar{s}_{n,t}(\theta) \bar{s}'_{n,t}(\theta) \in \mathbb{R}^{d_\theta \times d_\theta}$ is ULS$(p, q, \{\theta : \|\theta - \theta(u)\| < \epsilon\})$ for some $p \geq 1$ and $q > 0$ with $\bar{\omega}_t^*(\theta|u)$ being L_1-continuous at $\theta = \theta(u)$; and (iii) $\mathbb{E}[\|\bar{s}_{n,t}(\theta(t/n)) - s_{n,t}(\theta(t/n))\|^p]^{1/p} \leq C/n^{q_s}$ for some $p \geq 1$ and $q_s > 0$.*

The above assumption is almost identical to Assumption 1.5 except $s_{n,t}(\theta(t/n))$ has been replaced by $\bar{s}_{n,t}(\theta(t/n))$. The important difference appears in part (iii) which states that the former is well-approximated by the latter. In the case of Markov-type models, (iii) is automatically satisfied; for GAR-type models, we provide tools for its verification below. We then redefine variance and bias components as $S_n(u) = \frac{1}{n} \sum_{t=1}^n K_{n,t}(u) D_{n,t}(u)' \bar{s}_{n,t}(\theta(t/n))$ and

$$\bar{B}_n(u) = \frac{1}{n} \sum_{t=1}^n K_b(t/n - u) b_{n,t}, \ b_{n,t} = s_{n,t}(\theta(u)) - \bar{s}_{n,t}(\theta(t/n)),$$

where the latter can be decomposed into $\bar{B}_n(u) = B_n(u) + R_n(u)$ where $B_n(u)$ is defined

in eq. (1.4) and

$$R_n(u) = \frac{1}{n}\sum_{t=1}^{n} K_b(t/n - u)\, r_{n,t}, \quad r_{n,t} = s_{n,t}(\theta(t/n)) - \bar{s}_{n,t}(\theta(t/n)). \tag{1.12}$$

We then apply the existing theory to $S_n(u)$ and $B_n(u)$ using Assumption 1.10(i)-(ii) while $R_n(u) = O_p(n^{-q_s})$ under part (iii), and so is negligible if the bandwidth sequence is chosen such that $n^{q_s}b^2 \to \infty$:

Theorem 1.5. *Suppose that Assumptions 1.1-1.4, and 1.6-1.10 hold with $m = 0$. Then $\bar{B}_n(u) = B_n(u) + O_P(1/n^{q_s})$ where $B_n(u)$ satisfies eq. (1.10). In particular, under the additional restriction that $n^{q_s}b^2 \to \infty$, eq. (1.11) remains valid with $\Omega(u) = \mathbb{E}[\bar{\omega}_t^*(u)]$.*

Compared to the estimation of Markov-type models considered in the previous section, an additional bias term appears in the estimation of GAR models due to the additional approximation error in $\lambda_{n,t-1} - \lambda_{n,t-1}(\theta(u))$. In order to apply the above theory, it is useful with primitive conditions under which $\ell_{n,t}(\theta)$, $\omega_{n,t}(\theta)$ and $h_{n,t}(\theta)$ are ULS and part (iii) of Assumption 1.10 holds. To this end, observe that these are all functions of $Z_{n,t}(\theta) := \left(\lambda_{n,t}(\theta), \partial_\theta \lambda_{n,t}(\theta), \partial^2_{\theta\theta}\lambda_{n,t}(\theta)\right)$ where the first two components are defined above and $\partial^2_{\theta\theta}\lambda_{n,t}(\theta)$ is the matrix of second-order partial derivatives. These satisfy

$$\partial^2_{\theta\theta}\lambda_{n,t}(\theta) = \nabla^2_{\theta\theta} G_\lambda\left(\partial^2_{\theta\theta}\lambda_{n,t-1}(\theta), \partial_\theta \lambda_{n,t-1}(\theta), Y_{n,t-1}, \lambda_{n,t-1}(\theta), \theta\right)$$

with $\partial^2_{\theta\theta}\lambda_{n,t}(\theta) = 0$, for some function $\nabla^2_{\theta\theta}G_\lambda$. Importantly, $Z_{n,0}(\theta) = (\lambda_0, 0, 0)$ is fixed and so existing results for local stationarity do not apply, c.f. the discussion following Definition 1.1 and we instead develop new tools to show that $Z_{n,t}(\theta)$ is ULS. We can then apply Theorem 1.1 to show that $\ell_{n,t}(\theta)$ and its derivatives are also ULS. Observe that, for a suitably defined function G, $Z_{n,t}(\theta)$ satisfies $Z_{n,t}(\theta) = G(Y_{n,t-1}, Z_{n,t-1}(\theta); \theta)$. The following theorem states sufficient conditions for processes on this form to be ULS where we here allow data to also be parameter dependent:

Theorem 1.6. *Suppose that $W_{n,t}(\theta)$ is ULS($p_W, 1, \Theta$) with stationary approximation $W_t^*(\theta|u)$ satisfying $\mathbb{E}[\sup_{\theta\in\Theta}\|W_t^*(\theta|u)\|^{p_W}] < \infty$, and $Z_{n,t}(\theta) = G(W_{n,t-1}(\theta), Z_{n,t-1}(\theta); \theta)$ with*

$Z_{n,0} = z_0$ *where, for some* $\beta < 1$ *and* $r_W, r_\theta \geq 0$,

$$\begin{aligned}\left\|G\left(w,z;\theta\right)-G\left(w',z';\theta'\right)\right\| \leq & C\left(1+\left\|w\right\|^{r_W}+\left\|w'\right\|^{r_W}\right)\left\|w-w'\right\|+\beta\left\|z-z'\right\| \\ & +C\left(1+\left\|w\right\|^{r_\theta}+\left\|w'\right\|^{r_\theta}\right)\left\|\theta-\theta'\right\|.\end{aligned}$$

Then the following results hold:

(i) $Z_{n,t}\left(\theta\right)$ *is ULS*$(p_W/\left(r_W+1\right),1,\Theta)$ *with* $Z_t^*\left(\theta|u\right)=G\left(W_{t-1}^*\left(\theta|u\right),Z_{t-1}^*\left(\theta|u\right);\theta\right)$ *satisfying* $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_t^*\left(\theta|u\right)\right\|^{p_W/(r_W+1)}\right]<\infty$.

(ii) If $\mathbb{E}\left[\left\|W_{n,t}\left(\theta\right)-W_{n,t}\left(\theta'\right)\right\|^{p_W}\right]^{1/p_W}\leq C\left\|\theta-\theta'\right\|$, *then*

$$\mathbb{E}\left[\left\|Z_{n,t}\left(\theta\right)-Z_{n,t}\left(\theta'\right)\right\|^{\tilde{p}_Z}\right]^{1/\tilde{p}_Z}\leq C\left\|\theta-\theta'\right\|,\quad \tilde{p}_Z=p_W/\left(\max\left\{r_W,r_\theta\right\}+1\right).$$

(iii) For $Z_{n,t}=G\left(W_{n,t-1}\left(\theta\left(t/n\right)\right),Z_{n,t-1};\theta\left(t/n\right)\right)$,

$$\mathbb{E}\left[\left\|Z_{n,t}\left(\theta\left(t/n\right)\right)-Z_{n,t}\right\|^{p_W/r_\theta}\right]^{r_\theta/p_W}\leq C/n.$$

(iv) If $W_t^*\left(\theta|u\right)$ *is time-differentiable in the* L_{α_W} *sense and* $G\left(w,z;\theta\right)$ *is continuously differentiable with respect to both* w *and* z, *then* $Z_t^*\left(\theta|u\right)$ *is also time-differentiable in the* L_{α_Z} *sense where* $\alpha_Z=p_W\alpha_W/\left(p_W+r_W\alpha_W\right)$ *with time-derivative*

$$\begin{aligned}\partial_u Z_t^*\left(\theta|u\right)=&\partial_z G\left(W_{t-1}^*\left(\theta|u\right),Z_{t-1}^*\left(\theta|u\right);\theta\right)\partial_u Z_{t-1}^*\left(\theta|u\right)\\ &+\partial_w G\left(W_{t-1}^*\left(\theta|u\right),Z_{t-1}^*\left(\theta|u\right);\theta\right)\partial_u W_{t-1}^*\left(\theta|u\right).\end{aligned}$$

Part (i) of the theorem provides us with conditions under which $\lambda_{n,t}\left(\theta\right)$ and its derivatives are ULS and Lipschitz w.r.t. θ supposing that $Y_{n,t}$ is LS and G_λ and its derivatives are Lipschitz. The above can then be combined with Theorem 1.1 to show ULS of the likelihood and its derivatives. Parts (ii)-(iii) can be used to verify, e.g., $E\left[\left\|\lambda_{n,t}\left(\theta\left(t/n\right)\right)-\lambda_{n,t}\right\|^{p_\lambda}\right]^{1/p_\lambda}\leq C/n$ for some $p_\lambda\geq 1$. Suppose now that $s\left(Y_{n,t},\lambda_{n,t}\left(\theta\left(t/n\right)\right)\right)$ satisfies the conditions of Theorem 1.1. By the same arguments as used in the proof of this theorem, it then holds that $\mathbb{E}\left[\left\|s\left(Y_{n,t},\lambda_{n,t}\left(\theta\left(t/n\right)\right)\right)-s\left(Y_{n,t},\lambda_{n,t}\right)\right\|^{p}\right]^{1/p}\leq C/n$ for a suitable $p\geq 1$ thereby verifying part (iii) of Assumption 1.10; as an example of this, we refer the reader to the proof of

Corollary 1.5.

1.5 Examples

To demonstrate the usefulness of our general set-up, we here apply our theory to some particular models. Throughout this section, Assumption 1.1 is implicitly assumed. All proofs can be found in Appendix 1.8.3.

Example 1.1. (Cai, 2007; Kristensen, 2012) Consider the following d-dimensional tv-VAR(q) model,

$$Y_{n,t} = \sum_{i=1}^{q} \Phi_i\left(t/n\right) Y_{n,t-i} + \Sigma\left(t/n\right)\varepsilon_t = \theta\left(t/n\right) X_{n,t} + \Sigma\left(t/n\right)\varepsilon_t, \tag{1.13}$$

where $\varepsilon_t \in \mathbb{R}^d$ is i.i.d. with $\mathbb{E}\left[\varepsilon_t\right] = 0$ and $\mathbb{E}\left[\varepsilon_t\varepsilon_t'\right] = I_d$, $\Phi_i\left(\cdot\right) \in \mathbb{R}^{d\times d}$, $i = 1, ...p$, $\Sigma\left(\cdot\right) \in \mathbb{R}^{d\times d}$, $\theta\left(u\right) = \left(vec'\left(\Phi_1\left(u\right)\right), ..., vec'\left(\Phi_p\left(u\right)\right)\right)' \in \Theta = \mathbb{R}^{d^2q}$, and $X_{n,t} = \left(Y_{n,t-1}', ..., Y_{n,t-q}'\right)' \otimes I_d$. Under regularity conditions, its stationary approximation is given by

$$Y_t^*\left(u\right) = \sum_{i=1}^{q} \Phi_i\left(u\right) Y_{t-i}^*\left(u\right) + \Sigma\left(u\right)\varepsilon_t = \theta\left(u\right) X_t^*\left(u\right) + \Sigma\left(t/n\right)\varepsilon_t,$$

where $X_t^*\left(u\right) = \left(Y_{t-1}^*\left(u\right)', ..., Y_{t-q}^*\left(u\right)'\right)' \otimes I_d$, while its derivative process $\partial_t Y_t^*\left(u\right)$ takes the form

$$\partial_u Y_t^*\left(u\right) = \sum_{i=1}^{q} \Phi_i\left(u\right)\partial_t Y_t^*\left(u\right) + \sum_{i=1}^{q} \Phi_i^{(1)}\left(u\right) Y_{t-i}^*\left(u\right) + \Sigma^{(1)}\left(u\right)\varepsilon_t.$$

and we collect these in $\partial_u X_t^*\left(u\right) = \left(\partial_u Y_{t-1}^*\left(u\right)', ..., \partial_u Y_{t-q}^*\left(u\right)'\right)' \otimes I_d$. We estimate $\theta\left(u\right)$ by local least-squares, $\ell_{n,t}\left(\theta\right) = \left\|Y_{n,t} - \theta' X_{n,t}\right\|^2$. Applying our asymptotic theory, we obtain the following novel result for the estimation of time-varying VAR(p) models:

Corollary 1.3. *Suppose that $\theta\left(\cdot\right)$ and $\Sigma\left(\cdot\right)$ are twice continuously differentiable with $\Phi\left(v\right) = I_d - \sum_{i=1}^{q}\Phi_i\left(v\right)z^i$ having all its eigenvalues outside the unit circle, $v \in [0,1]$. Then the local linear estimator satisfies Theorem 1.3 with $H\left(u\right) = \mathbb{E}\left[X_t^*\left(u\right) X_t^*\left(u\right)'\right]$ and $\Omega\left(u\right) = \mathbb{E}\left[X_t^*\left(u\right)\Sigma\left(u\right)\Sigma\left(u\right)' X_t^*\left(u\right)'\right]$. If in addition $\mathbb{E}\left[\left\|\varepsilon_t\right\|^4\right] < \infty$, then the local constant estimator satisfies Theorem 1.4 with $\partial_\theta H\left(u\right) = 0$, and $\partial_u H\left(u\right) = 2\mathbb{E}\left[X_t^*\left(u\right)\partial_u X_t^*\left(u\right)'\right]$.*

Example 1.2. (Dahlhaus and Subba Rao, 2006; Fryzlewicz et al., 2008) Suppose $W_{n,t} = Y_{n,t}^2 \in \mathbb{R}_+$ solves the following tv-ARCH(q) model,

$$W_{n,t} = \lambda_{n,t}\varepsilon_t^2, \quad \lambda_{n,t} = \omega\left(t/n\right) + \sum_{i=1}^{q} \alpha_i\left(t/n\right) W_{n,t-i}, \tag{1.14}$$

where ε_t is i.i.d. with zero mean and unit variance. The corresponding stationary solution and derivative process are given by

$$\begin{aligned} W_t^*(u) =& \lambda_t^*(u)\varepsilon_t^2, \quad \lambda_t^*(u) = \omega(u) + \sum_{i=1}^{q} \alpha_i(u) W_{t-i}^*(u), \\ \partial_u W_t^*(u) =& \partial_u\lambda_t^*(u)\varepsilon_t^2, \ \partial_u\lambda_t^*(u) = \omega^{(1)}(u) + \sum_{i=1}^{q} \alpha_i(u)\partial_t W_{t-i}^*(u) + \sum_{i=1}^{q} \alpha_i^{(1)}\left(u\right) W_{t-i}^*\left(u\right). \end{aligned}$$

We estimate the time-varying parameter vector $\theta\left(u\right) = (\omega(u), \alpha_1(u), ..., \alpha_p(u))$ using our local polynomial estimator based on the Gaussian quasi-log likelihood,

$$\ell_{n,t}\left(\theta\right) = -\log\left(\lambda_{n,t}\left(\theta\right)\right) - \frac{W_{n,t}}{\lambda_{n,t}\left(\theta\right)}, \quad \lambda_{n,t}\left(\theta\right) = \omega + \sum_{i=1}^{q} \alpha_i W_{n,t-i}.$$

Corollary 1.4. *For the tv-ARCH(q) model given by (1.14), assume that (i)* $\mathbb{E}\left[\varepsilon_t^4\right] < \infty$*; (ii)* $\theta\left(\cdot\right)$ *is twice continuously differentiable with* $\sum_{i=1}^{q} \alpha_i\left(v\right) < 1$ *for all* $v \in [0,1]$*; and (iii)* $\theta\left(u\right) \in Int\left(\Theta\right)$ *where* $\Theta = \left\{\theta \in [\delta_L, \delta_U]^{q+1} \mid \sum_{i=1}^{q} \alpha_i \leq 1 - \delta\right\}$ *for some* $0 < \delta_L < \delta_U < \infty$ *and* $\delta > 0$*. Then the local linear and local constant estimators of the tvARCH model satisfy Theorems 1.3 and 1.4, respectively, with* $\Omega\left(u\right) = -Var\left(\varepsilon_t^2\right) H\left(u\right)$*,*

$$H\left(u\right) = -\mathbb{E}\left[\frac{\partial_\theta\lambda_t^*\left(u\right)\left(\partial_\theta\lambda_t^*\left(u\right)\right)'}{\lambda_t^*(u)^2}\right], \ \partial_{\theta_i} H\left(u\right) = 2\mathbb{E}\left[\frac{\partial_{\theta_i}\lambda_t^*\left(u\right)\partial_\theta\lambda_t^*\left(u\right)\left(\partial_\theta\lambda_t^*\left(u\right)\right)'}{\lambda_t^*(u)^3}\right],$$

$$\partial_u H\left(u\right) = 2\mathbb{E}\left[\frac{\partial_u\lambda_t^*\left(u\right)\partial_\theta\lambda_t^*\left(u\right)\left(\partial_\theta\lambda_t^*\left(u\right)\right)'}{\lambda_t^*(u)^3} + \frac{\partial_{\theta u}^2\lambda_t^*\left(u\right)\left(\partial_\theta\lambda_t^*\left(u\right)\right)'}{\lambda_t^*(u)^2}\right]$$

where $\partial_\theta\lambda_t^*\left(u\right) = \left(1, W_{t-1}^*\left(u\right), \ldots, W_{t-q}^*\left(u\right)\right)'$ *and* $\partial_{\theta u}^2\lambda_t^*\left(u\right) = \left(1, \partial_u W_{t-1}^*(u), \ldots, \partial_u W_{t-q}^*(u)\right)'$*.*

Comparing our conditions with the ones in Dahlhaus and Subba Rao (2006), we see that ours are substantially weaker: They require that $\mathbb{E}\left[\varepsilon_t^{12}\right]^{1/6} \sum_{j=1}^{q} \alpha_j\left(u\right) < 1 - \rho$ which rules out most empirically relevant situations. For example, if $\varepsilon_t \sim N\left(0,1\right)$ then their requirement becomes $\sum_{j=1}^{q} \alpha_j\left(u\right) < 0.22$. This strong condition is a by-product of their

proof strategy which requires mixing and stronger moment conditions of the derivative process. Furthermore, while their bias component for the local constant estimator involves the so-called second-order derivative process while ours only involves the first-order derivative.

Example 1.3. (Chen and Hong, 2016) Let $W_{n,t} = Y_{n,t}^2 \in \mathbb{R}_+$ solve the following tv-GARCH model,

$$W_{n,t} = \lambda_{n,t}\varepsilon_t^2, \quad \lambda_{n,t} = \omega\left(t/n\right) + \alpha\left(t/n\right) W_{n,t-1} + \beta\left(t/n\right)\lambda_{n,t-1}, \tag{1.15}$$

for $t = 1, 2, \ldots, n$, where ε_t is i.i.d. $(0, 1)$. We estimate $\theta\left(u\right) = (\omega(u), \alpha(u), \beta(u))'$ using the Gaussian log-likelihood which takes the same form as in Example 1.2 except that now $\lambda_{n,t}\left(\theta\right) = \omega + \alpha W_{n,t-1} + \beta\lambda_{n,t-1}\left(\theta\right)$ where $\lambda_{n,0}^2\left(\theta\right) = \lambda_0 > 0$. The stationary solution and its derivative process takes the form

$$W_t^*\left(u\right) = \lambda_t^*\left(u\right)\varepsilon_t^2, \quad \lambda_t^*\left(u\right) = \omega\left(u\right) + \alpha\left(u\right) W_{t-1}^*\left(u\right) + \beta\left(u\right)\lambda_{t-1}^*\left(u\right),$$

and $\partial_u W_t^*(u) = \partial_u \lambda_t^*(u)\varepsilon_t^2$ where

$$\partial_u\lambda_t^*(u) = \omega^{(1)}(u) + \alpha(u)\partial_u W_{t-1}^*(u) + \beta(u)\partial_u\lambda_{t-1}^*(u) + \alpha_i^{(1)}\left(u\right) W_{t-1}^*\left(u\right) + \beta^{(1)}\left(u\right)\lambda_{t-1}^*\left(u\right).$$

To state our asymptotic theory, we also need the stationary version of the derivative process w.r.t. θ, $\partial_\theta\lambda_t^*\left(u\right) = \left(1/\left(1-\beta\left(u\right)\right), \partial_\alpha\lambda_t^*\left(u\right), \partial_\beta\lambda_t^*\left(u\right)\right)'$ where

$$\partial_\alpha\lambda_t^*\left(u\right) = W_{t-1}^*\left(u\right) + \beta\left(u\right)\partial_\alpha\lambda_{t-1}^*\left(u\right), \ \partial_\beta\lambda_t^*\left(u\right) = \lambda_t^*\left(u\right) + \beta\left(u\right)\partial_\beta\lambda_{t-1}^*\left(u\right)\left(u\right),$$

and $\partial_{\theta u}^2\lambda_t^*\left(u\right) = \left(\beta^{(1)}\left(u\right)/\left(1-\beta\left(u\right)\right)^2, \partial_{\alpha u}^2\lambda_t^*\left(u\right), \partial_{\beta u}^2\lambda_t^*\left(u\right)\right)'$ where

$$\begin{aligned}
\partial_{\alpha u}^2\lambda_t^*\left(u\right) &= \partial_u W_{t-1}^*(u) + \beta^{(1)}\left(u\right)\partial_\alpha\lambda_{t-1}^*\left(u\right) + \beta\left(u\right)\partial_{\alpha u}^2\lambda_t^*\left(u\right) \\
\partial_{\beta u}^2\lambda_t^*\left(u\right) &= \partial_u\lambda_t^*\left(u\right) + \beta^{(1)}\left(u\right)\partial_\beta\lambda_{t-1}^*\left(u\right)\left(u\right) + \beta\left(u\right)\partial_{\beta u}^2\lambda_t^*\left(u\right).
\end{aligned}$$

Corollary 1.5. *For the tvGARCH model given by (1.15), assume that (i)* $\mathbb{E}\left[\varepsilon_t^4\right] < \infty$*; (ii)* $\theta\left(\cdot\right)$ *is twice continuously differentiable with* $\alpha\left(v\right) + \beta\left(v\right) < 1$ *for all* $v \in [0,1]$*; and (iii)* $\theta\left(u\right) \in Int\left(\Theta\right)$ *where* $\Theta = \left\{\theta = (\omega, \alpha, \beta)' \in [\delta_L, \delta_U]^3 \,|\alpha + \beta \leq 1 - \delta\right\}$ *for some* $0 < \delta_L < \delta_U < \infty$ *and* $\delta > 0$*. Then the local constant estimator of the tvGARCH model satisfies*

Theorem 1.4 with the relevant moments being on the same form as in Corollary 1.4 but now with $\lambda_t^*(u)$, $\partial_t \lambda_t^*(u)$, $\partial_\theta \lambda_t^*(u)$ *and* $\partial_{\theta u}^2 \lambda_t^*(u)$ *as defined above.*

Again, our conditions are substantially weaker compared to those found in the existing literature: Chen and Hong (2016) require $\mathbb{E}\left[\varepsilon_t^{16}\right] < \infty$, that the GARCH process and its derivative process are ϕ-irreducible, and that the bandwidth shrinks to zero at a very slow rate.

Example 1.4. (Agosto et al., 2016) Let $Y_{n,t} \in \{0, 1, 2, \ldots\}$ solve the following time-varying Poisson Autoregression (tvPAR),

$$Y_{n,t}|\mathcal{F}_{n,t-1} \sim \text{Poisson}\left(\lambda_{n,t}\right), \quad \lambda_{n,t} = \omega\left(t/n\right) + \sum_{i=1}^{q} \alpha_i\left(t/n\right) Y_{n,t-i}. \tag{1.16}$$

where $\mathcal{F}_{n,t-1} = \mathcal{F}\left\{Y_{n,t-i} : i = 1, 2, ...\right\}$, Poisson($\lambda$) denotes a Poisson distribution with intensity parameter λ. This model is a time-varying parameter version of the model considered in Agosto et al. (2016) who analyze the properties of $Y_{n,t}$ and of the MLE when $\theta(u) = \left(\omega(u), \alpha_1(u), ..., \alpha_p(u)\right)'$ is constant. We here apply our general theory to the local linear MLE where the log-likelihood function takes the form

$$\ell_{n,t}(\theta) := Y_{n,t} \log\left\{\lambda_{n,t}(\theta)\right\} - \lambda_{n,t}(\theta), \quad \lambda_{n,t}(\theta) = \omega + \sum_{i=1}^{q} \alpha_i Y_{n,t-i}.$$

Note here that the derivative process of $Y_{n,t}$ is not well-defined due to it being discrete-valued, and so existing results, such as the ones in Dahlhaus et al. (2017), cannot be used to analyze the local MLE. The following corollary provides the first asymptotic theory for local linear estimation of the tvPAR model:

Corollary 1.6. *For the tvPAR model given by (1.16), assume that (i)* $\theta(\cdot)$ *is twice continuously differentiable with* $\sum_{i=1}^{q} \alpha_i(v) < 1$ *for all* $v \in [0, 1]$*; and (ii)* $\theta(u) \in Int(\Theta)$ *where, for some* $0 < \delta_L < \delta_U < \infty$ *and* $\delta > 0$, $\Theta = \left\{\theta \in [\delta_L, \delta_U]^{q+1} | \sum_{i=1}^{q} \alpha_i \leq 1 - \delta\right\}$ *. Then the local linear estimator satisfies Theorem 1.3 with*

$$\Omega(u) = \mathbb{E}\left[\frac{\left(\partial_\theta \lambda_t^*(u)\right)\left(\partial_\theta \lambda_t^*(u)\right)'}{\lambda_t^*(u)}\right] = -H(u),$$

where $\lambda_t^*(u) = \omega(u) + \sum_{i=1}^{q} \alpha_i(u) Y_{t-i}^*(u)$ *and* $\partial_\theta \lambda_t^*(u) = \left(1, Y_{t-1}^*(u), ..., Y_{t-q}^*(u)\right)'$.

1.6 Simulation study

In this section, we examine the finite-sample performances of our estimators. Throughout, we use the Epanechnikov kernel and all results are based on 500 simulated data sets. The performance of the estimators is evaluated using the mean absolute deviation error (MADE), $MADE := \frac{1}{n}\sum_{t=1}^{n}\left|\hat{\theta}(t/n) - \theta(t/n)\right|$, as well as their bias, variance, and mean squared error.

The estimators were implemented as follows: First note that in most applications, we wish to estimate the full parameter path that generated data, say, $\theta(1/n), ..., \theta((n-1)/n)$. This involves $n-1$ optimization problems but observe that we will in general expect $\hat{\beta}(i/n)$ will be fairly close to $\hat{\beta}((i-1)/n)$, $i = 2, ..., n$. This motivates the following sequential procedures: Do a full parameter search to obtain $\hat{\beta}(1/n) = \arg\max_{\beta\in\mathcal{B}} L_n(\beta|1/n)$ and then use Newton's method for the remaining estimates: With $\hat{b}_{i,0} = \hat{\beta}((i-1)/n)$, compute

$$\hat{b}_{i,k+1} = \hat{b}_{i,k} - H_n^{-1}\left(\hat{b}_{i,k}|i/n\right) S_n\left(\hat{b}_{i,k}|i/n\right),$$

for $k = 1, 2, ...$, where $S_n(\beta|u)$ and $H_n(\beta|u)$ denote the score and hessian of $L_n(\beta|i/n)$, until convergence is achieved and set $\hat{\beta}(i/n)$ equal to the termination value. We found this method to work very well in practice. When $m > 0$, the initial computation of $\hat{b}_{i,k}$ is of dimension $(m+1)\dim(\theta)$ which may be a high-dimensional problem. To resolve this, we again propose a sequential procedure: First, compute the local constant estimator, $\hat{\theta}(u) = \arg\max_{\theta\in\Theta} L_n(\theta|1/n)$; second, compute the local linear estimator initialized at $\left(\hat{\theta}(u), 0\right)$, and so forth.

To select b, we employ a generalized version of the cross-validation method proposed in Richter and Dahlhaus (2017): As a first step, we compute the leave-one-out estimator,

$$\hat{\beta}_b(t_0/n) = \arg\max_{\beta\in\mathcal{B}} \sum_{t\neq t_0}^{n} K_b\left(\frac{t-t_0}{n}\right) \ell_{n,t}\left(D(t/n-u)\beta\right),$$

for $t_0 = 1, ..., n$, and then use as criterion the over-all global quasi-likelihood,

$$CV(b) = \sum_{t=1}^{n} \ell_{n,t}\left(D(t/n-u)\hat{\beta}_b(t/n)\right).$$

We then choose our bandwidth as the minimizer of $CV(b)$. Chu and Marron (1991) indicate that cross-validation may be severely affected when the model is misspecified so that the score function is no longer a martingale difference. This can be handled by using a "leave-$(2l+1)$-out" version of the above cross-validation method.

1.6.1 Time-varying ARCH

We first consider the time-varying ARCH(1) in eq. (1.14) where $\varepsilon \sim i.i.d.N(0,1)$ and

$$\omega(u) = -.5\cos(6\pi u) + .7, \quad \alpha(u) = .4\cos(6\pi u) + .45.$$

We estimate $\omega(u)$ and $\alpha(u)$ using both local Gaussian log-likelihood and the WLS method of Fryzlewicz et al. (2008) witht K chosen as the Epanechnikov kernel. The following results are based on 500 simulated data sets with sample sizes n =250, 500 and 1000. Table 1.1 reports the performance of the estimators based on cross-validated bandwidths. The local linear MLE performs best in terms of IMSE and MADE among the four estimators for all sample sizes. We also report MADE values for both local constant and local linear estimators. For all sample sizes, the bias for the QML estimator is always smaller than one for WLS estimator.

Table 1.1: Performance of the local constant (LC) and local linear (LL) estimators: ARCH(1)

		$\omega(u)$				$\alpha(u)$			
		WLS		ML		WLS		ML	
n		LC	LL	LC	LL	LC	LL	LC	LL
250	ISB	0.035	0.029	0.038	0.032	0.033	0.037	0.032	0.033
	IV	0.077	0.094	0.068	0.066	0.087	0.112	0.091	0.090
	IMSE	0.112	0.123	0.107	0.098	0.120	0.149	0.124	0.123
	MADE	0.264	0.263	0.264	0.236	0.271	0.293	0.269	0.272
500	ISB	0.015	0.012	0.014	0.014	0.017	0.018	0.014	0.018
	IV	0.042	0.045	0.042	0.035	0.056	0.068	0.062	0.056
	IMSE	0.058	0.057	0.056	0.049	0.073	0.087	0.076	0.074
	MADE	0.184	0.180	0.178	0.169	0.214	0.225	0.211	0.209
1000	ISB	0.008	0.008	0.006	0.007	0.009	0.009	0.006	0.008
	IV	0.024	0.024	0.022	0.019	0.035	0.041	0.037	0.035
	IMSE	0.032	0.031	0.028	0.026	0.044	0.049	0.043	0.043
	MADE	0.138	0.132	0.129	0.124	0.162	0.169	0.159	0.158

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

To investigate the performance of the estimators near the end of the sample, we plot the estimates of ω for n =1000 in Figure 1.1. As predicted by the theory, we observe that the local linear estimators enjoys smaller biases near the boundary.

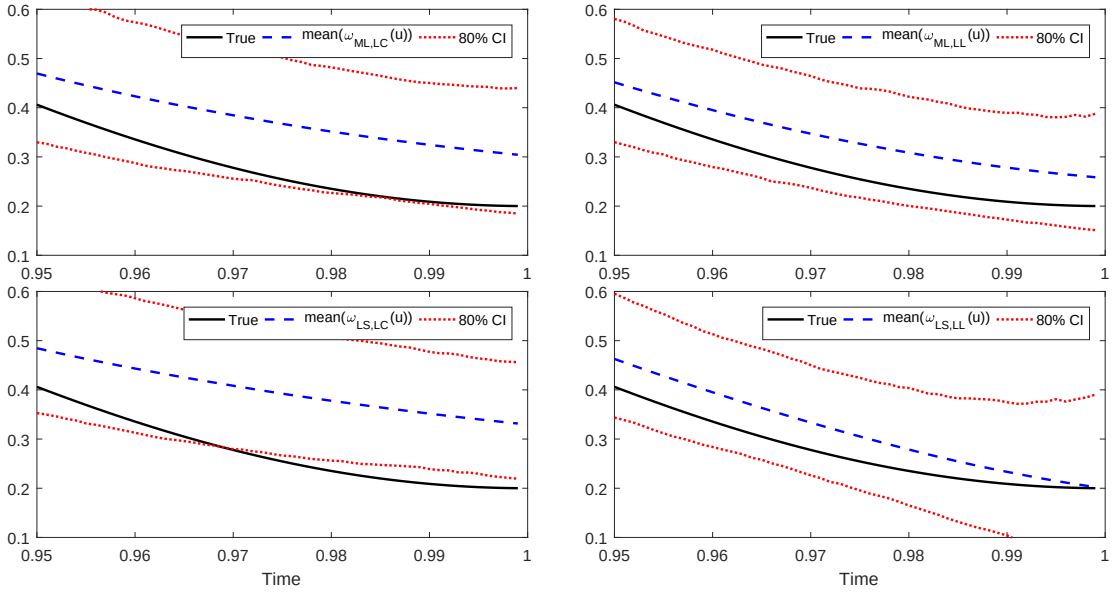

Figure 1.1: Pointwise means of local constant and local linear MLE's and LS estimators of ω in ARCH(1)

1.6.2 Poisson Autoregression

We here report simulation results for the local constant and local linear MLE's of the following PARX(1) model with an additional exogeneous regressor $X_{n,t}$,

$$\lambda_{n,t} = \omega\left(t/n\right) + \alpha\left(t/n\right) Y_{n,t-1} + \gamma\left(t/n\right) \exp\left(X_{t-1}\right),$$

where

$$\omega\left(u\right) = 0.7 - 0.5\sin\left(2\pi u\right), \quad \alpha\left(u\right) = 0.5 + 0.4\sin\left(2\pi u\right), \quad \gamma\left(u\right) = 1 + 0.5\sin\left(2\pi u\right),$$

and

$$X_{n,t} = \rho\left(t/n\right) X_{n,t-1} + \sigma\left(t/n\right)\varepsilon_t, \ \varepsilon_t \sim \text{i.i.d.} N\left(0,1\right).$$

We examine the performance of the MLE under two different data generating processes (DGP's) for the covariate X_t.

DGP1 Strictly stationary $X_{n,t}$: $\rho(u) = 0.5$, $\sigma(u) = 1$.

DGP2 Locally stationary $X_{n,t}$: $\rho(u) = 0.5 - 0.4\cos(\pi u)$, $\sigma(u) = 1 + 0.5\cos(2\pi u)$.

Table 1.2 reports the over-all performance of the estimators in terms of integrated squared bias, variance, MSE and MADE. The table shows that the variance of the local linear estimators is slightly smaller than the one of the local constant estimator. Otherwise, the performance of the estimators are similar. Overall, we find that the performance of the local linear estimator for DGP2 is better than the one for DGP1. Finally, similar to the case of the tvARCH model, the local linear estimator again enjoys better performance near the boundaries; we leave out the plots to save space.

Table 1.2: Performance of the local constant (LC) and local linear (LL) estimators: PARX(1)

		$\omega(u)$		$\alpha(u)$		$\gamma(u)$	
DGP	$n = 500$	LC	LL	LC	LL	LC	LL
1	ISB	0.006	0.014	0.002	0.003	0.003	0.003
	IVar	0.075	0.078	0.005	0.004	0.018	0.019
	IMSE	0.081	0.092	0.007	0.007	0.021	0.022
	med.(MADE)	0.202	0.208	0.061	0.063	0.112	0.111
2	ISB	0.008	0.015	0.002	0.002	0.002	0.002
	IVar	0.091	0.077	0.005	0.003	0.019	0.016
	IMSE	0.099	0.091	0.006	0.006	0.022	0.018
	med.(MADE)	0.205	0.205	0.061	0.059	0.109	0.098

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

1.7 Empirical application

The aim of this section is to analyze possible time-varying effects of various factors explaining US corporate default rates when modeled with a PARX model. The data set on defaults consists of monthly number of bankruptcies among Moody's rated industrial firms in the United States for the period 1982-2011 ($T = 360$ observations), collected from Moody's Credit Risk Calculator (CRC). Figure 1.2, which shows default counts and the corresponding autocorrelation function, reveals (i) high temporal dependence in default counts; (ii) existence of default clusters over time.

We follow Agosto et al. (2016) and model monthly number of bankrupcies, Y_t, by a

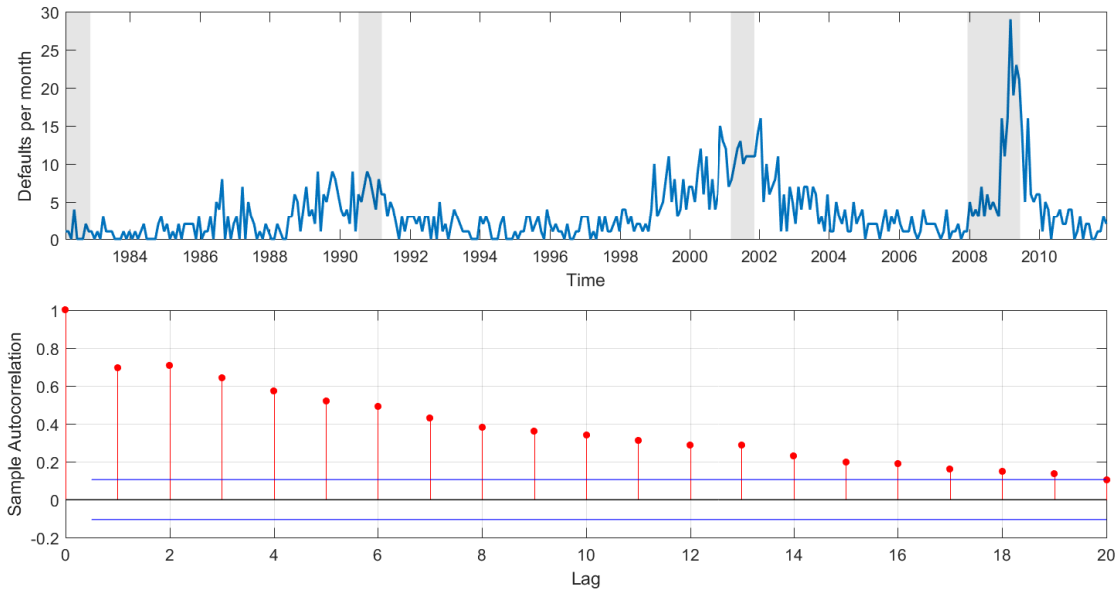

Figure 1.2: Number of defaults per month among Moody's rated US industrial firms in the period 1982-2011 (top) and autocorrelation function of defaults (bottom)

PARX model, but here allow for the possibility of time-varying parameters,

$$Y_{n,t}|\mathcal{F}_{n,t-1} \sim \text{Poisson}\left(\lambda_{n,t}\right), \quad t = 1, 2, \ldots, n;$$
$$\lambda_{n,t} = \omega\left(t/n\right) + \sum_{i=1}^{p} \alpha_i\left(t/n\right) Y_{n,t-i} + \gamma_{LI}\left(t/n\right) \exp\left(-LI_{n,t-1}\right),$$

where LI is the so-called Leading Index released by the Federal Reserve (LI). This can be seen as a leading indicator of economic activity. To select the number of lags, we first estimate the model with constant parameters and then use AIC and BIC for model selection. The results are reported in Table 1.3 from which we see that the preferred specification is the PARX(3) model.

Table 1.3: Model selection results for corporate defaults

	PARX(1)	PARX(3)	PARX(6)	GPARX(1,1)	GPARX(2,1)	GPARX(3,1)
logL	-811.6	-737.2	-723.6	-741.0	-734.0	-731.3
AIC	1629.2	1484.3	1463.2	1490.0	1478.0	1474.6
BIC	1640.9	1503.7	1494.3	1505.5	1497.4	1497.9
p-value of PIT	$< 10^{-4}$	0.0194	0.0151	0.0028	0.0051	0.0068

Agosto et al. (2016) found evidence of of two significant break when the Dot-com bubble burst in the late 1990's and again around the onset of the most recent financial crisis in 2008.

The aim here is to see whether this finding is supported by the nonparametric estimation for the time-varying parameters. We here focus on the tvPARX(6) model. Figure 1.3 shows the time-series of local linear estimates, $\left\{\hat{\theta}_t\right\}$, for tvPARX(6). These graphs provide some evidence of structural change. In particular, the impact of $\exp(-LI)$ on the default intensity is significant and dramatically changes over the whole estimation period. All together, we find substantial time-variation in the parameters that our local polynomial estimators are able to capture well.

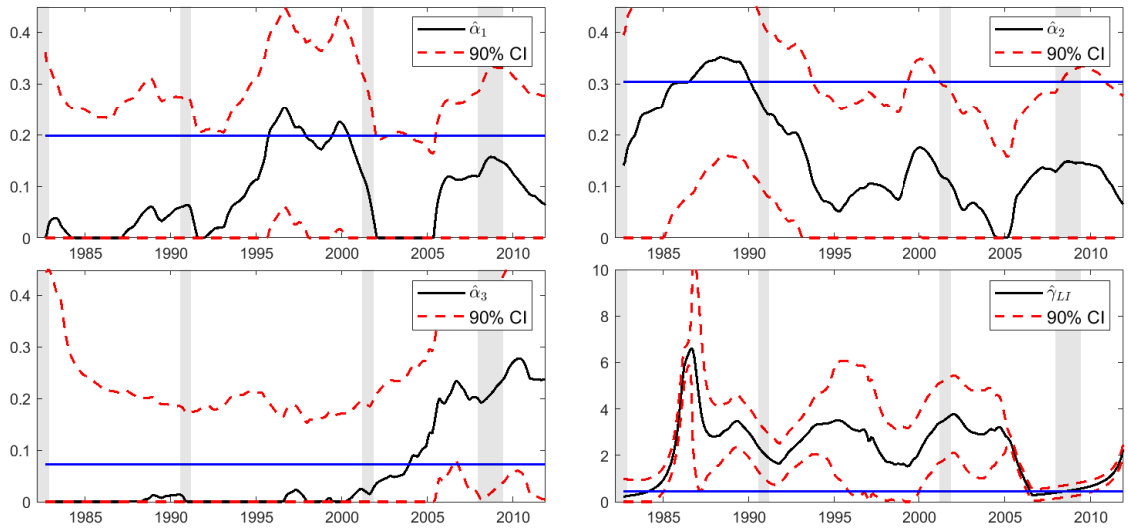

Figure 1.3: Local linear estimate of tvPARX(6) model

1.8 Appendix

1.8.1 Auxiliary results

In the following, assume that L satisfies: (i) $L(\cdot)$ has a compact support; (ii) for some $\Lambda < \infty$, $|L(v) - L(v')| \leq \Lambda |v - v'|$, $v, v' \in \mathbb{R}$. We denote $L_b(\cdot) \coloneqq L(\cdot/b)/b$.

Lemma 1.1. *The following hold as $b \to 0$ and $nb \to \infty$:*

(i) Suppose $\{W_{n,t}(\theta)\}$ is ULS(p, q, Θ) with its stationary approximation $\{W_t^*(\theta|u)\}$ being L_p continuous for some $p \geq 1, q > 0$ and Θ is compact Then, with $\mathcal{A}$ defined in Assumption 1.2,

$$\sup_{\alpha \in \mathcal{A}} \left\| \frac{1}{n} \sum_{t=1}^{n} L_b(t/n - u) W_{n,t}(D_b(t/n - u)\alpha) - \int L(v) \mathbb{E}\left[W_t^*(D(v)\alpha|u)\right] dv \right\| = o_p(1).$$

(ii) Suppose $\{W_{n,t}(\theta(t/n)), \mathcal{F}_{n,t}\}$ *is a martingale difference array; and, for some* $p \geq 1$ *and* $q, \epsilon > 0$, $V_{n,t}(\theta) = W_{n,t}(\theta) W'_{n,t}(\theta)$ *is ULS*$(p, q, \{\theta : \|\theta - \theta(u)\| < \epsilon\})$ *with its stationary approximation* $V_t^*(\theta|u)$ *being* L_p *continuous at* $\theta = \theta(u)$*; and* $v \mapsto \theta(v)$ *is continuous at* $v = u$*. Then*

$$\sqrt{\frac{b}{n}} \sum_{t=1}^{n} L_b(t/n - u) W_{n,t}(\theta(t/n)) \to^d N\left(0, \int L^2(v)\, dv \times \mathbb{E}\left[V_t^*(\theta(u)|u)\right]\right);$$
$$\sqrt{\frac{b}{n}} \sum_{t=1}^{n} L_b(t/n - cb) W_{n,t}(\theta(t/n)) \to^d N\left(0, \int_{-c}^{+\infty} L^2(v)\, dv \times \mathbb{E}\left[V_t^*(\theta(u)|u)\right]\right).$$

(iii) Suppose W_t^* *is a stationary and ergodic sequence with* $\sum_{s=0}^{\infty} \left|cov\left(W_t^*, W_{t+s}^*\right)\right| < \infty$*. Then, for any* $u \in (0,1)$

$$\left|\frac{1}{n} \sum_{t=1}^{n} L_b(t/n - u) W_t^* - \int L(v)\, dv \times \mathbb{E}[W_t^*]\right| = o_p\left(1/\sqrt{nb}\right).$$

Proof of Lemma 1.1. **Proof of (i).** We first show that for all $\theta \in \Theta$,

$$\frac{1}{n} \sum_{t=1}^{n} L_b(t/n - u) W_{n,t}(\theta) \to^p \int L(v)\, dv \times \mathbb{E}\left[W_t^*(\theta|u)\right].$$

Note that $L(v) = 0$ for $|v| \geq \bar{v}$ for some $\bar{v} > 0$. Then the Minkowski's inequality implies that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n} \sum_{t=1}^{n} L_b(t/n - u) \left\{W_{n,t}(\theta) - W_t^*(\theta|u)\right\}\right\|^p\right]^{1/p} \\
&\leq \frac{1}{n} \sum_{t=1}^{n} |L_b(t/n - u)| \, \mathbb{E}\left[\|W_{n,t}(\theta) - W_t^*(\theta|u)\|^p\right]^{1/p} \\
&\leq \frac{C}{n} \sum_{t=1}^{n} |L_b(t/n - u)| \left(b^q \left|\frac{t/n - u}{b}\right|^q + 1/n^q + \rho^t\right) \\
&\leq \frac{C}{n} \sum_{t=1}^{n} |L_b(t/n - u)| \times \left(b^q \bar{v}^q + 1/n^q + \rho^t\right) = O\left(b^q\right) + O\left(n^{-q}\right) + O\left(\frac{1}{\sqrt{nb}}\right),
\end{aligned}$$

where we have used that

$$\frac{1}{n} \sum_{t=1}^{n} |L_b(t/n - u)| \rho^{qt} \leq \frac{1}{\sqrt{nb}} \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(L^2\right)_b (t/n - u)} \sqrt{\sum_{t=1}^{n} \rho^{2qt}} = O\left(\frac{1}{\sqrt{nb}}\right).$$

Next, with $\bar{W}_t = W_t^*(\theta|u) - \mathbb{E}[W_t^*(\theta|u)]$, for sufficiently large n, $\frac{1}{n}\sum_{t=1}^{n} L_b(t/n-u)\bar{W}_t = \frac{1}{nb}\sum_{t=\underline{t}}^{\bar{t}} L_b(t/n-u)\bar{W}_t$, where $\bar{t} = [n(u+\bar{v}b)]$ and $\underline{t} = [n(u-\bar{v}b)]$. Here, $[x]$ denotes the integer part of any real number x. By summation by parts, we have, with $S_{n,t} = \sum_{j=\underline{t}}^{t} \bar{W}_j$,

$$\begin{aligned}\frac{1}{n}\sum_{t=\underline{t}}^{\bar{t}} L_b(t/n-u)\bar{W}_t &= \frac{1}{n}\sum_{t=\underline{t}}^{\bar{t}} L_b(t/n-u)(S_{n,t}-S_{n,t-1}) \\ &= \frac{1}{n}\sum_{t=\underline{t}}^{\bar{t}-1}[L_b(t/n-u) - L_b((t+1)/n-u)]S_{n,t} + \frac{1}{n}L_b(\bar{t}/n-u)S_{n,\bar{t}}.\end{aligned}$$

Since $\{\bar{W}_t\}$ is stationary, $S_{n,t}$ has the same distribution as $\tilde{S}_{n,t} = \sum_{j=1}^{t-\underline{t}+1}\bar{W}_j$. Thus, for some constant M, $\left|\frac{1}{n}\sum_{t=1}^{n} L_b(t/n-u)\bar{W}_t\right| \leq \frac{M}{nb}\sup_{t\leq \bar{t}-\underline{t}+1}\left|\tilde{S}_{n,t}\right|$. The ergodic theorem yields $\tilde{S}_{n,t}/t \to 0$ which in turn implies that $\frac{1}{n}\sum_{t=1}^{n} L_b(t/n-u)\bar{W}_t$ tends to zero almost surely. Finally, using the mean value theorem, there exists $v_{n,t} \in \left[\frac{t-1}{n}, \frac{t}{n}\right]$ so that with $\bar{L} = \sup_v L(v)$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{t=1}^{n} L_b(t/n-u) - \int L_b(x-u)\,dx\right| &= \left|\frac{1}{nb}\sum_{t=1}^{n} L_b(t/n-u) - \sum_{t=1}^{n}\int_{(t-1)/n}^{t/n} L_b(x-u)\,dx\right| \\ &\leq \frac{1}{nb}\sum_{t=1}^{n}|L_b(t/n-u) - L_b(v_{n,t}-u)| \\ &\leq \frac{1}{nb}\sum_{t=1}^{n}\Lambda\left|\frac{t/n-v_{n,t}}{b}\right| = O\left(\frac{1}{nb}\right),\end{aligned}$$

which shows that $\frac{1}{n}\sum_{t=1}^{n} L_b(t/n-u)\mathbb{E}[W_t^*(\theta|u)] = \int L_b(x-u)\,dx\mathbb{E}[W_t^*(\theta|u)] + O(1/nb)$.

For the uniform convergence, we note that by definition of $\mathcal{A}$, $D_b(v-u)\alpha \in \Theta$ for all $v \in \mathrm{supp}(L)$ and $\alpha \in \mathcal{A}$. Thus, $\frac{1}{n}\sum_{t=1}^{n} K_b(t/n-u)W_{n,t}(D_{n,t}(u)\alpha)$, where $D_{n,t}(u) = D_b(t/n-u)$, is well-defined for $\alpha \in \mathcal{A}$, and

$$\begin{aligned}\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}}\|W_{n,t}(D_{n,t}(u)\alpha) - W_t^*(D_{n,t}(u)\alpha|u)\|^p\right] &\leq \mathbb{E}\left[\sup_{\theta\in\Theta}\|W_{n,t}(\theta) - W_t^*(\theta|u)\|^p\right] \\ &\leq C\left(|t/n-u|^q + 1/n^q + \rho^t\right)^p.\end{aligned}$$

Using Hölder's inequality and Minkowski's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}}\left\|\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)\left\{W_{n,t}\left(D_{n,t}\left(u\right)\alpha\right)-W_t^*\left(D_{n,t}\left(u\right)\alpha|u\right)\right\}\right\|\right]\\
&\leq\frac{1}{n}\sum_{t=1}^{n}\left|L_b\left(t/n-u\right)\right|\mathbb{E}\left[\sup_{\alpha\in\mathcal{A}}\left\|W_{n,t}\left(D_{n,t}\left(u\right)\alpha\right)-W_t^*\left(D_{n,t}\left(u\right)\alpha|u\right)\right\|^p\right]^{1/p}\\
&\leq C\frac{b^q}{n}\sum_{t=1}^{n}\left|L_b\left(t/n-u\right)\right|\left(\left|\frac{t/n-u}{b}\right|^q+1/n^q+\rho^t\right)=O\left(b^q\right).
\end{aligned}$$

Next,

$$\begin{aligned}
&\sup_{\alpha\in\mathcal{A}}\left\|\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)\left\{W_t^*\left(D_{n,t}\left(u\right)\alpha|u\right)-\mathbb{E}\left[W_t^*\left(D_{n,t}\left(u\right)\alpha|u\right)\right]\right\}\right\|\\
&\leq\sup_{\theta\in\Theta}\left\|\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)\left\{W_t^*\left(\theta|u\right)-\mathbb{E}\left[W_t^*\left(\theta|u\right)\right]\right\}\right\|+o_p\left(1\right)
\end{aligned}$$

where $\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)\left\{W_t^*\left(\theta|u\right)-\mathbb{E}\left[W_t^*\left(\theta|u\right)\right]\right\}=o_P\left(1\right)$ for all $\theta\in\Theta$. Thus, the result will follow if we can show stochastic equicontinuity of $\theta\mapsto\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)W_t^*\left(\theta|u\right)$ but this follows from the assumption of $\theta\mapsto W_t^*\left(\theta|u\right)$ being L_p continuous: For a given $\theta\in\Theta$ and $\epsilon>0$ there exists $\delta>0$ so that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta':\|\theta-\theta'\|<\delta}\left\|\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)W_t^*\left(\theta|u\right)-\frac{1}{n}\sum_{t=1}^{n}L_b\left(t/n-u\right)W_t^*\left(\theta'|u\right)\right\|\right]\\
\leq&\frac{1}{n}\sum_{t=1}^{n}\left|L_b\left(t/n-u\right)\right|\mathbb{E}\left[\sup_{\theta':\|\theta-\theta'\|<\delta}\left\|W_t^*\left(\theta|u\right)-W_t^*\left(\theta'|u\right)\right\|\right]\\
=&\frac{\epsilon}{n}\sum_{t=1}^{n}\left|L_b\left(t/n-u\right)\right|=O\left(\epsilon\right).
\end{aligned}$$

Proof of (ii). Observe that $\sqrt{b/n}\sum_{t=1}^{n}L_b\left(t/n-u\right)W_{n,t}\left(\theta\left(t/n\right)\right)$ is a martingale with quadratic variation $Q_n=\frac{b}{n}\sum_{t=1}^{n}L_b^2\left(t/n-u\right)V_{n,t}\left(\theta\left(t/n\right)\right)$. To derive the limit of Q_n, write

$$\begin{aligned}
Q_n=&\frac{b}{n}\sum_{t=1}^{n}L_b^2\left(t/n-u\right)\mathbb{E}\left[V^*\left(\theta\left(t/n\right)|u\right)\right]\\
&+\frac{b}{n}\sum_{t=1}^{n}L_b^2\left(t/n-u\right)\left\{V_{n,t}\left(\theta\left(t/n\right)\right)-V_t^*\left(\theta\left(t/n\right)|u\right)\right\}\\
&+\frac{b}{n}\sum_{t=1}^{n}L_b^2\left(t/n-u\right)\left\{V_t^*\left(\theta\left(t/n\right)|u\right)-\mathbb{E}\left[V_t^*\left(\theta\left(t/n\right)|u\right)\right]\right\}.
\end{aligned}$$

For the first term, employing standard results for kernel averages together with the fact that $\theta \mapsto \mathbb{E}\left[V_t^*\left(\theta|u\right)\right]$ is continuous (because $V_t^*\left(\theta|u\right)$ is L_1-continuous),

$$\frac{b}{n}\sum_{t=1}^{n} L_b^2\left(t/n-u\right)\mathbb{E}\left[V_t^*\left(\theta\left(t/n\right)|u\right)\right] \to \int L^2\left(x\right)dx\mathbb{E}\left[V_t^*\left(\theta\left(u\right)|u\right)\right].$$

Applying arguments similar to those in the proof of Lemma 1.1(i) together with continuity of $v \mapsto \theta\left(v\right)$, L_1-continuity of $\theta \mapsto V_t^*\left(\theta|u\right)$ and L having compact support, we have for all n large enough,

$$\begin{aligned}&\frac{b}{n}\sum_{t=1}^{n} L_b^2\left(t/n-u\right)\mathbb{E}\left[\left\|V_{n,t}\left(\theta\left(t/n\right)\right)-V_t^*\left(\theta\left(t/n\right)|u\right)\right\|\right]\\ &\leq \frac{b}{n}\sum_{t=1}^{n} L_b^2\left(t/n-u\right)\sup_{\left\|\theta-\theta(u)\right\|<\epsilon}\mathbb{E}\left[\left\|V_{n,t}\left(\theta\right)-V_t^*\left(\theta|u\right)\right\|\right]=o\left(1\right),\end{aligned}$$

and

$$\begin{aligned}&\frac{b}{n}\sum_{t=1}^{n} L_b^2\left(t/n-u\right)\left\{V_t^*\left(\theta\left(t/n\right)|u\right)-\mathbb{E}\left[V_t^*\left(\theta\left(t/n\right)|u\right)\right]\right\}\\ &\leq \frac{b}{n}\sum_{t=1}^{n} L_b^2\left(t/n-u\right)\sup_{\left\|\theta-\theta(u)\right\|<\epsilon}\mathbb{E}\left[\left\|V_t^*\left(\theta|u\right)-\mathbb{E}\left[V_t^*\left(\theta|u\right)\right]\right\|\right]=o\left(1\right).\end{aligned}$$

The result now follows if the Lindeberg condition is satisfied, c.f. Brown (1971). But, as $nb \to \infty$, with $m_{n,t}\left(\theta\right)=\sqrt{b/n}L_b\left(t/n-u\right)W_t^*\left(\theta|u\right)$,

$$\begin{aligned}&\sum_{t=1}^{n}\left\|m_{n,t}\left(\theta\left(t/n\right)\right)\right\|^2 1\left(\left\|m_{n,t}\left(\theta\left(t/n\right)\right)\right\|>\varepsilon\right)\\ \leq&\sum_{t=1}^{n}\left(\left\|m_{n,t}\left(\theta\left(t/n\right)\right)\right\|^2-\left\|m_t^*\left(\theta\left(u\right)|u\right)\right\|^2\right)1\left(\left\|m_{n,t}\left(\theta\left(t/n\right)\right)\right\|>\varepsilon\right)\\ &+\sum_{t=1}^{n}\left\|m_t^*\left(\theta\left(u\right)|u\right)\right\|^2 1\left(\left\|m_{n,t}\left(\theta\left(t/n\right)\right)\right\|>\varepsilon,\left\|m_t^*\left(\theta\left(u\right)|u\right)\right\|\leq\varepsilon/\sqrt{2}\right)\\ &+\sum_{t=1}^{n}\left\|m_t^*\left(\theta\left(u\right)|u\right)\right\|^2 1\left(\left\|m_t^*\left(\theta\left(u\right)|u\right)\right\|>\varepsilon/\sqrt{2}\right).\end{aligned}$$

Recycling the arguments used in the analysis of Q_n, it follows that the first and third terms are $o_p\left(1\right)$. Similarly, the convergence of the second term is obtained with the following

inequality and Markov's inequality:

$$\sum_{t=1}^{n} \left\| m_t^* \left(\theta \left(u \right) | u \right) \right\|^2 1 \left(\left\| m_{n,t} \left(\theta \left(t/n \right) \right) \right\| > \varepsilon, \left\| m_t^* \left(\theta \left(u \right) | u \right) \right\| \leq \varepsilon / \sqrt{2} \right)$$
$$\leq \frac{\varepsilon^2}{2} \sum_{t=1}^{n} 1 \left(\left\| m_{n,t} \left(\theta \left(t/n \right) \right) \right\|^2 - \left\| m_t^* \left(\theta \left(u \right) | u \right) \right\|^2 > \varepsilon^2 / 2 \right).$$

Proof of (iii). Assume w.l.o.g. that $\mathbb{E}\left[W_t^* \right] = 0$ and then use

$$Var \left(A_n \right) \leq \frac{1}{n} \sum_{t_1, t_2 = 1}^{n} \left| L_b \left(t_1/n - u \right) \right| \left| L_b \left(t_2/n - u \right) \right| \left| cov \left(W_{t_1}^*, W_{t_2}^* \right) \right|$$
$$\leq \frac{\bar{L}}{\left(nb \right)^2} \sum_{t_1, t_2 = 1}^{n} \left| L \left(\frac{t_1/n - u}{b} \right) \right| \left| cov \left(W_{t_1}, W_{t_2} \right) \right| = O \left(\frac{1}{nb} \right).$$

□

1.8.2 Proofs: Main results

Proof of Theorem 1.1. We first note that $f \left(Z_t^* \left(\theta | u \right), \varepsilon_t; \theta \right)$ is stationary and ergodic because f is a measurable function of $\left(Z_t^* \left(\theta | u \right), \varepsilon_t \right)$. Moreover, with $p_Z = p/\left(r + 1 \right)$,

$$\mathbb{E} \left[\sup_{\theta \in \Theta} \left\| f \left(Z_{n,t} \left(\theta \right), \varepsilon_t; \theta \right) - f \left(Z_t^* \left(\theta | u \right), \varepsilon_t; \theta \right) \right\|^{p_Z} \right]^{1/p_Z}$$
$$\leq C \mathbb{E} \left[\left(1 + \left\| Z_{n,t} \left(\theta \right) \right\|^{pr/(r+1)} + \left\| Z_t^* \left(\theta | u \right) \right\|^{pr/(r+1)} \right) \left\| Z_{n,t} \left(\theta \right) - Z_t^* \left(\theta | u \right) \right\|^{p_Z} \right]^{1/p_Z}$$
$$\leq C \mathbb{E} \left[\left\| Z_{n,t} \left(\theta \right) - Z_t^* \left(\theta | u \right) \right\|^p \right]^{1/p} \leq C \left(\left| t/n - u \right|^q + 1/n^q + \rho^t \right).$$

where we have employed Hölder's inequality. □

Proof of Theorem 1.2. We apply Lemma 1.1(i) to $Q_n \left(\alpha | u \right) = \frac{1}{n} \sum_{t=1}^{n} K_b \left(t/n - u \right) \ell_{n,t} \left(D_{n,t} \alpha \right)$ and obtain $\sup_{\alpha \in \mathcal{A}} \left| Q_n \left(\alpha | u \right) - Q^* \left(\alpha | u \right) \right| = o_P \left(1 \right)$ for $Q^* \left(\alpha | u \right) = \int K \left(v \right) \mathbb{E} \left[\ell_t^* \left(D \left(v \right) \alpha | u \right) \right] dv$. Now, observe that for any $\alpha = \left(\alpha_1, ..., \alpha_{m+1} \right)$ with $\alpha_i \neq 0$ for some $i \geq 2$, the polynomial $v \mapsto D \left(v \right) \alpha$ is non-constant almost everywhere. Thus, for any $\alpha \neq \alpha^* = \left(\theta \left(u \right), 0, ..., 0 \right)$, $D \left(v \right) \alpha \neq \theta \left(u \right) = D \left(v \right) \alpha^*$ for almost all $v \in [0, 1]$ and so by Assumption 1.3(iii), for almost every v, $\mathbb{E} \left[\ell_t^* \left(D \left(v \right) \alpha | u \right) \right] < \mathbb{E} \left[\ell_t^* \left(\theta \left(u \right) | u \right) \right] = \mathbb{E} \left[\ell_t^* \left(D \left(v \right) \alpha^* | u \right) \right]$. Since $K \left(\cdot \right) \geq 0$ this in turn

implies that

$$Q^*(\alpha|u) = \int K(v)\,\mathbb{E}\left[\ell_t^*\left(D(v)\,\alpha|u\right)\right]dv < \int K(v)\,\mathbb{E}\left[\ell_t^*\left(D(v)\,\alpha_0|u\right)\right]dv = Q^*(\alpha^*|u).$$

Finally, by the dominated convergence theorem together with Assumption 1.3(ii) $\alpha \mapsto Q^*(\alpha|u)$ is continuous. This proves $\hat{\alpha} \to^p \alpha^*$, c.f. Theorem 2.1 in Newey and McFadden (1994). □

Proof of Theorem 1.3. From Theorem 1.2 we know that $\hat{\alpha} \to^p \alpha^* := (\theta(u), 0, ..., 0)$. It is easily checked that the limit is situated in the interior of $\mathcal{A}$ and so w.p.a.1. so will $\hat{\alpha}$. As a consequence, $\hat{\alpha}$ will satisfy (1.3) w.p.a.1. Adding and subtracting $S_n(u)$ and then rearranging yields

$$0 = \sqrt{nb}S_n(u) + H_n(\bar{\alpha}|u)\sqrt{nb}\left(\hat{\alpha} - \alpha_0 - H_n^{-1}(\bar{\alpha}|u)\left\{S_n(\alpha_0|u) - S_n(u)\right\}\right).$$

Here, $H_n^{-1}(\bar{\alpha}|u)$ is well-defined w.p.a.1 since, as shown below, it converges towards an invertible matrix. The claimed asymptotic result now follows if we can verify the claims of eqs. (1.6)-(1.7):

Proof of eq. (1.6). First note that $\sqrt{nb}S_n(u) = \sqrt{\frac{b}{n}}\sum_{t=1}^n L_b(t/n - u) \otimes s_{n,t}(\theta(t/n))$ with $L(u) = K(u)D(u)$. The result now follows from Lemma 1.1(ii) under Assumption 1.5.

Proof of eq. (1.7)(i). We can write $H_n(\beta|u) = \frac{1}{n}\sum_{t=1}^n L_b(t/n-u)\otimes h_{n,t}(D_{n,t}(u)\,\beta)$ with $L(u) = K(u)D(u)\,D(u)'$. Applying Lemma 1.1(i) in conjunction with Assumption 1.6, we then obtain $\sup_{\alpha\in\mathcal{B}(\epsilon)}\|H_n(\alpha|u) - \mathbb{K}_1 \otimes H(D(v)\,\alpha|u)\| = o_p(1)$, where $H(\theta|u) = \mathbb{E}\left[h_t^*(\theta|u)\right]$ is continuous w.r.t. θ and $\mathcal{B}(\epsilon) = \{\alpha : \|\alpha - \alpha^*\| < \epsilon\}$ for some small $\epsilon > 0$. Thus, given that $\bar{\alpha} \to^p \alpha^*$, $H_n(\bar{\alpha}|u) \to^p \mathbb{K}_1 \otimes H(\theta(u)|u)$. Finally, note here that since K is a probability density function, $\mathbb{K}_1$ is invertible, while $H(\theta(u)|u) = H(u)$ is invertible by assumption.

Proof of eq. (1.7)(ii). First observe that $D_{n,t}(u)\,\alpha_0 = \theta_u^*(t/n)$ where $\theta_u^*(t/n)$ was defined in (1.2). Now, employ the mean-value theorem twice to obtain that, for some $\bar{\theta}_{n,t}$

lying between $\theta_u^*(t/n)$ and $\theta(t/n)$ and some $u_{n,t} \in [t/n, u]$,

$$\begin{aligned} b_{n,t} =& h_{n,t}\left(\bar{\theta}_{n,t}\right)\left\{\theta_u^*(t/n) - \theta(t/n)\right\} = -h_{n,t}\left(\bar{\theta}_{n,t}\right)\frac{\theta^{(m+1)}(u_{n,t})}{(m+1)!}(t/n-u)^{m+1} \\ =& -(t/n-u)^{m+1} h_{n,t}\left(\theta\left(\frac{t}{n}\right)\right)\frac{\theta^{(m+1)}\left(\frac{t}{n}\right)}{(m+1)!} \\ &+\left\{h_{n,t}\left(\theta\left(\frac{t}{n}\right)\right)\theta^{(m+1)}\left(\frac{t}{n}\right) - h_{n,t}\left(\bar{\theta}_{n,t}\right)\theta^{(m+1)}(u_{n,t})\right\}\frac{\left(\frac{t}{n}-u\right)^{m+1}}{(m+1)!}. \end{aligned}$$

The first term is locally stationary and so by the same arguments as in the proof of Lemma 1.1(ii),

$$\begin{aligned} &\frac{b^{m+1}}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} h_{n,t}(\theta(t/n))\frac{\theta^{(m+1)}(t/n)}{(m+1)!} \\ &= b^{m+1}\left\{\mu_1 \otimes H(u)\frac{\theta^{(m+1)}(u)}{(m+1)!} + o_p(1)\right\}. \end{aligned}$$

Next, observe that for $|t/n-u| \leq Cb$, $\left\|\bar{\theta}_{n,t} - \theta(t/n)\right\| \leq \left\|\theta_u^*(t/n) - \theta(t/n)\right\| \leq \tilde{C}b^{m+1}$and so, using the ULS property of $h_{n,t}(\theta)$,

$$\begin{aligned} &\sup_{n,t}\mathbb{E}\left[\left\|h_{n,t}\left(\theta\left(\frac{t}{n}\right)\right) - h_{n,t}\left(\bar{\theta}_{n,t}\right)\right\|\right] \\ &\leq C\left(b^q\left|\frac{\frac{t}{n}-u}{b}\right|^q + 1/n^q\right) + \sup_{\|\theta-\theta'\|\leq\tilde{C}b^{m+1}}\mathbb{E}\left[\left\|h_t^*(\theta|u) - h_t^*\left(\theta'|u\right)\right\|\right] \to 0, \end{aligned}$$

as $n \to \infty$. Similarly, $\sup_{n,t}\left\|\theta^{(m+1)}\left(\frac{t}{n}\right) - \theta^{(m+1)}(u_{n,t})\right\| \to 0$ as $n \to \infty$ using the uniform continuity of $\theta^{(m+1)}(\cdot)$. These two results show that the remainder term is $o_p(1)$. □

Proof of Theorem 1.4. Proof proceeds exactly as the one of Theorem 1.3, but we now establish that $b_{n,t}$ satisfies eq. (1.9). Using a second-order expansion w.r.t. θ followed by a second-order Taylor expansion w.r.t. u, we obtain

$$\begin{aligned} b_{n,t} =& -h_{n,t}(\theta(t/n))\left[\frac{\theta^{(m+1)}(u)}{(m+1)!}\{t/n-u\}^{m+1} + \frac{\theta^{(m+2)}(u_{n,t})}{(m+2)!}\{t/n-u\}^{m+2}\right]. \\ &+\frac{1}{2}\sum_{i=1}^{d_\theta}\frac{\theta_i^{(m+1)}(u_{n,t})}{(m+1)!}\frac{\partial h_{n,t}\left(\bar{\theta}_{n,t}\right)}{\partial\theta_i}\frac{\theta^{(m+1)}(u_{n,t})}{(m+1)!}\{t/n-u\}^{2m+2}. \end{aligned}$$

For the first term, write

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} h_{n,t}(\theta(t/n)) \\
=&\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} [h_{n,t}(\theta(t/n)) - h_{n,t}(\theta(u))] \\
&+\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} h_t^*(\theta(u)|u) \\
&+\frac{b}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+2} \partial_u h_t^*(\theta(u)|u) \\
&+\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} \begin{bmatrix} h_{n,t}(\theta(u)) - h_t^*(\theta(u)|u) \\ -\partial_u h_t^*(\theta(u)|u)(t/n-u) \end{bmatrix}
\end{aligned}$$

where, by Lemma 1.1(iii) together with Assumption 1.9 and Lemma 1.1(i), respectively,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} [h_{n,t}(\theta(t/n)) - h_{n,t}(\theta(u))] \\
&= b\mu_2 \sum_{i=1}^{d_\theta} \theta_i^{(1)}(u) \partial_{\theta_i} H(u) + o_P(b), \\
&\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+1} h_t^*(\theta(u)|u) = \mu_1 H(u) + o_P\left(1/\sqrt{nb}\right), \\
&\frac{1}{n}\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left(\frac{t/n-u}{b}\right)^{m+2} \partial_u h_t^*(\theta(u)|u) = \mu_2 \partial_u H(u) + o_P(1),
\end{aligned}$$

while, using Assumption 1.7,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} |K_{n,t}(u)| \|D_{n,t}(u)\| \left|\frac{t/n-u}{b}\right|^{m+1} \mathbb{E}\left[\left\| \begin{array}{l} h_{n,t}(\theta(u)) - h_t^*(\theta(u)|u) \\ -\partial_u h_t^*(\theta(u)|u)(t/n-u) \end{array} \right\|\right] \\
\leq&\frac{1}{n}\sum_{t=1}^{n} |K_{n,t}(u)| \|D_{n,t}(u)\| \left|\frac{t/n-u}{b}\right|^{m+1} C\left(1/n^q + \rho^t\right) \\
&+\frac{1}{n}\sum_{t=1}^{n} |K_{n,t}(u)| \|D_{n,t}(u)\| \left|\frac{t/n-u}{b}\right|^{m+1} \mathbb{E}\left[\left\| \begin{array}{l} h_t^*(\theta(u)|t/n) - h_t^*(\theta(u)|u) \\ -\partial_u h_t^*(\theta(u)|u)(t/n-u) \end{array} \right\|\right] \\
=&O\left(n^{-q}\right) + O\left(1/\sqrt{nb}\right) + o(b).
\end{aligned}$$

For the second term and the third term ($i = 1, ..., d_\theta$), copying the arguments from the proof

of eq. (1.7)(ii),

$$\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left\{ \frac{t/n-u}{b} \right\}^{m+2} h_{n,t}\left(\bar{\theta}_{n,t}\right) \frac{\theta^{(m+2)}(u_{n,t})}{(m+2)!} = \mu_2 H(u) \frac{\theta^{(m+2)}(u)}{(m+2)!} + o_P(1),$$

$$\sum_{t=1}^{n} K_{n,t}(u) D_{n,t}(u)' \left\{ \frac{t/n-u}{b} \right\}^{2m+2} \frac{\theta_i^{(m+1)}(u_{n,t})}{(m+1)!} \frac{\partial h_{n,t}\left(\bar{\theta}_{n,t}\right)}{\partial \theta_i} \frac{\theta^{(m+1)}(u_{n,t})}{(m+1)!}$$
$$= \mu_{m+2} \frac{\theta_i^{(m+1)}(u)}{\{(m+1)!\}^2} \partial_{\theta_i} H(u) \theta^{(m+1)}(u) + o_P(1).$$

where the second result uses Assumption 1.8. Collecting terms yield the claimed result. □

Proof of Theorem 1.5. All arguments in the proofs of Theorems 1.3-1.4 remain valid except for the following two adjustments: First, we now have an additional bias component $R_n(u)$, as defined in eq. (1.12), which we have to show is negligible. Second, the variance component now takes the form $S_n(u) = \frac{1}{n}\sum_{t=1}^{n} K_b(t/n-u)\bar{s}_{n,t}$. But under Assumption 1.10(iii),

$$\mathbb{E}[\|R_n(u)\|] \leq \frac{1}{n}\sum_{t=1}^{n} K_b(t/n-u)\mathbb{E}[\|s_{n,t}(\theta(t/n)) - \bar{s}_{n,t}\|] = O(1/n^{q_s}),$$

and so $R_n(u) = O_p(1/n^{q_s}) = o_p(b^2)$ where the second equality follows from the added bandwidth condition $b^2 n^{q_s} \to \infty$. Moreover, it is easily checked that the arguments used in the proof of Lemma 1.1(iii) carries over to the redefined version of $S_n(u)$ under Assumption 1.10(i)-(ii). □

Proof of Theorem 1.6. With $p_Z = p_W/(r_W+1)$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\Theta} \|Z_{n,t}(\theta) - Z_t^*(\theta|t/n)\|^{p_Z}\right]^{1/p_Z} \\
=&\mathbb{E}\left[\sup_{\theta\in\Theta} \left\|G(W_{n,t-1}(\theta), Z_{n,t-1}(\theta);\theta) - G\left(W_{t-1}^*(\theta|t/n), Z_{t-1}^*(\theta|t/n);\theta\right)\right\|^{p_Z}\right]^{1/p_Z} \\
\leq& C\mathbb{E}\left[\sup_{\theta\in\Theta} \left(1 + \|W_{n,t-1}(\theta)\|^{p_Z r_W} + \left\|W_{t-1}^*(\theta|t/n)\right\|^{p_Z r_W}\right) \left\|W_{n,t-1}(\theta) - W_{t-1}^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z} \\
&+ \beta\mathbb{E}\left[\sup_{\theta\in\Theta} \left\|Z_{n,t-1}(\theta) - Z_{t-1}^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z}.
\end{aligned}$$

The first term is less than or equal to $C\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|W_{n,t-1}(\theta) - W_{t-1}^*(\theta|t/n)\right\|^{p_W}\right]^{1/p_W}$ for

some $C < \infty$ by applying Hölder's inequality. Then,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_{n,t}(\theta) - Z_t^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z} \\
\leq & C\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|W_{n,t-1}(\theta) - W_{t-1}^*(\theta|t/n)\right\|^{p_W}\right]^{1/p_W} + \beta\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_{n,t-1}(\theta) - Z_{t-1}^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z} \\
&\vdots \\
\leq & C\sum_{i=1}^{t}\beta^i\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|W_{n,t-i}(\theta) - W_{t-i}^*(\theta|t/n)\right\|^{p_W}\right]^{1/p_W} + \beta^t\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|z - Z_0^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z} \\
\leq & C\sum_{i=1}^{t}\beta^i\left((i+1)/n + \rho^{t-i}\right) + \beta^t\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|z - Z_0^*(\theta|t/n)\right\|^{p_Z}\right]^{1/p_Z} \leq C\left(1/n + \rho^t\right).
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|W_{n,t-i}(\theta) - W_{t-i}^*(\theta|t/n)\right\|^{p_W}\right]^{1/p_W} \\
\leq & \mathbb{E}\left[\sup_{\theta\in\Theta}\left\|W_{n,t-i}(\theta) - W_{t-i}^*(\theta|(t-i)/n)\right\|^{p_W}\right]^{1/p_W} \\
&+ \mathbb{E}\left[\left\|W_{t-i}^*(\theta|(t-i)/n) - W_{t-i}^*(\theta|t/n)\right\|^{p_W}\right]^{1/p_W} \\
\leq & C_1\left(1/n + \rho^{t-i}\right) + C_2(i/n).
\end{aligned}$$

The proof of $Z_t^*(\theta|u)$ being well-defined and stationary with $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_t^*(\theta|u)\right\|^{p_Z}\right] < \infty$ and $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_t^*(\theta|u) - Z_t^*(\theta|v)\right\|^{p_Z}\right] \leq C|u-v|^q$ proceeds in the same way.

To show the second part write, with $\tilde{p}_Z = p_W/r_\theta$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|Z_{n,t}(\theta) - Z_{n,t}\right\|^{\tilde{p}_Z}\right]^{1/\tilde{p}_Z} \\
&= \mathbb{E}\left[\sup_{\theta\in\Theta}\left\|G\left(W_{n,t-1}(\theta), Z_{n,t-1}(\theta); \theta\right) - G\left(W_{n,t-1}(\theta), Z_{n,t-1}; \theta(t/n)\right)\right\|^{\tilde{p}_Z}\right]^{1/\tilde{p}_Z} \\
&\leq C\mathbb{E}\left[\sup_{\theta\in\Theta}\left(1 + \left\|W_{n,t-1}(\theta)\right\|^{p_W}\right)\right]^{1/p_Z}\left\|\theta - \theta(t/n)\right\| + \beta\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|Z_{n,t-1}(\theta) - Z_{n,t-1}\right\|^{p_Z}\right]^{1/p_Z} \\
&\vdots \\
&\leq C\sum_{i=1}^{t}\beta^i\left\|\theta - \theta((t-i)/n)\right\|.
\end{aligned}$$

Substituting in $\theta = \theta(t/n)$ and using it is continuously differentiable,

$$\sum_{i=1}^{t} \beta^i \left\| \theta(t/n) - \theta((t-i)/n) \right\| \leq C \sum_{i=1}^{t} \beta^i i/n \leq C/n.$$

To show the final part, write

$$\begin{aligned}\partial_u Z_t^*(\theta|u) =& \partial_z G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \partial_u Z_{t-1}^*(\theta|u) \\ &+ \partial_w G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \partial_u W_{t-1}^*(\theta|u).\end{aligned}$$

By assumption, $\left\| \partial_z G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \right\| \leq \beta < 1$. Moreover, by applying Hölder's inequality, with $p_Z = p_W \alpha_W / (p_W + r_W \alpha_W)$,

$$\begin{aligned}&\mathbb{E}\left[\left\| \partial_w G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \partial_u W_{t-1}^*(\theta|u) \right\|^{p_Z} \right]^{1/p_Z} \\ &\leq C\mathbb{E}\left[\left(1 + 2\left\| W_{t-1}^*(\theta|u) \right\|^{p_W r_W \alpha_W/(p_W + r_W \alpha_W)}\right) \left\| \partial_u W_{t-1}^*(\theta|u) \right\|^{p_Z} \right]^{1/p_Z} \\ &\leq C\mathbb{E}\left[\left\| \partial_u W_{t-1}^*(\theta|u) \right\|^{\alpha_W} \right]^{1/\alpha_W} < \infty.\end{aligned}$$

It implies that $\partial_u Z_t^*(\theta|u)$ has a finite $\alpha_Z = p_W \alpha_W / (p_W + r_W \alpha_W)$-th moment. Also,

$$\begin{aligned}&\left\| Z_t^*(\theta|u+b) - Z_t^*(\theta|u) - \partial_u Z_t^*(\theta|u)\, b \right\| \\ =& \left\| G\left(W_{t-1}^*(\theta|u+b), Z_{t-1}^*(\theta|u+b); \theta\right) - G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) - \partial_u Z_t^*(\theta|u)\, b \right\| \\ \leq& \beta \left\| Z_{t-1}^*(\theta|u+b) - Z_{t-1}^*(\theta|u) - \partial_u Z_{t-1}^*(\theta|u)\, b \right\| \\ &+ C\left(1 + 2\left\| W_{t-1}^*(\theta|u) \right\|^{r_W}\right) \left\| W_{t-1}^*(\theta|u+b) - W_{t-1}^*(\theta|u) - \partial_u W_{t-1}^*(\theta|u)\, b \right\| \\ &+ b \left\| \partial_z G\left(\bar{W}_{t-1}^*(\theta|u), \bar{Z}_{t-1}^*(\theta|u); \theta\right) - \partial_z G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \right\| \partial_u Z_{t-1}^*(\theta|u) \\ &+ b \left\| \partial_w G\left(\bar{W}_{t-1}^*(\theta|u), \bar{Z}_{t-1}^*(\theta|u); \theta\right) - \partial_w G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \right\| \partial_w Z_{t-1}^*(\theta|u)\end{aligned}$$

where $\left(\bar{W}_{t-1}^*(\theta|u), \bar{Z}_{t-1}^*(\theta|u)\right)$ is situated on the line connecting $\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u)\right)$ and $\left(W_{t-1}^*(\theta|u+b), Z_{t-1}^*(\theta|u+b)\right)$. Since $\left\{ \left\| \partial_w G\left(\bar{W}_{t-1}^*(\theta|u), \bar{Z}_{t-1}^*(\theta|u); \theta\right) \right\|^{p_W/r_W} \right\}$ is uniformly integrable,

$$\mathbb{E}\left[\left\| \partial_w G\left(\bar{W}_{t-1}^*(\theta|u), \bar{Z}_{t-1}^*(\theta|u); \theta\right) - \partial_w G\left(W_{t-1}^*(\theta|u), Z_{t-1}^*(\theta|u); \theta\right) \right\|^{p_W/r_W} \right] \to 0$$

as $b \to 0$. This completes the proof. □

1.8.3 Proofs: Examples

Proof of Corollary 1.3. We apply our theory with $\Theta = \mathbb{R}^{d^2q}$ since the least-squares criterion used for estimation is concave in θ, c.f. the comments following Assumptions 1.1-1.3. We first show that $X_{n,t}$ is locally stationary with $p \geq 2$ moments when $\mathbb{E}\left[\|\varepsilon_t\|^2\right] < \infty$. Without loss of generality, we here only provide a proof for $Y_{n,t} = \Phi(t/n) Y_{n,t-1} + \Sigma(t/n) \varepsilon_t$ to be locally stationary under the following conditions: $\Phi(u)$ and $\Sigma(u)$ are twice continuously differentiable; and all eigenvalues of $\Phi(u)$ lie inside the unit circle for $u \in [0,1]$. We first verify the conditions of Theorems 1.7 for $G(x, e, \vartheta) := \Phi x + \Sigma e$ where $\vartheta = (\Phi, \Sigma)$ with Φ having all eigenvalues inside the unit circle. First,

$$\mathbb{E}\left[\|G(0, \varepsilon_t; u)\|^p\right] \leq \|\Sigma(u)\|^p \mathbb{E}\left[\|\varepsilon_t\|^p\right] < \infty;$$

second, for all $x, x' \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|G(x, \varepsilon_t; \vartheta) - G\left(x', \varepsilon_t; \vartheta\right)\right\|^p\right]^{1/p} \leq \left\|\Phi\left(x - x'\right)\right\| \leq \frac{\|\Phi(x - x')\|}{\|x - x'\|} \left\|x - x'\right\| \leq \rho \left\|x - x'\right\|,$$

where $\rho = \sup_{x \neq 0} \frac{\|\Phi x\|}{\|x\|} < 1$ since all eigenvalues of Φ lie inside the unit circle; and for all ϑ, ϑ',

$$\begin{aligned}\mathbb{E}\left[\left\|G(x, \varepsilon_t; \vartheta) - G\left(x, \varepsilon_t; \vartheta'\right)\right\|^p\right]^{1/p} &= \left\|\Phi - \Phi'\right\| \|x\| + \left\|\Sigma - \Sigma'\right\| \mathbb{E}\left[\|\varepsilon_t\|^p\right]^{1/p} \\ &\leq C(1 + \|x\|) \left\|\vartheta - \vartheta'\right\|.\end{aligned}$$

Next, we verify that the log-likelihood and its derivatives are ULS: Observe that

$$\left\|\frac{\partial \ell_{n,t}(\theta)}{\partial Y_{n,t}}\right\| \leq 2\left(\|Y_{n,t}\| - \|\theta\| \|X_{n,t}\|\right), \quad \left\|\frac{\partial \ell_{n,t}(\theta)}{\partial X_{n,t}}\right\| = 2\left(\|Y_{n,t}\| - \|\theta\| \|X_{n,t}\|\right) \|\theta\|,$$

and so 1.1 applies with $r = 1$. For the score function, observe that $\bar{s}_{n,t} = 2X_{n,t}\Sigma(t/n)\varepsilon_t$ which is a Martingale difference with $\omega_{n,t} = 4X_{n,t}\Sigma(t/n)\varepsilon_t\varepsilon_t'(\Sigma(t/n))' X_{n,t}'$. Then,

$$\bar{s}_{n,t} = s_{n,t}(\theta(t/n)), \quad \left\|\frac{\partial \omega_{n,t}}{\partial X_{n,t}}\right\| = 4\|\Sigma(t/n)\varepsilon_t\|^2 \|X_{n,t}\|,$$

and so 1.1 applies with $r = 1$. The hessian is also ULS by similar arguments. Thus, with $\mathbb{E}\left[\|\varepsilon_t\|^2\right] < \infty$, all conditions for Theorem 1.3 hold.

To analyze the local constant estimator, first note that our proof of local stationarity also implies that $Y_t^*(u)$ is a GMC(p) and so we can apply Proposition 2 in Wu and Shao (2004) to obtain that the process is short-range dependent. With $p = 4$ this in turn implies that $h_t^*(\theta|u) = X_t^*(u) X_t^*(u)'$ satisfies Assumption 1.9. The derivative process takes the form $\partial_u Y_t^*(u) = \Phi^{(1)}(u) Y_{t-1}^*(u) + \Phi(u) \partial_u Y_{t-1}^*(u) + \Sigma^{(1)}(u) \varepsilon_t$. The joint process $Z_t^*(u) = \left(\partial_u Y_t^*(u)', \partial_u Y_t^*(u)'\right)'$ solves another VAR model whose stability condition is satisfied due to all eigenvalues of $\Phi(u)$ lying inside the unit circle. It now follows by Propostion 2.5 in Dahlhaus et al. (2017) and the remarks following this that the derivative process of the hessian, $\partial_u h_t^*(\theta|u) = 2\partial_u X_t^*(u) X_t^*(u)'$, satisfies Assumption 1.7. Finally, we note that the third-order derivatives of the log-likelihood are zero and so Assumption 1.8 is trivially satisfied. Thus, with $\mathbb{E}\left[\|\varepsilon_t\|^4\right] < \infty$, Theorem 1.4 applies to the local constant estimator. □

Proof of Corollary 1.4. Verification of all our general conditions for the stationary version, including identification and existence of relevant moments, follow from Kristensen and Rahbek (2005). For the analysis of the local linear estimator, what remains is to show local stationarity of the log-likelihood function, the conditional variance of the score function and the hessian. First, it follows from, e.g., Dahlhaus and Subba Rao (2006) that $W_{n,t}$ is LS(1, 1) with $\sup_{n,t} \mathbb{E}[W_{n,t}] < \infty$ and $\mathbb{E}[W_t^*(u)] < \infty$. Thus, local stationarity of the log-likelihood and its derivatives can be shown by verifying the conditions of Theorem 1.1. We have $\ell_{n,t}(\theta) = \log(\lambda_{n,t}(\theta)) + W_{n,t}/\lambda_{n,t}(\theta)$ with $\lambda_{n,t}(\theta) = \theta' V_{n,t}$ and $V_{n,t} = (1, W_{n,t-1}, ..., W_{n,t-q})'$. Here, $\lambda_{n,t}(\theta)$ is trivially ULS$(1, 1, \Theta)$ while

$$\left|\frac{\partial \ell_{n,t}(\theta)}{\partial W_{n,t}}\right| = \frac{1}{\lambda_{n,t}(\theta)} \leq \frac{1}{\omega} \leq \frac{1}{\delta_L}, \quad \left\|\frac{\partial \ell_{n,t}(\theta)}{\partial \lambda_{n,t}(\theta)}\right\| \leq \frac{1}{\lambda_{n,t}(\theta)} + \frac{W_{n,t}}{\lambda_{n,t}^2(\theta)} \leq \frac{1}{\delta_L} + \frac{W_{n,t}}{\delta_L \lambda_{n,t}(\theta)},$$

where

$$\frac{W_{n,t}}{\lambda_{n,t}(\theta)} = \frac{\theta(t/n)' V_{n,t}}{\theta' V_{n,t}} \varepsilon_t^2 \leq \frac{\sup_u \omega(u) + \sum_{i=1}^p \sup_u \alpha_i(u)}{\delta_L} \varepsilon_t^2. \tag{1.17}$$

Thus, $\ell_{n,t}(\theta)$ satisfies the conditions of Theorem 1.1 with $r = 0$ and $q = 1$. Next, we verify Assumption 1.5 with the score function $s_{n,t}(\theta) = (1 - W_{n,t}/\lambda_{n,t}(\theta)) \partial_\theta \lambda_{n,t}(\theta) / \lambda_{n,t}(\theta)$.

Here, $\partial_\theta \lambda_{n,t}(\theta) = V_{n,t}$ is trivially ULS$(1,1,\Theta)$. The process $\{s_{n,t}(\theta(t/n)), \mathcal{F}_{n,t-1}\}$ is a MGD and $\omega_{n,t}(\theta)$ takes the form

$$\omega_{n,t}(\theta) = \frac{\partial_\theta \lambda_{n,t}(\theta)(\partial_\theta \lambda_{n,t}(\theta))'}{\lambda_{n,t}^2(\theta)}(1 - W_{n,t}/\lambda_{n,t}(\theta))^2.$$

By combining (1.17) with $\sup_{\{\theta:\|\theta-\theta(u)\|<\epsilon\}} \|\partial_\theta \lambda_{n,t}(\theta)/\lambda_{n,t}(\theta)\| \leq p/\delta_L$,

$$\begin{aligned}
\left\| \frac{\partial \omega_{n,t}(\theta)}{\partial W_{n,t}} \right\| &= \frac{2\|\partial_\theta \lambda_{n,t}(\theta)\|^2}{\lambda_{n,t}^3(\theta)} |1 - W_{n,t}/\lambda_{n,t}(\theta)| \leq C\left(1 + \varepsilon_t^2\right), \\
\left\| \frac{\partial \omega_{n,t}(\theta)}{\partial \lambda_{n,t}(\theta)} \right\| &= \frac{2\|\partial_\theta \lambda_{n,t}(\theta)\|^2}{\lambda_{n,t}^3(\theta)} \left|1 - \frac{3W_{n,t}}{\lambda_{n,t}(\theta)} + \frac{4W_{n,t}^2}{\lambda_{n,t}^2(\theta)}\right| \leq C\left(1 + \varepsilon_t^2 + \varepsilon_t^4\right), \\
\left\| \frac{\partial \omega_{n,t}(\theta)}{\partial(\partial_\theta \lambda_{n,t}(\theta))} \right\| &= \frac{2\|\partial_\theta \lambda_{n,t}(\theta)\|}{\lambda_{n,t}^2(\theta)}(1 - W_{n,t}/\lambda_{n,t}(\theta))^2 \leq C\left(1 + \varepsilon_t^2 + \varepsilon_t^4\right),
\end{aligned}$$

and so $\omega_{n,t}(\theta)$ satisfies the conditions of Theorem 1.1 with $r = 0$ and $q = 1$. The hessian takes the form

$$h_{n,t}(\theta) = \frac{\partial_\theta \lambda_{n,t}(\theta)\partial_\theta \lambda_{n,t}(\theta)'}{\lambda_{n,t}^2(\theta)}\left[\frac{2W_{n,t}}{\lambda_{n,t}(\theta)} - 1\right],$$

and recycling the inequalities established above it follows that the hessian is also ULS(1,1,Θ). This verifies the conditions for Theorem 1.3.

For the analysis of the local constant estimator, observe that $h_t^*(\theta(u)|u)$ is GMC(p) for some $p > 0$, and so Lemma 1 and Proposition 2 in Wu and Shao (2004) imply that the process is short-range dependent and so satisfies Assumption 1.9. Next, to verify Assumption 1.7, we apply Proposition 2.5(ii) in Dahlhaus et al. (2017): Under their assumptions, the derivative process takes the form

$$\partial_u h_t^*(\theta|u) = \frac{\partial h_t^*(\theta|u)}{\partial W_t^*(u)}\partial_u W_t^*(u) + \frac{\partial h_t^*(\theta|u)}{\partial \lambda_t^*(\theta|u)}\partial_u \lambda_t^*(\theta|u) + \frac{\partial h_t^*(\theta|u)}{\partial[\partial_\theta \lambda_t^*(\theta|u)]}\partial_{\theta u}\lambda_t^*(\theta|u),$$

where $\partial_u \lambda_t^*(\theta|u) = \theta' \partial_u V_t^*(u)$, $\partial_{\theta u}\lambda_t^*(\theta|u) = \partial_u V_t^*(u) = \left(0, \partial_u W_{t-1}^*(u), ..., \partial_u W_{t-q}^*(u)\right)'$, and the derivative process $\partial_u W_t^*(u)$ is given in Section 1.5. We note that the first partial derivative is bounded by a constant C and the remaining two partial derivatives are bounded by $C\left(1 + \varepsilon_t^2\right)$. Also, Proposition 3.1 in Subba Rao (2006) implies that $W_t^*(u)$ is time-differentiable in the L_1-sense at u. By employing the proof of Theorem 1.1, we have that $h_t^*(\theta(u)|u)$ is time-differentiable in the L_1-sense. Finally, the third-order derivatives takes

the form

$$\frac{\partial h_{n,t}(\theta)}{\partial \theta_i} = -2\frac{\partial_{\theta_i}\lambda_{n,t}(\theta)\,\partial_\theta\lambda_{n,t}(\theta)\,\partial_\theta\lambda_{n,t}(\theta)'}{\lambda_{n,t}^3(\theta)}\left[\frac{3W_{n,t}}{\lambda_{n,t}(\theta)}-1\right];$$

recycling the inequalities derived above, we conclude that these are ULS$(1,1)$. □

Proof of Corollary 1.5. We here verify the conditions of Theorem 1.5. First, using the results of Subba Rao (2006), we have that $(W_{n,t},\lambda_{n,t})$ is LS$(1,1)$ with $\sup_{n,t}\mathbb{E}[W_{n,t}]<\infty$ and $\sup_u \mathbb{E}[W_t^*(u)]<\infty$. Next, observe that the likelihood and score and take the same form as in the proof of Corollary 1.4. The volatility process is now given by $\lambda_{n,t}(\theta)=\omega+\alpha W_{n,t-1}+\beta\lambda_{n,t-1}(\theta)$, and its first and second-order derivatives w.r.t. θ take the form $\partial_\theta\lambda_{n,t}(\theta)=(\partial_\omega\lambda_{n,t}(\theta),\partial_\alpha\lambda_{n,t}(\theta),\partial_\beta\lambda_{n,t}(\theta))'$ and $\partial^2_{\theta\theta}\lambda_{n,t}(\theta)=\left(\partial^2_{\omega\beta}\lambda_{n,t}(\theta),\partial^2_{\alpha\beta}\lambda_{n,t}(\theta),\partial^2_{\beta\beta}\lambda_{n,t}(\theta)\right)'$ where

$$\partial_\omega\lambda_{n,t}(\theta)=1+\beta\partial_\omega\lambda_{n,t-1}(\theta),\ \partial_\alpha\lambda_{n,t}(\theta)=W_{n,t-1}+\beta\partial_\alpha\lambda_{n,t-1}(\theta),$$
$$\partial_\beta\lambda_{n,t}(\theta)=\lambda_{n,t-1}(\theta)+\beta\partial_\beta\lambda_{n,t-1}(\theta),$$

and

$$\partial^2_{\omega\beta}\lambda_{n,t}(\theta)=\partial_\omega\lambda_{n,t-1}(\theta)+\beta\partial^2_{\omega\beta}\lambda_{n,t-1}(\theta),\ \partial^2_{\alpha\beta}\lambda_{n,t}(\theta)=\partial_\alpha\lambda_{n,t-1}(\theta)+\beta\partial^2_{\alpha\beta}\lambda_{n,t-1}(\theta),$$
$$\partial^2_{\beta\beta}\lambda_{n,t}(\theta)=2\partial_\beta\lambda_{n,t-1}(\theta)+\beta\partial^2_{\beta\beta}\lambda_{n,t-1}(\theta).$$

We the proceed to show that $\lambda_{n,t}(\theta)$, $\partial_\theta\lambda_{n,t}(\theta)$ and $\partial^2_{\theta\theta}\lambda_{n,t}(\theta)$ are ULS and establish bounds for the following ratios: $\lambda_{n,t}/\lambda_{n,t}(\theta)$, $\|\partial_\theta\lambda_{n,t}(\theta)\|/\lambda_{n,t}(\theta)$ and $\left\|\partial^2_{\theta\theta}\lambda_{n,t}(\theta)\right\|/\lambda_{n,t}(\theta)$. Given that $\beta<1$, it is easily checked using Theorem 1.6(i) that $\lambda_{n,t}(\theta)$, $\partial_\theta\lambda_{n,t}(\theta)$ and $\partial^2_{\theta\theta}\lambda_{n,t}(\theta)$ are ULS$(1,1,\Theta)$. For example, $\lambda_{n,t}(\theta)=\omega+\alpha W_{n,t-1}+\beta\lambda_{n,t-1}(\theta)$ satisfies the conditions of Theorem 1.6 with $r_W=0$ and $r_\theta=1$. To establish the desired bound, first note that, with $\bar\omega=\sup_u\omega(u)$, $\bar\alpha=\sup_u\alpha(u)$ and $\bar\beta=\sup_u\beta(u)$,

$$\lambda_{n,t}\le\bar\omega+\bar\alpha Y^2_{n,t-1}+\bar\beta\lambda_{n,t-1}\le\cdots\le\frac{\bar\omega}{1-\bar\beta}+\sum_{i=1}^{t}\bar\beta^i W_{n,t-i}+\lambda_0.$$

We can now apply the same arguments as on p. 62 in Francq and Zakoïan (2004) to show

there exists constants $c < \infty$ and $0 < \rho < 1$ such that for all $\theta \in \Theta$ and some $r < 1$,

$$\frac{\lambda_{n,t}}{\lambda_{n,t}(\theta)} \leq c \sum_{i=0}^{t} \rho^{ri} \bar{W}_{n,t-i}^{r},$$

where $\bar{W}_{n,t} := \bar{\omega} + \bar{\alpha} W_{n,t-1}$ satisfies $\sup_{n,t} \mathbb{E}\left[\bar{W}_{n,t}^{r}\right] < \infty$. Similarly, again copying arguments from Francq and Zakoïan (2004),

$$\frac{\|\partial_\theta \lambda_{n,t}(\theta)\|}{\lambda_{n,t}(\theta)} \leq C, \frac{\left\|\partial_{\theta\theta}^2 \lambda_{n,t}(\theta)\right\|}{\lambda_{n,t}(\theta)} \leq C.$$

It now follows by the same arguments as in the proof of Corollary 1.4 that Assumptions 1.3-1.4 are satisfied.

Next, we verify Assumption 1.10 with

$$\bar{s}_{n,t}(\theta) = (1 - W_{n,t}/\lambda_{n,t}) \frac{\partial_\theta \lambda_{n,t}(\theta)}{\lambda_{n,t}(\theta)} = \left(1 - \varepsilon_t^2\right) \frac{\partial_\theta \lambda_{n,t}(\theta)}{\lambda_{n,t}(\theta)}.$$

The process $\{\bar{s}_{n,t}(\theta), \mathcal{F}_{n,t-1}\}$ is a MGD and $\omega_{n,t}(\theta)$ takes the form

$$\omega_{n,t}(\theta) = \frac{\partial_\theta \lambda_{n,t}(\theta) (\partial_\theta \lambda_{n,t}(\theta))'}{\lambda_{n,t}^2(\theta)} \left(1 - \varepsilon_t^2\right)^2.$$

By the same arguments as for the tvARCH,

$$\left\| \frac{\partial \omega_{n,t}(\theta)}{\partial \lambda_{n,t}(\theta)} \right\| = \frac{2 \|\partial_\theta \lambda_{n,t}(\theta)\|^2}{\lambda_{n,t}^3(\theta)} \left(1 - \varepsilon_t^2\right)^2, \quad \left\| \frac{\partial \omega_{n,t}(\theta)}{\partial (\partial_\theta \lambda_{n,t}(\theta))} \right\| = \frac{2 \|\partial_\theta \lambda_{n,t}(\theta)\|}{\lambda_{n,t}^2(\theta)} \left(1 - \varepsilon_t^2\right)^2,$$

and so $\omega_{n,t}(\theta)$ satisfies the conditions of Theorem 1 wirh $r = 0$ and $q = 1$. Furthermore, Assumption 1.10(iii) holds with $q_s = 1$ since it is easily checked using Theorem 1.6(iii) that $\mathbb{E}\left[|\lambda_{n,t} - \lambda_{n,t}(\theta(t/n))|\right] \leq C/n$.

The hessian takes the form

$$h_{n,t}(\theta) = -\frac{\partial_\theta \lambda_{n,t}(\theta) \partial_\theta \lambda_{n,t}(\theta)'}{\lambda_{n,t}^2(\theta)} \left[1 - 2\frac{W_{n,t}}{\lambda_{n,t}(\theta)}\right] + \frac{\partial_{\theta\theta}^2 \lambda_{n,t}(\theta)}{\lambda_{n,t}(\theta)} \left[1 - \frac{W_{n,t}}{\lambda_{n,t}(\theta)}\right],$$

and recycling the inequalities established above and again applying Theorem 1.6 it follows

that the hessian is also ULS$(1,1,\Theta)$ with stationary version

$$\begin{aligned} h_t^*\left(\theta\left(u\right)|u\right) &= -\frac{\partial_\theta \lambda_t^*\left(u\right)\partial_\theta \lambda_t^*\left(u\right)'}{\lambda_t^*\left(u\right)^2}\left[1-\frac{2W_t^*\left(u\right)}{\lambda_t^*\left(u\right)}\right]+\frac{\partial_{\theta\theta}^2\lambda_t^*\left(u\right)}{\lambda_t^*\left(u\right)}\left[1-\frac{W_t^*\left(u\right)}{\lambda_t^*\left(u\right)}\right] \\ &= -\frac{\partial_\theta \lambda_t^*\left(u\right)\partial_\theta \lambda_t^*\left(u\right)'}{\lambda_t^*\left(u\right)^2}\left(1-2\varepsilon_t^2\right)+\frac{\partial_{\theta\theta}^2\lambda_t^*\left(u\right)}{\lambda_t^*\left(u\right)}\left(1-\varepsilon_t^2\right). \end{aligned}$$

It can be shown that $h_t^*\left(\theta\left(u\right)|u\right)$ is GMC(p) for some $p>0$, and so Lemma 1 and Proposition 2 in Wu and Shao (2004) imply that it is short-range dependent and so satisfies Assumption 1.9.

Next, we verify Assumption 1.7: Since $h_{n,t}\left(\theta\right)$ is ULS$(1,1,\Theta)$,

$$\mathbb{E}\left[\left\|h_{n,t}\left(\theta\left(u\right)\right)-h_t^*\left(\theta\left(u\right)|t/n\right)\right\|\right]<C\left(1/n+\rho^t\right).$$

The derivative process of $h_t^*\left(\theta|u\right)$ takes the form

$$\begin{aligned} \partial_u h_t^*\left(\theta|u\right) =& \frac{\partial h_t^*\left(\theta|u\right)}{\partial W_t^*(u)}\partial_u W_t^*\left(u\right)+\frac{\partial h_t^*\left(\theta|u\right)}{\partial \lambda_t^*\left(\theta|u\right)}\partial_u \lambda_t^*\left(\theta|u\right) \\ &+\frac{\partial h_t^*\left(\theta|u\right)}{\partial\left[\partial_\theta \lambda_t^*\left(\theta|u\right)\right]}\partial_{\theta u}\lambda_t^*\left(\theta|u\right)+\frac{\partial h_t^*\left(\theta|u\right)}{\partial\left[\partial_{\theta\theta} \lambda_t^*\left(\theta|u\right)\right]}\partial_{\theta\theta u}\lambda_t^*\left(\theta|u\right). \end{aligned}$$

We note that the first partial derivative is bounded by a constant C and the remaining three partial derivatives are bounded by $C\left(1+\varepsilon_t^2\right)$. Also, Proposition 3.1 in Subba Rao (2006) implies that $W_t^*\left(u\right)$ is time-differentiable in the L_1-sense at u. We then obtain from Theorem 1.6(iv) that $\lambda_t^*\left(\theta|u\right)$, $\lambda_t^*\left(\theta|u\right)$, and $\partial_{\theta\theta}^2\lambda_t^*\left(\theta|u\right)$ are time-differentiable in the L_1-sense. By employing Theorem 1.6(iv) one more time, we find that $h_t^*\left(\theta\left(u\right)|u\right)$ is also L_1-differentiable.

Finally, the third-order derivatives takes the form

$$\begin{aligned} \frac{\partial h_{n,t}\left(\theta\right)}{\partial\theta_i} =& \frac{2\partial_{\theta_i}\lambda_{n,t}\left(\theta\right)\partial_\theta \lambda_{n,t}\left(\theta\right)\partial_\theta \lambda_{n,t}\left(\theta\right)'}{\lambda_{n,t}^3\left(\theta\right)}\left[1-3\frac{W_{n,t}}{\lambda_{n,t}\left(\theta\right)}\right] \\ &-\frac{\partial_{\theta\theta_i}^2\lambda_{n,t}\left(\theta\right)\partial_\theta \lambda_{n,t}\left(\theta\right)'+\partial_\theta \lambda_{n,t}\left(\theta\right)\partial_{\theta\theta_i}^2\lambda_{n,t}\left(\theta\right)'}{\lambda_{n,t}^2\left(\theta\right)}\left[1-2\frac{W_{n,t}}{\lambda_{n,t}\left(\theta\right)}\right] \\ &-\frac{\partial_{\theta_i}\lambda_{n,t}\left(\theta\right)\partial_{\theta\theta}^2\lambda_{n,t}\left(\theta\right)}{\lambda_{n,t}\left(\theta\right)^2}\left[1-2\frac{W_{n,t}}{\lambda_{n,t}\left(\theta\right)}\right]+\frac{\partial_{\theta\theta\theta_i}^3\lambda_{n,t}\left(\theta\right)}{\lambda_{n,t}\left(\theta\right)}\left[1-\frac{W_{n,t}}{\lambda_{n,t}\left(\theta\right)}\right] \end{aligned}$$

recycling the inequalities derived above, we conclude that these are ULS$(1,1)$. □

Proof of Corollary 1.6. We first show that the PAR process is locally stationary by verifying the conditions of Theorem 1.7. First, write the process as

$$Y_{n,t} = G\left(Y_{n,t-1}, ..., Y_{n,t-q}, \varepsilon_t; \theta\left(t/n\right)\right) := N_t\left(\omega\left(t/n\right) + \sum_{i=1}^{q} \alpha_i\left(t/n\right) Y_{n,t-i}\right),$$

where $\varepsilon_t := N_t(\cdot)$, $t = 1, 2, ...$, are i.i.d. copies of a Poisson process (see Agosto et al. (2016) for details). For any $x_0 \in {\mathbb{R}_+}^q$ and all $\theta \in \Theta$,

$$\mathbb{E}\left[\left|G\left(x_0, N_t; \theta\right)\right|\right] \leq \mathbb{E}\left[N_t\left(\omega + \sum_{i=1}^{q} \alpha_i x_{0,i}\right)\right] = \omega + \sum_{i=1}^{q} \alpha_i x_{0,i} < \infty;$$

and for all $x, x' \in {\mathbb{R}_+}^q$,

$$\mathbb{E}\left[\left\|G\left(x, N_t; \theta\right) - G\left(x', N_t; \theta\right)\right\|\right] \leq \mathbb{E}\left[\left\|N_t\left(\sum_{i=1}^{q} \alpha_i \left|x_i - x_i'\right|\right)\right\|\right] = \sum_{i=1}^{q} \alpha_i \left|x_i - x_i'\right|,$$

where $\sum_{i=1}^{q} \alpha_i < 1$. Finally,

$$\mathbb{E}\left[\left\|G\left(x, N_t; \theta\right) - G\left(x, N_t; \theta'\right)\right\|\right] = \left|\omega - \omega'\right| + \sum_{i=1}^{q} \left|\alpha_i - \alpha_i'\right| \mathbb{E}\left[N_t\left(x\right)\right] \leq C\left(1 + x\right)\left\|\theta - \theta'\right\|.$$

This shows that $X_{n,t} := \left(Y_{n,t-1}, ..., Y_{n,t-q}\right)$ is LS$(1,1)$ which in turn implies that $Y_{n,t}$ is LS$(1,1)$, c.f. Theorem 1.1. However, later we need the existence of higher-order moments, and so we demonstrate by induction that $E\left[\lambda_t^*\left(u\right)^r\right] < \infty$ for all $r < \infty$: First, $E[\lambda_t^*\left(u\right)] = \omega\left(u\right) + \sum_{i=1}^{q} \alpha_i\left(u\right) E\left[\lambda_t^*\left(u\right)\right]$ which has a well-defined solution while

$$\left(\lambda_t^*\left(u\right)\right)^r = \sum_{j=0}^{r} \binom{r}{j} \left(\sum_{i=1}^{q} \alpha_i\left(u\right) Y_{t-i}^*\left(u\right)\right)^j \omega^{r-j}\left(u\right),$$

and so

$$\begin{aligned} E[\left(\lambda_t^*\left(u\right)\right)^r] &= \sum_{j=0}^{r} \binom{r}{j} E\left[\left(\sum_{i=1}^{q} \alpha_i\left(u\right) Y_{t-i}^*\left(u\right)\right)^j\right] \left(\omega\left(u\right)\right)^{r-j} \\ &= \omega\left(u\right)^r + E\left[\left(\sum_{i=1}^{q} \alpha_i\left(u\right) Y_{t-i}^*\left(u\right)\right)^r\right] + E\left[p_{r-1}\left(X_t^*\left(u\right)\right)\right], \end{aligned}$$

with $p_{r-1}\left(x\right)$ being an $(r-1)$th order polynomial. By induction, $E\left[p_{r-1}\left(X_t^*\left(u\right)\right)\right] < \infty$,

and we are left with considering terms of the form, for some constants c_{ij},

$$E\left[\left(\sum_{i=1}^{q}\alpha_i\left(u\right)Y_{t-i}^{*}\left(u\right)\right)^r\right]=\sum_{i=1}^{q}\sum_{j=0}^{r}c_{ij}\alpha_i^j\left(u\right)E\left[Y_{t-i}^{*}\left(u\right)^j\right]$$
$$=\sum_{i=1}^{q}\alpha_i^r\left(u\right)E\left[Y_{t-i}^{*}\left(u\right)^r\right]+C_r=\sum_{i=1}^{q}\alpha_i^r\left(u\right)E\left[\lambda_t^{*}\left(u\right)^r\right]+C_r$$

where, again by induction, $C_r<\infty$. Collecting terms, $E\left[\left(\lambda_t^{*}\left(u\right)\right)^r\right]=\sum_{i=1}^{q}\alpha_i^r\left(u\right)E\left[\lambda_t^{*}\left(u\right)^r\right]+\tilde{C}_r$ which has a well-defined solution since $\sum_{i=1}^{q}\alpha_i^r\left(u\right)<1$. This in turn implies that $E\left[Y_t^{*}\left(u\right)^r\right]<\infty$ for all $r<\infty$. We can now apply Theorem 1.7 to obtain that $\lambda_{n,t}$ and $Y_{n,t}$ are $\mathrm{LS}(r,1)$ with $E[\lambda_{n,t}^r]<\infty$ and $E[Y_{n,t}^r]<\infty$.

Next, we observe that $\lambda_{n,t}\left(\theta\right)$, $\partial_\theta\lambda_{n,t}\left(\theta\right)$ and $\partial_{\theta\theta}^2\lambda_{n,t}\left(\theta\right)$ are on the same form as in the GARCH model, except that $Y_{n,t-1}^2$ has been replaced by $Y_{n,t-1}$. In particular, it is easily checked that $\lambda_{n,t}\left(\theta\right)$, $\partial_\theta\lambda_{n,t}\left(\theta\right)$ and $\partial_{\theta\theta}^2\lambda_{n,t}\left(\theta\right)$ are $\mathrm{ULS}(1,1,\Theta)$ and with all polynomial moments since $Y_{n,t-1}$ has all polynomial moments. Thus, it only remains to show that the log-likelihood and its derivatives w.r.t. θ satisfy the conditions of Theorem 1.1. First,

$$\left|\frac{\partial\ell_{n,t}\left(\theta\right)}{\partial Y_{n,t}}\right|=\left|\log\left\{\lambda_{n,t}\left(\theta\right)\right\}\right|\leq\max\left\{\left|\log\delta_L\right|,\lambda_{n,t}\left(\theta\right)\right\},$$

where $\lambda_{n,t}\left(\theta\right)$ has all relevant moments. Second,

$$\left|\frac{\partial\ell_{n,t}\left(\theta\right)}{\partial\lambda_{n,t}\left(\theta\right)}\right|=\frac{Y_{n,t}}{\lambda_{n,t}\left(\theta\right)}+1\leq\frac{Y_{n,t}}{\delta_L}+1,$$

where again the right-hand side has all relevant moments. The score function takes the form $s_{n,t}\left(\theta\right)=\left(Y_{n,t}/\lambda_{n,t}\left(\theta\right)-1\right)\partial_\theta\lambda_{n,t}\left(\theta\right)$ which satisfies the Martingale difference condition with conditional variance $\omega_{n,t}\left(\theta\right)=\partial_\theta\lambda_{n,t}\left(\theta\right)\partial_\theta\lambda_{n,t}\left(\theta\right)'/\lambda_{n,t}\left(\theta\right)$. As before, due to all polynomial moments existing, it is easily checked that the conditional variance satisfies the conditions of Theorem 1.1 and similarly for the hessian which is on the form

$$h_{n,t}\left(\theta\right)=\frac{Y_{n,t}}{\lambda_{n,t}^2\left(\theta\right)}\partial_\theta\lambda_{n,t}\left(\theta\right)\partial_\theta\lambda_{n,t}\left(\theta\right)'-\left(\frac{Y_{n,t}}{\lambda_{n,t}\left(\theta\right)}-1\right)\partial_{\theta\theta}^2\lambda_{n,t}\left(\theta\right).$$

The analysis of the third-order derivatives is similar and so is left out. □

1.8.4 Local stationarity of Markov processes

Dahlhaus et al. (2017) consider the following general class of nonlinear autoregressive distributed lag models,

$$Y_{n,t} = G\left(Y_{n,t-1}, \varepsilon_t, \theta\left(t/n\right)\right), \quad t = 1, \ldots, n,$$

where $G : \mathcal{Y} \times \mathcal{E} \times \Theta \mapsto \mathcal{Y}$ is some mapping, $\varepsilon_t \in \mathcal{E} \subseteq \mathbb{R}^{d_\varepsilon}$ is a sequence of i.i.d. errors, and $\theta\left(\cdot\right) \in \Theta$. We here develop generalized versions of their results concerning local stationarity of $Y_{n,t}$ with stationary approximation $Y_t^*\left(u\right)$ solving

$$Y_t^*\left(u\right) = G\left(Y_{t-1}^*\left(u\right), \varepsilon_t; \theta\left(u\right)\right), \quad u \in [0,1]. \tag{1.18}$$

In particular, which is in contrast to the existing literature, we do not require that $Y_{n,0} = Y_0^*\left(u\right)$ and instead allow $Y_{n,0}$ to be initialized at some arbitrary value.

Assumption 1.11. *(i) There exist $y_0 \in \mathcal{Y}$ and $p > 0$ such that $\sup_{\theta\in\Theta} \mathbb{E}\left[\left\|G\left(y_0, \varepsilon_t; \theta\right)\right\|^p\right] < \infty$; (ii) there exists $\rho < 1$ so that for all $y, y' \in \mathcal{Y}$,*

$$\mathbb{E}\left[\left\|G\left(y, \varepsilon_t; \theta\right) - G\left(y', \varepsilon_t; \theta\right)\right\|^p\right]^{1/p} \leq \rho \left\|y - y'\right\|;$$

(iii) there exist $\tilde{p} \geq 1$, $q > 0$ and $r \geq 0$ so that $\mathbb{E}\left[\left\|G\left(y, \varepsilon_t; \theta\right) - G\left(y, \varepsilon_t; \theta'\right)\right\|^{\tilde{p}}\right]^{1/\tilde{p}} \leq C\left(1 + \left\|y\right\|^r\right)\left\|\theta - \theta'\right\|^q$ for all $\theta, \theta' \in \Theta$, and (iv) $\mathbb{E}\left[\left\|Y_{n,0}\right\|^{\tilde{p}}\right] < \infty$.

Assumption 1.11(i) ensures that the process is well-behaved around y_0 while Assumption 1.11(ii) is the contraction condition implying that $\{X_t^*\left(u\right)\}$ is attracted with uniform rate towards "the centre" of its state space for any given starting point. Finally, Assumption 1.11(iii) allows us to bound the the difference between $Y_{n,t}$ and $Y_t^*\left(t/n\right)$. Compared to Dahlhaus et al. (2017), we here allow for $p \neq \tilde{p}$. In particular, assuming we can verify $\mathbb{E}\left[\left\|Y_t^*\left(u\right)\right\|^{\tilde{p}r}\right] < \infty$ this allows us to show higher-order local stationarity ($\tilde{p} > p$). This is, for example, used in Example 1.4.

Theorem 1.7. *Under Assumptions 1.11(i)-(ii), there exists a stationary and ergodic solution, $\{Y_t^*\left(u\right)\}$ to (1.18) which is GMC(p) with $\sup_{u\in[0,1]} \mathbb{E}\left[\left\|Y_t^*\left(u\right)\right\|^p\right] < \infty$. If furthermore*

1.11(iii)-(iv) hold, $\sup_{u\in[0,1]}\mathbb{E}\left[\left\|Y_t^*(u)\right\|^{\tilde{p}r}\right]<\infty$ *and* $\theta(\cdot)\in\Theta$ *is continuously differentiable, then* $Y_{n,t}$ *is LS*(p,q) *with* $\sup_{n,t}\mathbb{E}\left[\left\|Y_{n,t}\right\|^{\tilde{p}}\right]<\infty$ *so that, for some* $C<\infty$*,*

$$\mathbb{E}\left[\left\|Y_{n,t}-Y_t^*(t/n)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\leq C\left(\frac{1}{n^q}+\rho^t\right),\quad \mathbb{E}\left[\left\|Y_t^*(u)-Y_t^*(v)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\leq C\left|u-v\right|^q.$$

Proof of Theorem 1.7. The first part of the result follows from Proposition 4.4 in Dahlhaus et al. (2017). For the second part,

$$\begin{aligned}
&\mathbb{E}\left[\left\|Y_t^*(u)-Y_t^*(v)\right\|^p\right]^{1/\tilde{p}}\\
=&\mathbb{E}\left[\left\|G\left(Y_{t-1}^*(u),\varepsilon_t,\theta(u)\right)-G\left(Y_{t-1}^*(v),\varepsilon_t,\theta(v)\right)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
\leq&\mathbb{E}\left[\left\|G\left(Y_{t-1}^*(u),\varepsilon_t,\theta(u)\right)-G\left(Y_{t-1}^*(u),\varepsilon_t,\theta(v)\right)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
&+\mathbb{E}\left[\left\|G\left(Y_{t-1}^*(u),\varepsilon_t,\theta(v)\right)-G\left(Y_{t-1}^*(v),\varepsilon_t,\theta(v)\right)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
\leq&C\left(1+E\left[\left\|Y_{t-1}^*(u)\right\|^{r\tilde{p}}\right]^{1/\tilde{p}}\right)|u-v|^q+\rho\mathbb{E}\left[\left\|Y_{t-1}^*(u)-Y_{t-1}^*(v)\right\|^{\tilde{p}}\right]^{1/\tilde{p}},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left\|Y_{n,t}-Y_t^*(t/n)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
=&\mathbb{E}\left[\left\|G\left(Y_{n,t-1},\varepsilon_t,\theta(t/n)\right)-G\left(Y_{t-1}^*(t/n),\varepsilon_t,\theta(t/n)\right)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
\leq&\rho\mathbb{E}\left[\left\|Y_{n,t-1}-Y_{t-1}^*(t/n)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}\\
\leq&\rho\mathbb{E}\left(\left\|G\left(Y_{n,t-2},\varepsilon_{t-1},\theta\left((t-1)/n\right)\right)-G\left(Y_{t-2}^*(t/n),\varepsilon_t,\theta\left((t-1)/n\right)\right)\right\|\right)^{1/\tilde{p}}\\
&+\rho\mathbb{E}\left[\left\|G\left(Y_{t-2}^*(t/n),\varepsilon_t,\theta\left((t-1)/n\right)\right)-G\left(Y_{t-2}^*(t/n),\varepsilon_t,\theta(t/n)\right)\right\|^p\right]^{1/\tilde{p}}\\
\leq&\rho^2\mathbb{E}\left[\left\|Y_{n,t-2}-Y_{t-2}^*(t/n)\right\|^{\tilde{p}}\right]^{1/\tilde{p}}+\rho C\left(1+E\left[\left\|Y_{t-2}^*(t/n)\right\|^{r\tilde{p}}\right]^{1/\tilde{p}}\right)1/n^q.
\end{aligned}$$

Continuing the above two recursions yield the desired results. □

Chapter 2

Sieve Estimation of Optimal Transport with an Application to Conditional Vector Quantiles

2.1 Introduction

The Monge-Kantorovich's optimal transport problem is a class of stochastic optimization problems bringing new insights into mathematics and many applied sciences. The main object of this problem is to find the joint probability measure on $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}^d$, with given marginal probability measures, maximizing (minimizing) the average overall surplus (cost) generated by linking two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. In economics, matching models under transferable utility have a close relationship to the Monge-Kantorovich problem; just think of a social planner matching people and trying to maximize total welfare. Recently, it has been used as a method for investigating many other problems in econometrics and microeconomic theory, such as models of differentiated demand, incomplete econometric models, and quantile methods (see, for example, Galichon, 2016, 2017).

This chapter considers a class of Monge-Kantorovich problems whose solutions associate each point X to a single point Y with a measurable function $T : \mathcal{X} \to \mathcal{Y}$ such that $Y = T(X)$. I then propose a sieve M-estimation method for $T(\cdot)$ being called the deterministic optimal coupling or Monge transport. The objective function for the sieve M-estimation

is the dual problem of the original Monge-Kantorovich problem. This is because the optimal joint probability measure π for the Monge transport is identified with a Dirac delta function while the dual problem is the one minimizing (maximizing) over measurable and integrable functions. Under certain assumptions on the surplus (cost) function and marginal probability measures, one can guarantee the existence, uniqueness, and regularity of the solution of the dual problem, say $g : \mathcal{X} \to \mathbb{R}$, which is closely connected to T. For example, when the surplus function is $X'Y$, $T(X) = \nabla g(X)$ where ∇g is the gradient of g.

In this framework, the solution of the dual problem, $g(\cdot)$, is consistently estimated by a linear combination of sieve terms. As the number of sieve terms increases, the estimation error decreases, approaching zero in the limit. The theory impose very little structure on the underlying optimal transport problem being solved. I then derive convergence rates for sieve M-estimators of $g(\cdot)$ and its derivatives when $\Phi(X, Y) = X'Y$. In particular, in contrast to the existing literature on nonparametric estimation of the solution to the optimal transport problem, the conditions under which we derive our results are more easily verified. Moreover, the derived convergence rates for sieve estimators are the same as the optimal rate in the context of regression and density estimations (Stone, 1982).

As a generalization of the Monge-Kantorovich problem, we consider the conditional Monge-Kantorovich problem having an application to the conditional vector quantiles. In one dimension, the quantile function is the inverse of the cumulative distribution function, and it is monotone. We follow the definition of the conditional vector quantiles in Carlier et al. (2016) and Chernozhukov et al. (2017). They define the vector quantiles in terms of solution for the Monge-Kantorovich problem with the objective function $X'Y$, which includes the original one-dimensional quantile as a special case. Implementation solving the linear programming problem from the primal optimal transport problem still suffers from the crossing problem. Simulation studies and empirical applications for the Engel curve demonstrate that the sieve method avoids this problem. The sieve method also provides a stable monotone estimate by adding the convexity constraint on the function space. Our estimator can also capture the nonlinear effect of conditioning variables on vector quantiles. It decreases the possibility of misspecification.

After Monge (1781) introduced the optimal transport problem, Kantorovich developed

the problem of optimal coupling, which is a class of joint distribution connected to Monge's work. For the deterministic form of the optimal transport map, Brenier (1991) and McCann (1995) show that, for the quadratic surplus function, the optimal transport mapping is unique: a gradient of a convex function. Carlier (2003) and Villani (2008) include the uniqueness result for the case of a general surplus. Caffarelli (1992) developed the smoothness result of the unique optimal transportation map for the quadratic surplus. This is extended to the general surplus case by Ma et al. (2005). In this chapter, we use slightly stronger but simpler conditions, introduced by Lindenlaub (2017), compared to those in Ma et al. (2005). This property implies that the conventional sieve M-estimation method is applicable. For the general theory of optimal transport, see Villani (2003), Villani (2008), or De Philippis and Figalli (2014).

Our convergence rate result is closely related to the work by Gunsilius (2018). This chapter adapts result from Gunsilius (2018) to construct the second variation on the dual problem of the Monge-Kantorovich problem. Gunsilius (2018) examines the convergence rate for the kernel estimator of the solution to the optimal transport problem. For the result, Poincaré inequality in probability is required to ensure the optimum is well-separated, but this inequality is a high-level condition. Many popular probability measures including those for normal, exponential, and uniform distribution, satisfy Poincaré inequality. However, it is not easy to verify sufficient conditions for unknown distribution except for log-concave distribution (Bobkov, 1999). We show that Poincaré inequality is satisfied under the smoothness condition of the densities of two variables and in the constrained case, in which the true parameter function is under some restrictions.

This chapter is organized as follows. In Section 2.2, we introduce the Monge-Kantorovich's optimal transport problem and review the characteristics of optimal transport under certain regularity conditions on the probability measures and the surplus (cost) function. Section 2.3 proposes the consistent sieve M-estimation method for the solution of the dual problem. Section 2.4 presents the convergence rate result of the sieve M-estimator. In Section 2.5, we extend the theory to cover the conditional Monge-Kantorovich problem and then apply our general theory to the conditional vector quantiles. We present the results of simulation studies for the conditional vector quantiles in Section 2.6. Section 2.7 revisits the empirical

study presented in Carlier et al. (2016). Section 2.8 restates our findings and presents our conclusion. All auxiliary lemmas and proofs have been relegated to the Appendix.

2.2 Monge-Kantorovich problem and Monge transport

Inspired by Monge (1781), who formulated the transport problem in Euclidean space minimizing a specific cost function over two centuries ago, optimal transport is now one of the most active research areas in mathematics. In economics, the optimal transport is related to the cost minimization or the surplus maximization. We consider the problem of assigning workers to jobs: assume that there are workers and occupations and their matching generate a quantity of output. As an economics application of the optimal transport problem, we can think of a social planner problem deciding which workers to assign to which occupations to maximize the total output. For this problem the equilibrium matching between and occupations, and wage function for workers can be expressed with solution of the Monge-Kantorovich's optimal transport problem.

The Monge-Kantorovich problem is the stochastic optimization problem depending on two parts of marginal probability measures and the surplus function. Let $X \in \mathcal{X} \in \mathbb{R}^d$ and $Y \in \mathcal{Y} \in \mathbb{R}^d$ have the probability measures P_X and P_Y respectively. Denoting $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d \cup \{-\infty\}$ as the objective criterion on $\mathcal{X} \times \mathcal{Y}$, the Monge-Kantorovich problem is formulated as follows:

$$\sup_{\pi \in \mathcal{M}(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} \Phi(X, Y)\, d\pi(X, Y) \tag{2.1}$$

where $\mathcal{M}(P_X, P_Y)$ is the set of all joint measures admitting P_X and P_Y as marginals on $\mathcal{X}$ and $\mathcal{Y}$ respectively. The pair (X, Y) having a joint measure $\pi \in \mathcal{M}(P_X, P_Y)$ is called a coupling of (P_X, P_Y),[1] and (X, Y) runs over all possible couplings of (P_X, P_Y) in this problem. Such couplings, achieving the supremum, are called optimal couplings. One important feature is that the Monge-Kantorovich problem always has a solution but the optimal coupling is not necessarily deterministic.[2] For example, let $\mathcal{X} = \{0\} \times [-1, 1]$ and

[1] By extension, π is also called a coupling of (P_X, P_Y).

[2] The search of deterministic optimal transport map is called the Monge problem. Then, instead of introducing the possibility of randomization, one can maximize over all deterministic transport maps:

$\mathcal{Y} = \{-1, 1\} \times [-1, 1]$ with uniform marginal distributions $\mathcal{U}(\mathcal{X})$ and $\mathcal{U}(\mathcal{Y})$ respectively. Then, for $\Phi(X, Y) = X'Y$, there is a unique optimal coupling at which one half of the mass at $(0, a)$ matches with $(-1, a)$ and the other half with $(1, a)$.

Another important feature of the Monge-Kantorovich problem is that it is a linear programming problem. The objective function in (2.1) is linear with respect to π, so the primal problem (2.1) admits the following dual problem:

$$\inf_{g \in L_1(P_X),\ h \in L_1(P_Y)} \mathbb{E}_X[g(X)] + \mathbb{E}_Y[h(Y)] \quad \text{s.t. } g(X) + h(Y) \geq \Phi(X, Y), \tag{2.2}$$

where $L_r(P)$ is the space of functions for which rth power is integrable with respect to P. If Φ is upper semicontinuous and bounded from above, both the primal and the dual problem have solutions, and the values of two problems are equal (see, e.g., Theorem 5.10 in Villani, 2008 or Theorem 1 in Chiappori et al., 2010). This duality provides two interpretations of the total surplus. From an economic intuition, the primal problem (2.1) is a social planner problem maximizing total welfare by pairs. On the other hand, the dual problem (2.2) is a decentralized problem, offering a breakdown of the total welfare at the individual level. In the case of the worker-firm matching model, $g(X)$ and $h(Y)$ can be interpreted as the equilibrium wage and profit that worker X and firm Y receive at equilibrium, respectively.

We note that, for arbitrary pair (g, h) satisfying the constraint in (2.2), h can be improved by $h_1(Y) = \sup_{x \in \mathcal{X}} \{\Phi(x, Y) - g(x)\}$. A pair (g, h) is said to be tight if

$$g(X) = \sup_y \{\Phi(X, y) - h(y)\}, \quad h(Y) = \sup_x \{\Phi(x, Y) - g(x)\}.$$

If (g, h) is tight in the worker-firm matching model, then it is impossible for the worker to raise the wage without losing the firm's profit. Thus, it is logical to restrict our attention to tight pairs in the dual problem (2.2). We can then reconstruct h in terms of g and rewrite

---

$$\max_{T(\cdot)} \mathbb{E}_X[\Phi(X, T(X))] \quad \text{s.t.} \quad T(P_X) = P_Y.$$

Monge problem is a social planner problem finding the deterministic matching function that maximizes the average overall surplus. In the Kantorovich approach, it is not required to impose that all the mass sharing same value, x, should go to the one value, $y = T(x)$.

(2.2), so the only unknown is g:

$$\inf_{g \in L_1(P_X)} Q(g),\ Q(g) = \mathbb{E}_X[g(X)] + \mathbb{E}_Y\left[\sup_{x \in \mathcal{X}} \{\Phi(x, Y) - g(x)\}\right]. \tag{2.3}$$

From now on, we consider a class of the dual Monge-Kantorovich problem whose solutions associate each X to one Y with a measurable function $T : \mathcal{X} \to \mathcal{Y}$ such that $Y = T(X)$. In the context of optimal transport, $T(\cdot)$ is called the deterministic optimal coupling or Monge transport.[3] Notice that the primal problem always has a solution π, but the dual problem does not. However, the optimal joint probability measure π for the Monge transport is not easy to compute because it is identified with a Dirac delta function. As we mentioned, the dual problem is the one optimizing minimizing over measurable and integrable functions. Once a solution g for the dual problem (2.3) exists, we can easily apply the sieve M-estimation method.

We are now ready to state the conditions under which the solution of the dual problem uniquely exists:

Assumption 2.1. *(i) P_X and P_Y have compact and convex supports $\mathcal{X}$ and $\mathcal{Y}$ in $\mathbb{R}^d$;*

(ii) P_X is absolutely continuous with respect to the Lebesgue measure.

Assumption 2.2. *(i) (Lipschitz condition) $\Phi : \mathcal{X} \times Y \to \mathbb{R}$ is twice continuously differentiable and there exists $c > 0$ such that for every $X_1, X_2 \in \mathcal{X}$,*

$$\sup_{y \in \mathcal{Y}} |\Phi(X_1, y) - \Phi(X_1, y)| \leq c \|X_1 - X_2\|,$$

where $\|\cdot\|$ is the Euclidean norm;

(ii) (Twist condition) For any fixed $X \in \mathcal{X}$ and $Y_1 \neq Y_2 \in \mathcal{Y}$, $\nabla_X \Phi(X, Y_1) \neq \nabla_X \Phi(X, Y_2)$.[4]

[3] It is equivalent to any of the following conditions: (i) If $X \sim P_X$, the $T(X) \sim P_Y$; (ii) Equality $P_X\left(T^{-1}(B)\right) = P_Y(B)$ holds for every subset B of $\mathcal{Y}$; (iii) When P_X and P_Y have respective densities f_X and f_Y, and when T is smooth, the so-called Monge-Ampère equation holds:

$$f_X(X) = |\det \nabla T(X)| f_Y(T(X)),$$

where $\nabla_X T$ stands for the Jacobian matrix of T.

[4] For example, $\Phi(X, Y) = X'AY$ satisfies Assumption 2.2(ii) by imposing the invertibility of A (see, Chapter 3).

Assumption 2.2(ii) implies the injectivity of $\nabla_X \Phi(X,Y)$ for each fixed X, which guarantees uniqueness of optimal transport map. Under the above assumptions, the following result holds:

Lemma 2.1. *(Theorem 1 in Carlier, 2003) Let Assumptions 2.1 and 2.2 hold. Then, there exists a solution g_0, which is unique up to an additive constant, such that the map $T : \mathcal{X} \to \mathcal{Y}$, satisfying $\nabla g_0(X) \coloneqq \nabla_X \Phi(X,Y)|_{Y=T(X)}$, is the unique optimal transport map, sending P_X onto P_Y.*

We find that the solution of the dual problem, g, is closely connected to T. For example, when $\Phi(X,Y) = X'Y$, $T(X) = \nabla g(X)$. Furthermore, a constant can be freely added to or subtracted from g in (2.3). We can determine g_0 uniquely by imposing a location normalization such as $\int_{\mathcal{X}} g(X)\, dX = 0$ or $g(X^*) = 0$ for a fixed $X^* \in \mathcal{X}$.

2.3 Consistent sieve estimation of Monge transport

In the dual Monge-Kantorovich problem (2.3), a infinite-dimensional function g is a unknown parameter of interest. Since $Q(g) = Q(g + c)$ for any constant c, we look at the solution among those such that $\int_{\mathcal{X}} g(X)\, dX = 0$, which provides a simple way to derive the convergence rates of our sieve M-estimator. Hence, we denote $g_0 \in \mathcal{G}$ as the true unknown infinite-dimensional parameter, where $\mathcal{G}$ is a linear subspace of the space of real-valued functions with $\mathbb{E}\left[g(X)^2\right] < \infty$ and $\int_{\mathcal{X}} g(X)\, dX = 0$.

We apply the sieve M-estimation method of Chen and Shen (1998). To describe the method, let $\{X_i\}_{i=1}^m$ and $\{Y_j\}_{j=1}^n$ be d-dimensional independent and identically distributed (i.i.d.) sequences of observations with unknown marginal probability measure P_X and P_Y, respectively. We allow the different numbers of observations for X and Y but here set $m = n$ without loss of generality. Let $\{p_j(X),\ j = 1, 2, \ldots\}$ denote a sequence of known basis functions that can approximate any $g \in \mathcal{G}$. Then, for a finite-dimensional linear sieve

$$\mathcal{G}_n = \left\{ g_n : \mathcal{X} \to \mathbb{R},\ g_n(X) = \sum_{j=1}^{k_n} a_j p_j(X) :\ \int_{\mathcal{X}} g_n(X)\, dX = 0,\ a_1, \ldots, a_{k_n} \in \mathbb{R} \right\}, \quad (2.4)$$

with $\dim(\mathcal{G}_n) = k_n \to \infty$ slowly as $n \to \infty$, we estimate the unknown sieve coefficients of

g_n:

$$\hat{g}_n = \arg\min_{g_n \in \mathcal{G}_n} Q_n(g_n) = \arg\min_{g_n \in \mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n}\left[g_n(X_i) + \sup_{x\in\mathcal{X}}\{\Phi(x, Y_i) - g_n(x)\}\right] \tag{2.5}$$

is a sieve M-estimator of g_0. Notice from Lemma 2.1 that the assignment function for Y, $T(X)$, satisfies $\nabla g(X) = \nabla_X \Phi(X,Y)|_{Y=T(X)}$. Therefore, the Monge transport T can be estimated from $\hat{g}_n$: $\nabla \hat{g}_n(X) = \nabla_X \Phi(X,Y)|_{Y=\hat{T}_n(X)}$. .[5]

Computing exact $\sup_{x\in\mathcal{X}}\{x'Y_i - g_n(x)\}$ is easy for the small dimension of k_n for $\mathcal{G}_n$, but it is burdensome for the large value of k_n. In this case, we consider the set of grid points or samples on $\mathcal{X}$, $\mathcal{X}_n$, and then compute

$$h_n(Y_i, g_n) = \max_{x\in\mathcal{X}_n}\left\{x'Y_i - g_n(x)\right\}.$$

Since the supports for X and Y are bounded, h_n approximates h well as the number of grid points for $\mathcal{X}_n$ increases (see Lemma 2.3).

We note that $Q_n(g)$ is convex on $\mathcal{G}_n$ and $\mathcal{G}_n$ is convex. The pointwise convergence of $Q_n(g)$ to $Q(g)$ is easily checked, and hence Theorem 2.7 in Newey and McFadden (1994) and Theorem 3.1 in Chen (2007) are applicable to show that $\hat{g}_n$ is consistent. In the following we let $\|g\|_\infty \equiv \sup_{X\in\mathcal{X}}|g(X)|$ and $\|g\|_{2,leb} \equiv \left\{\int_{\mathcal{X}}[g(X)]^2 dX\right\}^{1/2}$ denote its L_∞ norm and L_2 norm with respect to the Lebesgue measure of $\mathcal{X}$, respectively. Then we have the following consistency result:

Theorem 2.1. *Suppose that Assumptions 2.2 and 2.1 hold. Then,* $\|\hat{g}_n - g_0\|_\infty = o_p(1)$.

2.4 Convergence rates for sieve M-estimators

In this section, we investigate the convergence rate of the sieve M estimator $\hat{g}_n$ of the unknown solution g_0. Before investigating the convergence rate results of $\hat{g}_n$, I shall state the regular property of g_0. To apply a sieve estimation method, we will rely on the smoothness of the solution to the optimal transport problem as well as uniqueness. Notice that the convergence order of sieve estimator depends on the order of differentiability of solution

[5]When $\Phi(X,Y) = X'Y$, the derivative of $\hat{g}_n$ is a sieve estimator of the Monge transport T.

function. To obtain the results, we introduce the Hölder class of functions. Let $0 < \gamma \leq 1$. A real-valued function g on $\mathcal{X}$ is said to be Hölder continuous with exponent γ if there is a positive number c such that $|g(X_1) - g(X_2)| \leq c\|X_1 - X_2\|^\gamma$ for all $X_1, X_2 \in \mathcal{X} \subset \mathbb{R}^d$. Then the Hölder space $C^{m,\gamma}(\mathcal{X})$ is the space of m-times continuously differentiable functions whose partial derivatives up to order m are Hölder continuous with exponent γ.

The regularity of g_0 depends on (i) the Ma-Trudinger-Wang (MTW) condition (Ma et al., 2005) from the theory of optimal transport, and (ii) the the regularity of marginal densities f_X and f_Y for X and Y, respectively. The MTW condition is not simple and involves fourth-order condition on Φ. Lindenlaub (2017) provides a necessary conditions for MTW condition to hold but much simpler one based on the concept of supermodularity, which is known as the Spence-Mirrlees condition: The twice continuously differentiable $\phi : \mathbb{R}^2 \to \mathbb{R}$ is supermodular if $\partial^2\phi/\partial x\partial y \geq 0$. The strict supermodularlity obtains whenever the the strict inequality holds.

Assumption 2.3. *P_X and P_Y have compact and convex supports $\mathcal{X}$ and $\mathcal{Y}$ in $\mathbb{R}^d$ and respective f_X and f_Y such that*

(i) f_X and f_Y are bounded away both from zero and infinity on $\mathcal{X}$ and $\mathcal{Y}$; and

(ii) $f_X \in C^{m,\gamma}(\mathcal{X})$ and $f_Y \in C^{m,\gamma}(\mathcal{Y})$ for some $m \in \mathbb{N}$ and $\gamma \in (0,1)$.

Assumption 2.4. *$\Phi(X,Y) = \sum_{k=1}^d \phi_k(X_k, Y_k)$ is a real-valued function defined on $\mathcal{X}' \times \mathcal{Y}$, where $\mathcal{X}'$ is a open subset in $\mathbb{R}^d$ with $\mathcal{X} \subset \mathcal{X}'$, such that*

(i) ϕ_k is four-times continuously differentiable and strictly supermodular; and

(ii) Both $\partial^2\phi_k/\partial X_k\partial Y_k$ and $\log\left(\partial^2\phi_k/\partial X_k\partial Y_k\right)$ are supermodular.

Under the above assumptions, the following result holds:

Lemma 2.2. *(Theorem 12.51 in Villani, 2008) Suppose that Assumptions 2.3 and 2.4 hold. Then, $g_0 \in C^{m+2,\gamma}(\mathcal{X})$.*

As a result of Lemma 2.2, $g_0 \in \mathcal{G} = \left\{g \in C^{m+2,\gamma}(\mathcal{X}) : \int_{\mathcal{X}} g(X)\, dX = 0\right\}$. This is called the interior regularity and developed for general surplus function Φ in a large literature.[6] We note that a larger domain $\mathcal{X}'$ is not required for $\Phi(X,Y) = X'Y$ (see Caffarelli, 1996).

[6]See for instance Trudinger and Wang (2009), Figalli et al. (2013), De Philippis and Figalli (2014), and Chen and Wang (2016).

Let $\mathcal{G}_{n\ell}$ be a univariate linear sieve space[7] for ℓth dimension for $1 \leq \ell \leq d$ and $\mathcal{G}_n$ be the tensor product of $\mathcal{G}_{n1}, \ldots, \mathcal{G}_{nd}$. Then, it is standard to construct linear sieves with $\prod_{\ell=1}^{d} g_{n\ell}(x_\ell)$, where $g_{n\ell} \in \mathcal{G}_{n\ell}$ for $1 \leq \ell \leq d$, to approximate a multivariate function $g_0 \in \mathcal{G}$. We define the sieve approximation errors to $g_0 \in \mathcal{G} \subset C^{m+2,\gamma}(\mathcal{X})$ in $L_\infty(\mathcal{X})$-norm and $L_2(\mathcal{X})$-norm as

$$\rho_{\infty n} \equiv \inf_{g \in \mathcal{G}_n} \|g - g_0\|_\infty, \quad \rho_{2n} \equiv \inf_{g \in \mathcal{G}_n} \|g - g_0\|_{2,leb}.$$

Then, by letting $\dim(\mathcal{G}_n) = k_n$ and $\dim(\mathcal{G}_{n\ell}) = J_n$ for all $\ell \in \{1, \ldots, n\}$, we have that $\rho_{\infty n} = O\left(J_n^{-(m+2+\gamma)}\right) = O\left(k_n^{-(m+2+\gamma)/d}\right)$ if

- $\mathcal{G}_{n\ell} = \text{Pol}(J_n) = \left\{\sum_{k=1}^{J_n} a_k x_\ell^k, x_\ell \in \mathcal{X}_\ell : a_k \in \mathbb{R}\right\}$;
- $\mathcal{G}_{n\ell} = \text{TriPol}(J_n) = \left\{a_0 + \sum_{k=1}^{J_n} \left[a_k \cos(2k\pi x_\ell) + b_k \sin(2k\pi x_\ell)\right], x_\ell \in \mathcal{X}_\ell : a_k, b_k \in \mathbb{R}\right\}$; or
- $\mathcal{G}_{n\ell} = \text{Spl}(r, J_n) = \left\{\sum_{k=0}^{r-1} a_k x_\ell^k + \sum_{j=1}^{J_n} b_j \left[\max\{x_\ell - t_j, 0\}\right]^{r-1}, x_\ell \in \mathcal{X}_\ell : a_k, b_j \in \mathbb{R}\right\}$ with $r \geq m + 3$.

We consider the dual problem with the surplus function $\Phi(X, Y) = X'Y$ throughout the remaining chapter. Notice that the estimation procedure and convergence rate result can be easily applied to the problem with $\Phi(X, Y) = X'AY$ where A is invertible.[8] Let $\|\cdot\|_2$ denote the $L_2(P_X)$-norm. We employ the result of Theorem 1 in Shen and Wong (1994) (or Theorem 3.2 in Chen, 2007) to address how well one may estimate g_0 and its αth partial derivatives simultaneously in the $L_2(P_X)$-norm loss. That is, we bound

$$\|\hat{g}_n(X) - g_0(X)\|_2 \quad \text{and} \quad \|\partial^\alpha \hat{g}_n(X) - \partial^\alpha g_0(X)\|_2$$

where $\partial^\alpha = \partial^{[\alpha]} / \partial X_1^{\alpha_1} \cdots \partial X_d^{\alpha_d}$ given a d-tuple α of nonnegative integers and $[\alpha] = \alpha_1 + \cdots + \alpha_d$.

Different from other sieve M-estimation problems, it is not easy to establish that the optimum of the optimal transport problem is well-separated, i.e., $[Q(g) - Q(g_0)]^{1/2} \geq$

[7] See Section 2.3.1 in Chen (2007) for commonly used finite-dimensional linear sieves and their approximation error rates.

[8] This specification is not covered by Assumption 2.2, but we can interpret the Monge-Kantorovich problem as assigning from $\mathcal{X}$ to $A\mathcal{Y} \coloneqq \{AY : Y \in \mathcal{Y}\}$.

$M \|g - g_0\|_2$ for some positive M (See Condition C1 in Shen and Wong, 1994). When we consider $Q(g)$ as a functional from $\mathcal{G}$ to $\mathbb{R}$, it is convex on $\mathcal{G}$. Nevertheless, the second-order directional derivative of Q has not been developed.[9] Without any result on the second-order directional derivative of Q, the conventional theory in Huang (2001) is not applicable.

Gunsilius (2018) develops the second variation in a neighborhood of the optimum g_0 recently:

$$Q(g) - Q(g_0) = \frac{1}{2}\mathbb{E}_X\left[\|\nabla(g(X) - g_0(X))\|^2\right] + o\left(\|g - g_0\|^2\right) \quad (\|g - g_0\| \to 0).$$

Gunsilius (2018) derives the convergence rate for the kernel-weighted M-estimator of g_0 by imposing a high-level requirement that the probability measure for X satisfies Poincaré inequality:

$$\mathbb{E}_X\left[\|\nabla(g(X) - g_0(X))\|^2\right] \geq c\mathrm{Var}_X(g - g_0).$$

Many popular probability measures including those for normal, exponential, and uniform distribution, satisfy Poincaré inequality. However, it is not easy to verify this inequality for unknown distribution except for log-concave distribution (Bobkov, 1999).

We show that Poincaré inequality is satisfied without any further high-level condition. The main advantage of our setting is that we only consider the space of functions whose integrals are same. The smoothness condition of the density of one variable among two implies Poincaré inequalities with the Lebesgue measure and those with probability measure follow in the shrunk function space.

Other conditions for the convergence rate results are straightforward. Let us consider the space

$$\mathcal{F}_n = \left\{\sup_{x\in\mathcal{X}}[\Phi(x,y) - g(x)] - \sup_{x\in\mathcal{X}}[\Phi(x,y) - \Pi_n g_0(x)] : g \in \mathcal{G}_n\right\},$$

where Π_n is a $L_2(\mathcal{X})$ projection mapping from $\mathcal{G}$ to $\mathcal{G}_n$. Under the Assumption 2.3(i), $M_1\|g\|_{2,leb} \leq \|g\|_2 \leq M_2\|g\|_{2,leb}$ for some positive M_1 and M_2. We also note that

$$\left|\sup_{x\in\mathcal{X}}[\Phi(x,y) - g(x)] - \sup_{x\in\mathcal{X}}[\Phi(x,y) - \Pi_n g_0(x)]\right| \leq \sup_{x\in\mathcal{X}}|g(x) - \Pi_n g_0(x)| = \|g - \Pi_n g_0\|_\infty.$$

[9] Chartrand et al. (2009) derive the first-order directional (Gâteaux) derivative of Q.

Then, the existing results on L_∞-metric entropy of the space $\mathcal{F}_n$ for small $\varepsilon > 0$, $H(\varepsilon, \mathcal{F}_n)$, can be employed for finite-dimensional linear sieves:

$$H(\varepsilon, \mathcal{F}_n) \leq H(\varepsilon, \mathcal{G}_n) \leq Ck_n \log(1/\varepsilon),$$

Here k_n controls the effective size of the approximating space $\mathcal{G}_n$. Then we can develop the convergence rate of our estimator $\hat{g}_n$:

Theorem 2.2. *Suppose that Assumptions 2.3 holds for the dual Monge-Kantorovich problem with $\Phi(X,Y) = X'Y$. Then $\|\hat{g}_n - g_0\|_2 = O_p\left(\sqrt{k_n/n} + \|\Pi_n g_0 - g_0\|_2\right)$.*

Corollary 2.1. *Suppose that Assumptions 2.3 holds for the dual Monge-Kantorovich problem with $\Phi(X,Y) = X'Y$. Then $\|\partial^\alpha \hat{g}_n - \partial^\alpha g_0\|_2 = O_p\left(k_n^{[\alpha]/d}\|\hat{g}_n - g\|_2\right)$.*

For $\ell = 1, \ldots, d$ and $p = m + 2 + \gamma$ where $g_0 \in C^{m+2,\gamma}(\mathcal{X})$, if

- $\mathcal{G}_{n\ell} = \mathrm{Pol}(J_n)$, $p > d$, and $J_n^{3d}/n \to 0$;
- $\mathcal{G}_{n\ell} = \mathrm{TriPol}(J_n)$, $p > d/2$, and $J_n^{2d}/n \to 0$; or
- $\mathcal{G}_{n\ell} = \mathrm{Spl}(r, J_n)$, $r \geq m+3$, $p > d/2$, and $J_n^{2d}/n \to 0$,

then $\|\hat{g} - g_0\|_2 = O_p\left(\sqrt{J_n^d/n} + J_n^{-p}\right)$ and $\|\partial^\alpha \hat{g}_n - \partial^\alpha g_0\|_2 = O_p\left(J_n^{[\alpha]}\sqrt{J_n^d/n} + J_n^{-(p-[\alpha])}\right)$. By choosing $J_n = O\left(n^{1/(2p+d)}\right)$, $\|\hat{g} - g_0\|_2 = O_p\left(n^{-p/(2p+d)}\right)$, which is the same as the optimal rate in the context of regression and density estimations (see Stone, 1982). Also, the convergence rate of the sieve estimate of Monge transport for Y_ℓ $(1 \leq \ell \leq d)$, T_ℓ, can be easily derived with same order of J_n:

$$\left\|\hat{T}_\ell - T_\ell\right\|_2 = \left\|\frac{\partial \hat{g}_n}{\partial X_\ell} - \frac{\partial g_0}{\partial X_\ell}\right\|_2 = O_p\left(n^{-(p-1)/(2p+d)}\right).$$

2.5 Application to (conditional) vector quantiles

The optimal transport problem can be applied to quantiles. In dimension 1, there are three types of definitions for quantile map $Q_Y : [0,1] \to \mathbb{R}$:

1. Inverse of a CDF, F_Y: $Q_Y(u) = F_Y^{-1}(u) = \inf\{y : F_Y(y) \geq u\}$;

2. $Q_Y(u) = \arg\min_q \mathbb{E}[\rho_x(Y-q)]$ where $\rho_x(z) = uz^+ + (1-u)z^-$;

3. Nondecreasing transport map T such that $F_U\left(T^{-1}(y)\right) = F_Y(y)$ where $U \sim \mathrm{U}([0,1])$.

When we try to generalize this concept to the multivariate case, the first two definitions cannot be employed to the case when Y is multivariate. However, the mapping can be interpreted as a gradient of a convex function. Carlier et al. (2016) and Chernozhukov et al. (2017), using the last definition, define the vector quantile based on the optimal transport theory:

Definition 2.1 (Vector quantile)**.** Vector quantile associated with $Y \sim F_Y$ is the unique gradient of a convex function $Q_Y(U) = \nabla g_0(U)$ such that $F_U\left(\nabla g_0^{-1}(Y)\right) = F_Y(Y)$ where $U \sim \mathrm{U}\left([0,1]^d\right)$.

We note that, in dimension 1, this definition coincides with the classical notion of a quantile. Chernozhukov et al. (2017) define empirical quantiles and show that this is consistent. However, besides consistency, further asymptotic behaviors of empirical vector quantiles are an open problem. The above definition leads to the following dual problem:[10]

$$\inf_{g\in\mathcal{G}} \mathbb{E}_Y\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g(u)\right\}\right],\ \mathbb{E}[g(U)] = \int_{[0,1]^d} g(U)\,dU = 0.$$

We do not observe the value of $U \in \mathcal{U}$. We aim to estimate quantile function $Q_Y : \mathcal{U} \mapsto \mathcal{Y}$ given the distribution for U, which is usually $\mathrm{U}\left([0,1]^d\right)$. Using the optimal transport theory, there exists a measurable map $U \mapsto Q_Y(U)$ from $\mathcal{U}$ to $\mathbb{R}$, such that the map $U \mapsto Q_Y(U)$ is the unique gradient of convex function, g_0. This implies that[11]

$$\left(Q_Y(U) - Q_Y\left(\bar{U}\right)\right)'\left(U - \bar{U}\right) \geq 0, \quad \text{for all } U, \bar{U} \in [0,1]^d.$$

Whenever $U \sim \mathrm{U}\left([0,1]^d\right)$, the random vector $Q_Y(U)$ has the distribution function $F_Y(\cdot)$,

[10] Brenier (1991): If $\Phi(X,Y) = X'Y$ in the Euclidean space, F_X is absolutely continuous, and F_X, F_Y have finite second order moments, then there is a unique optimal Monge coupling between F_X and F_Y. McCann (1995) extends the result without moment conditions, but he does not refer to the Kantorovich or dual problem. Therefore, the Monge-Kantorovich problem might not be well defined.

[11] A function f is convex if and only if $f''(U) := Q'_{Y|Z}(U,Z) \geq 0$ for all U. In this case, if $U \leq \bar{U}$, $Q_{Y|Z}(U,Z) \leq Q_{Y|Z}\left(\bar{U},Z\right)$. Indeed, in the scalar case the requirement that the transform is the gradient of a convex map reduces to the requirement that transform is nondecreasing.

that is,

$$F_Y(Y) = \int 1\{Q_Y(U) \leq Y\}\, dU, \quad \text{for all } Y \in \mathbb{R}.$$

Moreover, there exists a random variable $V = U$ such that almost surely $Y = Q_Y(V)$ and $V \sim U\left([0,1]^d\right)$.

The sieve M-estimation method and its convergence rate result in Section 2.3 and 2.4 are directly applicable for vector quantiles. Since $U \sim U\left([0,1]^d\right)$, we obtain desirable convergence rate results with suitable order of sieves and Assumption 2.3 for P_Y.

To extend the notion of the vector quantiles to the conditional vector quantiles, we consider a random vector (Y, U, Z) defined on a complete probability space. The random vector Z is a vector covariate, taking values in $\mathcal{Z} \subset \mathbb{R}^{d_Z}$. We further assume that U is uniformly distributed on $[0,1]^d$ and independent of Z. Then, the joint measure of (U, Z), P_{UZ}, satisfies that $P_{UZ}(A \times B) = P_U(A) P_Z(B)$ for any $A \in \mathcal{U}$ and $B \in \mathcal{Y}$.

The conditional vector quantile function also can be characterized and even defined with a solution to the dual optimal transportation problem. Then we consider the dual optimal transport problem in Carlier et al. (2016):

$$\inf_g \left\{ \mathbb{E}_{UZ}[g(U,Z)] + \mathbb{E}_{YZ}\left[\sup_{u \in [0,1]^d} \left\{ u'Y - g(u,Z) \right\} \right] \right\}$$

This problem can be written as

$$\inf_{g \in L_1(P_U P_Z)} \left\{ \mathbb{E}_Z \left[\mathbb{E}_{U|Z}[g(U,Z)|Z] + \mathbb{E}_{Y|Z}\left[\sup_{u \in [0,1]^d} \left\{ u'Y - g(u,Z) \right\} |Z \right] \right] \right\}$$
$$= \inf_{g \in L_1(P_U P_Z)} \mathbb{E}_{YZ}\left[\sup_{u \in [0,1]^d} \left\{ u'Y - g(u,Z) \right\} \right], \ \mathbb{E}_{U|Z}[g(U,Z)|Z] = 0 \quad (2.6)$$

Let $\left\{(Y_i', Z_i')'\right\}_{i=1}^n$ be $(d + d_Z)$-dimensional independent and identically distributed (i.i.d.) sequences of observations with unknown joint probability measure P_{YZ} . We consider $g_0 \in \mathcal{G}$ as the true unknown infinite-dimensional parameter, where $\mathcal{G}$ is a linear subspace of the space of real-valued functions with $\mathbb{E}\left[g(U,Z)^2\right] < \infty$ and $\mathbb{E}_{U|Z}[g(U,Z)|Z] = \int_{[0,1]^d} g(U,Z)\, dU = 0$ for all $Z \in \mathcal{Z}$. Let $\{p_j(U,Z),\ j = 1, 2, \ldots\}$ denote a sequence of known basis functions

that can approximate any $g \in \mathcal{G}$. Then, for a finite-dimensional linear sieve

$$\mathcal{G}_n = \left\{ g_n : \mathcal{U} \times \mathcal{Z} \to \mathbb{R},\ g_n(U,Z) = \sum_{j=1}^{k_n} a_j p_j(U,Z) :\ \int_{[0,1]^d} g(U,Z)\, dU = 0 \right\},$$

with $\dim(\mathcal{G}_n) = k_n \to \infty$ slowly as $n \to \infty$, we estimate the unknown sieve coefficients of g_n:

$$\hat{g}_n = \arg\min_{g_n \in \mathcal{G}_n} \frac{1}{n} \sum_{i=1}^{n} \sup_{u \in [0,1]^d} \left\{ u' Y_i - g_n(u, Z_i) \right\}$$

is a sieve M-estimator of g_0 and $\nabla_U \hat{g}(U,Z)$ is a sieve estimator of conditional vector quantile $Q_{Y|Z}$.

We are now ready to state the conditions for the convergence rate results of sieve M-estimator $\hat{g}_n$ and its partial derivatives:

Assumption 2.5. *For each $Z \in \mathcal{Z}$, the conditional probability measure $P_{Y|Z}$ has a compact and convex support $\mathcal{Y}_Z \subset \mathbb{R}^d$ and admits a density $f_{Y|Z}$ such that*

(i) $f_{Y|Z}$ is bounded away both from zero and infinity on $\mathcal{Y}_Z$; and

(ii) $f_{Y|Z} \in C^{m,\gamma}(\mathcal{Y}_Z)$ for some $m \in \mathbb{N}$ and $\gamma \in (0,1)$.

Assumption 2.6. *$g_0(U,Z) \in C^{m_Z,\gamma_Z}(\mathcal{Z})$ for each $U \in \mathcal{U}$.*

We note that under Assumption 2.5, $g_0(U,Z) \in C^{m+2,\gamma}(\mathcal{U})$ for each $Z \in \mathcal{Z}$. It is a simple application of Lemma 2.2. If Z takes only a finite number of values, we estimate g_0 and its derivative from the observations with the same value of Z. Assumption 2.5 is sufficient to apply Theorem 2.2 and Corollary 2.1.

For continuous random variable Z, a smoothness of g_0 with respect to Z is required for our sieve M-estimator. It is a high-level requirement, but we ensure from the above two conditions that $g_0 \in C^{\min\{m+2,m_Z\},\min\{\gamma,\gamma_Z\}}(\mathcal{U} \times \mathcal{Z})$.

Theorem 2.3. *Suppose that Assumptions 2.5 and 2.6 hold. Then, the sieve estimator, $\hat{g}_n$, exists uniquely with probability approaching one as $n \to \infty$, and*

$$\begin{aligned} \left\| \hat{Q}_{Y_\ell|Z}(U|Z) - Q_{Y_\ell|Z}(U|Z) \right\|_2 &= \left\| \frac{\partial \hat{g}_n}{\partial u_\ell} - \frac{\partial g_0}{\partial u_\ell} \right\|_2 \\ &= O_p\left(k_n^{1/(d+d_Z)} \sqrt{k_n/n} + k_n^{1/(d+d_Z)} \left\| \Pi_n g_0 - g_0 \right\|_2 \right). \end{aligned}$$

In contrast to Carlier et al. (2016), who implement vector quantile regression with discretization and linear programming problem, we employ the sieve method to the conditional optimal transport problem. Our estimators have several attractive features that are not shared by linear programming problem. We can uniformly control the convergence rate of our estimator as typical sieve estimators under the smoothness condition of g. Also, sieve methods allow imposing shape constraints easily. The solution to the optimal transport problem is a convex function in theory, but it still possible to suffer from the crossing problem in the implementation of vector quantiles. To obtain a stable monotone estimate, we can simply impose the convexity constraint on the function space. Our estimation method can also capture the nonlinear effect of conditioning variable on vector quantiles. It decreases the possibility of misspecification.

Alternatively, we can consider the following optimal transportation problem: for each $Z \in \mathcal{Z}$

$$\inf_{g_Z \in \mathcal{G}} \mathbb{E}_{Y|Z} \left[\sup_{u \in \mathcal{U}} \left\{ u'Y - g_Z(u) \right\} \right], \ \mathbb{E}\left[g_Z(U) \right] = 0, \tag{2.7}$$

where $U \mapsto g_Z(U)$ are lower semicontinuous. We can easily check that $g_0(U, Z) = g_{Z0}(U)$, where $g_{Z0}(U)$ is a unique solution to the problem (2.7), with probability 1. Furthermore, the infimum in (2.7) coincides with the infimum in (2.6). Local polynomial approach with this problem provides another useful method. We estimate the unknown function $g_Z(U)$ at a fixed point Z and the estimated function changes with Z. The asymptotic properties of an estimator for g_Z will depend on only Assumption 2.5 but require additional conditions for local polynomial method.

2.6 Simulation study

In our procedure, the smoothing parameter is the number of terms in the sieve approximation of the unknown function. In empirical analysis, we select this parameter by using information criteria (IC). That is, $k_n = \hat{k}_n = \hat{J}_X^d \times \hat{J}_Z^{d_Z}$ is selected to minimizing

$$\min_k \left\{ 2 \sum_{i=1}^n \sup_{x \in \mathcal{X}} \left\{ \Phi(x, Y_i) - \hat{g}_{n,k}(x, Z_i) \right\} + C_n k \right\},$$

where $\hat{g}_{n,k}$ denotes the sieve M estimate given k basis functions. When $C_n = 2$ or $\log(n)$, the IC becomes Akaike information criterion (AIC) and Bayesian information criterion (BIC) respectively. Compared to the cross-validation method,[12] the IC is a computationally simple and useful way to select the number of basis functions. In this section, we see the performance of the estimators across different sieve number of terms.

In both simulations and empirical parts, we use Bernstein polynomials to approximate each of the unknown functions. We note that Bernstein polynomial is very convenient to enforce several restrictions. We impose the zero mean constraints in estimation routine. Also, it is possible to obtain a stable monotone estimate by adding the convexity constraint on the function space (See the appendix for imposing the convexity constraint.).

We assess the performance of our estimation method for in various distributions. The performance of the estimators is evaluated using the mean absolute deviation error, MADE $= \int_{\mathcal{X}\times\mathcal{Z}} |\hat{g}(X,Z) - g_0(X,Z)| f_X(X) f_Z(Z) dXdZ$, as well as their bias, standard deviation, and mean squared error.

We first consider a two-dimensional Gaussian copula, $C : [0,1]^2 \to [0,1]$. Th Gaussian copula is defined as

$$C(Y_1, Y_2; \rho) \coloneqq \Phi\left(\Phi_1^{-1}(Y_1), \Phi_2^{-1}(Y_2); \rho\right),$$

where Φ_1 and Φ_2 are standard univariate normal CDFs and $\Phi(\cdot,\cdot;\rho)$ denotes the joint CDF of the bivariate normal distribution with unit variances and covariance ρ. So, ρ is the dependence parameter of the copula measuring the dependence between the standard univariate normal marginals Φ_1 and Φ_2.

We estimate $Q_Y(U)$ for the simulated Gaussian copula with $\rho = 0.5$ using Bernstein polynomial. Our estimation method is different to the conventional estimation method for

[12] To select the number of basis, $k = J^d$, we can also employ the cross-validation method: As a first step, we compute the leave-one-out estimator,

$$\hat{g}_{i_0}^k = \arg \inf_{g_z \in \mathcal{G}_{nk}} \sum_{i \neq i_0}^{n} \sup_{x \in \mathcal{X}} \{\Phi(x, Y_i) - g(x, Z_i)\},$$

for $i = 1, \ldots, n$, and then use as criterion the over all global sample analog,

$$CV(k) = \sum_{i=1}^{n} \sup_{x \in \mathcal{X}} \left\{\Phi(x, Y_i) - \hat{g}_i^k(x, Z_i)\right\}.$$

We then choose the number of basis as the minimizer of $CV(k)$. The optimal value for k is then easily evaluated for a grid of natural numbers.

quantile functions. Instead of estimating it for each u, we estimate the whole function. In earlier section, we derived the uniform convergence rate of $Q_Y(U) = \nabla_U g(U)$.

For model 2, Table 2.1 reports the simulation results with a different number of sieve terms. The sieve estimator with three sieve terms is the best for all sample sizes. Estimation performance is mainly affected by the variance and imposing convexity constraint gives us more stable estimate with smaller variance

Table 2.1: Performance of the sieve M-estimators with different order of sieves

$Nobs = 200$	No convexity constraint				Convexity constraint			
J_U	ISB	IV	IMSE	IMADE	ISB	IV	IMSE	IMADE
$X \sim \mathrm{U}\left([0,1]^2\right)$, Y: Gaussian copula with linear correlation parameters $\rho = -0.5$								
2	0.108	0.165	0.276	5.798	0.108	0.165	0.276	5.796
3	0.012	0.193	0.213	5.112	0.013	0.189	0.209	5.073
4	0.012	0.217	0.236	5.360	0.012	0.199	0.218	5.151

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

All values are multiplied by 100

Next, we examine the finite-sample performances of our estimator for the conditional vector quantiles. We consider the following one-dimensional conditional quantiles:

$$Y_1 = 5Z\varepsilon_1, \quad Z, \varepsilon_1 \sim \mathrm{U}([0,1]);$$
$$Y_2 = \sin(2\pi Z) + \left(1 + Z^2\right)\varepsilon_2/4, \quad \varepsilon_2 \sim \mathrm{N}(0,1), \ Z \sim \mathrm{U}([0,1]).$$

The above two models are smooth location-scale models. We are interested in estimating the uth conditional quantile of Y, $Q_U(Z) = a(Z) + b(Z) Q_{\varepsilon,U}$ where $Q_{\varepsilon,U}$ is the uth quantile of ε. Figure 2.1 illustrates true quantile functions conditional on Z over regular grid points between 0.025 and 0.975 with an increment 0.025. We find that the quantile function in the first model is linear and monotone in Z, but the second one is nonlinear and not monotone.

For model 1, Table 2.2 reports the integrated squared error (IBias2), variance (IVar), MSE (IMSE) and mean absolute deviation error (IMADE) with a different number of sieve terms. The results are based on 500 simulated data sets with sample sizes $Nobs$ =200 and 500. We observe that all values decrease as n increases as we expect. The sieve estimator for model 1 with two sieve terms is the best for all sample sizes. This is because the model 1 has a simple specification, and it does not require many sieve terms. The biases are already

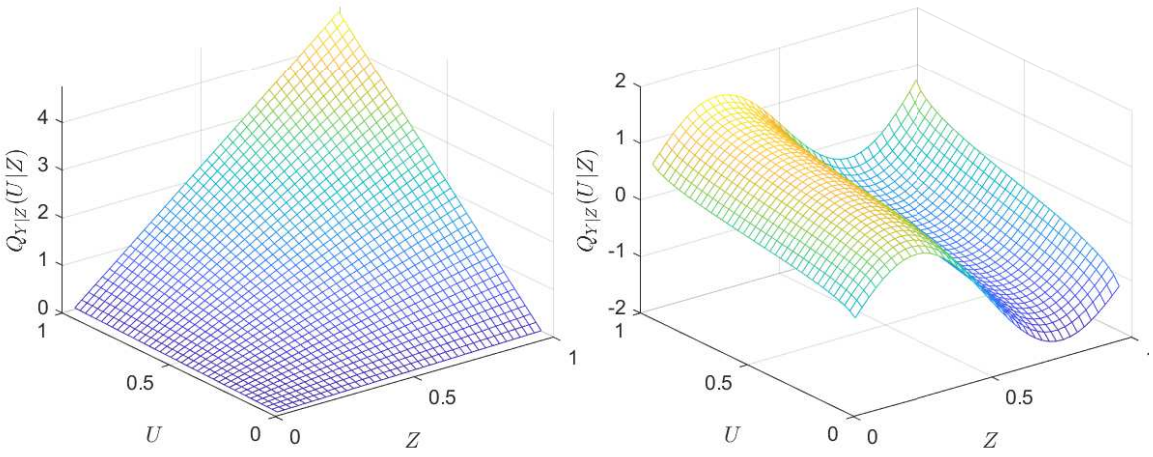

Figure 2.1: True conditional quantile functions for model 1 (left) and model 2 (right)

tiny with small number of sieve terms, and estimation performance is mainly affected by the variance. Also, we find that imposing convexity constraint gives us more stable estimate with smaller variance, but the difference of performance decreases as n increases.

Table 2.2: Performance of the sieve M-estimators with different order of sieves: Model 1

	No convexity constraint				Convexity constraint			
$J_U = J_Z$	ISB	IV	IMSE	IMADE	ISB	IV	IMSE	IMADE
				$Nobs = 200$				
2	0.048	1.213	1.261	7.455	0.048	1.209	1.257	7.452
3	0.062	1.928	1.990	9.209	0.063	1.869	1.932	9.090
4	0.079	2.517	2.596	10.477	0.083	2.329	2.412	10.121
5	0.100	3.064	3.164	11.590	0.100	2.712	2.812	10.938
				$Nobs = 500$				
2	0.012	0.504	0.516	4.764	0.012	0.503	0.515	4.760
3	0.013	0.782	0.796	5.774	0.014	0.777	0.790	5.756
4	0.016	1.024	1.040	6.623	0.017	0.971	0.988	6.462
5	0.021	1.260	1.281	7.357	0.021	1.150	1.171	7.056

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

All values are multiplied by 100

For model 2, Table 2.3 reports the simulation results with a different number of sieve terms. The sieve estimator with five sieve terms is the best for all sample sizes. Since the model 2 is nonlinear and not monotone in Z, it requires more sieve terms than model 1. Estimation performance for model 2 is mainly affected by the bias. In particular, model 1 has a simpler form than model 2, but the estimation performs better for model 2.

As a simple example for two-dimensional CVQR, we consider the simplest multivariate

Table 2.3: Performance of the sieve M-estimators with different order of sieves: Model 2

	No convexity constraint				Convexity constraint			
$J_U = J_Z$	ISB	IV	IMSE	IMADE	ISB	IV	IMSE	IMADE
				$Nobs = 200$				
3	0.685	0.441	1.126	8.304	0.685	0.417	1.103	8.236
4	0.474	0.585	1.059	8.093	0.540	0.505	1.045	8.042
5	0.053	0.669	0.723	6.283	0.069	0.553	0.622	5.826
6	0.030	0.786	0.817	6.726	0.057	0.619	0.676	6.110
				$Nobs = 500$				
3	0.634	0.177	0.811	7.161	0.635	0.171	0.806	7.143
4	0.423	0.235	0.658	6.596	0.479	0.208	0.687	6.705
5	0.034	0.261	0.295	3.987	0.040	0.225	0.265	3.778
6	0.010	0.310	0.320	4.195	0.030	0.257	0.287	3.966

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

All values are multiplied by 100

model with the bivariate standard normal model:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & Z \\ Z & 1 \end{pmatrix}\right).$$

If (Y_1, Y_2) are independent, that is $Z = 0$, then $Y_1 = Q_1(U_1)$ and $Y_2 = Q_2(U_2)$, which is equivalent to two single-dimensional quantiles. An increase in Y_1 is not associated to the change in Y_2. Assume (Y_1, Y_2) are not independent. In the case that $Z > 0$, an increase in U_1 not only increases Y_1, but also Y_2. In Figure 2.2, we plot the quantiles with different values of conditioning variable Z.

The simulation results are presented in table 2.4. The integrated squared bias, variance, mean squared error and MADE are the sum of results for $Q_{Y_1|Z}$ and $Q_{Y_2|Z}$ conditional on $Z = 0$. We only report the results with four set of sieve terms. By adding one more variable, the dimension of the tensor product spaces increases by product the dimension of the additional variable. Estimation performance for this model is mainly affected by the variance, and it requires a smaller number of sieve terms for Z.

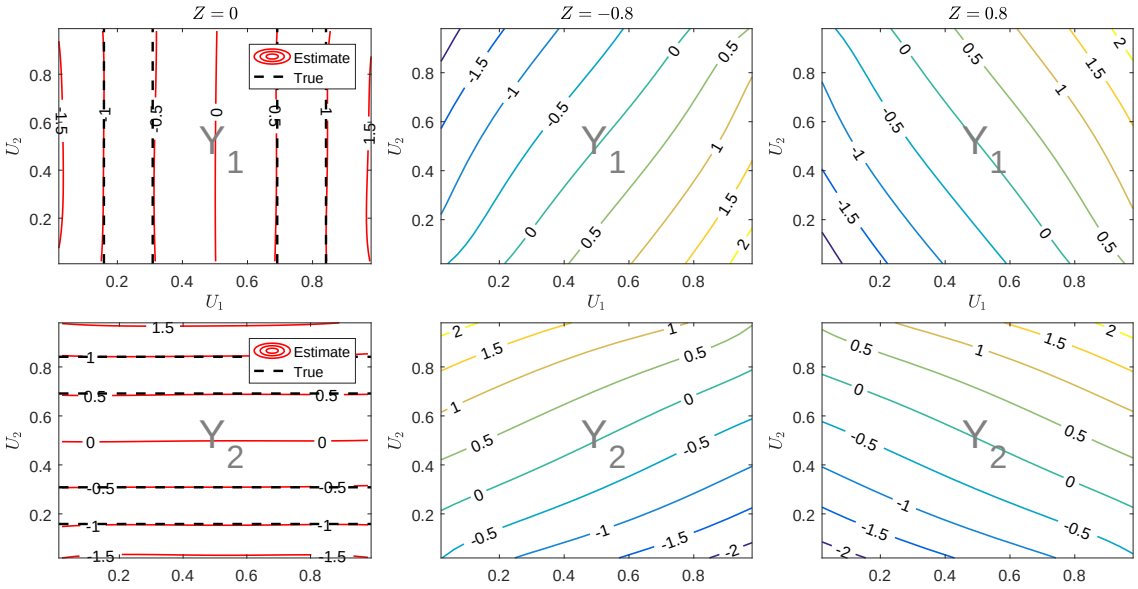

Figure 2.2: Vector quantiles for Y_1 and Y_2 conditional on $Z = z$: $Y_1, Y_2 \sim N(0,1)$, $cov(Y_1, Y_2) = Z$

Table 2.4: Performance of the sieve M-estimators with different order of sieves: Model 3

	$Nobs = 200$				$Nobs = 500$			
J_Z	ISB	IV	IMSE	IMADE	ISB	IV	IMSE	IMADE
$J_U = 2$								
1	6.031	2.111	8.142	26.782	6.022	0.864	6.886	22.720
2	4.325	4.301	8.632	30.389	4.291	1.789	6.080	24.332
3	4.566	4.880	9.446	31.758	4.518	2.022	6.540	25.029
$J_U = 3$								
1	5.691	2.611	8.302	27.264	5.699	1.051	6.750	22.790

Integrated squared bias (ISB), variance (IV), and mean squared errors (IMSE)

All values are multiplied by 100

2.7 Empirical application

We demonstrate the use of our sieve estimation method to the optimal transport problem on a standard application of quantile regression. The data are taken from the Engel's data on household expenditures. Engel's data set is richer and classifies household expenses in nine broad categories, but we focus on a two-dimensional dependent variable. We choose food expenditure and housing expenditure as Y_1 and Y_2. We take the total expenditure as a conditioning variable.

We first run a pair of one-dimensional conditional vector quantiles, Y_1 on Z and Y_2 on Z. We plot the results in Figure 2.3; the curves drawn here are $U \mapsto Q_{Y_\ell|Z}(U, Z)$, $\ell = 1, 2$,

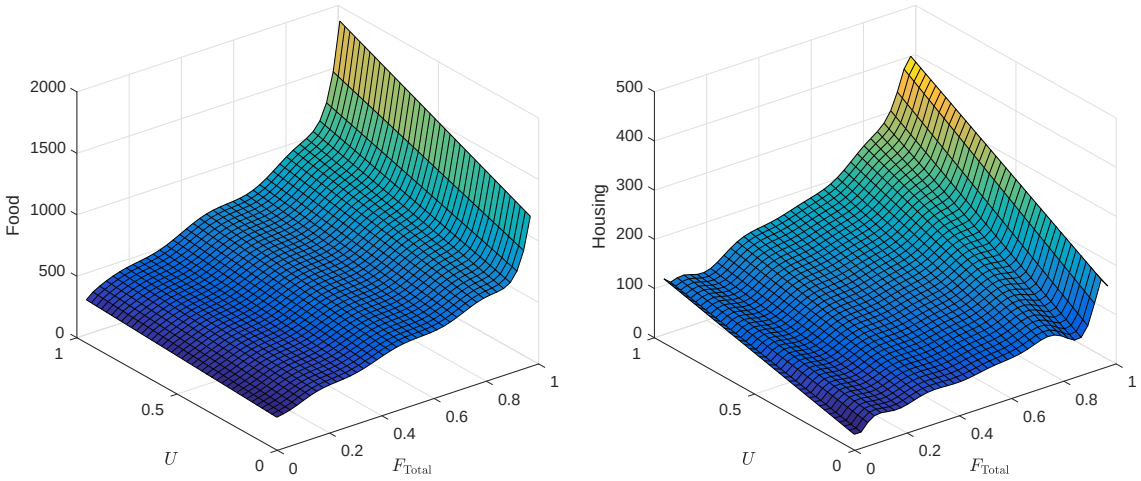

Figure 2.3: One-dimensional vector quantile regression conditional on total expenditure

for values of U and the cdf of Z. We can find that our estimation does not suffer from the "crossing problem." However, it does not convey information about the joint conditional dependence in Y_1 and Y_2 given X.

The two-dimensional vector quantile yields the curves drawn in Figure 2.4, which are $(U_1, U_2) \to Q_{Y_j|Z}(U_1, U_2, Z))$, $j = 1, 2$.

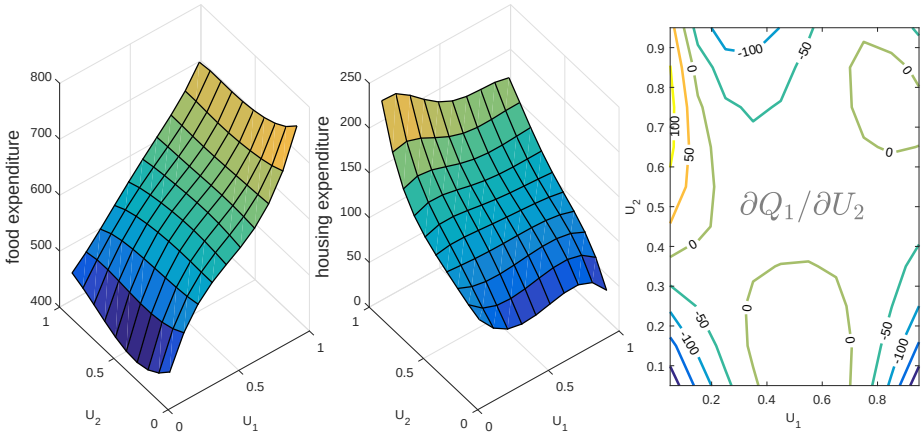

Figure 2.4: Two-dimensional vector quantile regression conditional on median value of total expenditure

As Carlier et al. (2016) mentioned, the two-dimensional vector quantile may be used to check if Y_1 and Y_2 are local complements or substitutes. Two graphs on the right-hand side in Figure 2.5 shows that, at a median level of income, the food and housing expenditure are

local substitutes, which is also consistent to the result in Carlier et al. (2016). However, we can find that those relationships are not so strong.

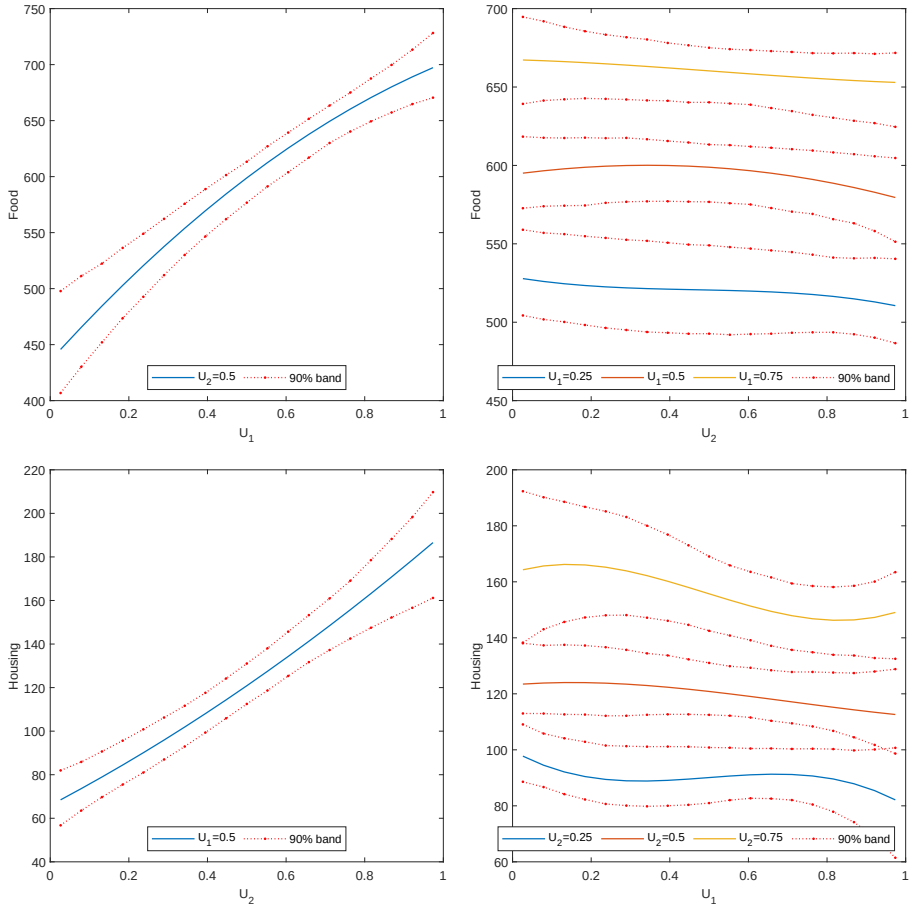

Figure 2.5: 90% Bootstrap confidence bands of two-dimensional quantiles for food and housing expenditure

2.8 Conclusion

In this chapter, we examined the statistical properties of sieve M-estimation in the optimal transport problem. We provide identification and regularity of a solution to the optimal transport problem and establish the nonparametric convergence rate. We also present Monte Carlo simulation results with a different choice of smoothing parameters and the performance of the sieve M-estimator.

I highlight two directions for future work. The first direction is to develop the inference theory on the entire optimal transport mapping and its linear functionals. We here only

present the convergence rate of the sieve estimator. One possible inference method is to use the weighted bootstrap method. We will be able to check the conditions for the validity of the weighted bootstrap method easily. Then, we can use this result to construct uniform confidence bands for linear functionals and how to test shape restrictions.

The second direction is to extend the original optimal transport problem to allow for more general structures. Compared to the development of the optimal transport problem with flexible output functions, our asymptotic theory applies to the simplest output function $X'Y$. To broaden the applicability of this theory further, it will be essential to extend the well-separateness property to more flexible output functions. These are challenging problems for future research.

2.9 Appendix

2.9.1 Bernstein polynomials with convex constraints

Without any constraint on $g(X)$, the unique solution (up to constant) for the dual Kantorovich problem with $\Phi(X,Y) = X'Y$ is convex. However, the estimator of $g(X)$, $\hat{g}_n(X)$, might be nonconvex at the values close to the boundary of $\mathcal{X}$. To obtain more stable estimator, we can give a convexity restriction without loss of generality. Among many several linear approximating space, we consider the following Bernstein polynomial sieve space:

$$\mathcal{G}_n = \left\{ g_n : \mathcal{X} \to \mathbb{R} : g_n(X) = \sum_{j_1,\cdots,j_d=0}^{J_n} a_{j_1\cdots j_d} \left[\prod_{\ell=1}^{d} g_{j_\ell}(X_\ell) \right] : \right.$$
$$\left. g_{j_\ell}(X_\ell) = \frac{1}{\bar{X}_\ell - \underline{X}_\ell} \binom{J_n}{j_\ell} (X_\ell - \underline{X}_\ell)^{j_\ell} \left(\bar{X}_\ell - X_\ell\right)^{J_n - j_\ell} \right\},$$

for $J_n = 1, 2, \ldots$, where $g_{j_\ell}$is the Bernstein basis polynomial. Let $g(x)$ be a real-valued convex function in $\mathcal{F} = \left\{ g \in C^{m+2,\gamma}(\mathcal{X}) : 2g\left(\frac{X_1+X_2}{2}\right) \leq g(X_1) + g(X_2),\ \forall X_1, X_2 \in \mathcal{X} \right\}$. Notice that we are not assuming that the true function $g(X)$ has derivatives of any order. For the simplicity, we consider the one-dimensional sieve space ($d = 1$):

$$\mathcal{F}_n = \{ g_n(X) \in \mathcal{G}_n :\ A_{J_n} \alpha_{J_n} \geq 0 \}$$

where

$$A_{J_n}\alpha_{J_n} \equiv \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ & & \ddots & & & \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}_{(J_n-1)\times(J_n+1)} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{J_n} \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Since the second derivatives of $g_n(X)$ can be written as

$$g_n^{(2)}(X) = J_n(J_n-1)\sum_{j=0}^{J_n-2}(\alpha_{j+2}-2\alpha_{j+1}+\alpha_j)\,g_j(X, J_n-2),$$

the above restriction ensures $g_n^{(2)}(\cdot) \geq 0$ for all n.

2.9.2 Auxiliary Lemma

Computing exact $\sup_{x\in\mathcal{X}}\{\Phi(x, Y_i) - g_n(x)\}$ is easy for the small dimension of k_n for $\mathcal{G}_n$, but it is burdensome for the large value of J_n. In this case, we consider the set of grid points or samples on $\mathcal{X}$, $\mathcal{X}_n$, and then compute

$$h_n(Y_i, g) = \max_{x\in\mathcal{X}_n}\{\Phi(x, Y_i) - g_n(x)\}.$$

The following lemma shows that the error caused by the set of grid points is negligible with large sample.

Lemma 2.3. *Suppose that $|g - g_0|_\infty \leq \varepsilon$ for continuous functions g and g_0 on $\mathcal{X}$. Under Assumption 2.1 and 2.2.(i),* $\sup_y |h(y) - h_{J_n}(y)| = \varepsilon + O_p\left(J_n^{-1}\right)$ *where*

$$h(Y) = \sup_{x\in\mathcal{X}}[\Phi(x, Y) - g(x)].$$

$$h_{J_n}(Y) = \max_{x_j}[\Phi(x_j, Y) - g_n(x_j)], \quad x_j \in \{x_1, \ldots, x_{J_n}\}.$$

Proof of Lemma 2.3. Since

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g_n\left(x\right)\right] &\leq \sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g\left(x\right)\right]+\sup_{x\in\mathcal{X}}\left|g_n\left(x\right)-g\left(x\right)\right| \\ &= \sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g\left(x\right)\right]+\varepsilon; \\ \sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g_n\left(x\right)\right] &\geq \sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g\left(x\right)\right]-\sup_{x\in\mathcal{X}}\left|g_n\left(x\right)-g\left(x\right)\right| \\ &= \sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g\left(x\right)\right]+\varepsilon;\end{aligned}$$

We note that

$$\sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g_n\left(x\right)\right]\geq\max_{x_j}\left[\Phi\left(x_j,Y\right)-g_n\left(x_j\right)\right].$$

Let $\tilde{x}$ be the vector such that

$$\Phi\left(\tilde{x},Y\right)-g_n\left(\tilde{x}\right)=\sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g_n\left(x\right)\right].$$

Since there is $J_n\in\mathbb{N}$ such that $\left\|\tilde{x}-\tilde{x}_{J_n}\right\|\coloneqq\min_{j\in\{1,\ldots,J_n\}}\left\{\left\|\tilde{x}-x_j\right\|\right\}\leq C/J_n$,

$$\begin{aligned}&\sup_{y\in\mathcal{Y}}\left\{\min_j\left\{\Phi\left(\tilde{x},Y\right)-g_n\left(\tilde{x}\right)-\left[\Phi\left(x_j,Y\right)-g_n\left(x_j\right)\right]\right\}\right\} \\ \leq&\sup_{y\in\mathcal{Y}}\left|\Phi\left(\tilde{x},Y\right)-\Phi\left(\tilde{x}_{J_n},Y\right)\right|+\left|g_n\left(\tilde{x}\right)-g_n\left(\tilde{x}_{J_n}\right)\right|\leq C'/J_n.\end{aligned}$$

□

2.9.3 Proofs of main results

Proof of Theorem 2.1. We first note that both $Q_n\left(g\right)$ and $Q\left(g\right)$ are continuous under the sup norm since

$$\left|\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y_i\right)-g_1\left(x\right)\right\}-\sup_{x\in\mathcal{X}}\left[\Phi\left(x,Y\right)-g_2\left(x\right)\right]\right|\leq\sup_{x\in\mathcal{X}}\left|g_1\left(x\right)-g_2\left(x\right)\right|.$$

Let $\bar{\mathcal{G}}_k=\mathcal{G}_k\cap\{g:\left\|g-g_0\right\|_\infty\leq 2\varepsilon\}$ and $\mathcal{B}_k$ be its boundary. Since $Q_n\left(g\right)$ converges to $Q\left(g\right)$ in probability uniformly on $\bar{\mathcal{G}}_k$ for all k, as a result of Theorem 3.1 in Chen (2007), the minimand $\tilde{g}_n$ of $Q_n\left(g\right)$ on $\bar{\mathcal{G}}_n$ is consistent for g_0. The remaining proof comes from the one of Theorem 2.7 in Newey and McFadden (1994): The event that $\tilde{g}_n$ is within ε of g_0, so that

$Q_n(\tilde{g}_n) \leq \min_{g \in \mathcal{B}_n} Q_n(g)$, occurs with probability approaching one.In this event, for any g in $\mathcal{G}_n \cap \{g : \|g - g_0\|_\infty > 2\varepsilon\}$, there is a linear convex combination $\lambda\tilde{g}_n + (1-\lambda) g$ that lies in $\mathcal{B}_n$, so that $Q_n(\tilde{g}_n) \leq Q_n(\lambda\tilde{g}_n + (1-\lambda) g)$. It follows from the convexity of Q_n that $Q_n(\lambda\tilde{g}_n + (1-\lambda) g) \leq \lambda Q_n(\tilde{g}_n) + (1-\lambda) Q_n(g)$, which implies that $\tilde{g}_n$ is the minimand over $\mathcal{G}_n$. □

Proof of Lemma 2.2 . It follows from the proof of Proposition 1 in Lindenlaub (2017) that Φ satisfies the MTW(0) condition. Then, we can employ Theorem 12.51 in Villani (2008) to obtain the regularity in the Lemma. □

Proof of Theorem 2.2. We first show that there exists a constant $\kappa > 0$ such that $Q(g) - Q(g_0) \geq \kappa \mathbb{E}_X\left[|g(X) - g_0(X)|^2\right]$ for any $g \in \mathcal{G}$. We note that $g - g_0$ is also in $\mathcal{G}$, i.e., $\int_{\mathcal{X}} [g(X) - g_0(X)]\, dX = 0$. Lemma 2 in Gunsilius (2018) implies that

$$Q(g) - Q(g_0) = \frac{1}{2}\mathbb{E}_X\left[\|\nabla(g(X) - g_0(X))\|^2\right] + o\left(\|g - g_0\|_2^2\right) \quad (\|g - g_0\|_2 \to 0).$$

Therefore, it suffices to show that $\mathbb{E}_X\left[\|\nabla(g(X) - g_0(X))\|^2\right] \geq 2\kappa \|g - g_0\|_2^2$. Since $f_X(X)$ is bounded away from zero and infinite,

$$\begin{aligned} \mathbb{E}_X\left[|g(X) - g_0(X)|^2\right] &\leq \bar{f}_X \int_{\mathcal{X}} |g(X) - g_0(X)|^2 \, dX, \\ &= \bar{f}_X \int_{\mathcal{X}} |g(X) - g_0(X) - \bar{g}|^2 \, dX, \end{aligned} \tag{2.8}$$

where $\bar{f}_X = \sup_{x \in \mathcal{X}} f_X(x) < \infty$ and $\bar{g} = \left\{\int_{\mathcal{X}} [g(X) - g_0(X)]\, dX\right\} / \bar{\mathcal{X}} = 0$ with the volume of $\mathcal{X}$, $\bar{\mathcal{X}}$. The original Poincaré inequality implies that there exists a constant $C > 0$ such that

$$\begin{aligned} \int_{\mathcal{X}} |g(X) - g_0(X) - \bar{g}|^2 \, dX &\leq C \int_{\mathcal{X}} \|\nabla(g(X) - g_0(X))\|^2 \, dX \\ &\leq \frac{C}{\underline{f}_X} \mathbb{E}_X\left[\|\nabla(g(X) - g_0(X))\|^2\right], \end{aligned} \tag{2.9}$$

where $\underline{f}_X = \inf_{x \in \mathcal{X}} f_X(x) > 0$. Combining two inequalities (2.8) and (2.9), we obtain the

desired result:

$$\mathbb{E}_X\left[\left\|\nabla\left(g\left(X\right)-g_0\left(X\right)\right)\right\|^2\right]\geq\frac{\underline{f}_X}{C\bar{f}_X}\mathbb{E}_X\left[\left|g\left(X\right)-g_0\left(X\right)\right|^2\right].$$

Now we check Conditions 3.7 and 3.8 in Chen (2007). Since

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g\left(x\right)\right\}\leq\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g_0\left(x\right)\right\}+\sup_{x\in\mathcal{X}}\left|g\left(x\right)-g_0\left(x\right)\right|;\\
&\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g\left(x\right)\right\}\geq\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g_0\left(x\right)\right\}-\sup_{x\in\mathcal{X}}\left|g\left(x\right)-g_0\left(x\right)\right|,\\
&\left|\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g\left(x\right)\right\}-\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y\right)-g_0\left(x\right)\right\}\right|\leq\sup_{x\in\mathcal{X}}\left|g\left(x\right)-g_0\left(x\right)\right|,
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left|g\left(X_i\right)+\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y_i\right)-g\left(x\right)\right\}-g_0\left(X_i\right)-\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y_i\right)-g_0\left(x\right)\right\}\right|\\
\leq&2\sup_{x\in\mathcal{X}}\left|g\left(x\right)-g_0\left(x\right)\right|=2\left\|g-g_0\right\|_\infty.
\end{aligned}$$

By Lemma 2 in Chen and Shen (1998) for any $p>0$, we have $\left\|g-g_0\right\|_\infty\leq c\left\|g-g_0\right\|_2^{2p/(2p+d)}$. Hence

$$\begin{aligned}
&\mathbb{E}\left[\left|g\left(X_i\right)+\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y_i\right)-g\left(x\right)\right\}-g_0\left(X_i\right)-\sup_{x\in\mathcal{X}}\left\{\Phi\left(x,Y_i\right)-g_0\left(x\right)\right\}\right|^2\right]\\
\leq&4c\left\|g-g_0\right\|_2^{4p/(2p+d)}.
\end{aligned}$$

So Condition 3.7 is satisfied for all $\varepsilon\leq1$. On the other hand, using Lemma 2 in Chen and Shen (1998) again we see that Condition 3.8 is then satisfied with $s=2p/\left(2p+d\right)$, $U\left(X_i,Y_i\right)=1$ and any value $\gamma\geq2$.

To apply Theorem 3.2 in Chen (2007), it remains to compute the metric entropy with bracketing $H_{[]}\left(w,\mathcal{F}_n,\left\|\cdot\right\|_2\right)$ of the class $\mathcal{F}_n=\left\{\ell\left(g,\left(X_i,Y_i\right)\right)-\ell\left(g_0,\left(X_i,Y_i\right)\right):\left\|g-g_0\right\|_2\leq\delta,g\in\mathcal{G}_n\right\}$. By definition,

$$\left\|g_0-\pi_ng_0\right\|=\mathbb{E}\left[\left\|g_0-\Pi_ng_0\right\|^2\right]^{1/2}\leq c\left\|g_0-\Pi_ng_0\right\|_\infty.$$

Then, for all $0<w\leq\delta<1$, $H_{[]}\left(w,\mathcal{F}_n,\left\|\cdot\right\|_2\right)\leq\log N\left(w,\mathcal{G}_n,\left\|\cdot\right\|_\infty\right)$. Since $\log N\left(w,\mathcal{G}_n,\left\|\cdot\right\|_\infty\right)\leq$

$k_n \log(1 + C/w)$ by Lemma 2.5 in van de Geer (2009), δ_n solves

$$\frac{1}{\sqrt{n}\delta_n^2} \int_{b\delta_n^2}^{\delta_n} \sqrt{H_{[]}(w, \mathcal{F}_n, \|\cdot\|_2)} dw \leq \frac{1}{\sqrt{n}\delta_n^2} \int_{b\delta_n^2}^{\delta_n} \sqrt{k_n \log(1 + C/w)} dw \leq \frac{1}{\sqrt{n}\delta_n^2} \sqrt{k_n}\delta_n \leq C'.$$

The solution is $\delta_n \approx \sqrt{k_n/n}$. The statement of theorem follows from Theorem 3.2 in Chen (2007). □

Proof of Corollary 2.1. Note that Bernstein inequalities from approximation theory imply that $\|\partial^\alpha g\|_2 = O\left(k_n^{[\alpha]/d}\right) \|g\|_2$ for all $g \in \mathcal{G}_n$. Then, we have that for $i = 1, \ldots, d$ and $g \in \mathcal{G}_n$,

$$\begin{aligned} \|\partial^\alpha (\Pi_n g_0) - \partial^\alpha g_0\|_2 &\leq O\left(k_n^{[\alpha]/d}\right) \|\Pi_n (g_0 - g)\|_2 + \|\partial^\alpha g - \partial^\alpha g_0\|_2 \\ &\leq O\left(k_n^{[\alpha]/d}\right) \|g_0 - g\|_2 + \|\partial^\alpha g - \partial^\alpha g_0\|_2 \end{aligned}$$

Since the above inequality holds uniformly in $g \in \mathcal{G}_n$, we choose g such that

$$\|g_0 - g\|_2 = O_p\left(\|\hat{g}_n - g\|_2\right)$$

and

$$\|\partial^\alpha g - \partial^\alpha g_0\|_2 = O_p\left(\|\hat{g}_n - g\|_2 \times k_n^{[\alpha]/d}\right).$$

By similar arguments to the above, we have that

$$\begin{aligned} \|\partial^\alpha \hat{g}_n - \partial^\alpha g_0\|_2 &\leq \|\partial^\alpha \hat{g}_n - \partial^\alpha (\Pi_n g_0)\|_2 + \|\partial^\alpha (\Pi_n g_0) - \partial^\alpha g_0\|_2 \\ &\leq O\left(k_n^{[\alpha]/d}\right) \|\hat{g}_n - g_0\|_2 + \|\partial^\alpha (\Pi_n g_0) - \partial^\alpha g_0\|_2 \end{aligned}$$

and the result follows. □

Proof of Theorem 2.3. We first show that there exists a constant $\kappa > 0$ such that

$$\begin{aligned}
&Q(g) - Q(g_0) \\
&\coloneqq \mathbb{E}_{YZ}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g(u,Z)\right\}\right] - \mathbb{E}_{YZ}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g_0(u,Z)\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}_{Y|Z}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g(u,Z)\right\}|Z\right] - \mathbb{E}_{Y|Z}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g_0(u,Z)\right\}|Z\right]\right] \\
&\geq \kappa\mathbb{E}_{UZ}\left[\left(g(U,Z) - g_0(U,Z)\right)^2\right]
\end{aligned}$$

for $g \in \mathcal{G}$. We note that $g - g_0$ is also in $\mathcal{G}$, i.e., $\int_{\mathcal{U}}\left[g(U,z) - g_0(U,z)\right]dU = 0$ for all $z \in \mathcal{Z}$. Lemma 2 in Gunsilius (2018) implies that

$$\begin{aligned}
&Q(g(\cdot,z)) - Q(g_0(\cdot,z)) \\
&\coloneqq \mathbb{E}_{Y|Z=z}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g(u,Z)\right\}|Z=z\right] - \mathbb{E}_{Y|Z=z}\left[\sup_{u\in\mathcal{U}}\left\{u'Y - g_0(u,Z)\right\}|Z=z\right] \\
&= \frac{1}{2}\int_{\mathcal{U}}\left\|\nabla\left(g(U,z) - g_0(U,z)\right)\right\|^2 dU + o\left(\left\|g(U,z) - g_0(U,z)\right\|_2^2\right) \quad (\|v\| \to 0).
\end{aligned}$$

We can easily check from the original Poincaré inequality that there exists a constant $C > 0$ such that

$$\int_{\mathcal{U}}\left|g(U,z) - g_0(U,z)\right|^2 dU \leq C\int_{\mathcal{U}}\left\|\nabla\left(g(U,z) - g_0(U,z)\right)\right\|^2 dU$$

It implies that

$$\begin{aligned}
Q(g) - Q(g_0) &= \mathbb{E}_Z\left[Q(g(\cdot,Z)) - Q(g_0(\cdot,Z))\right] \\
&\geq \frac{1}{2C}\mathbb{E}\left[\left|g(U,z) - g_0(U,z)\right|^2\right] + o\left(\left\|g(U,z) - g_0(U,z)\right\|_2^2\right)
\end{aligned}$$

Now we check Conditions 3.7 and 3.8 in Chen (2007). Since

$$\begin{aligned}
\sup_{u\in\mathcal{U}}\left\{u'y - g(u,z)\right\} &\leq \sup_{u\in\mathcal{U}}\left\{u'y - g_0(u,z)\right\} + \sup_{u\in\mathcal{U}}\left|g(u,z) - g_0(u,z)\right|; \\
\sup_{u\in\mathcal{U}}\left\{u'y - g(u,z)\right\} &\geq \sup_{u\in\mathcal{U}}\left\{u'y - g_0(u,z)\right\} - \sup_{u\in\mathcal{U}}\left|g(u,z) - g_0(u,z)\right|,
\end{aligned}$$

$$\left|\sup_{u\in\mathcal{U}}\left\{u'y-g\left(u,z\right)\right\}-\sup_{u\in\mathcal{U}}\left\{u'y-g_0\left(u,z\right)\right\}\right|\leq\sup_{u\in\mathcal{U}}\left|g\left(u,z\right)-g_0\left(u,z\right)\right|,$$

it follows from Lemma 2 in Chen and Shen (1998) that

$$\mathbb{E}_{YZ}\left[\left|\sup_{u\in\mathcal{U}}\left\{u'Y-g\left(u,Z\right)\right\}-\sup_{u\in\mathcal{U}}\left\{u'Y-g_0\left(u,Z\right)\right\}\right|^2\right]\leq 4c\left\|g-g_0\right\|_2^{4p/(2p+d+d_Z)}.$$

So Condition 3.7 is satisfied for all $\varepsilon\leq 1$. On the other hand, using Lemma 2 in Chen and Shen (1998) again we see that Condition 3.8 is then satisfied with $s=2p/\left(2p+d+d_Z\right)$, $U\left(Y_i,Z_i\right)=1$ and any value $\gamma\geq 2$. □

Chapter 3

Multidimensional Matching as Optimal Transport Problem

3.1 Introduction

The empirical analysis of matching between workers and jobs affected by technological progress represents an important area in the labour market. Each worker has different levels of multiple skills and each job demands different levels of worker's multiple skills. When their matching generates a quantity of output, how could we sort workers into jobs to maximize the total output? This social planner problem is an application of the optimal transport problem. Agents' multiple characteristics are not usually perfectly correlated and neglecting these multidimensional heterogeneity is problematic. In stead of aggregating multivariate characteristics into one-dimensional index, Lindenlaub (2017) extends the scalar notion of positive assortative matching (PAM) to the multidimensional one and develops a framework based one the optimal transport having studies the existence, uniqueness, and purity of multidimensional matching under transferable utility.

This chapter follows the setting in Lindenlaub (2017) and extends her results. Adding to the uniqueness and smoothness of equilibrium function, we derive the equilibrium wage and matching functions in terms of the solution to the dual Monge-Kantorovich problem in Chapter 2. The results are not in closed form unlike Lindenlaub (2017), who solves for the assignment and wage function explicitly when both workers and jobs characteristic variables

follow bivariate normal distributions. However, our results are more general in the sense that we allow for any multidimensional variables having probability densities.

Any model of matching based on optimal transport will not be exploitable because it will generate far too strong predictions. The equilibrium wage and assignment are deterministic, but some matchings will never hold. Hence, we need to regularize the matching model. The standard approach is to allow for a class of unobserved heterogeneity or search frictions, but we introduce measurement error in the equilibrium functions to keep the model in line with the theory of optimal transport. Notice that we could avoid a situation in which the unobserved heterogeneity affects the assignment by assuming that it involves in non-interaction terms.

We then study equilibrium wage and matching system with exogenous worker's multiple skills. The specification for output generated by the matching involves unknown finite-dimensional parameters for which there are nonparameteric equilibrium wage and matching functions. In our analysis of the semiparametric matching model, we first provide identification under the theory of optimal transport, which is the uniqueness and continuously differentiability of the solution to the dual Monge-Kantorovich problem. Moreover, we employ the sieve minimum distance estimation of conditional moment restrictions for the nonparametric wage function and finite-dimensional parameter in the output function.

The estimation procedure is comparable to those of Lindenlaub (2017), who transforms two-dimensional skills data to make them close to a bivariate normal random variable and then uses the closed form solution for the equilibrium wage function in estimation. Parametric models lead to more efficient estimation if they are correctly specified. However, the transformed data does not fully follow the specifically aimed distributions. Another one is that even if both worker's cognitive and manual skills and job's skill demands follow the bivariate normal distribution, the model with measurement errors in Lindenlaub (2017) may still be misspecified. The equilibrium wage and matching equations depend on the productivity correlation under the bivariate normality, but Lindenlaub (2017) uses a correlation of error contaminated job's skill demands, which is different from the productivity correlation.

As an empirical study, the data in Lindenlaub (2017) is revisited to estimate the production functions quantifying technological progress in the U.S. between 1990 and 2010. Our

estimation results using transformed data are similar to the results of Lindenlaub (2017): worker-job complementarities in manual skills strongly decreased, where as complementarities in cognitive skills increased. However, the reformulated model does not explain the wage polarization (declining lower tail but expanding upper tail wage inequality) in the U.S. For this reason, despite of the consistency of our estimation method, we need to consider a more flexible environment such as search friction and higher dimensional skills.

This chapter is organized as follows. Section 3.2 introduces the multidimensional matching as an application of the optimal transport problem. Section 3.3 specifies the model with measurement errors and obtains the identification. Section 3.4 presents the sieve minimum distance procedure. Section 3.5 revisits the empirical study presented in Lindenlaub (2017).

3.2 Multidimensional matching: An optimal transport approach

One case of interest is found in the marriage market in Becker (1973). In one-dimensional case, with scalar "ability indices" of men and women, he defines the positive assortative matching (PAM) given a joint surplus, $\Phi(x, y)$, to be shared:

$$\partial^2 \Phi(x, y) / \partial x \partial y \geq 0.$$

Matching between men and women is described by the function $T : \mathcal{X} \to \mathcal{Y}$. For one-dimensional variables, x and y, the PAM means that high-type workers match high-type firms. The matching function is defined by

$$T(x) = F_Y^{-1}(F_X(x))$$

where F_X and F_Y are the cumulative distribution functions (CDFs) for X and Y, respectively.

We assume that d-dimensional heterogeneities are observable where workers and firms are characterized by $X = (X_1, \ldots, X_d) \in \mathcal{X} \subset \mathbb{R}^d$ and $Y = (Y_1, \ldots, Y_d) \in \mathcal{Y} \subset \mathbb{R}^d$, respectively. Then, assortativity involves properties of the first derivative of the matching function, given

by $\nabla Y(X) = (\nabla Y(X))_{ij} = \partial Y_i / \partial X_j$. Lindenlaub (2017) develops a theoretical framework that generalizes the scalar notion of PAM to the multidimensional one as follows:

Definition 3.1 (Multidimensional Assortative Matching)**.** The sorting pattern is PAM (NAM) if for all X,

$$\frac{\partial Y_\ell}{\partial X_\ell} > (<)\, 0 \text{ for } \ell = 1, \ldots, d, \quad \text{and} \quad \det(\nabla Y(X)) > 0.$$

We further assume that every firm produces a single homogeneous good by combining all inputs. If worker X works for firm characterized by Y, this generates a quantity of output $\Phi(X, Y)$. We consider the social planner's problem maximizing the total output assigning workers to firms. This leads to the Monge-Kantorovich problem:

$$\sup_{\pi \in \mathcal{M}(P_X, P_Y)} \int_{\mathcal{X} \times \mathcal{Y}} \Phi(X, Y)\, d\pi(X, Y) \tag{3.1}$$

where $\mathcal{M}(P_X, P_Y)$ is the set of all joint measures admitting P_X and P_Y as marginals on $\mathcal{X}$ and $\mathcal{Y}$ respectively. In most cases, the above problem leads to the dual problem:

$$\inf_{g \in \mathcal{G}} \mathbb{E}_X[g(X)] + \mathbb{E}_Y\left[\sup_{x \in \mathcal{X}} \{\Phi(x, Y) - g(x)\}\right], \quad \text{s.t.} \int_{\mathcal{X}} g(X)\, dX = 0. \tag{3.2}$$

We now consider the firm's profit maximization problem. By assuming that the wage function, w, is not a function of skills demand, the firm's problem can be written as

$$\max_{x \in \mathcal{X}} \{\Phi(x, y) - w(x)\}.$$

We aim to find the equilibrium assignment, $T : \mathcal{X} \to \mathcal{Y}$, and wage function, $w : \mathcal{X} \to \mathbb{R}_+$. Lindenlaub (2017) provides the sufficient conditions on the technology under which sorting is obtained. The existence of a unique deterministic equilibrium depends on well-established results in the optimal transport literature. Adding to Proposition 1 in Lindenlaub (2017), the equilibrium wage and matching functions are obtained in terms of the solution to the dual optimal transport problem by applying Lemma 2.1 in Chapter 2:

Theorem 3.1. *Suppose that Assumptions 2.1 and 2.2 hold. Then, there exists the unique*

solution, $g_0(x)$, for (3.2). Furthermore, the equilibrium wage function is given by $w(X) = g_0(X) + c$, where c is the constant of integration, and the map $T : \mathcal{X} \to \mathcal{Y}$, satisfying $\nabla_X g_0(X) \coloneqq \nabla_X \Phi(X, Y)|_{Y=T(X)}$, is the unique equilibrium assignment.

To apply this multidimensional matching framework to the technological change, we specify the technology to $X'AY$ with a $d \times d$ matrix A. We assume that there are sets of firms and workers with $d = 2$ without loss of generality. Every worker is endowed with a bundle of cognitive and manual skills, $X = (X_C, X_M) \in \mathcal{X} = \mathbb{R}^2$. Points in $\mathcal{X}$ represent worker types. In turn, each firm is endowed with both cognitive and manual skill demands, $Y = (Y_C, Y_M) \in \mathcal{Y} = \mathbb{R}^2$. Coordinate Y_C (respectively manual Y_M) corresponds to the productivity or skill requirement of cognitive task C (respectively manual task M). Points in $\mathcal{Y}$ represent firm types. We denote the marginals of (X_C, X_M) and (Y_C, Y_M) by P_X and P_Y respectively.

We consider the social planner's problem with the following technology:

$$\begin{aligned}\Phi(X, Y; \theta) &= X'AY + X'b \\ &\coloneqq \begin{pmatrix} X_C & X_M \end{pmatrix} \begin{pmatrix} \alpha_{CC} & \alpha_{CM} \\ \alpha_{MC} & \alpha_{MM} \end{pmatrix} \begin{pmatrix} Y_C \\ Y_M \end{pmatrix} + \begin{pmatrix} X_C & X_M \end{pmatrix} \begin{pmatrix} \beta_C \\ \beta_M \end{pmatrix}.\end{aligned}$$

Here the elements of A represent the level of worker-job complementabilities or substitutabilities. The diagonal elements capture within-task complementarity and the off-diagonal elements indicate between-task complementarity. We note from Proposition 2 in Lindenlaub (2017) that if A is a diagonal matrix with all positive principal minors, there exists a unique optimal matching function T satisfying PAM.

Assumption 3.1. *A is invertible.*

Let $\tilde{Y} = AY$. Since the assignment is unaffected by non-interaction terms, the original problem (3.1) and its dual problem can be rewritten as

$$\begin{aligned}&\sup_{\pi \in \mathcal{M}\left(P_X, P_{\tilde{Y}}\right)} \mathbb{E}_\pi \left[X'\tilde{Y}\right], \\ &\inf_{g \in \mathcal{G}} \mathbb{E}_X \left[g(X)\right] + \mathbb{E}_{\tilde{Y}} \left[\sup_{x \in \mathcal{X}} \left\{X'\tilde{Y} - g(x)\right\}\right], \quad \text{s.t.} \int_{\mathcal{X}} g(X)\, dX = 0. \end{aligned} \tag{3.3}$$

We can interpret this problem as assigning from $\mathcal{X}$ to $A\mathcal{Y} \coloneqq \{AY : Y \in \mathcal{Y}\}$.

Lindenlaub (2017) derives T and w in closed form under the assumption that x and y follow bivariate standard normal distributions. Using the explicit form of T and w, she estimates the production function, $\Phi(X, Y; \theta)$, to investigate how technology in the U.S. has evolved over time. The following statement follows from the existing results in the optimal transport literature and generalizes the results in Lindenlaub (2017) by allowing X and Y to be non-Gaussian. The equilibrium matching and wage, $Y^* = (Y_C^*, Y_M^*)$ and w^*, are derived in terms of the solution of the dual Monge-Kantorovich problem:

Corollary 3.1. *Suppose that Assumptions 2.1 and 3.1 are hold. Then, there exists the unique convex solution, $g_0(X; A)$, for (3.3), and the equilibrium assignment and wage function are given by*

$$A \begin{pmatrix} Y_C^*(X) \\ Y_M^*(X) \end{pmatrix} = \nabla_X g_0(X; A) \;\Rightarrow\; \begin{pmatrix} Y_C^*(X) \\ Y_M^*(X) \end{pmatrix} = A^{-1} \begin{pmatrix} \partial g_0(X; A)/\partial X_C \\ \partial g_0(X; A)/\partial X_M \end{pmatrix},$$
$$w^*(X) = g_0(X; A) + \beta_C X_C + \beta_M X_M + c,$$

where c is the constant of integration.

Proof. We note from Proposition 1.(i) in Lindenlaub (2017) that

$$\nabla w^*(X) - \begin{bmatrix} \beta_C \\ \beta_M \end{bmatrix} = \nabla_X X' \tilde{Y}\Big|_{\tilde{Y}=T(X)} = \nabla g_0(X; A).$$

It implies that $w^*(X) = g_0(X; A) + \beta_C X_C + \beta_M X_M + c$. □

This specification follows from the existing results in the optimal transport literature (see, Villani, 2008; Lindenlaub, 2017). Also, the result coincides with the interpretation of $g_0(X)$ and $h_0(Y)$, which are the equilibrium payoffs that worker x and firm y receive at equilibrium. We add non-interaction terms using the first-order condition of firm's problem. In the dual problem (3.2), we may impose the constraint, $g(0) = 0$, among other constraints such as zero mean of $g(X)$. Then, c equals the wage of the worker having a pair of cognitive and manual skills, $(0, 0)$.

Lindenlaub (2017) derives $T = \nabla g_0$ and w in closed form under the assumption that X and Y follow bivariate standard normal distributions.[1] However, in most cases, X and Y are not bivariate normally distributed. To employ the closed forms of equilibrium functions, she transformed the original data. Figure 3.1 illustrates how she derives the equilibrium assignment and wage function explicitly from transformed data. To align the data with the model that features standard Gaussian distribution, the empirical distributions are transformed into Gaussian copulas. Each one-dimensional variables are converted into Gaussian variables using the inverse transform method, and their dependence is modeled using Gaussian copula. If the transformed data follow the bivariate normal distribution, this transformation provides a way of studying dependency independent of the marginals by removing the marginal characteristics. However, joint distribution of two normally distributed random variables does not always follow the bivariate normal. When transformed data, $\tilde{X}$ and $\tilde{Y}$, do not follow bivariate normal distribution, the model based on the closed form expression for $g\left(\tilde{X}; A\right)$ can be misspecified.

3.3 Model and identification

Let $\{(w_i, X_i', Y_i')\}_{i=1}^n$ represent an independent and identically distributed (i.i.d.) sequence of n matched observations on the worker i's wage w_i, cognitive and manual skills $X_i = (X_{Ci}, X_{Mi})'$, and on the matched job's skill demands $Y_i = (Y_{Ci}, Y_{Mi})'$. Any model of

[1] Let $X \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho_X \\ \rho_X & 1 \end{pmatrix}\right)$ and $Y \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho_Y \\ \rho_Y & 1 \end{pmatrix}\right)$. Then, a unique equilibrium matching function T^* and wage function w^* for the technology $\Phi(X,Y) = \alpha_C(X_C Y_C + \gamma X_M Y_M) + \beta_C X_C + \beta_M X_M$ are as follows:

$$Y = T^*(X; \gamma, \rho_X, \rho_Y), \quad w = w^*(X; \theta, \rho_X, \rho_Y)$$

where

$$T^*(X; \gamma, \rho_X, \rho_Y) = \begin{pmatrix} J_{11}(\gamma, \rho_X, \rho_Y) & J_{12}(\gamma, \rho_X, \rho_Y) \\ J_{21}(\gamma, \rho_X, \rho_Y)/\gamma & J_{22}(\gamma, \rho_X, \rho_Y)\gamma \end{pmatrix} X,$$

$$w^*(X; \theta, \rho_X, \rho_Y) = \frac{1}{2}\alpha_C X' \begin{pmatrix} J_{11}(\gamma, \rho_X, \rho_Y) & J_{12}(\gamma, \rho_X, \rho_Y) \\ J_{21}(\gamma, \rho_X, \rho_Y) & J_{22}(\gamma, \rho_X, \rho_Y) \end{pmatrix} X + \beta_C X_C + \beta_M X_M + c$$

with the constant of integration c and

$$\begin{pmatrix} J_{11}(\gamma, \rho_X, \rho_Y) & J_{12}(\gamma, \rho_X, \rho_Y) \\ J_{21}(\gamma, \rho_X, \rho_Y) & J_{22}(\gamma, \rho_X, \rho_Y) \end{pmatrix} =$$

$$\frac{1}{\sqrt{1 + 2\gamma\left(\rho_X\rho_Y + \sqrt{1-\rho_X^2}\sqrt{1-\rho_Y^2} + \gamma^2\right)}} \begin{pmatrix} 1 + \gamma\frac{\sqrt{1-\rho_Y^2}}{\sqrt{1-\rho_X^2}} & \gamma\left(\rho_Y - \rho_X\frac{\sqrt{1-\rho_Y^2}}{\sqrt{1-\rho_X^2}}\right) \\ \gamma\left(\rho_Y - \rho_X\frac{\sqrt{1-\rho_Y^2}}{\sqrt{1-\rho_X^2}}\right) & \gamma\left(\gamma + \frac{\sqrt{1-\rho_Y^2}}{\sqrt{1-\rho_X^2}}\right) \end{pmatrix}.$$

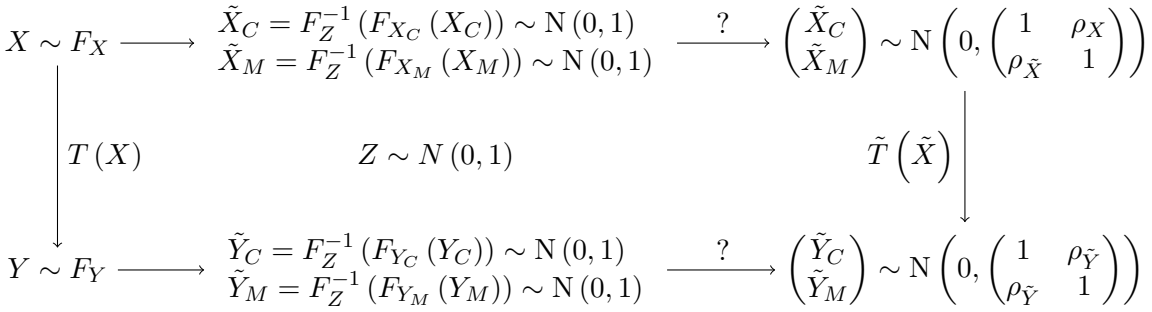

Figure 3.1: Lindenlaub (2017)

matching based on optimal transport will not be exploitable because it will generate far too strong predictions. Some matchings will never hold. Hence, we need to regularize the matching model. The standard approach is to allow for a class of unobserved heterogeneity or search frictions, but we introduce measurement error in the equilibrium functions to keep the model in line with the theory of optimal transport.[2][3] By doing this, it keeps the empirical model in line with the theory:

$$\begin{aligned} w_i &= w_i^* + \varepsilon_{wi} = g(X_i; A) + \beta_C X_{Ci} + \beta_M X_{Mi} + c + \varepsilon_{wi}; \\ AY_i &= A(Y_i^* + \varepsilon_{Yi}) = \nabla g(X_i; A) + A \begin{pmatrix} \varepsilon_{Ci} \\ \varepsilon_{Mi} \end{pmatrix} \end{aligned} \tag{3.4}$$

with the exogeneity of X:

$$\mathbb{E}[\varepsilon_{wi}|X_i] = \mathbb{E}[\varepsilon_{Ci}|X_i] = \mathbb{E}[\varepsilon_{Mi}|X_i] = 0 \tag{3.5}$$

Let $\theta = \left(vec(A')', b', c\right)' = (\alpha_{CC}, \alpha_{CM}, \alpha_{MC}, \alpha_{MM}, \beta_C, \beta_M, c)'$ denote a vector of unknown finite-dimensional parameters and $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^7$. We

[2]Based on the bivariate normality of X and Y, the closed form expression for $g(X)$ involves the productivity correlation under the bivariate normality, but Lindenlaub (2017) uses a correlation of error contaminated Y_C and Y_M, which is different from the productivity correlation. The estimation result with data shows that the estimated variances of measurement errors, ε_C and ε_M are greater than one, which is not desirable in the setting. If the measurement error is introduced in the assignment equation, then the productivity correlation should be reformulated.

[3]Notice that we could avoid a situation in which the unobserved heterogeneity affects the assignment by assuming that it involves in non-interaction terms.

denote $Z_i = (w_i, X_i', Y_i')'$ and $\rho \equiv (\rho_1, \rho_2, \rho_3)' \in \mathbb{R}^3$, where

$$\begin{aligned}
\rho_1(w_i, X_i; \beta_C, \beta_M, c, g) &= w_i - (g(X_i) + \beta_C X_{Ci} + \beta_M X_{Mi} + c), \\
\rho_2(Y_i, X_i; \alpha_{CC}, \alpha_{CM}, g) &= \alpha_{CC} Y_{Ci} + \alpha_{CM} Y_{Mi} - \partial g(X_i)/\partial X_C, \\
\rho_3(Y_i, X_i; \alpha_{MC}, \alpha_{MM}, g) &= \alpha_{MC} Y_{Ci} + \alpha_{MM} Y_{Mi} - \partial g(X_i)/\partial X_M.
\end{aligned}$$

For each observation i, the model (3.4) satisfies (3.5), which we rewrite as

$$\mathbb{E}[\rho(Z_i; \theta_0, g_0) | X_i] = 0, \tag{3.6}$$

where (θ_0, g_0) is the true but unknown parameter.

Here, we estimate parameters using a standard SMD estimator, which does not require the closed form of function g. The parameters (θ, g), where $\theta = (A, b, c)$ and g is an unknown convex function, are identified from Corollary 3.1 and the exogeneity of X:

Theorem 3.2. *Suppose that Assumptions 2.1, 3.1, and (3.6) hold. Then, θ and g are identified.*

Proof of Theorem 3.2. We first note from $\mathbb{E}[\varepsilon_w | X] = 0$ that the convex function $\tilde{g}(X) = \mathbb{E}[w|X] = g(X) + Xb + c$ is easily identified. Since $\nabla \tilde{g}(X)$ is Then, the remaining exogeneities of X, $\mathbb{E}[\varepsilon_C | X] = \mathbb{E}[\varepsilon_M | X] = 0$, and the invertibility of A imply that A and b are identified. Finally, the specification for $\tilde{g}$ and the normalization for g identify c and $g(X) = \tilde{g}(X) - \beta_C X_C - \beta_M X_M - c$:

$$\int_{\mathcal{X}} (\tilde{g}(X) - \beta_C X_C - \beta_M X_M - c)\, dX = 0.$$

□

Assumption 2.1 and 3.1 imply that there exists a unique deterministic equilibrium with unique convex g, which follows from the theory of optimal transport (see, e.g., Villani, 2008; De Philippis and Figalli, 2014). Assumption 3.1 guarantees the injectivity of $Y \mapsto \nabla_X \Phi(X, Y)$ for each fixed X, which can be viewed as a generalization of the Spence-Mirrlees condition. Furthermore, the convexity of g implies that $\mathbb{E}\left[\nabla g(X) \nabla g(X)'\right]$ has full rank,

which implies that A and b are identified.

3.4 Sieve minimum distance estimation

We consider the model (3.4) with diagonal $A = \alpha_C \text{diag}(1, \gamma)$ and apply the sieve minimum distance (SMD) estimation method of Ai and Chen (2003) for semiparametric conditional moment restrictions. Notice that γ represents the relative level of complementarities across cognetive and manual tasks. First we approximate the unknown function $g \in \mathcal{G}$ by $g_n \in \mathcal{G}_n$ where $\mathcal{G}_n$ is an approximating function space becoming dense in $\mathcal{G}$ as $n \to \infty$. Then for given (θ, g_n) in the parameter space $\Theta \times \mathcal{G}_n$, we estimate the conditional moment function $m(X, \theta, g) = \mathbb{E}[\rho(Z; \theta, g) | X]$ nonparametrically by $\hat{m}(X, \theta, g)$. Finally, we estimate the θ and the unknown sieve coefficient of g_n jointly by applying the SMD procedure:

$$\min_{(\theta,g)\in\Theta\times\mathcal{G}_n} \sum_{i=1}^{n} \hat{m}(X_i; \theta, g)' \left[\hat{\Sigma}_0(X_i)\right]^{-1} \hat{m}(X_i; \theta, g),$$

where $\hat{\Sigma}_0(X)$ is a consistent estimator of the optimal weighting matrix $\Sigma_0(X)$.

We estimate the parameter using the three-step procedure proposed in Ai and Chen (2003), which is summarized in the table.

Algorithm: Computing the Sieve MD Estimator of $\Phi(X, Y^*; \theta)$
Obtain an initial consistent sieve MD estimator $\left(\tilde{\theta}_n, \tilde{g}_n\right)$ by $\min_{(\theta,g)} \sum_{i=1}^{n} \hat{m}(X_i; \theta, g)' \hat{m}(X_i; \theta, g)$, where $\hat{m}(X_i; \theta, g)$ is the sieve least square estimator of $\mathbb{E}[\rho(Z; \theta, g) \| X]$.
Obtain a consistent estimator $\hat{\Sigma}_0(X)$ of $\Sigma_0(X) = \text{Var}[\rho(Z; \theta, g) \| X]$ using $\left(\tilde{\theta}_n, \tilde{g}_n\right)$ and sieve LS estimation.
Obtain the optimally weighted sieve MD estimator $\left(\hat{\theta}_n, \hat{g}_n\right)$ by $\min_{(\theta,g)} \sum_{i=1}^{n} \hat{m}(X_i; \theta, g)' \left[\hat{\Sigma}_0(X_i)\right]^{-1} \hat{m}(X_i; \theta, g)$.

When $\rho(Z; \theta, g) - \rho(Z; \theta_0, g_0)$ does not depend on Y, we can apply the sieve GLS procedure with the replacement of $\hat{m}(X; \theta, g)$ by $\rho(Z_i; \theta, g)$:

$$\min_{(\theta,g)} \sum_{i=1}^{n} \rho(Z_i; \theta, g)' \left[\hat{\Sigma}_0(X_i)\right]^{-1} \rho(Z_i; \theta, g).$$

If we change ρ_2 to $\tilde{\rho}_2(Y_M, X; \gamma, g) = Y_M - (\partial g(X)/\partial X_M)/\gamma$, $\tilde{\rho}(Z; \theta, g) - \tilde{\rho}(Z; \theta_0, g_0)$ with $\tilde{\rho}_2$ does not depend on Y. However, $\tilde{\rho}_2$ may not satisfy the pointwise Hölder continuity in γ, which is the typically imposed sufficient condition in the literature such as Ai and Chen (2003).

For each fixed (X, θ, g), it is required to estimate $m(X, \theta, g)$ and $\Sigma_0(X)$. Let $p_{0j}(X)$, $j = 1, \ldots, J_n$, be a sequence of known basis functions approximating any square integrable function of X well as $J_n \to \infty$. With $p^{J_n}(X) = (p_{o1}(X), \ldots, p_{0J_n}(X))'$ and $P = \left(p^{J_n}(X_1), \ldots, p^{J_n}(X_n)\right)'$. Then the sieve LS estimators of $m(X, \theta, g)$ and $\Sigma_0(X)$ are

$$\hat{m}(X, \theta, g) = \sum_{i=1}^{n} \rho(Z_i; \theta, g)\, p^{J_n}(X_i)' \left(P'P\right)^{-1} p^{J_n}(X),$$
$$\hat{\Sigma}_0(X) = \sum_{i=1}^{n} \rho\left(Z_i; \tilde{\theta}, \tilde{g}_n\right) \rho\left(Z_i; \tilde{\theta}, \tilde{g}_n\right)' p^{J_n}(X_i)' \left(P'P\right)^{-1} p^{J_n}(X),$$

where $\left(\tilde{\theta}, \tilde{g}_n\right)$ is the SMD estimator obtained in the first step.

We use the weighted bootstrap method as an inference method. Ma and Kosorok (2005) and Chen and Pouzo (2009) established results for a semiparametric M-estimation with or without nonparametric endogeneity, respectively. We employ there results to obtain distributional approximation for $\hat{\theta}$. To describe this method, consider an i.i.d. sample of positive weights, $\{\pi_i\}_{i=1}^n$, satisfying $\mathbb{E}[\pi_i] = 1$, $\mathrm{Var}(\pi_i) = 1$, and is independent of the data. We define the weighted bootstrap estimator $\hat{\theta}_b$ as the solution to the weighted minimization problem.

3.5 Empirical application to U.S. sorting and wage inequality shifts

We apply the sieve MD procedure to estimate the production using the national longitudinal survey of youth (NLSY) and O*NET[4] data. The NLSY and O*NET data are used to construct a two-dimensional vector of workers' cognitive and manual skills as well as the

[4]U.S. Department of Labour Occupational Characteristics Database

cognitive and manual skill requirements of firms.[5] To assess the effect of technological changes on wage inequality, we compare the data sets based on two cohorts: the first starting in 1979 (NLSY79) and the second starting in 1997 (NLSY97). We focus on employed workers between the ages of 27 and 29 in 1990 to 1991 and 2009 to 10 (from the NLSY79 and NLSY97 respectively). The wage, w, is the hourly rate adjusted by the CPI.

Two-dimensional skill demand is constructed from the O*NET, which contains information on skill requirements for occupations. We use a dataset for (Y_C, Y_M) constructed by Sanders (2016). We then match individual's education and training to their corresponding occupation in the NLSY. The matched value of (X_C, X_M) is (Y_C, Y_M) from O*NET. It is important to note that the workers' skills are independent of their occupation. Table 3.1 illustrates the summary statistics of both workers' skills and firms' skill demand.

Table 3.1: Summary statistics of skills and skill demand

	1990/91 ($n = 2984$)				2009/10 ($n = 4495$)			
	X_C	X_M	Y_C	Y_M	X_C	X_M	Y_C	Y_M
Mean	0.3596	-0.2912	0.0135	-0.1189	0.5667	-0.6601	0.0468	-0.2509
SD	0.7423	0.9923	0.8490	1.0240	0.7556	0.8358	0.9280	0.9656
Min	-2.0595	-1.7004	-2.0622	-1.6949	-2.3019	-1.8116	-2.5200	-1.6597
Max	2.1649	2.1855	2.0925	2.1895	1.9160	2.1838	3.0504	2.1351

To align the data with the model that features standard Gaussian distribution, the empirical distributions are transformed into Gaussian copulas. Each one-dimensional variable is converted into Gaussian variables using the inverse transform method, and their dependence is modeled using Gaussian copula. We employ Mardia's test[6] (Mardia, 1970) to check the bivariate normality for each two-dimensional variable. Table 3.2 shows that every $\tilde{X}$ and $\tilde{Y}$ for the two periods did not follow the bivariate normal distribution. Using SME procedure

[5] In this exercise, we rely on the data constructed by Lindenlaub (2017).

[6] Create the $n \times n$ matrix:

$$C = (c_{ij}) = X^* S^{-1} (X^*)',$$

where the ith row of X^*, $X_i^* = \tilde{X}_i - \bar{\tilde{X}}$, and define multivariate measures of skewness and kurtosis as follows:

$$b_1 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^3, \quad b_2 = \frac{1}{n} \sum_{i=1}^{n} c_{ii}^2$$

Under multidimensional normality, the limiting distributions of

$$\frac{nb_1}{6}, \quad \frac{\sqrt{n}\left(b_2 - d\left(d+2\right)\right)}{\sqrt{8d\left(d+2\right)}}$$

are a chi-square distribution with $d\left(d+1\right)\left(d+2\right)/6$ degrees of freedom and a $\mathrm{N}\left(0,1\right)$ distribution, respectively.

is recommended because it does not require the multivariate normality of variables.

Table 3.2: Multivariate normality: Mardia statistics (p-value) of transformed data

	1990/91 ($n_{90} = 2984$)		2009/10 ($n_{00} = 4495$)	
	$\tilde{X}$	$\tilde{Y}$	$\tilde{X}$	$\tilde{Y}$
Skewness	4.58 (0.333)	100.09 (0.000)	16.34 (0.003)	145.14 (0.000)
Kurtosis	4.44 (0.000)	0.29 (0.774)	14.42 (0.000)	1.98 (0.048)

We estimate the model for each period separately. We apply the SMD procedure using a finite-dimensional Bernstein polynomial sieve to construct the approximating space $\mathcal{G}_n$ of $\mathcal{G}$. The estimation results for technology parameters are given in Table 3.3. We report the results with Bernstein(6) and Bernstein(10) as $\mathcal{G}_n$, and $p^{J_n}\left(\tilde{X}\right)$ for $\hat{m}$ and $\hat{\Sigma}$ consisting of Bernstein(3) and Bernstein(5). There is a substantial decrease in δ, which represents the relative complementarities across tasks. This shows that technological advances have replaced workers for manual tasks but increased strong complementarities between skills and job attributes in cognitive tasks. In this two-dimensional world, the cognitive dimension becomes much more important in sorting. The parameters for non-interaction terms have no impact on the assignment and the curvature of the wage function. The increase in λ means that the productivity of cognitive skill also increases. We note that this productivity, unlike δ, is independent of a firm's cognitive skill demand. Overall, our results are consistent with the results in Lindenlaub (2017). However, their magnitudes are different. For both two period, the estimate of α is much bigger with a sieve MD estimation.

Table 3.3: Estimates of technology parameters

	1990/91		2009/10	
	ML	SMD	ML	SMD
α_C	0.446 (0.019)	3.777 (0.599)	0.712 (0.015)	2.734(0.632)
γ	0.964 (0.058)	0.298 (0.040)	-0.300 (0.024)	0.170 (0.053)
β_C	1.737 (0.014)	1.655 (0.089)	2.079 (0.012)	2.003 (0.090)
β_M	-0.361 (0.014)	-0.347 (0.088)	0.116 (0.012)	0.068 (0.082)
R^2	0.070	0.082	0.052	0.054

Standard errors in the parentheses

We now consider the effect of technological changes on wage inequality. Wage polarization is defined as the wage growth in the bottom and upper tails relative to the median. This phenomenon characterizes the U.S. labour market until the late 2000s. Figure 3.2 plots how the wages relative to the median wage change between 1990/91 and 2009/10 by wage

percentile. We find that there is a increases in upper tail wage inequality and a decrease in lower tail wage inequality. However, our estimated models with both transformed and untransformed data do not account for wage polarization. This result is inconsistent with the result presented in Lindenlaub (2017), who shows that the estimated technological changes, α_C and γ, could possibly trigger the wage polarization.

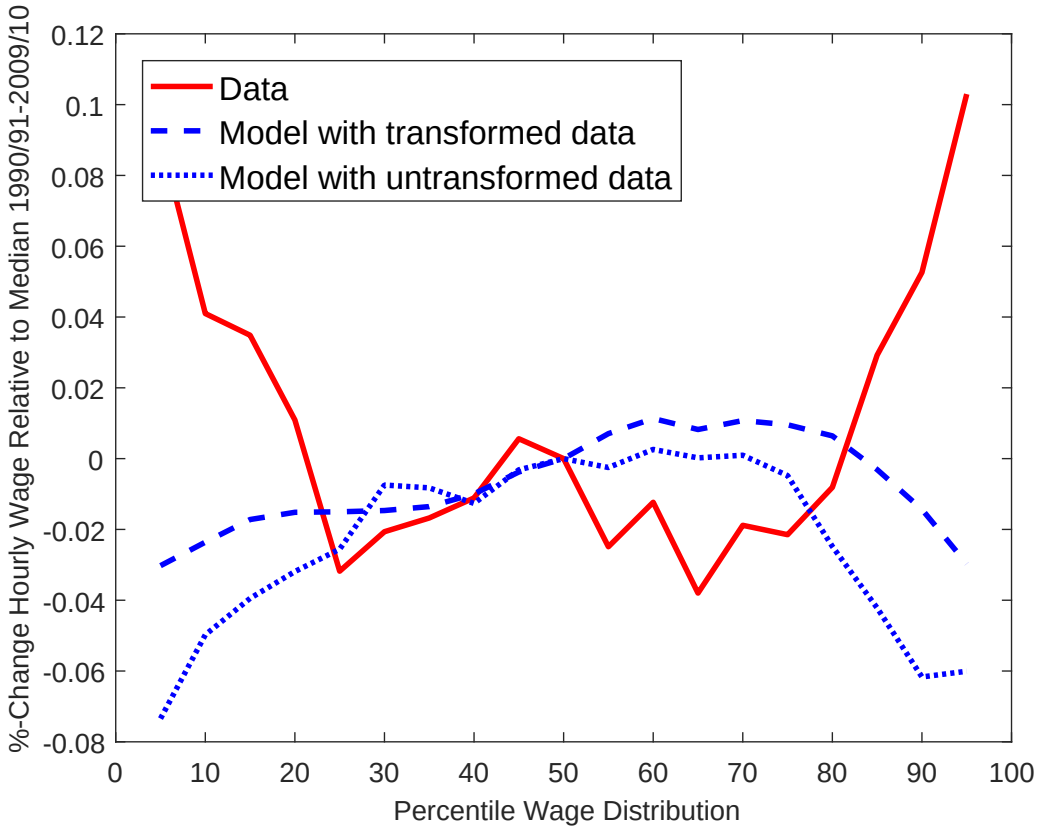

Figure 3.2: Wage polarization (data and model)

In sum, the matching model with two-dimensional heterogeneity is not sufficient to explain the phenomena in the labor market fully. We could think of possible extensions. I derive the identification result for the technology with between-task complementarities, but the estimation is based on the bi-linear technology to compare with the results in Lindenlaub (2017). Also, there might be another heterogeneity, e.g. interpersonal skill, affecting the assignment. Finally, in our framework, all workers with same cognitive-manual skills matched firms with the same skill demand without randomness. It suggests that we could consider unobserved heterogeneity.

3.6 Conclusion

In this chapter, we studied multidimensional matching models as an optimal transport problem. Based on the identification and regularity of a solution to the optimal transport problem, we propose to apply the SMD procedure to estimate the model. Our application to multidimensional matching model in the U.S. shows that worker-job complementarities in manual skills strongly decreased, whereas complementarities in cognitive skills increased. This phenomenon is consistent with that found in Lindenlaub (2017), but the magnitudes between two methods are quite different. We surmise that it might be due to the risk of misspecification.

I would like to highlight two potential directions for future work. The first direction is to introduce the measurement error in workers' characteristics as well. Ben-Moshe (2019) analyzes the classical linear regression model with measurement errors in all the variables. I hope to study the extension of this to the nonparametric regression models.

We can also think of other ways of introducing randomness in the assignment, for example, search frictions and unobserved heterogeneity. Although Eeckhout and Kircher (2011) develop the theory in matching with search frictions, it is restricted to the one-dimensional case. It will be challenging to extend the optimal transport approach to multidimensional setting.

Bibliography

Agosto, A., G. Cavaliere, D. Kristensen, and A. Rahbek (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *Journal of Empirical Finance 38, Part B*, 640 – 663.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy 81*(4), 813–846.

Ben-Moshe, D. (2019). Linear errors-in-variables and dependent factor models.

Bobkov, S. G. (1999, 10). Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab. 27*(4), 1903–1921.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics 44*(4), 375–417.

Brown, B. (1971). Martingale central limit theorems. *Annals of Mathematical Statistics 42*, 59–66.

Caffarelli, L. A. (1992). The regularity of mappings with a convex potential. *J. Amer. Math. Soc. 5*(1), 99–104.

Caffarelli, L. A. (1996). Boundary regularity of maps with convex potentials–II. *Annals of Mathematics 144*(3), 453–496.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics 136*(1), 163 – 188.

Carlier, G. (2003). *Duality and existence for a class of mass transportation problems and economic applications*, pp. 1–21. Tokyo: Springer Japan.

Carlier, G., V. Chernozhukov, and A. Galichon (2016, 06). Vector quantile regression: An optimal transport approach. *Ann. Statist. 44*(3), 1165–1192.

Chartrand, R., B. Wohlberg, K. R. Vixie, and E. M. Bollt (2009). A gradient descent solution to the monge-kantorovich problem. *Applied Mathematical Sciences 3*(22), 1071–1080.

Chen, B. and Y. Hong (2016). Detecting for smooth structural changes in garch models. *Econometric Theory 32*, 740–791.

Chen, S. and X.-J. Wang (2016). Strict convexity and C1,α regularity of potential functions in optimal transportation under condition A3w. *Journal of Differential Equations 260*(2), 1954 – 1974.

Chen, X. (2007). Chapter 76 large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6 of *Handbook of Econometrics*, pp. 5549 – 5632. Elsevier.

Chen, X. and D. Pouzo (2009, sep). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics 152*(1), 46–60.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica 66*(2), 289–314.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017, 02). Monge-kantorovich depth, quantiles, ranks and signs. *Ann. Statist. 45*(1), 223–256.

Chiappori, P.-A., R. J. McCann, and L. P. Nesheim (2010, Feb). Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Economic Theory 42*(2), 317–354.

Chu, C.-K. and J. Marron (1991). Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics 19*, 1906–1918.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist. 25*(1), 1–37.

Dahlhaus, R., S. Richter, and W. B. Wu (2017). Towards a general theory for non-linear locally stationary processes. Unpublished manuscript.

Dahlhaus, R. and S. Subba Rao (2006). Statistical inference for time-varying arch processes. *Ann. Statist. 34*(3), 1075–1114.

De Philippis, G. and A. Figalli (2014). The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc. (N.S.) 51*(4), 527–580.

Eeckhout, J. and P. Kircher (2011). Identifying sorting-in theory. *The Review of Economic Studies 78*(3), 872–906.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics 21*, 196–216.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*(429), 141–150.

Figalli, A., Y.-H. Kim, and R. J. McCann (2013, Sep). Hölder continuity and injectivity of optimal maps. *Archive for Rational Mechanics and Analysis 209*(3), 747–795.

Francq, C. and J.-M. Zakoïan (2004). Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli 10*, 605–637.

Fryzlewicz, P., T. Sapatinas, and S. Subba Rao (2008). Normalized least-squares estimation in time-varying arch models. *Ann. Statist. 36*(2), 742–786.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A. (2017). A survey of some recent applications of optimal transport methods to econometrics. *The Econometrics Journal 20*(2), C1–C11.

Gunsilius, F. F. (2018). On the convergence rate of potentials of brenier maps.

Han, H. and D. Kristensen (2014). Asymptotic theory for the qmle in garch-x models with stationary and nonstationary covariates. *Journal of Business & Economic Statistics 32*(3), 416–429.

Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica*, 173–197.

Kristensen, D. (2012). Non-parametric detection and estimation of structural change. *The Econometrics Journal 15*(3), 420–461.

Kristensen, D. and A. Rahbek (2005). Asymptotics of the qmle for a class of arch(q) models. *Econometric Theory null*, 946–961.

Lindenlaub, I. (2017). Sorting multidimensional types: Theory and application. *The Review of Economic Studies 84*(2), 718–789.

Loader, C. (2006). *Local regression and likelihood*. Springer Science & Business Media.

Ma, S. and M. R. Kosorok (2005, sep). Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis 96*(1), 190–217.

Ma, X.-N., N. S. Trudinger, and X.-J. Wang (2005, Aug). Regularity of potential functions of the optimal transportation problem. *Archive for Rational Mechanics and Analysis 177*(2), 151–183.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*(3), 519–530.

McCann, R. J. (1995, 11). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J. 80*(2), 309–323.

Monge, G. (1781). Memoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111 – 2245.

Richter, S. and R. Dahlhaus (2017). Cross validation for locally stationary processes.

Robinson, P. (1989). Nonparametric estimation of time-varying parameters. In P. Hackl (Ed.), *Statistical Analysis and Forecasting of Economic Structural Change*, pp. 253–264. Springer Berlin Heidelberg.

Sanders, C. (2016). Skill accumulation, skill uncertainty, and occupational choice.

Shen, X. and W. H. Wong (1994, 06). Convergence rate of sieve estimates. *Ann. Statist. 22*(2), 580–615.

Stone, C. J. (1982, 12). Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*(4), 1040–1053.

Subba Rao, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Adv. in Appl. Probab. 38*(4), 1155–1172.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*, 559–567.

Trudinger, N. S. and X.-J. Wang (2009). On the second boundary value problem for Monge-Ampère type equations and optimal transportation. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) 8*(1), 143–174.

van de Geer, S. A. (2009). *Empirical Processes in M-Estimation (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Villani, C. (2003). *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*, Volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer.

Wu, W. B. and X. Shao (2004). Limit theorems for iterated random functions. *J. Appl. Probab. 41*(2), 425–436.