4-2014

Multivariate Autoregressive Time Series Using Schweppe Weighted Wilcoxon Estimates

Jaime Burgos
Western Michigan University, jaime.burgosmariot@gmail.com

MULTIVARIATE AUTOREGRESSIVE TIME SERIES USING
SCHWEPPE WEIGHTED WILCOXON ESTIMATES

by

Jaime Burgos

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Statistics
Western Michigan University
April 2014

Doctoral Committee:

Jeffrey T. Terpstra, Ph.D., Chair
Joseph W. McKean, Ph.D.
Joshua Naranjo, Ph.D.
Edward A. Roth, Ph.D.

MULTIVARIATE AUTOREGRESSIVE TIME SERIES USING
SCHWEPPE WEIGHTED WILCOXON ESTIMATES

Jaime Burgos, Ph.D.

Western Michigan University, 2014

The increasing needs of forecasting techniques has led to the popularity of the vector autoregressive model in multivariate time series analysis, which has become of typical use across different fields due to its simplicity in application. The traditional method for estimating the model parameters is the least squares minimization, due to the linear nature of the model and its similarity with multivariate linear regression. However, since least squares estimates are sensitive to outliers, more robust techniques have become of interest. This manuscript investigates a robust alternative by obtaining the estimates using a weighted Wilcoxon dispersion with Schweppe-type weights. The first section introduces the typical definition of a vector autoregressive model, along with popular estimation methods and weighting schemes. In section two, the proposed estimator is shown to be asymptotically multivariate normal, centered about the true model parameters, at a rate of $n^{-\frac{1}{2}}$. Section three follows with an in depth discussion of the derivation of the main theoretical results. After that, in section four, a Monte Carlo study is presented to evaluate the performance of alternative estimators compared against the least squares estimates. The study results suggest that the Schweppe-weighted Wilcoxon estimates will generally have best performance. This result is most noticeable under the presence of additive outliers or when the series is closer to non-stationarity. In the last section, the estimation methods are applied to quadrivariate financial time series and results are compared. The applied example results indicate that estimates that use weights are better at detecting outliers by reducing their influence on the fit. This work provides a high efficiency robust alternative to the estimation problem of the vector autoregressive model parameters in multivariate time series analysis.

Keywords: Asymptotic normality, High-breakdown estimates, Rank-based estimates, Vector autoregressive, Schweppe weights, Multivariate times series, Innovation outliers, Additive outliers, U-Statistics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 Time Series Model

A widely used model in multivariate time series analysis is the stationary m-variate vector autoregressive model of order p. Through this manuscript, we refer it as the $\text{VAR}_m(p)$. The model, with an intercept, is generally expressed as

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{Y}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{Y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t; \quad t = 1, \ldots, T \\ &\stackrel{\text{def}}{=} \boldsymbol{\phi}_0 + \boldsymbol{\Phi} \boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \end{aligned} \tag{1}$$

where $p \in \{1, 2, \ldots\}$; $\boldsymbol{Y}_t, \boldsymbol{\phi}_0, \boldsymbol{\varepsilon}_t \in \mathbb{R}^m$; $m \in \{2, 3, \ldots\}$; $\boldsymbol{X}_{t-1} = (\boldsymbol{Y}'_{t-1}, \boldsymbol{Y}'_{t-2}, \ldots, \boldsymbol{Y}'_{t-p})' \in \mathbb{R}^{mp}$; $\boldsymbol{\Phi}_i \in \mathbb{R}^{m \times m}$, $i = 1, 2, \ldots, p$; $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p) \in \mathbb{R}^{m \times mp}$; $\boldsymbol{X}_0$ is an observable process random vector independent of $\boldsymbol{\varepsilon}_t$; and T is the number of realizations in the time series. The process has a stationary solution if and only if

$$\det(x^p \mathbb{I}_m - x^{p-1} \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p) = 0 \quad \Rightarrow \quad |x| < 1, \tag{2}$$

where $\det(\cdot)$ is the determinant operator, $\mathbb{I}_m$ represents the dimension m identity matrix, x may be complex valued, and $|\cdot|$ is the modulus operator on the complex plane (Brockwell & Davis, 2002). Furthermore, $\boldsymbol{\varepsilon}_t$ are assumed to be independent with an identical continuous distribution function $\text{F}(\cdot)$ that satisfies

$$\text{V}[\boldsymbol{\varepsilon}_t] = \boldsymbol{\Omega} \;\; p.d., \tag{3}$$

where *p.d.* stands for positive definite. Under assumptions (1), (2) and (3), $\{\boldsymbol{Y}_t\}$ is causal, invertible (Brockwell & Davis, 2002), ergodic (Krengel, 1985), and geometrically absolutely regular (g.a.r.) (Terpstra & Rao, 2002).

1.2 Parameter Estimation

Estimates for the parameters in (1) are typically obtained by minimizing, with respect to $\boldsymbol{\phi}_0$ and $\boldsymbol{\Phi}$, a dispersion function. The most common selection, generalized with weights and for a multivariate setting, is the L_2 dispersion function,

$$\mathrm{D}_2(\boldsymbol{\phi}_0, \boldsymbol{\Phi}) = \sum_{t=1}^{T} b_t \| \,\boldsymbol{\varepsilon}_{\mathrm{t}}(\boldsymbol{\phi}_0, \boldsymbol{\Phi})\|^2, \tag{4}$$

where $b_t > 0$ denotes a weight function, that may depend on both the design $\boldsymbol{X}_{t-1}$ and the response $\boldsymbol{Y}_t$. $\boldsymbol{\varepsilon}_{\mathrm{t}}(\boldsymbol{\phi}_0, \boldsymbol{\Phi}) = \boldsymbol{Y}_t - \boldsymbol{\phi}_0 - \boldsymbol{\Phi}\boldsymbol{X}_{t-1}$, and $\| \cdot \|$ denotes the Euclidean norm. Typically $b_t \equiv 1$, which results in a component wise Ordinary Least Squares (OLS) estimation. The asymptotic theory for the OLS estimates can be found in many time series textbooks, including Lütkepohl (1993) and Fuller (1996).

As with most least squares estimates, outlying observations can yield unreliable estimates and predictions (Rousseeuw & Leroy, 1987; Hettmansperger & McKean, 2011). A number of robust estimators for the parameters in (1) have been studied. These include a functional least squares approach by Heathcote and Welsh (1988); an extension of the RA-estimates proposed by Bustos and Yohai (1986), Li and Hui (1989), and Ben, Martinez and Yohai (1999); GR-estimates proposed by Terpstra and Rao (2002); weighted-L_1 estimates proposed by Reber, Terpstra and Chen (2008); an extension of least trimmed squares by Croux and Joosens (2008); and an extension of MM-estimates by Muler and Yohai (2013).

More robust estimators can be obtained by using different dispersion functions. For instance, a multivariate generalization of the weighted-L_1 dispersion function

$$\mathrm{D}_1(\boldsymbol{\phi}_0, \boldsymbol{\Phi}) = \sum_{t=1}^{T} b_t \| \,\boldsymbol{\varepsilon}_{\mathrm{t}}(\boldsymbol{\phi}_0, \boldsymbol{\Phi})\|, \tag{5}$$

where $b_t > 0$ denotes a weight function, that may depend on both the design $\boldsymbol{X}_{t-1}$ and the response $\boldsymbol{Y}_t$. Using Mallows weights (Mallows, 1975), the asymptotic distribution of the weighted-L_1 estimates is described by Reber *et al.* (2008).

Another robust estimator can be obtained from a multivariate generalization of the weighted-Wilcoxon dispersion function

$$\mathrm{D}_{\mathrm{WW}}(\boldsymbol{\Phi}) = \sum_{i<j}^{T} b_{ij} \| \,\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\phi}_0, \boldsymbol{\Phi}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\phi}_0, \boldsymbol{\Phi})\|, \tag{6}$$

where $b_{ij} = b_{ji} > 0$ denotes a weight function, that may depend on both the design and the response points at the i^{th} and j^{th} realizations. Similar to the univariate case, $\boldsymbol{\phi}_0$ cannot be estimated using (6) due to intercept invariance property of the dispersion function. Hettmansperger and McKean

(2011) describe methods for post-estimation of $\boldsymbol{\phi}_0$ for the univariate case, which can be generalized to the multivariate setting. Using Mallows weights, the asymptotic theory of multivariate weighted-Wilcoxon (WW) estimates, known as Multivariate Generalized Rank (GR) estimates, was introduced by Terpstra and Rao (2002).

1.3 Weighting Schemes

The estimates obtained from (4), (5) and (6) are influenced by the selection of the weight functions. The weighting schemes used throughout this manuscript can be grouped into three classes: non-random weights, Mallows weights and Schweppe weights (Handschin, Schweppe, Kohlas & Fiechter, 1975).

1.3.1 Non-random Weights

Non-random weights do no explicitly depend on the time series realizations. The most common case is selecting the weights constant to one. Here $b_t \equiv 1$ and $b_{ij} \equiv 1$, thus weighing each realization equally. Another example of non-random weights would be selecting $b_t = t$ and $b_{ij} = ij$, giving more weight to realizations as they are more recent.

1.3.2 Mallows Weights

Mallows weights depend only on the design point. A simple example of Mallows weights for (4), (5) and (6) are the Boldin (Boldin, 1994) and Theil weights (Theil, 1950). These can be generalized to the multivariate setting by defining them as

$$b_t = \|\boldsymbol{X}_{t-1}\|^{-1} \; if \; \boldsymbol{X}_{t-1} \neq \boldsymbol{0} \quad \text{and} \quad b_{ij} = \|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^{-1} \; if \; \boldsymbol{X}_{i-1} \neq \boldsymbol{X}_{j-1}, \tag{7}$$

and when $\boldsymbol{X}_{t-1} = \boldsymbol{0}$ or $\boldsymbol{X}_{i-1} = \boldsymbol{X}_{j-1}$, the largest observed finite weight is used for completeness and computational feasibility.

Another popular Mallows weighting scheme can be found in Chang, McKean, Naranjo and Sheather (1999). These can be generalized to the multivariate setting by defining them as

$$b_t = \min\left\{1, \left(\frac{c}{\mathrm{d}^2(\boldsymbol{X}_{t-1})}\right)^{k/2}\right\} \quad \text{and} \quad b_{ij} = b_i b_j, \tag{8}$$

where $\mathrm{d}(\cdot)$ denotes the Mahalanobis distance based on a robust measure of center and covariance and c and k are tuning constants. Typically, $c = \chi^2_{1\text{-}\alpha}(mp)$ is the $100(1-\alpha)^{th}$ percentile of a chi-squared distribution with mp degrees of freedom and $k = 2$.

1.3.3 Schweppe Weights

Schweppe weights depend on both design and response points. An example of Schweppe weights are the High Breakdown weights defined in Chang *et al.* (1999). These can be generalized to the multivariate setting by defining them as

$$b_t = \psi\left(\frac{b}{a_t}\right) \quad \text{and} \quad b_{ij} = \psi\left(\frac{b^2}{a_i a_j}\right), \tag{9}$$

where $a_t = \frac{\mathrm{d}(\hat{\boldsymbol{\varepsilon}}_t)}{\psi(c/\mathrm{d}^2(\boldsymbol{X}_{t-1}))}$; $\hat{\boldsymbol{\varepsilon}}_t$ denotes the t^{th} residual vector based on an initial robust estimate, and $\psi(x) = 1, x, -1$ when $x \geq 1, -1 < x < 1, x \leq -1$, respectively. The tuning parameter $b = \mathrm{med}_{\mathrm{i}}\{a_i\} + 3\,\mathrm{MAD}_{\mathrm{i}}\{a_i\}$.

Another type of Schweppe weights are defined in Terpstra, McKean and Naranjo (2001). These can be generalized to the multivariate setting by defining them as

$$b_t = 1 - \mathrm{I}(\mathrm{d}^2(\hat{\varepsilon}_t) > c_1)\,\mathrm{I}(\mathrm{d}^2(\boldsymbol{X}_{t-1}) > c_2)(1 - h_t) \quad \text{and} \quad b_{ij} = b_i b_j, \tag{10}$$

where h_t is the Mallows weight defined in (8) at the t^{th} realization and $\mathrm{I}(\cdot)$ is an indicator function yielding 1 if the logical statement is true, and 0 otherwise. The tuning parameters c_1, c_2 are typically $\chi^2_{1\text{-}\alpha}(m)$ and $\chi^2_{1\text{-}\alpha}(mp)$, respectively. These weights make use of an indicator function to downweight only the *bad* leverage points.

2 Theoretical Results

2.1 The Estimate of $\mathbf{\Phi}$

We propose the estimator of $\mathbf{\Phi}$ as a value that minimizes the weighted Wilcoxon dispersion function using Schweppe weights. Under the definition in (6), the dispersion function $\mathrm{D}_{\mathrm{WW}}(\mathbf{\Phi})$ is non-negative, piecewise linear, and convex. Thus, $\mathrm{D}_{\mathrm{WW}}(\mathbf{\Phi})$ has a minimum and an estimate can be obtained.

For notational convenience, the $\mathrm{VAR}_m(p)$ model in (1) is rewritten in a notation where $\mathbf{\Phi}$ is vectorized. Let $\mathrm{vec}(\cdot)$ denote a columns stacking operation that transforms a $r \times c$ matrix into a $rc \times 1$ vector. Also, let $\otimes$ denote the Kronecker product. Then,

$$\boldsymbol{Y}_t = \boldsymbol{\phi}_0 + \mathbf{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t = \boldsymbol{\phi}_0 + (\boldsymbol{X}'_{t-1} \otimes \mathbb{I}_m)\,\mathrm{vec}(\mathbf{\Phi}) + \boldsymbol{\varepsilon}_t \stackrel{\text{def}}{=} \boldsymbol{\phi}_0 + \boldsymbol{x}'_{t-1}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t; \;\; t = 1, \ldots, n,$$

where $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{\Phi})$ and $\boldsymbol{x}'_{t-1} = \boldsymbol{X}'_{t-1} \otimes \mathbb{I}_m$. More details on the use of $\mathrm{vec}(\cdot)$ to rewrite models can be found in Lütkepohl (1993, Appendix A.12). Furthermore, we redefine $\boldsymbol{\varepsilon}_{\mathrm{t}}(\boldsymbol{\phi}_0, \boldsymbol{\beta}) = \boldsymbol{Y}_t - \boldsymbol{\phi}_0 - \boldsymbol{x}'_{t-1}\boldsymbol{\beta}$, so that the dispersion is now viewed as a function of $\boldsymbol{\beta}$ instead of $\mathbf{\Phi}$. Hence, our proposed estimator, denoted as $\hat{\boldsymbol{\beta}}_n$, is such that $\mathrm{D}_{\mathrm{WW}}(\hat{\boldsymbol{\beta}}_n) = \min_{\boldsymbol{\beta}} \mathrm{D}_{\mathrm{WW}}(\boldsymbol{\beta})$. Recall that, $\mathrm{D}_{\mathrm{WW}}(\boldsymbol{\beta})$ is invariant to $\boldsymbol{\phi}_0$ and as a consequence it can not be directly estimated as part of minimization. Therefore, we further simplify the notation by removing $\boldsymbol{\phi}_0$ from expressions in $\mathrm{D}_{\mathrm{WW}}(\boldsymbol{\beta})$.

2.2 Asymptotic Distribution

We start by stating a list of assumptions under which the asymptotic theory holds. The results rely on assumptions

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{O}_p(1), \tag{W1}$$

$$\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \boldsymbol{\nabla}\, \mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}) \textit{ is continuous } \forall\, (i, j, \boldsymbol{\theta}), \tag{W2}$$

$$\|\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta})\| \leq B_D < \infty \;\forall\, (i, j, \boldsymbol{\theta}), \tag{W3}$$

$$\mathrm{b}(\boldsymbol{X}_{i-1}, \boldsymbol{\varepsilon}_i, \boldsymbol{X}_{j-1}, \boldsymbol{\varepsilon}_j) = \mathrm{b}(\boldsymbol{X}_{j-1}, \boldsymbol{\varepsilon}_j, \boldsymbol{X}_{i-1}, \boldsymbol{\varepsilon}_i) \;\forall\, (i, j), \tag{W4}$$

$$\sup_{i<j} \mathrm{E}[\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \|\varepsilon_j - \varepsilon_i\|^{-1}]^{2+\delta} < \infty \textit{ for some } \delta > 0, \tag{E1}$$

$$\sup_{i<j} \mathrm{E}[b_{ij}\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \|\varepsilon_j - \varepsilon_i\|^{-1}] < \infty, \tag{E2}$$

$$\sup_{i<j} \mathrm{E}[b_{ij}\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \|\varepsilon_j - \varepsilon_i\|^{-1}]^{2+\delta} < \infty \textit{ for some } \delta > 0, \tag{E3}$$

$$\sup_{i<j} \mathrm{E}[\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|]^{2+\delta} < \infty \textit{ for some } \delta > 0, \tag{E4}$$

$$\sup_{i<j} \mathrm{E}[b_{ij}\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|]^{2+\delta} < \infty \textit{ for some } \delta > 0, \tag{E5}$$

$$\sup_{i<j, i<k} \mathrm{E}[b_{ij} b_{ik}\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|\|\boldsymbol{X}_{k-1} - \boldsymbol{X}_{i-1}\|]^{2+\delta} < \infty \textit{ for some } \delta > 0, \text{ and} \tag{E6}$$

$$\mathbb{E}_{\mathrm{i.j}}[b_{ij}\|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \|\varepsilon_j - \varepsilon_i\|^{-1}]^{1+\delta} < \infty \textit{ for some } \delta > 0. \tag{E7}$$

Since the main result to obtain is the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$, we start by denoting the true parameter vector for the $\mathrm{VAR}_m(p)$ as $\boldsymbol{\beta}_0$. Following traditional methods of proof, we first need to establish asymptotic linearity (AL), asymptotic uniform linearity (AUL), and asymptotic uniform quadradicity (AUQ). Thus, we define

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) &= -\boldsymbol{\nabla}\, \mathrm{D}_{\mathrm{WW}}(\boldsymbol{\beta}), \\
\mathrm{D_n}(\boldsymbol{\Delta}) &= n^{-1}\, \mathrm{D}_{\mathrm{WW}}(\boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\boldsymbol{\Delta}), \\
\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) &= -\frac{\partial}{\partial \boldsymbol{\Delta}}\, \mathrm{D_n}(\boldsymbol{\Delta}) = n^{-\frac{3}{2}}\, \boldsymbol{S}(\boldsymbol{\beta}_0 + n^{-\frac{1}{2}}\boldsymbol{\Delta}), \text{ and} \\
\mathrm{Q_n}(\boldsymbol{\Delta}) &= \mathrm{D_n}(\boldsymbol{0}) - \boldsymbol{\Delta}'\, \boldsymbol{S}_{\mathrm{n}}(\boldsymbol{0}) + \boldsymbol{\Delta}'\boldsymbol{C}\boldsymbol{\Delta},
\end{aligned}$$

where $\boldsymbol{\nabla}$ is the gradient operator, $\boldsymbol{\Delta} \in \mathbb{R}^{m^2 p}$ is arbitrary but fixed, and $\boldsymbol{C}$ is a fixed $m^2p \times m^2p$ positive definite matrix.

In the multivariate setting, AL, AUL and AUQ refer to

$$\begin{aligned}
\text{AL}: &\quad \| \boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - [\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) - 2\boldsymbol{C}\boldsymbol{\Delta}]\| = o_p(1),\\
\text{AUL}: &\quad \sup_{\|\boldsymbol{\Delta}\|\le \mathrm{c}} \| \boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - [\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) - 2\boldsymbol{C}\boldsymbol{\Delta}]\| = o_p(1)\ \forall\, c>0, \text{ and}\\
\text{AUQ}: &\quad \sup_{\|\boldsymbol{\Delta}\|\le \mathrm{c}} |\,\mathrm{D}_{\mathrm{n}}(\boldsymbol{\Delta}) - \mathrm{Q}_{\mathrm{n}}(\boldsymbol{\Delta})| = o_p(1)\ \forall\, c>0,
\end{aligned}$$

for some $\boldsymbol{C}$ matrix. Theorem 2.1 establishes these results and its proof is provided in Section 3.

Theorem 2.1. *Let* $\mathbb{E}_{\mathrm{i.j}}[\cdot]$ *be the Product Expectation functional defined in Appendix* A.1*,* $\boldsymbol{J}[\cdot]$ *the Jacobian operator,*

$$\begin{aligned}
\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{x}) &= \frac{\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i - \boldsymbol{x}}{\|\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i - \boldsymbol{x}\|}, \text{ and}\\
\boldsymbol{C} &= \frac{1}{4}\,\mathbb{E}_{\mathrm{i.j}}[b_{ij}(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\,\boldsymbol{J}[-\boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})](\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})'].
\end{aligned}$$

Then, AL, AUL*, and* AUQ *hold under model assumptions* (1)-(3)*,* W1-W4*,* E1-E3 *and* E7*.*

With AUL and AUQ established, we proceed to derive the asymptotic distribution of $\boldsymbol{S}_{\mathrm{n}}(\mathbf{0})$. Theorem 2.2 establishes this result and its proof is provided in Section 3.

Theorem 2.2. *Let* $\mathbb{E}_{\mathrm{i.j.k}}[\cdot]$ *be the Product Expectation functional defined in Appendix* A.1*,*

$$\begin{aligned}
\boldsymbol{u}(\boldsymbol{x}) &= \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \text{ and}\\
\boldsymbol{\Sigma} &= \mathbb{E}_{\mathrm{i.j.k}}[b_{ij}b_{ik}(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\,\boldsymbol{u}(\boldsymbol{\varepsilon}_k - \boldsymbol{\varepsilon}_i)'(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{i-1})'].
\end{aligned}$$

Then, $\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}(\mathbf{0}, \boldsymbol{\Sigma})$*, under model assumptions* (1)-(3)*,* W1-W4*, and* E4-E5*.*

Now, by combining Theorems 2.1 and 2.2 with the Jaeckel (1972) convexity argument, we proceed to state our main result. Theorem 2.3 establishes this result and its proof is provided in Section 3.

Theorem 2.3. *Under assumptions of Theorems* 2.1 *and* 2.2*, we have that*

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}\left(\mathbf{0}, \frac{1}{4}\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}\right).$$

For practicality of Theorem 2.3, estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{C}$ are needed. Note that, $\boldsymbol{\Sigma}$ and $\boldsymbol{C}$ are actually functionals defined on the distribution function of $\boldsymbol{Z}_t = (\boldsymbol{\varepsilon}_t', \boldsymbol{X}_{t-1}')'$. Then, we let $\hat{\boldsymbol{\Sigma}}_n$ and $\hat{\boldsymbol{C}}_n$ be the corresponding von Mises statistics. Thus, by Lemma 3.1 and assumptions E3 and E6, it follows from Theorem 1 part (c) of Denker and Keller (1983, p.507) that $\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma} = \boldsymbol{o}_p(1)$ and $\hat{\boldsymbol{C}}_n - \boldsymbol{C} = \boldsymbol{o}_p(1)$. Finally, we state our practical result as Theorem 2.4.

Theorem 2.4. *Under assumptions of Theorems* 2.1, 2.2, *and assumption E6, we have that*

$$
\begin{aligned}
&2n^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_n^{-\frac{1}{2}}\hat{\boldsymbol{C}}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}(\mathbf{0}, \mathbb{I}_{m^2p}) \text{ and} \\
&4n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)'\hat{\boldsymbol{C}}_n\hat{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{C}}_n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \chi^2(m^2p),
\end{aligned}
$$

where $\hat{\boldsymbol{\Sigma}}_n$ and $\hat{\boldsymbol{C}}_n$ are the von Mises statistics of $\boldsymbol{\Sigma}$ and $\boldsymbol{C}$, respectively.

3 Technical Details

The following property of the $\mathrm{VAR}_m(p)$ and proof is given as Lemma 2.1 in Terpstra *et al.* (2002). This property is used in multiple proofs and it is stated here, using the notation of this manuscript, for completeness.

Lemma 3.1. *Consider the* $\mathrm{VAR}_m(p)$ *model defined in* (1)-(3) *and let* $\boldsymbol{Z}_t = (\boldsymbol{\varepsilon}_t', \boldsymbol{X}_{t-1}')'$. *Then,* $\{\boldsymbol{Z}_t\}$ *and* $\{\boldsymbol{X}_t\}$ *are strictly stationary g.a.r. processes.*

The following inequality which we find useful in some proofs is given as Lemma 2.2 in Terpstra *et al.* (2002). We present it here as Lemma 3.2, simplified in order to avoid extending the notation used in this manuscript.

Lemma 3.2. *Let* $\boldsymbol{Z}_t$ *be a g.a.r. process so that* $\beta(n) = \rho^n$ *for some* $\rho \in (0,1)$, $\mathbb{E}_{\mathrm{i.j}}[\cdot]$ *be the Product Expectation functional defined in Appendix* A.1, $i < j$, $g_n = \mathrm{g_n}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$ *be a measurable function (possibly depending on n) that satisfies* $\mathrm{E}\,|g_n|^q < \infty$ *and* $\mathbb{E}_{\mathrm{i.j}}\,|g_n|^q < \infty$ *for some* $q > 1$, *and* $M_q = \max\{\mathrm{E}[|g_n|^q]^{\frac{1}{q}}, \mathbb{E}_{\mathrm{i.j}}[|g_n|^q]^{\frac{1}{q}}\}$. *Then,*

$$|\,\mathrm{E}[g_n] - \mathbb{E}_{\mathrm{i.j}}[g_n]| \leq 4M_q \left(\rho^{\frac{q-1}{q}}\right)^{j-i}.$$

The expression for $\boldsymbol{S}(\boldsymbol{\beta})$ is also common to several proofs and is obtained by using rules of derivatives with respect to vectors and the definition of $\mathrm{D_{WW}}(\boldsymbol{\beta})$. Starting from the definition of $\boldsymbol{S}(\boldsymbol{\beta})$, we have that

$$\begin{aligned}
\boldsymbol{S}(\boldsymbol{\beta}) &= -\boldsymbol{\nabla}\,\mathrm{D_{WW}}(\boldsymbol{\beta}) = -\frac{\partial}{\partial\boldsymbol{\beta}}\,\mathrm{D_{WW}}(\boldsymbol{\beta}) = -\frac{\partial}{\partial\boldsymbol{\beta}}\sum_{i<j}^{n} b_{ij}\|\,\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})\| \\
&= -\sum_{i<j}^{n} b_{ij}\|\,\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})\|^{-1}\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta}))'(\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})) \\
&= \sum_{i<j}^{n} b_{ij}\|\,\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})\|^{-1}(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})) \\
&\overset{\mathrm{def}}{=} \sum_{i<j}^{n} b_{ij}(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_{\mathrm{j}}(\boldsymbol{\beta}) - \boldsymbol{\varepsilon}_{\mathrm{i}}(\boldsymbol{\beta})),
\end{aligned}$$

where $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$.

Establishing asymptotic results is simplified under the assumption that b_{ij} has no additional stochastic components besides $\boldsymbol{X}_{i-1}$, $\boldsymbol{\varepsilon}_i$, $\boldsymbol{X}_{j-1}$, and $\boldsymbol{\varepsilon}_j$. For notation purpose, we let $\boldsymbol{\theta}$ denote the ν dimensional vector of parameters used in the calculation of the weights. We also denote $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ as the corresponding vector of true parameters and estimators, respectively. Furthermore, we denote

$\mathrm{b_{ij}}(\hat{\boldsymbol{\theta}})$ as a function of $\boldsymbol{X}_{i-1}$, $\boldsymbol{X}_{j-1}$, $\hat{\boldsymbol{\varepsilon}}_i$, and $\hat{\boldsymbol{\varepsilon}}_j$ from an initial fit, and other stochastic components. Similarly, we denote $\mathrm{b_{ij}}(\boldsymbol{\theta}_0)$ as a function of $\boldsymbol{X}_{i-1}$, $\boldsymbol{\varepsilon}_i$, $\boldsymbol{X}_{j-1}$, $\boldsymbol{\varepsilon}_j$, and no other stochastic components.

Lemma A.1 shows that, under model assumptions (1)-(3), W1-W3, and E1, AL can be established using $b_{ij} = \mathrm{b_{ij}}(\boldsymbol{\theta}_0)$. Analogously, Lemma A.2 shows that, under model assumptions (1)-(3), W1-W4, and E4, the asymptotic distribution of $\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{0})$ can be obtained using $b_{ij} = \mathrm{b_{ij}}(\boldsymbol{\theta}_0)$.

3.1 Proof of Theorem 2.1

Heiler and Willers (1988) have shown that AL, AUL, and AUQ are equivalent in the context of linear regression. Their proof implies that linearity in parameters of the regression model and convexity of the dispersion function are sufficient conditions for this result to hold. Since the $\mathrm{VAR}_m(p)$ and $\mathrm{D_{WW}}(\cdot)$ satisfy these conditions, it suffices to establish AL.

To begin, let $\boldsymbol{\lambda} \in \mathbb{R}^{m^2 p}$ be arbitrary but fixed and $T_n = \boldsymbol{\lambda}'(\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - \boldsymbol{S}_{\mathrm{n}}(\boldsymbol{0}))$. Thus, by the Cramér-Wold Theorem it suffices to show that $T_n + 2\boldsymbol{\lambda}'\boldsymbol{C}\boldsymbol{\Delta} = o_p(1)$. Then, it follows by definitions that

$$
\begin{aligned}
T_n = \boldsymbol{\lambda}'(\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - \boldsymbol{S}_{\mathrm{n}}(\boldsymbol{0})) &= n^{-\frac{3}{2}} \sum_{i<j}^{n} b_{ij} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}(\varepsilon_j - \varepsilon_i - \boldsymbol{d}_{ijn}) - \boldsymbol{u}(\varepsilon_j - \varepsilon_i)) \\
&\stackrel{\text{def}}{=} n^{-\frac{3}{2}} \sum_{i<j}^{n} b_{ij} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})), \qquad (11)
\end{aligned}
$$

where $\boldsymbol{d}_{ijn} = (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})' n^{-\frac{1}{2}} \boldsymbol{\Delta}$ and $\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{x}) = \boldsymbol{u}(\varepsilon_j - \varepsilon_i - \boldsymbol{x})$. Next, we let

$$
\begin{aligned}
U_n^{\mathrm{AL}} &= n^{-\frac{3}{2}} \sum_{i<j}^{n} b_{ij} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})] \boldsymbol{d}_{ijn} \\
&= \left(\frac{n-1}{n}\right) \left(\frac{2}{n(n-1)}\right) \sum_{i<j}^{n} 2^{-1} b_{ij} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})](\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})' \boldsymbol{\Delta} \\
&\stackrel{\text{def}}{=} \left(\frac{n-1}{n}\right) \left(\frac{2}{n(n-1)}\right) \sum_{i<j}^{n} \mathrm{h_{AL}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j),
\end{aligned}
$$

where $\boldsymbol{J}[\cdot]$ is the Jacobian operator and $\boldsymbol{Z}_t = (\boldsymbol{\varepsilon}_t', \boldsymbol{X}_{t-1}')'$.

Lemma A.3 shows that, under model assumptions (1)-(3), E3 and E7, $T_n - U_n^{\mathrm{AL}} = o_p(1)$. Hence, it suffices to show that $U_n^{AL} + 2\boldsymbol{\lambda}'\boldsymbol{C}\boldsymbol{\Delta} = o_p(1)$. Furthermore, under assumption W4, U_n^{AL} is a U-statistic with symmetric kernel $\mathrm{h_{AL}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$. Thus, by Lemma 3.1 and assumption E3, it follows from Theorem 1 part (c) of Denker and Keller (1983, p.507) that $U_n^{\mathrm{AL}} - \mathbb{E}_{\mathrm{i.j}}[\mathrm{h_{AL}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] = o_p(1)$.

Finally, we have that

$$
\begin{aligned}
\mathbb{E}_{\mathrm{i.j}}[\mathrm{h}_{\mathrm{AL}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] &= \mathbb{E}_{\mathrm{i.j}}[2^{-1} b_{ij} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})](\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})' \boldsymbol{\Delta}] \\
&= -2\boldsymbol{\lambda}' \left(\frac{1}{4}\, \mathbb{E}_{\mathrm{i.j}}[b_{ij} (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{J}[-\, \boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})](\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})'] \right) \boldsymbol{\Delta} \\
&= -2\boldsymbol{\lambda}' \boldsymbol{C} \boldsymbol{\Delta},
\end{aligned}
$$

which completes our proof. □

3.2 Proof of Theorem 2.2

To obtain the asymptotic distribution of $\boldsymbol{S}_{\mathrm{n}}(\mathbf{0})$, we follow the approach for U-statistics under absolute regularity proposed by Denker and Keller (1983). To begin, let $\boldsymbol{\lambda} \in \mathbb{R}^{m^2 p}$ be arbitrary but fixed. Then, it follows from definitions that

$$
\begin{aligned}
\boldsymbol{\lambda}' \boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) &= n^{-\frac{3}{2}} \sum_{i<j}^{n} b_{ij} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i) \\
&= \left(\frac{n-1}{n} \right) n^{\frac{1}{2}} \left(\frac{2}{n(n-1)} \right) \sum_{i<j}^{n} 2^{-1} b_{ij} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i) \\
&\stackrel{\mathrm{def}}{=} \left(\frac{n-1}{n} \right) n^{\frac{1}{2}} U_n,
\end{aligned}
$$

where, under assumption W4, U_n is a U-statistic with symmetric kernel $\mathrm{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) = 2^{-1} b_{ij} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)$ and $\boldsymbol{Z}_t = (\boldsymbol{\varepsilon}_t', \boldsymbol{X}_{t-1}')'$. Next, we obtain

$$
\mathbb{E}_{\mathrm{i.j}}[\mathrm{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] = \iint 2^{-1} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1}) \left(\iint b_{ij}\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) \right) \mathrm{dG}(\boldsymbol{X}_{i-1})\, \mathrm{dG}(\boldsymbol{X}_{j-1}),
$$

where $\mathrm{G}(\cdot)$ denotes the distribution function of $\boldsymbol{X}_t$. Note that, by definition of $\boldsymbol{u}(\cdot)$, and assumption W4, it follows that

$$
\begin{aligned}
&\iint b_{ij}\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) = -\iint b_{ij}\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j), \\
&\iint b_{ij}\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) = 0, \text{ and} \\
&\mathbb{E}_{\mathrm{i.j}}[\mathrm{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] = 0.
\end{aligned}
$$

Next, following the approach, we let

$$\mathrm{h}_1(\boldsymbol{Z}_i) = \iint \mathrm{h}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j)\, \mathrm{dG}(\boldsymbol{X}_{j-1}) \text{ and}$$
$$\sigma_n^2 = \mathrm{E}\left[\sum_{i=1}^{n} \mathrm{h}_1(\boldsymbol{Z}_i)\right]^2.$$

By model assumptions (1) and (2), we have

$$\begin{aligned}\sigma_n^2 &= \mathrm{E}\left[\sum_{i=1}^{n} \mathrm{h}_1(\boldsymbol{Z}_i)\right]^2 \\ &= n\,\mathrm{E}[\mathrm{h}_1^2(\boldsymbol{Z}_i)] + 2\sum_{k=2}^{n}(n-(k-1))\,\mathrm{E}[\mathrm{h}_1(\boldsymbol{Z}_1)\,\mathrm{h}_1(\boldsymbol{Z}_k)]. \qquad (12)\end{aligned}$$

Focusing attention to the expectation inside the summation and by model assumption (3), we have that

$$\begin{aligned}&\mathrm{E}[\mathrm{h}_1(\boldsymbol{Z}_1)\,\mathrm{h}_1(\boldsymbol{Z}_k)] \\ &= \int \mathrm{h}_1(\boldsymbol{Z}_1)\,\mathrm{h}_1(\boldsymbol{Z}_k)\,\mathrm{dH_k}(\boldsymbol{Z}_1, \boldsymbol{Z}_k) \\ &= \int \mathrm{h}_1(\boldsymbol{Z}_1)\left(\iint 2^{-1} b_{kj} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{k-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j)\,\mathrm{dG}(\boldsymbol{X}_{j-1})\right)\,\mathrm{dH_k}(\boldsymbol{Z}_1, \boldsymbol{Z}_k) \\ &= \iint \mathrm{h}_1(\boldsymbol{Z}_1)\left(\iint 2^{-1} b_{kj} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{k-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j)\,\mathrm{dG}(\boldsymbol{X}_{j-1})\right)\,\mathrm{d\tilde{H}_k}(\boldsymbol{Z}_1, \boldsymbol{X}_{k-1})\,\mathrm{dF}(\boldsymbol{\varepsilon}_k) \\ &= \int \mathrm{h}_1(\boldsymbol{Z}_1) \int 2^{-1} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{k-1})\left(\iint b_{kj}\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j)\right)\,\mathrm{dG}(\boldsymbol{X}_{j-1})\,\mathrm{d\tilde{H}_k}(\boldsymbol{Z}_1, \boldsymbol{X}_{k-1}),\end{aligned}$$

where $\mathrm{H_k}(\cdot)$ denotes the joint distribution function of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_k$, and $\tilde{\mathrm{H}}_\mathrm{k}(\cdot)$ denotes the joint distribution function of $\boldsymbol{Z}_1$ and $\boldsymbol{X}_{k-1}$. Note that, by definition of $\boldsymbol{u}(\cdot)$ and assumption W4, it follows that

$$\begin{aligned}&\iint b_{kj}\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j) = -\iint b_{kj}\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j), \\ &\iint b_{kj}\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_k)\,\mathrm{dF}(\boldsymbol{\varepsilon}_j) = 0, \text{ and} \\ &\mathrm{E}[\mathrm{h}_1(\boldsymbol{Z}_1)\,\mathrm{h}_1(\boldsymbol{Z}_k)] = 0 \;\forall\; k \in \{2, 3, \ldots, n\}.\end{aligned}$$

Next, we focus attention back to equation (12), and have that

$$\begin{aligned} n^{-1}\sigma_n^2 &= \mathrm{E}[\mathrm{h}_1^2(\boldsymbol{Z}_i)] \\ &= 4^{-1}\boldsymbol{\lambda}'\,\mathbb{E}_{\mathrm{i.j.k}}[b_{ij}b_{ik}(\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i)\,\boldsymbol{u}(\boldsymbol{\varepsilon}_k-\boldsymbol{\varepsilon}_i)'(\boldsymbol{x}_{k-1}-\boldsymbol{x}_{i-1})']\boldsymbol{\lambda} \\ &= 4^{-1}\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}. \end{aligned}$$

Thus, by Lemma 3.1 and assumption E5, it follows from Theorem 1 part (c) of Denker and Keller (1983, p.507) that

$$n^{\frac{1}{2}}U_n \xrightarrow{D} \mathrm{N}(0,\boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}).$$

Finally, since $\boldsymbol{\lambda}$ is arbitrary, it follows that

$$\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}(\mathbf{0},\boldsymbol{\Sigma}),$$

which completes our proof. □

3.3 Proof of Theorem 2.3

To obtain the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$, we follow the approach for regression coefficients by minimization of dispersion functions proposed by Jaeckel (1972). To begin, let $\boldsymbol{\Delta} = n^{\frac{1}{2}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$ and $\tilde{\boldsymbol{\Delta}}_n = \min_{\boldsymbol{\Delta}} \mathrm{Q_n}(\boldsymbol{\Delta})$. Then, by minimization of $\mathrm{Q_n}(\boldsymbol{\Delta})$ and Theorem 2.2, we have that

$$\begin{aligned} \tilde{\boldsymbol{\Delta}}_n &= n^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0) \\ &= \frac{1}{2}\boldsymbol{C}^{-1}\,\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}\left(\mathbf{0},\frac{1}{4}\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}\right). \end{aligned}$$

Next, we let $\hat{\boldsymbol{\Delta}}_n = n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)$. Thus, by Theorem 2.1 and the Jaeckel (1972) convexity argument, we have that $\hat{\boldsymbol{\Delta}}_n - \tilde{\boldsymbol{\Delta}}_n = o_p(1)$. Finally, it follows by definitions that

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}_{\mathrm{m^2p}}\left(\mathbf{0},\frac{1}{4}\boldsymbol{C}^{-1}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}\right),$$

which completes our proof. □

4 Monte Carlo Study

4.1 The Process

In this section, we study the behavior of several estimates, in particular the WW-estimates, via Monte Carlo simulation. For the sake of simplicity and computation time only the $\mathrm{VAR}_2(1)$ is considered. Thus, we define the core process as

$$\begin{aligned} \boldsymbol{Y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t; \quad t = 1, \ldots, T, \\ \boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} (1-\rho)\, \mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) + \rho\, \mathrm{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_\rho), \end{aligned}$$

where $\rho \in [0, 1)$, and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$, $\boldsymbol{\Sigma}_\rho$ are both positive definite. In addition, we define the observed process as

$$\begin{aligned} \boldsymbol{Y}_t^* = \boldsymbol{Y}_t + \boldsymbol{Z}_t, \\ \boldsymbol{Z}_t \overset{\text{iid}}{\sim} (1-\gamma)\boldsymbol{\delta}_0 + \gamma\, \mathrm{N}_2(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma), \end{aligned}$$

where $\gamma \in [0, 1)$, $\boldsymbol{\delta}_0$ represents a bivariate point mass distribution at $\mathbf{0}$, $\boldsymbol{\mu}_\gamma \in \mathbb{R}^2$, and $\boldsymbol{\Sigma}_\gamma$ is positive definite.

Note that when $\gamma = 0$ and $\rho > 0$, the observed process reduces to the core process, representing Fox's (Fox, 1972) Type II or Innovation Outlier (IO) model. As discussed by Rousseeuw & Leroy (Rousseeuw & Leroy, 1987, p.275), this model produces *good* leverage points in the sense that they have relatively little impact on estimates. When $\rho = 0$ and $\gamma > 0$, the observed process yields Fox's Type I or Additive Outlier (AO) model. This model produces *bad* leverage points which can have a significant impact on most estimates. When $\gamma > 0$ and $\rho > 0$, the observed process corresponds to a combination of the IO and AO models, denoted I&AO for convenience. In the remaining case, when $\gamma = 0$ and $\rho = 0$, the observed process is bivariate normal.

4.2 The Estimators

For interpretation convenience, we group estimators according to their weighting schemes as: no weights, Mallows, and Schweppe labeled as 1-3, respectively. This Monte Carlo study will simulate, compute, and compare the performance of the following estimators

OLS(1) - OLS estimate based on the dispersion in (4) with $b_t \equiv 1$.

L1(1) - L_1-estimate based on the dispersion in (5) with $b_t \equiv 1$.

WIL(1) - Wilcoxon-estimate based on the dispersion in (6) with $b_{ij} \equiv 1$.

BL2(2) - Weighted L_2-estimate based on the dispersion in (4) with b_t defined in (7).

BL1(2) - Weighted L_1-estimate based on the dispersion in (5) with b_t defined in (7).

THL(2) - WW estimate based on the dispersion in (6) with b_t defined in (7).

ML2(2) - Weighted L_2-estimate based on the dispersion in (4) with b_t defined in (8).

ML1(2) - Weighted L_1-estimate based on the dispersion in (5) with b_t defined in (8).

GR(2) - WW-estimate based on the dispersion in (6) with b_{ij} defined in (8).

HBL2(3) - Weighted L_2-estimate based on the dispersion in (4) with b_t defined in (9).

HBL1(3) - Weighted L_1-estimate based on the dispersion in (5) with b_t defined in (9).

HBR(3) - WW-estimate based on the dispersion in (6) with b_{ij} defined in (9).

TMNL2(3) - Weighted L_2-estimate based on the dispersion in (4) with b_t defined in (10).

TMNL1(3) - Weighted L_1-estimate based on the dispersion in (5) with b_t defined in (10).

TMNR(3) - WW-estimate based on the dispersion in (6) with b_{ij} defined in (10).

For the estimators based on Schweppe weights, we obtained initial residuals using ML1-estimates. We used the Minimum Covariance Determinant (Rousseeuw & Leroy, 1987) to obtain the robust measures of center and covariance required by $d(\cdot)$.

4.3 The Simulation Settings

We selected the simulation settings to account for the impact on estimates due to the degree of stationarity of the time series, the level of contamination due to innovation outliers, the level of contamination due to additive outliers, and the magnitude of the additive outlier.

Stationarity. Stationarity plays a critical role in time series, thus the following three coefficient matrices are considered:

$$\mathbf{\Phi}_1 \in \left\{ \begin{bmatrix} 0.10 & 0.03 \\ 0.01 & 0.05 \end{bmatrix}, \begin{bmatrix} 0.30 & -0.20 \\ -0.10 & 0.40 \end{bmatrix}, \begin{bmatrix} 1.20 & -0.50 \\ 0.60 & 0.30 \end{bmatrix} \right\},$$

representing three degrees of stationarity. The processes associated to these $\mathbf{\Phi}_1$ matrices have roots with their largest modulus at 0.1, 0.5 and 0.8, respectively. Therefore, we denoted them as *very stationary*, *moderate stationary*, and *close to non-stationary process*, respectively. Note that the coefficient matrices satisfy the stationarity assumption in (2).

Contamination. First, we evaluate the sensitivity of the estimates to innovative outliers by considering $\rho \in \{0, 0.05, 0.1\}$. Next, we evaluate the sensitivity of the estimates to additive outliers by considering $\gamma \in \{0, 0.05, 0.1\}$. Finally, we evaluate the impact of the magnitude of the additive outlier by considering $\boldsymbol{\mu}_\gamma \in \{(10, 13)', (100, 130)'\}$; denoted as *close* and *far* from the core process, respectively.

The remaining parameters of the simulation are fixed. We generated 1000 replicates of size $T = 100$, and set

$$\boldsymbol{\phi}_0 = \mathbf{0}, \quad \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} = \mathbb{I}_2, \quad \mathbf{\Sigma}_\rho = 16\, \mathbb{I}_2 \quad , \text{ and } \quad \mathbf{\Sigma}_\gamma = \mathbb{I}_2.$$

The estimates were computed using R (R Core Team, 2013).

4.4 Simulation Results

We estimate the efficiencies of estimators based on the trace of the empirical MSE matrix. For comparison purpose, results have the actual trace of the empirical MSE matrix in the OLS entry, while the other estimators are presented relative to OLS so that entries larger than 1 indicate a more efficient estimator. The results are presented in terms of the estimators and the corresponding 1-step forecasts.

4.4.1 Efficiency of Estimators

Table 1 shows results for the *very stationary* setting. Under multivariate normality, the OLS estimates prove to be the most efficient followed by the Wilcoxon and the HBR estimates at 0.96 efficiency. Under the AO model, the OLS estimates deteriorate and are out-performed by all estimates based on robust dispersions. In particular, the Wilcoxon, Theil, and HBR estimates show the best all around performance. Also of note, the High Breakdown L2 is the only variation of the weighted L2 dispersion that out-performs the OLS estimates. Under the IO model, the OLS is resistant, yet the robust dispersions perform better. Unweighted dispersions perform best, followed by Swcheppe weighted estimates, and last are the Mallows weighted estimates. Under the I&AO

model, the AO effect is mitigated, yet the OLS estimates performance drops. Similar to the AO model, the estimates based on robust dispersions have an overall better performance when compared to the OLS estimates. In particular, the Wilcoxon, HBR, Theil, and L_1 estimates show the best overall performance.

Table 1: Efficiency of Estimators - Very Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	0.04	0.16	13.75	0.26	24.35	0.04	0.12	8.46	0.17	15.88	0.04	0.10	5.58	0.15	10.60
L1(1)	0.77	3.93	315.41	6.59	393.40	1.10	4.05	273.39	5.37	459.20	1.50	3.89	198.38	5.41	371.37
WIL(1)	0.96	4.44	389.17	7.21	639.61	1.34	4.32	314.56	5.50	487.99	1.66	4.05	214.39	5.29	383.11
BL2(2)	0.77	0.67	0.67	0.61	0.62	0.59	0.54	0.59	0.53	0.51	0.53	0.54	0.50	0.51	0.48
BL1(2)	0.59	2.58	222.27	4.51	330.09	0.66	2.30	1.35	3.11	26.41	0.79	1.66	124.82	3.15	137.93
THL(2)	0.92	4.15	357.66	6.79	557.66	1.13	3.82	279.41	4.89	404.27	1.32	3.59	198.78	4.70	325.98
ML2(2)	0.95	0.59	0.54	0.62	0.55	0.66	0.48	0.42	0.45	0.44	0.55	0.43	0.34	0.43	0.35
ML1(2)	0.74	2.70	207.49	4.37	314.19	0.79	2.19	143.16	2.84	217.71	0.88	1.96	100.53	2.79	161.62
GR(2)	0.92	3.19	240.40	5.14	338.41	0.96	2.54	160.49	3.16	233.54	1.00	2.17	106.81	2.88	160.18
HBL2(3)	0.98	3.04	211.38	3.77	199.10	1.03	2.46	143.90	2.46	147.17	1.07	2.21	98.11	2.26	105.27
HBL1(3)	0.75	2.92	218.33	5.19	363.65	0.86	2.62	163.57	3.72	280.79	1.02	2.50	126.22	3.93	231.41
HBR(3)	0.96	4.20	281.81	7.21	466.14	1.27	4.35	236.49	5.77	399.43	1.61	4.38	197.82	5.98	347.91
TMNL2(3)	0.98	0.86	0.62	0.91	0.61	0.92	0.99	0.69	0.93	0.69	0.99	1.09	0.77	1.04	0.74
TMNL1(3)	0.75	2.96	210.25	5.07	321.78	0.92	2.87	167.97	3.87	259.80	1.19	2.90	138.77	4.18	227.67
TMNR(3)	0.94	3.47	248.75	5.88	351.36	1.13	3.33	196.40	4.32	293.78	1.39	3.30	156.90	4.50	243.40

Table 2 and 3 shows results for the *moderate stationary* and *close to non-stationary* settings, respectively. Results remain similar to those under the *very stationary* setting. In particular, HBR estimates show the best all around performance under the *moderate stationary*. On the other hand, under the *close to non-stationary* setting, the HBR estimates have a blind spot when additive outliers that are *close* to the core process are present. Under the *close to non-stationary* setting with additive outliers that are *close* to the core process, the High Breakdown weighted L_1 has the best performance.

Table 2: Efficiency of Estimators - Moderate Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	0.04	0.19	10.27	0.28	18.61	0.04	0.14	6.57	0.21	11.64	0.04	0.13	4.31	0.18	8.33
L1(1)	0.80	2.18	107.42	3.03	191.64	1.17	1.98	75.85	2.58	125.56	1.60	1.94	51.58	2.35	97.44
WIL(1)	0.98	2.30	114.21	3.08	199.83	1.39	2.04	77.98	2.59	129.29	1.76	1.96	52.45	2.32	97.88
BL2(2)	0.78	0.82	0.64	0.72	0.64	0.61	0.75	0.61	0.64	0.52	0.60	0.78	0.53	0.72	0.54
BL1(2)	0.58	1.72	5.08	2.82	200.17	0.72	2.53	101.62	2.68	126.91	0.87	2.52	44.04	2.66	2.67
THL(2)	0.94	3.29	145.81	3.82	209.65	1.21	3.02	104.61	3.18	136.90	1.48	2.97	77.31	2.98	106.26
ML2(2)	0.97	0.80	0.51	0.74	0.55	0.72	0.71	0.42	0.60	0.43	0.66	0.68	0.35	0.66	0.41
ML1(2)	0.77	3.27	173.52	4.44	277.50	0.88	2.94	127.26	3.65	183.05	1.03	2.87	97.97	3.56	157.78
GR(2)	0.94	3.82	202.23	4.87	296.78	1.05	3.30	141.11	3.75	190.60	1.18	3.06	102.94	3.52	158.51
HBL2(3)	0.99	3.49	176.42	3.44	183.99	1.11	3.02	126.08	2.68	121.07	1.24	2.80	92.33	2.60	102.28
HBL1(3)	0.78	3.38	187.06	4.62	322.73	0.94	3.20	146.24	4.19	234.44	1.19	3.40	122.82	4.09	219.09
HBR(3)	0.97	3.68	235.94	4.28	395.39	1.34	3.59	198.77	3.74	319.73	1.74	3.61	178.06	3.46	285.92
TMNL2(3)	0.99	0.96	0.57	0.91	0.61	0.98	1.11	0.66	0.94	0.63	1.10	1.26	0.68	1.12	0.73
TMNL1(3)	0.78	2.96	175.74	3.89	285.44	1.01	2.81	148.00	3.32	216.15	1.35	2.91	130.58	3.22	212.13
TMNR(3)	0.96	3.41	206.48	4.27	308.45	1.23	3.23	168.63	3.58	236.78	1.57	3.29	144.71	3.47	223.95

Table 3: Efficiency of Estimators - Close to Non-Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	0.01	0.32	6.32	0.52	11.48	0.01	0.22	4.80	0.38	7.80	0.01	0.16	3.34	0.28	5.54
L1(1)	0.77	3.67	8.02	2.03	13.70	1.16	4.91	6.11	2.42	9.53	1.58	5.80	4.40	3.12	6.83
WIL(1)	0.96	2.12	8.02	1.48	14.71	1.36	2.48	6.03	1.50	9.63	1.73	2.92	4.31	1.67	6.88
BL2(2)	0.77	1.92	0.75	1.52	0.67	0.68	1.80	0.69	1.48	0.64	0.67	1.53	0.67	1.42	0.60
BL1(2)	0.57	10.37	196.09	11.16	243.38	0.78	7.75	171.46	9.61	3.09	1.00	8.28	172.94	8.54	174.57
THL(2)	0.91	10.93	199.32	6.74	137.39	1.27	9.69	193.47	6.11	120.55	1.62	8.44	171.89	5.79	92.10
ML2(2)	0.95	2.49	0.84	1.80	0.75	0.84	1.97	0.81	1.56	0.71	0.83	1.62	0.76	1.41	0.69
ML1(2)	0.73	14.45	293.41	15.12	424.57	0.98	11.95	298.08	12.35	379.81	1.27	10.53	255.57	10.40	352.86
GR(2)	0.91	14.64	335.56	10.80	464.84	1.17	10.96	324.43	7.16	380.53	1.46	8.70	265.76	5.44	342.76
HBL2(3)	0.98	11.77	298.47	6.88	288.07	1.24	8.83	272.92	5.03	238.50	1.47	6.58	219.36	3.81	198.46
HBL1(3)	0.75	16.56	314.24	20.21	507.43	1.07	14.98	338.87	18.56	482.83	1.44	13.55	295.97	14.14	454.68
HBR(3)	0.96	12.11	386.91	5.13	622.49	1.38	10.45	424.09	3.89	594.39	1.80	8.80	369.76	3.49	535.52
TMNL2(3)	0.98	2.76	0.98	1.94	0.83	1.05	2.61	1.17	1.94	0.95	1.15	2.25	1.15	1.81	1.02
TMNL1(3)	0.75	14.90	300.29	15.87	437.88	1.09	13.76	325.26	14.61	425.18	1.50	12.44	290.34	12.75	409.07
TMNR(3)	0.94	15.40	346.02	11.52	480.29	1.31	13.37	369.74	9.02	439.75	1.73	11.19	316.56	7.36	418.20

Simulation results suggest that Schweppe weighted estimators can achieve comparable efficiencies across different degrees of stationarity and the presence of outliers. Additive outliers seem to have the largest effect on estimators and are best handled by Schweppe weighted robust dispersions.

4.4.2 Efficiency of Forecasts

Tables 4-6 shows results for the forecast performance across the simulation settings. Results indicate that there is little carry over from the efficiency of estimators. In general, estimates based on robust dispersions out-perform the OLS estimates. Forecast efficiencies are most sensitive to the presence of additive outliers. Under the *very stationary* setting, all estimates based on robust dispersions have similar performance. Under the *moderate stationary* setting, Schweppe weighted robust dispersions have best performance, in particular the HBR estimates. Under the *close to non-stationary* setting, the weighted Wilcoxon dispersions have best performance when the outlier is *far* from the core process. On the other hand, the weighted L_1 dispersions have best performance when the outlier is *close* from the core process.

Table 4: Efficiency of Forecasts - Very Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	1.96	2.19	17.40	2.20	24.32	3.36	3.55	25.20	3.48	43.94	4.69	5.10	22.28	5.36	26.07
L1(1)	1.00	1.07	8.56	1.08	12.03	1.01	1.04	7.75	1.06	11.04	1.02	1.06	5.36	1.07	4.40
WIL(1)	1.00	1.08	8.62	1.09	12.06	1.01	1.04	7.77	1.06	11.07	1.02	1.06	5.35	1.06	4.40
BL2(2)	0.99	0.96	0.74	0.95	0.63	0.99	0.96	0.72	0.89	0.53	0.98	0.97	0.64	0.91	0.47
BL1(2)	0.99	1.06	8.49	1.06	11.90	1.00	1.03	4.99	1.03	9.00	1.01	1.04	5.28	1.06	4.35
THL(2)	1.00	1.07	8.61	1.08	11.98	1.01	1.04	7.74	1.05	11.00	1.01	1.06	5.35	1.07	4.40
ML2(2)	1.00	0.94	0.60	0.92	0.55	0.99	0.94	0.63	0.88	0.47	0.97	0.96	0.42	0.89	0.35
ML1(2)	1.00	1.06	8.53	1.06	11.82	1.01	1.03	7.64	1.03	10.89	1.00	1.05	5.28	1.06	4.32
GR(2)	1.00	1.06	8.58	1.07	11.81	1.01	1.03	7.62	1.03	10.92	1.00	1.05	5.30	1.06	4.32
HBL2(3)	1.00	1.06	8.57	1.07	11.52	1.00	1.03	7.56	1.02	10.64	1.00	1.05	5.27	1.04	4.28
HBL1(3)	1.00	1.06	8.52	1.06	11.91	1.01	1.03	7.69	1.03	10.98	1.00	1.05	5.32	1.07	4.35
HBR(3)	1.00	1.07	8.60	1.08	11.96	1.01	1.04	7.69	1.05	11.11	1.02	1.07	5.39	1.07	4.37
TMNL2(3)	1.00	0.98	0.68	0.98	0.60	1.00	0.99	0.98	0.97	0.80	1.00	1.03	0.91	1.00	0.72
TMNL1(3)	1.00	1.06	8.53	1.07	11.81	1.00	1.03	7.65	1.04	10.99	1.01	1.06	5.36	1.06	4.37
TMNR(3)	1.00	1.07	8.60	1.07	11.80	1.01	1.03	7.64	1.04	11.06	1.01	1.06	5.35	1.06	4.37

Table 5: Efficiency of Forecasts - Moderate Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	2.14	2.19	17.81	2.43	25.29	3.73	4.23	16.53	4.09	25.13	4.84	6.09	25.23	5.24	40.86
L1(1)	1.00	1.05	8.07	1.11	11.91	1.00	1.05	5.10	1.06	6.48	1.01	1.04	4.55	1.06	7.52
WIL(1)	1.00	1.05	8.12	1.11	11.95	1.00	1.05	5.11	1.06	6.50	1.01	1.04	4.54	1.06	7.53
BL2(2)	1.00	0.95	0.71	0.94	0.58	0.99	0.98	0.61	0.95	0.49	0.98	0.98	0.71	0.97	0.67
BL1(2)	0.99	1.03	5.88	1.12	11.82	1.00	1.06	5.21	1.06	6.53	1.00	1.04	4.59	1.06	4.51
THL(2)	1.00	1.06	8.19	1.13	11.91	1.00	1.06	5.16	1.07	6.51	1.01	1.04	4.62	1.07	7.55
ML2(2)	1.00	0.97	0.65	0.96	0.56	0.99	0.97	0.46	0.96	0.42	0.99	0.95	0.56	0.98	0.49
ML1(2)	0.99	1.06	8.15	1.13	12.02	1.00	1.06	5.28	1.07	6.55	1.00	1.04	4.68	1.07	7.77
GR(2)	1.00	1.06	8.19	1.13	12.04	1.00	1.06	5.27	1.07	6.55	1.01	1.04	4.67	1.07	7.73
HBL2(3)	1.00	1.05	8.13	1.12	11.74	1.00	1.06	5.23	1.06	6.42	1.01	1.03	4.64	1.07	7.62
HBL1(3)	1.00	1.05	8.24	1.13	12.09	1.00	1.06	5.30	1.07	6.59	1.00	1.04	4.73	1.07	7.84
HBR(3)	1.00	1.06	8.27	1.13	12.17	1.00	1.07	5.31	1.07	6.62	1.01	1.05	4.71	1.07	7.88
TMNL2(3)	1.00	0.98	0.72	0.99	0.62	1.00	1.01	0.71	1.00	0.66	1.01	1.00	1.01	1.02	0.89
TMNL1(3)	1.00	1.05	8.18	1.12	12.01	1.00	1.06	5.29	1.06	6.56	1.01	1.04	4.72	1.07	7.84
TMNR(3)	1.00	1.05	8.21	1.12	12.05	1.00	1.06	5.29	1.07	6.59	1.01	1.05	4.69	1.07	7.83

Table 6: Efficiency of Forecasts - Close to Non-Stationary $\mathbf{\Phi}_1$ matrix

ρ	0					0.05					0.1				
γ	0	0.05		0.1		0	0.05		0.1		0	0.05		0.1	
$\boldsymbol{\mu}_\gamma$	-	Close	Far	Close	Far	-	Close	Far	Close	Far	-	Close	Far	Close	Far
OLS(1)	2.13	3.55	19.27	4.73	32.47	3.54	5.39	26.86	6.53	54.54	4.97	6.40	30.02	8.06	41.75
L1(1)	1.00	1.55	2.89	1.39	4.82	1.00	1.47	2.38	1.48	4.56	1.00	1.32	2.06	1.43	2.45
WIL(1)	1.00	1.34	2.86	1.19	4.90	1.01	1.30	2.35	1.21	4.66	1.00	1.23	2.03	1.22	2.45
BL2(2)	0.99	1.34	0.86	1.29	0.73	0.99	1.26	0.81	1.23	0.70	0.98	1.12	0.83	1.16	0.76
BL1(2)	0.99	1.74	9.34	2.03	15.02	1.00	1.56	7.50	1.91	6.20	0.99	1.34	5.95	1.63	8.57
THL(2)	1.00	1.73	9.10	1.84	12.86	1.01	1.55	7.40	1.75	13.72	1.00	1.34	5.92	1.54	7.61
ML2(2)	1.00	1.47	0.82	1.45	0.70	1.00	1.30	0.76	1.29	0.78	0.99	1.15	0.86	1.16	0.77
ML1(2)	0.99	1.76	9.60	2.09	15.69	1.00	1.58	7.68	1.95	16.57	1.00	1.36	6.06	1.65	8.82
GR(2)	1.00	1.77	9.66	2.03	15.85	1.01	1.57	7.67	1.83	16.51	1.00	1.35	6.06	1.55	8.85
HBL2(3)	1.00	1.75	9.63	1.94	15.47	1.01	1.55	7.61	1.73	16.07	1.00	1.33	6.03	1.47	8.71
HBL1(3)	1.00	1.77	9.58	2.12	15.71	1.00	1.59	7.72	1.99	16.59	1.00	1.36	6.09	1.67	8.84
HBR(3)	1.00	1.75	9.69	1.81	15.94	1.01	1.57	7.71	1.66	16.69	1.00	1.34	6.11	1.46	8.89
TMNL2(3)	1.00	1.50	0.95	1.48	0.77	1.00	1.38	1.07	1.39	0.98	0.99	1.21	1.17	1.25	1.03
TMNL1(3)	1.00	1.76	9.59	2.10	15.65	1.00	1.59	7.69	1.96	16.56	1.00	1.36	6.08	1.66	8.83
TMNR(3)	1.00	1.76	9.67	2.04	15.84	1.01	1.58	7.69	1.87	16.55	1.00	1.36	6.08	1.59	8.86

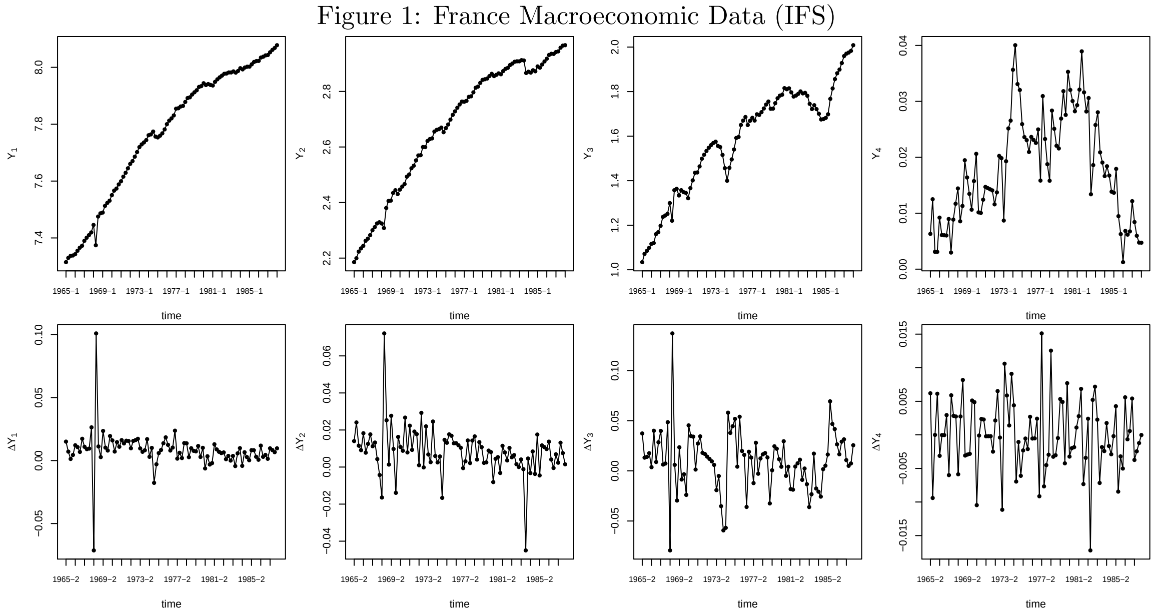
Simulation results suggest that, even in the presence of additive outliers, Schweppe weighted Wilcoxon estimates can achieve comparable efficiencies on forecasts as well.

5 Time Series Example

5.1 Data Description

We present an illustrative data example in order to demonstrate how the estimation method can affect the estimates and residual analysis. We use the quadravariate dataset made popular by Ooms (1994) for estimating $\text{VAR}_m(p)$ models under the presence of outliers for this illustration. The dataset was originally extracted from International Financial Statistics (IFS) and contains quarterly macroeconomic data for France that starts at the first quarter of 1965 and ends at the first quarter of 1988. The multivariate measurement ($\boldsymbol{Y}$) includes the log GDP in 1980 prices (Y_1), the log private consumption deflated by consumer price index (Y_2), the log gross fixed capital formation deflated by the GDP deflator (Y_3), and the first difference of the log consumer price index (Y_4). Figure 1 shows the time series realization on the top row and the differenced series on the bottom row.

Figure 1: France Macroeconomic Data (IFS)



Following Ooms' recommendation, we fitted first order differences of the time series to a $\text{VAR}_4(2)$ model with no intercept. We evaluate all estimators described in 4.2 and preserve the last four realizations to evaluate their forecast performance.

5.2 Estimation Results

Outlier Detection. We used a visual approach to evaluate the estimators ability to detect outliers. The approach involves a time series plot of $\text{d}(\hat{\boldsymbol{\varepsilon}}_t)$ that has special markers to identify leverage points.

Leverage points satisfy $\mathrm{d}^2(\boldsymbol{X}_{t-1}) > \chi^2_{1\text{-}\alpha}(mp)$ and are identified with open circles. A horizontal critical line is drawn at $\sqrt{\chi^2_{1\text{-}\alpha}(m)}$, with $\alpha = 0.05$, to separate outlying realizations. Thus, open circles above the critical line indicate *bad* leverage points. A more detailed description of this type of plot can be found in Terpstra *et al.* (2001).

We present the plots, across weights and for each dispersion function, in Figures 2-4. The weighted estimates identified the same set of outliers with more consistency. Focusing attention on the outliers identified as *bad* leverage points, realizations 1968-2 through 1969-1 are identified by weighted estimates while the least squares estimate and L_1-estimate did not detect the 1968-4 realization. The Wilcoxon estimate only detected realizations 1968-2 and 1968-3. These results, coupled with the insights from Section 4, suggest that the IFS series may contain additive outliers since weighted estimates offer better protection against fitting *bad* leverage points produced under this type of contamination.

Figure 2: Least Squares Residual Plots

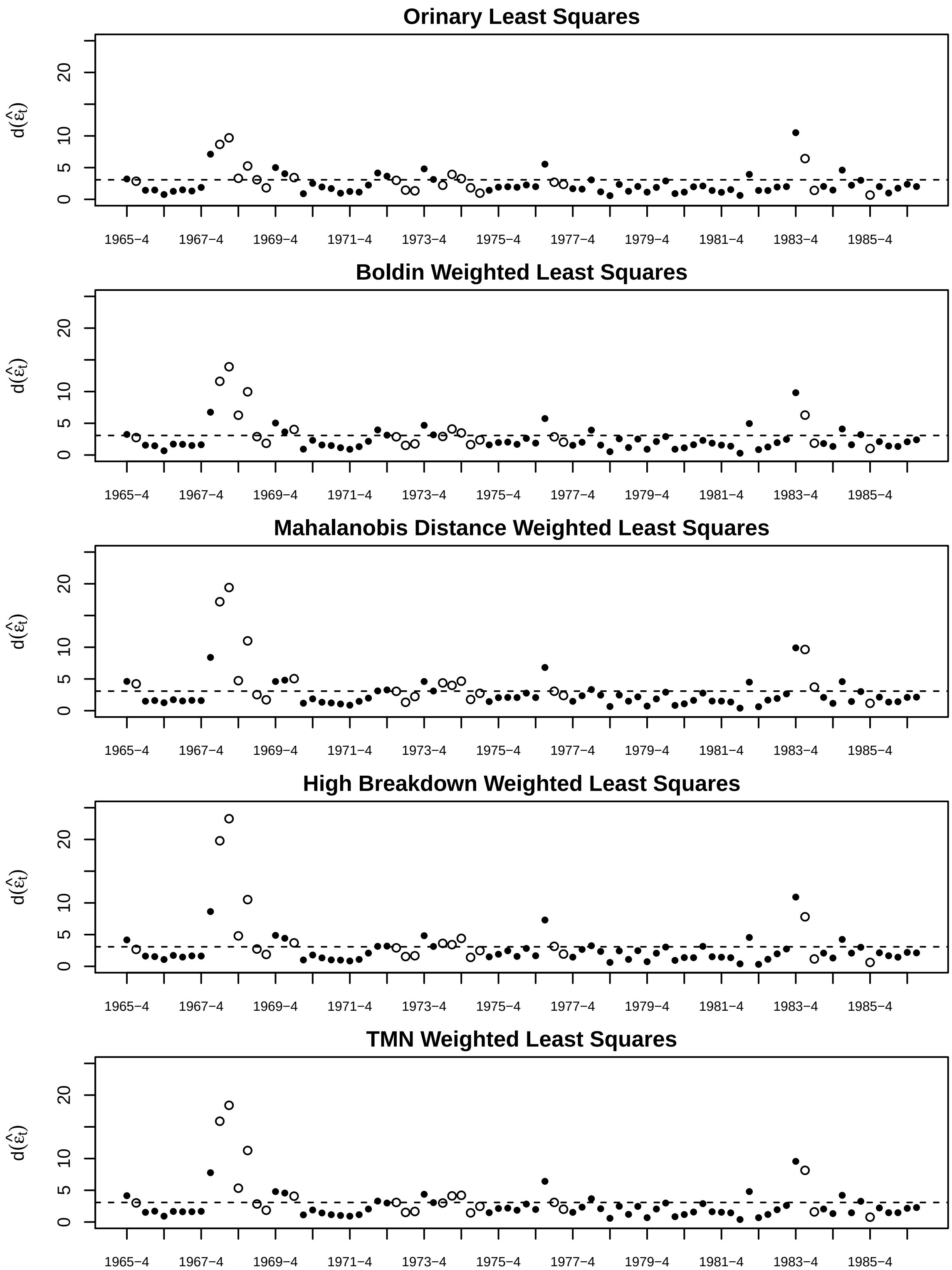


Figure 3: L1 Residual Plots

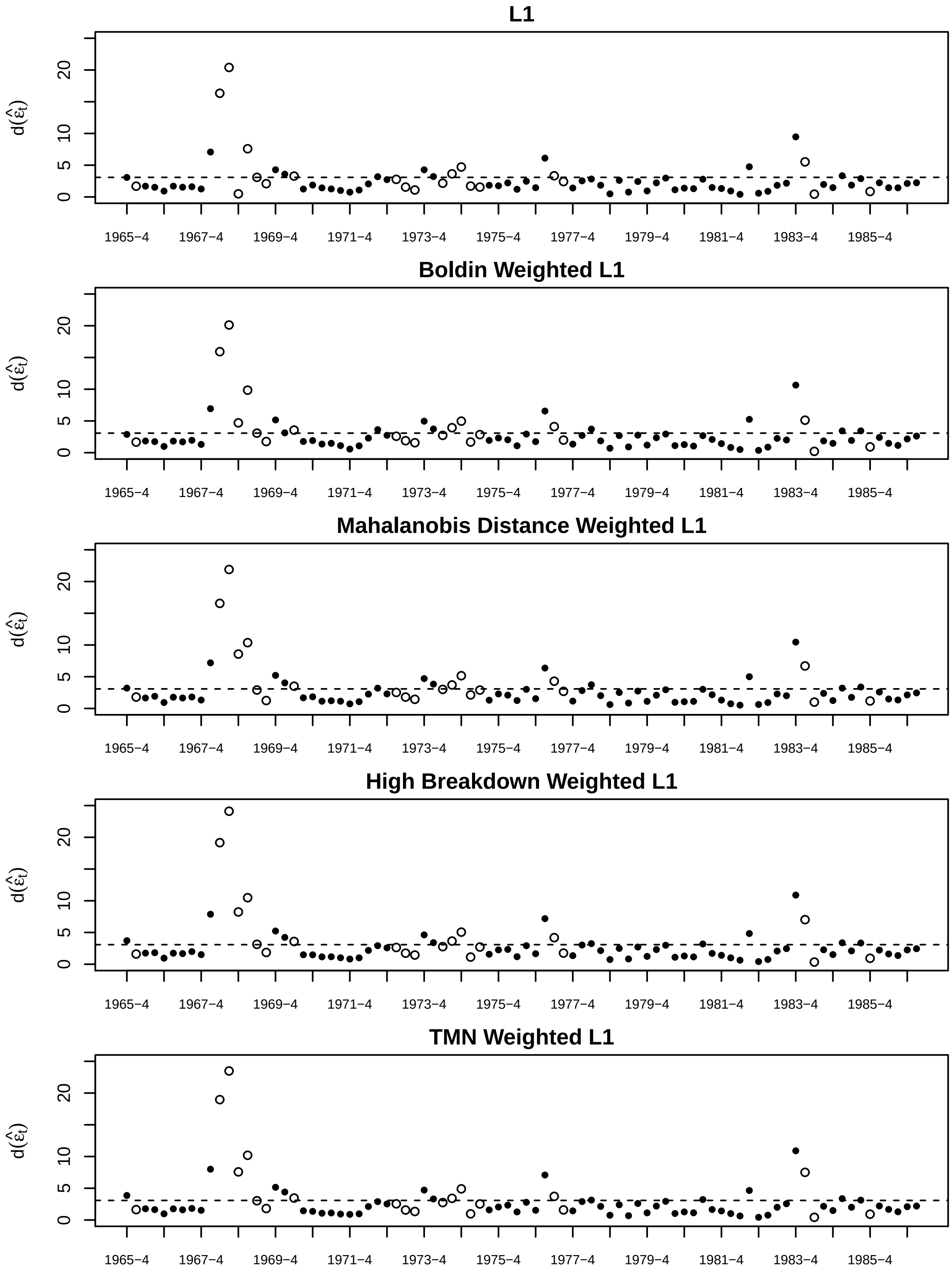
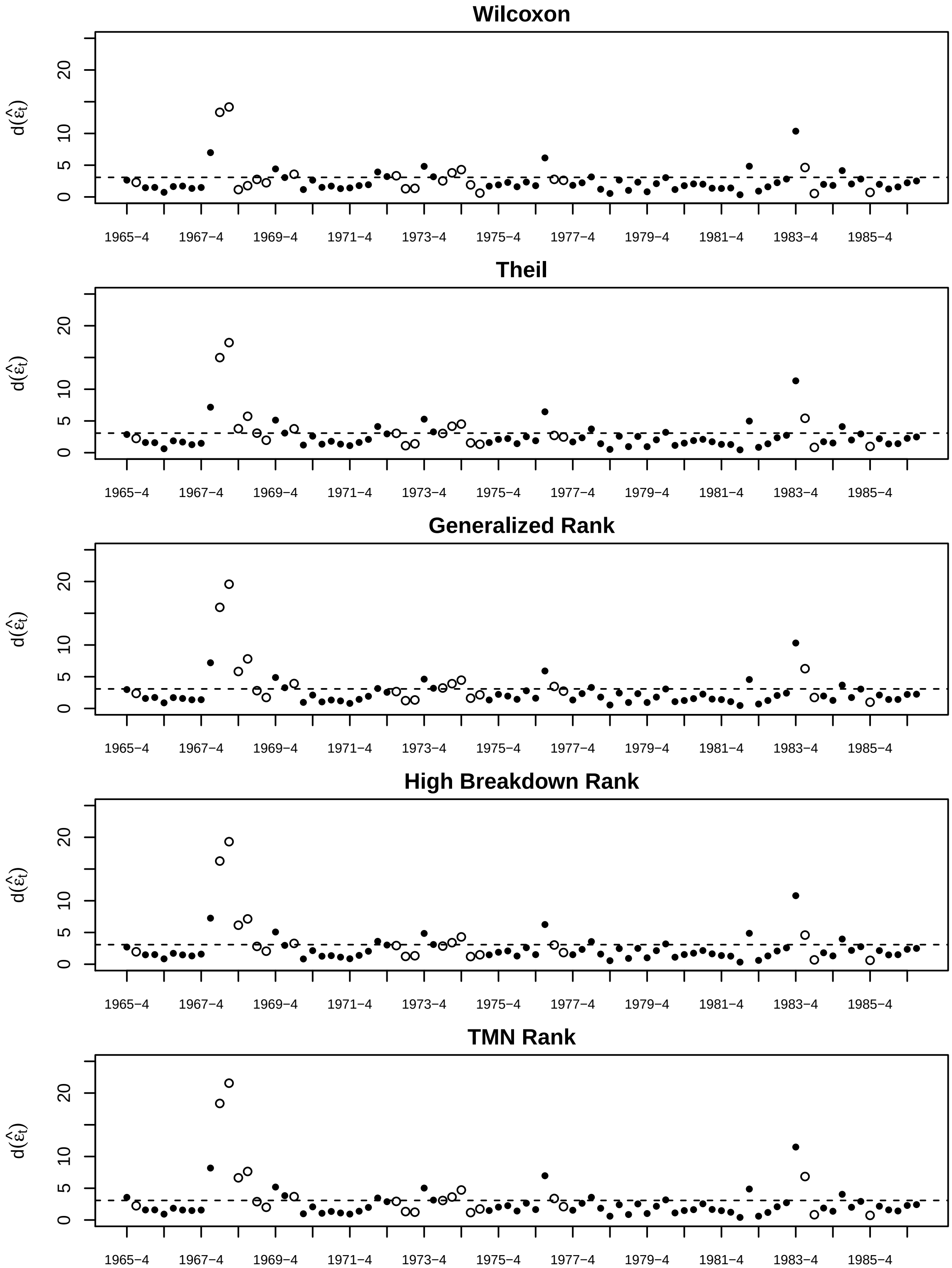


Figure 4: Rank Norm Residual Plots



Parameter Estimates. We present the estimated $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ matrices, across weights and for each dispersion function, in Tables 7 - 9. Inspection of these matrices indicate that the sign and magnitude of the estimates are influenced by the presence of outliers. All estimates yielded results that satisfy (2).

Table 7: Least Squares Estimates for IFS Series

Estimate	$\mathbf{\Phi}_1$				$\mathbf{\Phi}_2$			
OLS(1)	−0.519	0.707	−0.088	−0.272	0.002	0.349	−0.040	−0.195
	−0.177	0.304	0.038	−0.003	−0.050	0.226	0.095	0.033
	−1.417	1.024	0.359	−0.236	0.021	0.331	−0.063	0.066
	0.072	0.052	−0.059	−0.372	−0.043	0.015	0.011	−0.297
BL2(2)	−0.250	0.418	−0.074	−0.038	0.198	0.374	−0.022	−0.001
	0.044	0.051	0.049	0.125	0.172	0.250	0.092	0.085
	−0.977	0.619	0.343	−0.062	0.235	0.364	0.065	0.285
	0.084	0.032	−0.056	−0.346	−0.016	−0.008	0.015	−0.284
ML2(2)	−0.108	0.384	−0.047	0.065	0.164	0.383	−0.005	0.054
	0.273	−0.077	0.054	0.149	0.145	0.218	0.139	0.188
	−0.802	0.399	0.353	0.034	0.293	0.546	0.021	0.347
	0.065	0.022	−0.048	−0.294	0.066	−0.074	0.019	−0.291
HBL2(3)	0.148	0.242	−0.038	0.068	0.183	0.245	0.035	0.115
	0.328	−0.107	0.056	0.133	0.180	0.180	0.147	0.194
	−0.608	0.287	0.416	0.192	0.546	0.175	0.078	0.435
	0.075	0.025	−0.055	−0.343	−0.008	−0.002	0.009	−0.303
TMNL2(3)	−0.068	0.366	−0.042	0.029	0.216	0.281	0.015	0.076
	0.241	−0.065	0.062	0.170	0.189	0.182	0.149	0.186
	−0.945	0.440	0.416	0.111	0.534	0.294	0.054	0.321
	0.070	0.040	−0.057	−0.320	−0.013	−0.007	0.010	−0.285

Table 8: L1 Estimates for IFS Series

Estimate	$\mathbf{\Phi}_1$				$\mathbf{\Phi}_2$			
L1(1)	0.137	0.206	−0.002	−0.013	0.182	0.171	0.015	0.040
	0.336	−0.089	0.096	−0.039	0.061	0.153	0.097	0.118
	−0.712	0.503	0.462	−0.062	0.193	0.211	0.051	0.069
	0.093	0.041	−0.057	−0.315	−0.048	0.033	0.006	−0.378
BL1(2)	0.224	0.123	0.004	0.008	0.275	0.188	−0.004	0.043
	0.342	−0.180	0.106	0.011	0.283	0.178	0.050	0.057
	−0.795	0.437	0.435	0.118	0.559	0.204	0.072	0.348
	0.033	0.063	−0.048	−0.312	−0.055	0.014	0.017	−0.386
ML1(2)	0.220	0.181	0.008	0.023	0.199	0.232	0.007	0.079
	0.423	−0.230	0.103	−0.062	0.124	0.267	0.078	0.078
	−0.702	0.234	0.352	0.124	0.744	0.188	0.087	0.270
	−0.047	0.069	−0.039	−0.287	0.057	−0.044	0.020	−0.370
HBL1(3)	0.250	0.160	0.001	0.026	0.201	0.189	0.026	0.103
	0.352	−0.230	0.108	0.010	0.259	0.178	0.096	0.042
	−0.934	0.297	0.459	0.153	0.869	0.134	0.090	0.364
	0.036	0.081	−0.058	−0.308	−0.067	0.023	0.004	−0.373
TMNL1(3)	0.218	0.178	0.004	0.032	0.223	0.188	0.020	0.097
	0.365	−0.251	0.112	0.001	0.293	0.156	0.096	0.007
	−1.024	0.325	0.463	0.153	0.828	0.205	0.085	0.324
	0.035	0.094	−0.060	−0.302	−0.076	0.024	0.003	−0.349

Table 9: Wilcoxon Estimates for IFS Series

Estimate	$\mathbf{\Phi}_1$				$\mathbf{\Phi}_2$			
WIL(1)	−0.164	0.204	−0.052	0.007	−0.149	0.213	−0.004	0.060
	0.005	−0.053	0.051	0.083	−0.251	0.180	0.099	0.152
	−1.268	0.507	0.393	0.207	−0.373	0.195	0.022	0.447
	0.145	0.029	−0.060	−0.385	−0.013	0.014	0.013	−0.321
THL(2)	−0.026	0.161	−0.038	0.025	0.013	0.207	0.003	0.096
	0.094	−0.130	0.067	0.041	−0.046	0.181	0.092	0.111
	−0.937	0.363	0.368	0.323	−0.097	0.263	0.099	0.475
	0.144	0.019	−0.053	−0.401	0.027	−0.018	0.016	−0.327
GR(2)	0.128	0.125	−0.026	0.064	0.075	0.224	0.014	0.081
	0.343	−0.284	0.060	−0.004	0.008	0.167	0.124	0.128
	−0.689	0.173	0.371	0.085	0.301	0.279	0.081	0.379
	0.063	0.032	−0.046	−0.291	0.103	−0.073	0.018	−0.328
HBR(3)	0.133	0.097	−0.028	0.056	0.025	0.203	0.024	0.090
	0.160	−0.164	0.064	0.101	0.016	0.171	0.120	0.143
	−0.947	0.311	0.427	0.257	0.319	0.121	0.072	0.444
	0.142	0.029	−0.058	−0.379	0.007	0.014	0.014	−0.332
TMNR(3)	0.122	0.137	−0.020	0.045	0.062	0.207	0.019	0.108
	0.258	−0.258	0.070	0.042	0.031	0.164	0.121	0.127
	−0.963	0.243	0.438	0.194	0.391	0.169	0.087	0.385
	0.090	0.062	−0.056	−0.333	0.017	0.007	0.008	−0.328

Forecast Performance. We compared forecasting performance of the estimators, by adding the squared forecast errors of the four outcomes into a single statistic, at each forecast step. We refer to this statistic as the Total Squared Error (TSE). For comparison purpose, we present the actual TSE in the OLS entry, while the other estimates are presented relative to OLS, such that entries larger than 1 indicate a more efficient estimate. We present these results in Table 10. Results suggest that ML1 and ML2 have an overall better performance for the IFS series.

Table 10: Forecast Performance for IFS Series

Estimator	1-step	2-step	3-step	4-step
OLS(1)	1.424	1.357	0.269	5.742
L1(1)	1.11	1.02	0.80	1.46
WIL(1)	1.05	0.82	0.36	0.94
BL2(2)	1.03	0.70	0.73	1.39
BL1(2)	0.85	0.69	0.71	1.50
THL(2)	1.05	0.73	0.39	1.14
ML2(2)	1.37	1.19	1.62	1.40
ML1(2)	1.27	1.34	1.11	1.34
GR(2)	1.33	0.96	0.46	1.12
HBL2(3)	0.89	0.98	1.84	1.52
HBL1(3)	0.71	0.82	1.40	1.64
HBR(3)	0.90	0.81	0.50	1.13
TMNL2(3)	0.90	0.97	1.90	1.52
TMNL1(3)	0.73	0.81	1.39	1.66
TMNR(3)	0.94	0.83	0.53	1.18

References

Ben, M. G., Martinez, E. J., & Yohai, V. J. (1999). Robust estimation in vector autoregressive moving-average models. *J. Time Ser. Anal.*, *20*(4), 381–399.

Boldin, M. V. (1994). On median estimates and tests in autoregressive models. *Math. Methods Statist.*, *3*(2), 114–129.

Brockwell, P. J. & Davis, R. A. (2002). *Introduction to time series and forecasting* (Second ed.). New York: Springer-Verlag.

Bustos, O. H. & Yohai, V. J. (1986). Robust estimates for ARMA models. *J. Amer. Statist. Assoc.*, *81*(393), 155–168.

Chang, W. H., McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1999). High-breakdown rank regression. *J. Amer. Statist. Assoc.*, *94*(445), 205–219.

Croux, C. & Joossens, K. (2008). Robust estimation of the vector autoregressive model by a least trimmed squares procedure. In P. Brito (Ed.), *COMPSTAT 2008* (pp. 489–501). Physica-Verlag HD.

Denker, M. & Keller, G. (1983). On u-statistics and v. mise statistics for weakly dependent processes. *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *64*(4), 505–522.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society B*, *34*(3), 350–363.

Fuller, W. A. (1996). *Introduction to statistical time series* (Second ed.). Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons, Inc.

Handschin, E., Schweppe, F., Kohlas, J., & Fiechter, A. (1975). Bad data analysis for power system state estimation. *Power Apparatus and Systems, IEEE Transactions on*, *94*(2), 329–337.

Heathcote, C. R. & Welsh, A. H. (1988). Multivariate functional least squares. *J. Multivariate Anal.*, *25*(1), 45–64.

Heiler, S. & Willers, R. (1988). Asymptotic normality of R-estimates in the linear model. *Statistics*, *19*(2), 173–184.

Hettmansperger, T. P. & McKean, J. W. (2011). *Robust nonparametric statistical methods* (Second ed.)., volume 119 of *Monographs on Statistics and Applied Probability*. Boca Raton, FL: CRC Press.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, *43*, 1449–1458.

Krengel, U. (1985). *Ergodic theorems*, volume 6 of *de Gruyter Studies in Mathematics*. Berlin: Walter de Gruyter & Co.

Li, W. K. & Hui, Y. V. (1989). Robust multiple time series modelling. *Biometrika*, *76*(2), 309–315.

Lütkepohl, H. (1993). *Introduction to multiple time series analysis* (Second ed.). Berlin: Springer-Verlag.

Mallows, C. L. (1975). "On Some Topics in Robustness". Unpublished Memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Muler, N. & Yohai, V. J. (2013). Robust estimation for vector autoregressive models. *Computational Statistics & Data Analysis*, *65*(0), 68 – 79. Special issue on Robust Analysis of Complex Data.

Ooms, M. (1994). *Empirical vector autoregressive modeling*, volume 407 of *Lecture Notes in Economics and Mathematical Systems*. Berlin: Springer-Verlag.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reber, J. C., Terpstra, J. T., & Chen, X. (2008). Weighted L_1-estimates for a VAR(p) time series model. *J. Nonparametr. Stat.*, *20*(5), 395–411.

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Terpstra, J. T., McKean, J. W., & Naranjo, J. D. (2001). Weighted Wilcoxon estimates for autoregression. *Aust. N. Z. J. Stat.*, *43*(4), 399–419.

Terpstra, J. T. & Rao, M. B. (2002). On the asymptotic distribution of a multivariate GR-estimate for a VAR(p) time series. *Statist. Probab. Lett.*, *60*(2), 219–230.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. III. *Nederl. Akad. Wetensch., Proc.*, *53*, 1397–1412 = Indagationes Math. 12, 467–482 (1950).

A Appendix

A.1 Lemmas

Lemma A.1. *Let* $\boldsymbol{S}_{\mathrm{n}}(\cdot)$ *and* $\hat{\boldsymbol{S}}_{\mathrm{n}}(\cdot)$ *denote the gradient evaluated with* $\mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}_0)$ *and* $\mathrm{b}_{\mathrm{ij}}(\hat{\boldsymbol{\theta}})$ *as defined in Section* 3*. Then, under model assumptions* (1)-(3)*,* W1-W3*, and* E1*, we have that*

$$\|[\hat{\boldsymbol{S}}_{\mathrm{n}}(\boldsymbol{\Delta}) - \hat{\boldsymbol{S}}_{\mathrm{n}}(\mathbf{0})] - [\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - \boldsymbol{S}_{\mathrm{n}}(\mathbf{0})]\| = o_p(1).$$

Proof. To begin, on the left hand side of the proposition, we evaluate the expression inside the norm to obtain,

$$\begin{aligned}
&[\hat{\boldsymbol{S}}_{\mathrm{n}}(\boldsymbol{\Delta}) - \hat{\boldsymbol{S}}_{\mathrm{n}}(\mathbf{0})] - [\boldsymbol{S}_{\mathrm{n}}(\boldsymbol{\Delta}) - \boldsymbol{S}_{\mathrm{n}}(\mathbf{0})] \\
&= n^{-\frac{3}{2}} \sum_{i<j}^{n} (\mathrm{b}_{\mathrm{ij}}(\hat{\boldsymbol{\theta}}) - \mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}_0))(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i - \boldsymbol{d}_{ijn}) - \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)) \\
&\stackrel{\mathrm{def}}{=} n^{-\frac{3}{2}} \sum_{i<j}^{n} (\mathrm{b}_{\mathrm{ij}}(\hat{\boldsymbol{\theta}}) - \mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}_0))(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})),
\end{aligned}$$

where $\boldsymbol{d}_{ijn} = (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})' n^{-\frac{1}{2}} \boldsymbol{\Delta}$ and $\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{x}) = \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i - \boldsymbol{x})$. Next, by assumption W2 and the Mean Value Theorem, we have that

$$\begin{aligned}
&n^{-\frac{3}{2}} \sum_{i<j}^{n} [\mathrm{b}_{\mathrm{ij}}(\hat{\boldsymbol{\theta}}) - \mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}_0)](\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})) \\
&= n^{-2} \sum_{i<j}^{n} (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\mathbf{0}))\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)' n^{\frac{1}{2}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\stackrel{\mathrm{def}}{=} \boldsymbol{V}_n(n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)),
\end{aligned}$$

where $\boldsymbol{\theta}^* \in \{\lambda\hat{\boldsymbol{\theta}}+(1-\lambda)\boldsymbol{\theta}_0 : \lambda \in [0,1]\}$. Given assumption W1, it suffices to prove that $\|\boldsymbol{V}_n\| = o_p(1)$. By Theorems A.1 and A.2, and under assumption W3, it follows that

$$\begin{aligned}
\|\boldsymbol{V}_n\| &\leq n^{-2}\sum_{i<j}^{n}\|\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1}\|\|\,\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn})-\boldsymbol{u}_{\mathrm{ij}}(\mathbf{0})\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)\| \\
&\leq 2n^{-2}\sum_{i<j}^{n}\|\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1}\|\|(\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1})'\boldsymbol{\Delta}n^{-\frac{1}{2}}\|\|\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i\|^{-1}\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)\| \\
&\leq (2\|\boldsymbol{\Delta}\|B_D n^{-\frac{1}{2}})n^{-2}\sum_{i<j}^{n}\|\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1}\|^2\|\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i\|^{-1} \\
&= (2m\|\boldsymbol{\Delta}\|B_D n^{-\frac{1}{2}})n^{-2}\sum_{i<j}^{n}\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|^2\|\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i\|^{-1} \\
&= (m\|\boldsymbol{\Delta}\|B_D n^{-\frac{1}{2}})\left(\frac{n-1}{n}\right)\left(\frac{2}{n(n-1)}\right)\sum_{i<j}^{n}\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|^2\|\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i\|^{-1} \\
&\stackrel{\text{def}}{=} (m\|\boldsymbol{\Delta}\|B_D n^{-\frac{1}{2}})\left(\frac{n-1}{n}\right)\left(\frac{2}{n(n-1)}\right)\sum_{i<j}^{n}\mathrm{h}_{\mathrm{V}}(\boldsymbol{Z}_i,\boldsymbol{Z}_j) \\
&\stackrel{\text{def}}{=} (m\|\boldsymbol{\Delta}\|B_D n^{-\frac{1}{2}})\left(\frac{n-1}{n}\right)U_n^V,
\end{aligned}$$

where U_n^V is a U-statistic with symmetric kernel $\mathrm{h}_{\mathrm{V}}(\boldsymbol{Z}_i,\boldsymbol{Z}_j)$. Note that the first 2 factors of this expression have order relations of $o(1)$ and $O(1)$, respectively. Thus, it suffices to prove that $U_n^V = O_p(1)$. Finally, given Lemma 3.1 and assumption E1, it follows from Theorem 1 part (c) of Denker and Keller (1983, p.507) that $U_n^V = O_p(1)$, which completes the proof. □

Lemma A.2. *Let* $\boldsymbol{S}_{\mathrm{n}}(\cdot)$ *and* $\hat{\boldsymbol{S}}_{\mathrm{n}}(\cdot)$ *denote the gradient evaluated with* $\mathrm{b}_{\mathrm{ij}}(\boldsymbol{\theta}_0)$ *and* $\mathrm{b}_{\mathrm{ij}}(\hat{\boldsymbol{\theta}})$ *as defined in Section* 3. *Then, under model assumptions* (1)-(3), W1-W4, *and* E4, *we have that*

$$\|\,\hat{\boldsymbol{S}}_{\mathrm{n}}(\mathbf{0})-\boldsymbol{S}_{\mathrm{n}}(\mathbf{0})\| = o_p(1).$$

Proof. To begin, on the left hand side of the proposition, we evaluate the expression inside the norm, using assumption W2 and the Mean Value Theorem, to obtain,

$$
\begin{aligned}
\hat{\boldsymbol{S}}_{\mathrm{n}}(\mathbf{0})-\boldsymbol{S}_{\mathrm{n}}(\mathbf{0}) &= n^{-\frac{3}{2}}\sum_{i<j}^{n}\left[\mathrm{b_{ij}}(\hat{\boldsymbol{\theta}})-\mathrm{b_{ij}}(\boldsymbol{\theta}_0)\right](\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i)\\
&= n^{-2}\sum_{i<j}^{n}(\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i)\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)'n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\
&= n^{-2}\sum_{i<j}^{n}(\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1})\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i)([\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)]'+\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)')n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\
&\overset{\mathrm{def}}{=} (\boldsymbol{V}_{1n}+\boldsymbol{V}_{2n})n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0), \qquad (13)
\end{aligned}
$$

where $\boldsymbol{\theta}^* \in \{\lambda\hat{\boldsymbol{\theta}}+(1-\lambda)\boldsymbol{\theta}_0 : \lambda \in [0,1]\}$. Given assumption W1, it suffices to prove that $\|\boldsymbol{V}_{1n}\| = \|\boldsymbol{V}_{2n}\| = o_p(1)$. Focusing on $\boldsymbol{V}_{1n}$, by definition of $\boldsymbol{u}(\cdot)$ and Theorem A.2, we have that

$$
\begin{aligned}
\|\boldsymbol{V}_{1n}\| &\le n^{-2}\sum_{i<j}^{n}\|\boldsymbol{x}_{j-1}-\boldsymbol{x}_{i-1}\|\|\,\boldsymbol{u}(\boldsymbol{\varepsilon}_j-\boldsymbol{\varepsilon}_i)\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\|\\
&\le m^{\frac{1}{2}}n^{-2}\sum_{i<j}^{n}\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\|.
\end{aligned}
$$

By taking expectations, and using model assumptions (1)-(3), and Lemma 3.2, we have that

$$
\begin{aligned}
\mathrm{E}[\|\boldsymbol{V}_{1n}\|] &\le m^{\frac{1}{2}}n^{-2}\sum_{i<j}^{n}\mathrm{E}[\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\|]\\
&\le m^{\frac{1}{2}}n^{-2}\sum_{i<j}^{n}\mathbb{E}_{\mathrm{i.j}}[\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\|]+o(1)\\
&= m^{\frac{1}{2}}\left(\frac{n(n-1)}{2n^2}\right)\mathbb{E}_{\mathrm{i.j}}[\|\boldsymbol{X}_{j-1}-\boldsymbol{X}_{i-1}\|\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\|]+o(1).
\end{aligned}
$$

By assumption W1 and the Squeeze Theorem, we have that $\boldsymbol{\theta}^*-\boldsymbol{\theta}_0 = \boldsymbol{o}_p(1)$. Which, by assumption W2 and the Continuous Mapping Theorem, implies that $\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\| = o_p(1)$. Furthermore, by assumption W3, it is easy to show that $\|\,\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}^*)-\boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)\| \le 2B_D$. Thus, by model assumptions (1)-(3) and the Lebesgue Dominated Convergence Theorem, the expectation on the right hand side (RHS) is $o(1)$. Note that the remaining factor on the RHS is $O(1)$, which implies that $\mathrm{E}[\|\boldsymbol{V}_{1n}\|] = o(1)$. Thus, using Markov's Inequality, it follows that $\|\boldsymbol{V}_{1n}\| = o_p(1)$.

Next, we focus attention back to $\boldsymbol{V}_{2n}$, from equation (13). Let $\boldsymbol{\lambda}_1 \in \mathbb{R}^{m^2p}$, $\boldsymbol{\lambda}_2 \in \mathbb{R}^{\nu}$ be arbitrary but fixed, and $U_n^{V2} = \boldsymbol{\lambda}_1'\boldsymbol{V}_{2n}\boldsymbol{\lambda}_2$. Thus, by the Cramér-Wold Theorem it suffices to show that

$U_n^{V2} = o_p(1)$. By definitions, it follows that

$$\begin{aligned}
U_n^{V2} &= n^{-2} \sum_{i<j}^{n} \boldsymbol{\lambda}_1' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \boldsymbol{\lambda}_2 \\
&= \left(\frac{n-1}{n}\right) \left(\frac{2}{n(n-1)}\right) \sum_{i<j}^{n} 2^{-1} \boldsymbol{\lambda}_1' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \boldsymbol{\lambda}_2 \\
&\stackrel{\text{def}}{=} \left(\frac{n-1}{n}\right) \left(\frac{2}{n(n-1)}\right) \sum_{i<j}^{n} \mathrm{h}_{\mathrm{V2}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j).
\end{aligned}$$

Note that, under assumption W4, U_n^{V2} is a U-statistic with symmetric kernel $\mathrm{h}_{\mathrm{V2}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$. Thus, by Lemma 3.1 and assumption E4, it follows from Theorem 1 part (c) of Denker and Keller (1983, p.507) that $U_n^{V2} - \mathbb{E}_{\mathrm{i.j}}[\mathrm{h}_{\mathrm{V2}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] = o_p(1)$. Next, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathrm{i.j}}[\mathrm{h}_{\mathrm{V2}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] \\
&= \iint 2^{-1} \boldsymbol{\lambda}_1' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1}) \left(\iint \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \,\mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) \right) \boldsymbol{\lambda}_2 \,\mathrm{dG}(\boldsymbol{X}_{i-1})\, \mathrm{dG}(\boldsymbol{X}_{j-1}),
\end{aligned}$$

where $\mathrm{G}(\cdot)$ denotes the distribution function of $\boldsymbol{X}_t$. Note that, by definition of $\boldsymbol{u}(\cdot)$ and assumption W4, it follows that

$$\begin{aligned}
&\iint \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \,\mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) = - \iint \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \,\mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j), \\
&\iint \boldsymbol{u}(\boldsymbol{\varepsilon}_j - \boldsymbol{\varepsilon}_i)\, \boldsymbol{D}_{\mathrm{ij}}(\boldsymbol{\theta}_0)' \,\mathrm{dF}(\boldsymbol{\varepsilon}_i)\, \mathrm{dF}(\boldsymbol{\varepsilon}_j) = 0, \text{ and} \\
&\mathbb{E}_{\mathrm{i.j}}[\mathrm{h}_{\mathrm{V2}}(\boldsymbol{Z}_i, \boldsymbol{Z}_j)] = 0.
\end{aligned}$$

Finally, since $U_n^{V2} = o_p(1)$, $\|\boldsymbol{V}_{1n}\| = \|\boldsymbol{V}_{2n}\| = o_p(1)$, which completes the proof. □

Lemma A.3. *Let* T_n *be as defined in* (11) *and let*

$$U_n^{\mathrm{AL}} = n^{-\frac{3}{2}} \sum_{i<j}^{n} b_{ij} \boldsymbol{\lambda}' (\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})\, \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})] \boldsymbol{d}_{ijn}$$

where $\boldsymbol{J}[\cdot]$ *is the Jacobian operator. Then, under model assumptions* (1)-(3)*, and* E2*, we have the following,*

$$T_n - U_n^{\mathrm{AL}} = o_p(1).$$

Proof. To begin, note that

$$
\begin{aligned}
|T_n - U_n^{\mathrm{AL}}| &\le n^{-\frac{3}{2}} \sum_{i<j}^{n} |b_{ij} \boldsymbol{\lambda}'(\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1})(\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0}) - \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})]\boldsymbol{d}_{ijn})| \\
&\le n^{-\frac{3}{2}} \|\boldsymbol{\lambda}\| \sum_{i<j}^{n} b_{ij} \|\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1}\| \| \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0}) - \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})]\boldsymbol{d}_{ijn}\| \\
&= n^{-\frac{3}{2}} \|\boldsymbol{\lambda}\| \sum_{i<j}^{n} b_{ij} \|\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1}\| \|\boldsymbol{d}_{ijn}\| \|\boldsymbol{d}_{ijn}\|^{-1} \| \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{d}_{ijn}) - \boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0}) - \boldsymbol{J}[\boldsymbol{u}_{\mathrm{ij}}(\boldsymbol{0})]\boldsymbol{d}_{ijn}\| \\
&\stackrel{\text{def}}{=} n^{-\frac{3}{2}} \|\boldsymbol{\lambda}\| \sum_{i<j}^{n} b_{ij} \|\boldsymbol{x}_{j-1} - \boldsymbol{x}_{i-1}\| \|\boldsymbol{d}_{ijn}\| \, \eta(\boldsymbol{d}_{ijn}) \\
&\le n^{-2} m \|\boldsymbol{\lambda}\| \|\boldsymbol{\Delta}\| \sum_{i<j}^{n} b_{ij} \|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \, \eta(\boldsymbol{d}_{ijn}).
\end{aligned}
$$

By taking expectations, and using assumptions E3, E7, and Lemma 3.2, we have that

$$
\begin{aligned}
\mathrm{E}[|T_n - U_n^{\mathrm{AL}}|] &\le n^{-2} m \|\boldsymbol{\lambda}\| \|\boldsymbol{\Delta}\| \sum_{i<j}^{n} \mathrm{E}[b_{ij} \|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \, \eta(\boldsymbol{d}_{ijn})] \\
&\le n^{-2} m \|\boldsymbol{\lambda}\| \|\boldsymbol{\Delta}\| \sum_{i<j}^{n} \mathbb{E}_{\mathrm{i.j}}[b_{ij} \|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \, \eta(\boldsymbol{d}_{ijn})] + o(1) \\
&= m \|\boldsymbol{\lambda}\| \|\boldsymbol{\Delta}\| \left(\frac{n(n-1)}{2n^2} \right) \mathbb{E}_{\mathrm{i.j}}[b_{ij} \|\boldsymbol{X}_{j-1} - \boldsymbol{X}_{i-1}\|^2 \, \eta(\boldsymbol{d}_{ijn})] + o(1).
\end{aligned}
$$

By model assumptions (1)-(3) and the definition of the Multivariate Derivative, we have that $\boldsymbol{d}_{ijn} = \boldsymbol{o}_p(1)$ and $\eta(\boldsymbol{d}_{ijn}) = o_p(1)$. Furthermore, with Theorem A.3, it is easy to show that $\eta(\boldsymbol{d}_{ijn}) \le (m^{\frac{1}{2}} + 1)\|\varepsilon_j - \varepsilon_i\|^{-1}$. Thus, by assumption E2 and the Lebesgue Dominated Convergence Theorem, the expectation on the right hand side (RHS) is $o(1)$. Note that the remaining factor on the RHS is $O(1)$, which implies that $\mathrm{E}[|T_n - U_n^{\mathrm{AL}}|] = o(1)$. Finally, using Markov's Inequality, it follows that $T_n - U_n^{\mathrm{AL}} = o_p(1)$, which completes the proof. □

A.2 Support Definitions and Theorems

Definition A.1. *Let $g : \mathbb{R}^{m \times k} \to \mathbb{R}$ and $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k$ random m-dimensional vectors with marginal distribution functions $\mathrm{F}_1(\cdot), \cdots, \mathrm{F}_k(\cdot)$, respectively. The Expected Value of $g(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k)$ with respect to the product distribution is the functional*

$$
\mathbb{E}_{1.\cdots.\mathrm{k}}[g(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k)] = \int \cdots \int g(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_k) \, \mathrm{dF}_1(\boldsymbol{x}_1) \cdots \mathrm{dF}_{\mathrm{k}}(\boldsymbol{x}_k)
$$

and is called the the Product Expectation of $g(\boldsymbol{X}_1, \cdots, \boldsymbol{X}_k)$.

Theorem A.1. *Let* $\boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^n$ *and* $\|\cdot\|$ *be a norm. Then,*

$$\left\| \frac{\boldsymbol{x}-\boldsymbol{a}}{\|\boldsymbol{x}-\boldsymbol{a}\|} - \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right\| \leq 2\frac{\|\boldsymbol{a}\|}{\|\boldsymbol{x}\|}.$$

Theorem A.2. *Let* $\boldsymbol{A}$ *be an arbitrary matrix,* $\mathbb{I}_m$ *the* m *dimensional identity matrix,* $\|\cdot\|$ *be the euclidean norm, and* $\otimes$ *denote the Kronecker product. Then,*

$$\|\boldsymbol{A} \otimes \mathbb{I}_m\| = m^{\frac{1}{2}} \|\boldsymbol{A}\|.$$

Theorem A.3. *Let* $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^n$ *(a column vector),* $\|\cdot\|$ *be the euclidean norm,* $\mathrm{J}[\cdot]$ *the Jacobian operator, and* $\boldsymbol{u}(\boldsymbol{x}) = (\boldsymbol{a}-\boldsymbol{x})\|\boldsymbol{a}-\boldsymbol{x}\|^{-1}$. *Then,*

$$\boldsymbol{J}[\boldsymbol{u}(\boldsymbol{x})] = -\frac{1}{\|\boldsymbol{a}-\boldsymbol{x}\|}\left[\mathbb{I}_n - \frac{(\boldsymbol{a}-\boldsymbol{x})(\boldsymbol{a}-\boldsymbol{x})'}{\|\boldsymbol{a}-\boldsymbol{x}\|^2}\right], \text{ and}$$

$$\|\boldsymbol{J}[\boldsymbol{u}(\boldsymbol{x})]\| \leq (n^{\frac{1}{2}}+1)\|\boldsymbol{a}-\boldsymbol{x}\|^{-1}.$$