Massachusetts Institute of Technology
Department of Economics
Working Paper Series

RATE-ADAPTIVE GMM ESTIMATORS FOR LINEAR TIME SERIES MODELS

Guido Kuersteiner

Working Paper 03-17
October 2002

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

RATE-ADAPTIVE GMM ESTIMATORS FOR LINEAR TIME SERIES MODELS

By Guido M. Kuersteiner*

First Draft: November 1999
This Version: October 2002

Abstract

In this paper we analyze Generalized Method of Moments (GMM) estimators for time series models as advocated by Hansen and Singleton. It is well known that these estimators achieve efficiency bounds if the number of lagged observations in the instrument set goes to infinity. However, to this date no data dependent way of selecting the number of instruments in a finite sample is available. This paper derives an asymptotic mean squared error (MSE) approximation for the GMM estimator. The optimal number of instruments is selected by minimizing a criterion based on the MSE approximation. It is shown that the fully feasible version of the GMM estimator is higher order adaptive. In addition a new version of the GMM estimator based on kernel weighted moment conditions is proposed. The kernel weights are selected in a data-dependent way. Expressions for the asymptotic bias of kernel weighted and standard GMM estimators are obtained. It is shown that standard GMM procedures have a larger asymptotic bias and MSE than optimal kernel weighted GMM. A bias correction for both standard and kernel weighted GMM estimators is proposed. It is shown that the bias corrected version achieves a faster rate of convergence of the higher order terms of the MSE than the uncorrected estimator.

Key Words: time series, feasible GMM, number of instruments, rate-adaptive kernels, higher order adaptive, bias correction

---

**Massachusetts Institute of Technology, Dept. of Economics, 50 Memorial Drive E52-371A, Cambridge, MA 02142, USA. Email: gkuerste@mit.edu. Web: http://web.mit.edu/gkuerste/www/.* I wish to thank Xiaohong Chen, Ronald Gallant, Jinyong Hahn, Jerry Hausman, Whitney Newey and Ken West for helpful comments. Stavros Panageas provided excellent research assistance. Comments by a co-editor and three anonymous referees led to substantial improvements of the paper. Financial support from NSF grant SES-0095132 is gratefully acknowledged. All remaining errors are my own.

1. Introduction

In recent years GMM estimators have become one of the main tools in estimating economic models based on first order conditions for optimal behavior of economic agents. Hansen (1982) established the asymptotic properties of a large class of GMM estimators. Based on first order asymptotic theory it was subsequently shown by Chamberlain (1987), Hansen (1985) and Newey (1988) that GMM estimators based on conditional moment restrictions can be constructed to achieve semiparametric efficiency bounds.

The focus of this paper is the higher order asymptotic analysis of GMM estimators for the time series case. In the cross-sectional literature it is well know that using a large number of instruments can result in substantial second order bias of GMM estimators, thus putting limits to the implementation of efficient procedures. Similar results are obtained in this paper for the time series case. In addition, fully feasible, second order optimal implementations of efficient GMM estimators for time series models are developed.

In independent sampling situations feasible versions of efficient GMM estimators were implemented amongst others, by Newey (1990). In a time series context examples of first order efficient estimators are Hayashi and Sims (1983), Stoica, Soderstrom and Friedlander (1985), Hansen and Singleton (1991,1996) and Hansen, Heaton and Ogaki (1996). Under special circumstances and in a slightly different context, Kuersteiner (2002) constructs a feasible, efficient GMM estimator for autoregressive models where the number of instruments is allowed to increase at the same rate as the sample size. More generally however, such expansion rates do not lead to consistent estimates. In fact, to this date no analysis of the optimal expansion rate for the number of instruments for efficient GMM procedures depending on an infinite dimensional instrument set has been provided in the context of time series models. In this paper a data dependent selection rule for the number of instruments is obtained and a fully feasible version of GMM estimators for linear time series models is proposed.

Several moment selection procedures, applicable to time series data, were proposed in the literature. Andrews (1999) considers selection of valid instruments out of a finite dimensional instrument set containing potentially invalid instruments. Hall and Inoue (2001) propose an information criterion based on the asymptotic variance-covariance matrix of the GMM estimator to select relevant moment conditions from a finite set of potential moments. Both approaches do not directly apply to the case of infinite dimensional instrument sets considered here. Linton (1995) analyses the optimal choice of bandwidth parameters for kernel estimates of the partially linear regression model based on minimizing the asymptotic MSE of the estimator. Xiao and Phillips (1996) apply similar ideas to determine the optimal bandwidth in the estimation of the residual spectral density in a Whittle likelihood based regression set up. More recently Linton (1997) extended his procedure to the determination of the optimal bandwidth for an efficient semiparametric instrumental variables estimator. Donald and Newey (2001) use similar arguments to determine the optimal number of base functions in polynomial approximations to the optimal instrument. They analyze higher order asymptotic expansions of the estimators around their true parameter values. While the first order asymptotic terms typically do not depend on the estimation of infinite dimensional nuisance parameters as shown in Andrews (1994) and Newey (1994) this is not the case for higher order terms of the expansions.

In this paper we will obtain expansions similar to the ones of Donald and Newey (2001) for the case of GMM estimators for models with lagged dependent right hand side variables. This set up is important for the analysis of intertemporal optimization models which are characterized by first order conditions of maximization. One particular area of application is asset pricing models.

Minimizing the asymptotic approximation to the MSE with respect to the number of lagged instruments leads to a feasible GMM estimator for time series models. The trade of is between more asymptotic efficiency as measured by the asymptotic covariance matrix and bias.

Full implementation of the procedure requires the specification of estimators for the criterion function used to determine the optimal number of instruments. It is established that a plug-in estimator for the optimal number of instruments leads to a GMM estimator that is fully feasible and achieves the same asymptotic distribution as the infeasible optimal estimator. We also propose a new kernel weighted version of GMM. It is shown that the asymptotic bias and MSE can be reduced if suitable kernel weights are applied to the moment conditions. For this purpose a new rate-adaptive kernel that adjusts its smoothness to the smoothness of the underlying model is introduced. In addition, a data-dependent way to pick the optimal kernel is proposed.

Finally, a semiparametric correction of the asymptotic bias term is proposed. The bias corrected version of the GMM estimator achieves a faster optimal rate of convergence of the higher order terms.

The paper is organized as follows. Section 2 presents the time series models and introduces notation. Section 3 introduces the kernel weighted GMM estimator, contains the analysis of higher order asymptotic MSE terms and derives a selection criterion for the optimal number of instruments. Section 4 discusses implementation of the procedure, in particular consistent estimation of the criterion function for optimal bandwidth selection. Section 5 analyzes the asymptotic bias of the kernel weighted GMM estimator and introduces a data-dependent procedure to select the optimal kernel. Section 6 discusses non-parametric bias correction. Section 7 contains a small Monte Carlo experiment. The proofs are collected in Appendix A. Auxiliary Lemmas are collected in Appendix B which is available upon request.

2. Linear Time Series Models

We consider the linear time series framework of Hansen and Singleton (1996). Let $y_t \in \mathbb{R}^p$ be a strictly stationary stochastic process. It is assumed that economic theory imposes restrictions in the form of a structural econometric equation on the process y_t. In order to describe this structural equation we partition $y_t = [y_{t,1}, y'_{t,2}, y'_{t,3}]$. Here, $y_{t,1}$ is the scalar left hand side variable, $y_{t,2}$ are the included and $y_{t,3}$ are the excluded contemporaneous endogenous variables. The vector X_t is defined to contain, possibly a subset, of the lagged dependent variables $y_{t-1}, ..., y_{t-r}$ where r is known and fixed. The structural equation then takes the form

$$y_{t,1} = \alpha_0 + \beta'_0 y_{t,2} + \beta'_1 X_t + \varepsilon_t. \tag{2.1}$$

The structural model also imposes restrictions on the innovations ε_t. More specifically, ε_t is strictly stationary with $E\varepsilon_t = 0$ and follows a Moving-Average (MA) process of order $m-1$ for $m \geq 1$, where again, m is assumed known and finite. We denote the autocovariance function of ε_t by $\gamma_j^\varepsilon = E\varepsilon_t\varepsilon_{t-j}$ with $\gamma_j^\varepsilon = 0$ for $|j| \geq m$.

Letting $\beta = [\beta'_0, \beta'_1]' \in \mathbb{R}^d$ and collecting all the regressors in x_t where $x'_t = [y'_{t,2}, X'_t]$ we can write (2.1) as $y_{t,1} = \alpha_0 + \beta' x_t + \varepsilon_t$. An alternative representation of (2.1) is obtained by setting $a(L, \beta) = a_0 + a_1 L + ... + a_r L^r$ with $1 \times p$ vectors a_i such that $a(L, \beta) y_t = \alpha_0 + \varepsilon_t$. Note that a_i are subject to exclusion and normalization restrictions implied by β.

In addition to the structural equation (2.1) we also assume that the reduced form of y_t admits a representation as a vector autoregressive moving average (VARMA) process $A(L)y_t = A(1)\mu_y + B(L)u_t$ such that there exists an infinite moving average representation

$$y_t = \mu_y + A^{-1}(L)B(L)u_t. \tag{2.2}$$

Here, $\mu_y \in \mathbb{R}^p$ is a constant and u_t is a strictly stationary and conditionally homoskedastic martingale difference sequence.

In order to completely relate model (2.1) to the generating process (2.2) we define additional $p-1 \times p$ matrices of lag polynomials $A_1(L)$ and $B_1(L)$ such that $A_1(L)y_t = B_1(L)u_t + \alpha_1$. The matrices $A_1(L)$ and $B_1(L)$ satisfy $[a'(L, \beta), A'_1(L)]' = A(L)$, $[b'(L). B'_1(L)]' = B(L)$ and $[\alpha_0, \alpha'_1]' = A(1)\mu_y$. It follows from this representation that the structural innovations ε_t are related to the reduced form innovations by $\varepsilon_t = b(L)u_t$ where $b(L) = b_0 + b_1 L + ... + b_{m-1}L^{m-1}$ and $b_i \in \mathbb{R}^p$ are $1 \times p$ vectors of constants.

We assume that $A(L)$ and $B(L)$ have all their roots outside the unit circle and that all elements of $A(L)$ and $B(L)$ are finite order polynomials in L. Economic theory is assumed to provide restrictions on the polynomials $a(L, \beta)$ and $b(L)$ such that their degrees are known to the investigator. No restrictions are assumed to be known about the polynomials $A_1(L)$ and $B_1(L)$. In particular their degrees in L are unknown, although assumed finite. The investigator is concerned with inference regarding the parameter vector $\beta = (\beta_0', \beta_1')$ while $b(L), A_1(L)$ and $B_1(L)$ are treated as nuisance parameters.

The economic model (2.1) implies moment restrictions of the form

$$E(\varepsilon_{t+m}y_{t-j}) = 0 \text{ for all } j \geq 0. \tag{2.3}$$

These moment restrictions are the basis for the formulation of GMM estimators. Alternatively, the moment restrictions (2.3) are often implied by economic theory and then lead to the formulation of a structural model of the form (2.1). A well known example is asset pricing models.

In addition to the structural restrictions of Equation (2.1) we impose the following formal Assumptions on u_t and $A(L), B(L)$ and $b(L)$.

Assumption A. *Let $u_t \in \mathbb{R}^p$ be strictly stationary and ergodic, with $E(u_t|\mathcal{F}_{t-1}) = 0$, $E(u_tu_t'|\mathcal{F}_{t-1}) = \Sigma$ where Σ is a positive definite symmetric matrix of constants. Let u_t^i be the i-th element of u_t and $\text{cum}_{i_1,...,i_k}(t_1, ..., t_{k-1})$ the k-th order cross cumulant of $u_{t+t_1}^{i_1}, ..., u_t^{i_k}$. Assume that*

$$\sum_{t_1=-\infty}^{\infty} \cdots \sum_{t_{k-1}=-\infty}^{\infty} |\text{cum}_{i_1,...,i_k}(t_1, ..., t_{k-1})| < \infty \text{ for } k \leq 8.$$

Assumption B. *The lag polynomial $C(L)$ with coefficient matrices C_j is defined as $C(L) = A^{-1}(L)B(L)$ where $A(L)$ and $B(L)$ are $p \times p$ matrices of finite order polynomials in L such that $\det A(z) \neq 0$ and $\det B(z) \neq 0$ for $|z| \leq 1$. Moreover, assume $b(z) \neq 0$ for $|z| \leq 1$. Let p_a be the degree of the polynomial $A(L)$ and let λ_1 be the root of maximum modulus of $\det(z^{p_a}A(z^{-1})) = 0$. Let $f_\varepsilon(\lambda) = (2\pi)^{-1} b(e^{i\lambda})'\Sigma b(e^{-i\lambda})$ which can equivalently be written as $f_\varepsilon(\lambda) = (2\pi)^{-1}\sigma_\varepsilon^2 \left|\theta(e^{i\lambda})\right|^2$ for some constant σ_ε^2 and lag polynomial $\theta(L) = 1 - \theta_1 L - ... - \theta_{m-1}L^{m-1}$. Let $p_{\tilde{b}}$ be the degree of the polynomial $\tilde{B}(L) = \theta(L)B(L)$ and let $\tilde{\lambda}_1$ be the root of maximum modulus of $\det(z^{p_{\tilde{b}}}\tilde{B}(z^{-1})) = 0$. Define $\lambda = \max(\lambda_1, \tilde{\lambda}_1)$. Assume that $\lambda \in (0,1)$. Define the infinite dimensional instrument vector $z_{t,\infty} = (y_t', y_{t-1}', ...)'$ and let $P' = \text{Cov}(x_{t+m}, z_{t,\infty})'$. Assume that P has full column rank.*

Remark 1. *The column rank assumption for P is needed for identification (see Kuersteiner (2001) for an extensive discussion of this point). Assumption (B) guarantees that $f_\varepsilon(\lambda) \neq 0$ for $\lambda \in [-\pi, \pi]$. Then $1/f_\varepsilon(\lambda)$ exists and corresponds to the spectral density of an AR(m-1) model.*

The fact that u_t is a martingale difference sequence arises naturally in rational expectations models. In our context it is needed together with the conditional homoskedasticity assumption to guarantee that the optimal GMM weight matrix is of a sufficiently simple form. This allows us to construct estimates of the bias terms converging fast enough for bias correction and optimal number of instruments selection.

The conditional homoskedasticity condition $E(u_t u_t'|\mathcal{F}_{t-1}) = Eu_t u_t'$ is restrictive as it rules out time changing variances. Relaxing this restriction results in more complicated GMM weight matrices of the type analyzed in Kuersteiner (1997, 2001). In principle the higher order moment restriction implied by conditional homoskedasticity could be used in addition to the conditions (2.3). The resulting estimator is however nonlinear and will not be considered here.

The summability assumption for the cumulants limits the temporal dependence of the innovation process. Andrews (1991) shows for $k = 4$ that the summability condition on the cumulants is implied by a strong mixing assumption for u_t. The cumulant summability condition used here is similar but slightly stronger than the second part of Condition A in Andrews (1991). What is needed both in Andrews (1991) and here are restrictions on the eighth-moment dependence of the underlying process u_t.

Infeasible efficient GMM estimation for β is based on exploiting all the implications of the moment restriction (2.3). In our context this is equivalent to choosing all lagged observations as instruments. An infeasible estimator of β based on $z_{t,\infty}$, where $z_{t,\infty}$ is defined in Assumption (B), is used as a reference point around which we expand feasible versions of the estimator.

For this purpose let $\Omega = \sum_{l=-m+1}^{m-1} \gamma_l^\varepsilon \Omega(l)$ with $\Omega(l) = \mathrm{Cov}(z_{t,\infty}, z_{t-l,\infty}')$ and $D = P'\Omega^{-1}P$ where P is defined in Assumption (B). A detailed analysis of these infinite dimensional matrices can be found in Kuersteiner (2001) and Appendix (B.2). The infeasible estimator of β is given by

$$\beta_{n,\infty} = D^{-1}P'\Omega^{-1}\frac{1}{n}\sum_{t=1}^{n-m}\left(y_{t+m,1} - \mu_y^1\right)\left(z_{t,\infty} - \mathbf{1}_\infty \otimes \mu_y\right)$$

where $\mathbf{1}_\infty$ is an infinite dimensional vector containing the element 1 and μ_y^1 is the first element of μ_y.

Let $d_0 = P'\Omega^{-1}\frac{1}{\sqrt{n}}\sum(z_{t,\infty} - \mathbf{1}_\infty \otimes \mu)\varepsilon_t$ almost surely such that $\sqrt{n}\left(\beta_{n,\infty} - \beta_0\right) = D^{-1}d_0$. It can be shown that $D^{-1}d_0 \xrightarrow{p} N(0, D^{-1})$ as $n \to \infty$ under the assumptions made about y_t and ε_t.

For any fixed integer M, let $z_{t,M} = (y_t', y_{t-1}', ...y_{t-M+1}')'$ be a finite dimensional vector of instruments. An approximate version $\beta_{n,M}$ of $\beta_{n,\infty}$ is then based on $D_M = P_M'\Omega_M^{-1}P_M$ and $z_{t,M}$ where P_M and Ω_M are defined in the same way as P and Ω with $z_{t,\infty}$ replaced by $z_{t,M}$. It then follows that $\sqrt{n}\left(\beta_{n,M} - \beta\right) - D^{-1}d_0 \xrightarrow{p} 0$ as $n, M \to \infty$. The last statement is no longer true, at least not without specifying the rate at which M goes to infinity, once $\beta_{n,M}$ is replaced by a feasible estimator $\hat{\beta}_{n,M}$ where $\hat{\beta}_{n,M}$ is defined in the same way as $\beta_{n,M}$ but with P_M and Ω_M replaced by estimates $\hat{P}_M$ and $\hat{\Omega}_M$. We call $\hat{\beta}_{n,\hat{M}}$ a fully feasible estimator if M is a function of the data alone. A more detailed definition of $\hat{\beta}_{n,M}$ is given in Equation (3.2) while data-dependent selection of M is discussed in Section 4.

3. Kernel Weighted GMM

In this paper a generalized class of GMM estimators based on kernel weighted moment restrictions is introduced. Conventional GMM estimators are based on using the first M of the moment restrictions (2.3). More generally one can consider non-random weights $k(j, M)$ such that

$$k(j, M)E\varepsilon_{t+m}y_{t-j-1} = 0.$$

The function $k(j, M)$ is a generalized kernel weight. For the special case where $k(j, M) = k(j/M)$, $k(j, M)$ is a standard kernel function. The truncated kernel is $k(j/M) = \{|j/M| \leq 1\}$ where we use $\{.\}$ to denote the indicator function. The general kernel weighted approach therefore covers the standard GMM procedure as a special case when the truncated kernel is used. In Section 5 it is shown that many kernel functions reduce the higher order bias of GMM and that there always exists a kernel function that dominates the traditional truncated kernel in terms of higher order MSE.

We now describe the kernel weighted GMM estimator $\hat{\beta}_{n,M}$. Define the matrix

$$k_M = \operatorname{diag}(k(0, M), ..., k(M-1, M))'$$

having kernel weight $k(j-1, M)$ in the j-th diagonal element and zeros otherwise. Let $K_M = (k_M \otimes I_p)$ where I_p the p-dimensional identity matrix. An instrument selection matrix $S_M(t) = \operatorname{diag}(\{t \geq 1\}, ...\{t \geq M\})$ is introduced to exclude instruments for which there is no data in the sample. The vector of available instruments is denoted by $\bar{z}_{t,M} = (S_M(t) \otimes I_p)(z_{t,M} - \mathbf{1}_M \otimes \bar{y})$ where $\bar{y} = n^{-1}\sum_{t=1}^n y_t$.

An estimate of the weight matrix Ω_M is obtained as follows. We define $\hat{\Omega}_M(l) = \frac{1}{n}\sum_t \bar{z}_{t,M}\bar{z}'_{t-l,M}$. The optimal weight matrix is then given by

$$\hat{\Omega}_M = \sum_{l=-m+1}^{m-1} \hat{\gamma}^{\varepsilon}(l)\hat{\Omega}_M(l) \tag{3.1}$$

where $\hat{\gamma}^{\varepsilon}(l) = \frac{1}{n}\sum_{t=r-m+1}^{n-m} \hat{\varepsilon}_t\hat{\varepsilon}_{t-l}$ and $\hat{\varepsilon}_t = a(L, \tilde{\beta}_{n,M})(y_{t+m} - \bar{y})$ for some consistent first stage estimator $\tilde{\beta}_{n,M}$. For M fixed and possibly small, it is well known that such an estimator can be obtained from standard inefficient GMM procedures where $\Omega_M = I_{Mp}$.

Let Z_M be the matrix of stacked instruments $Z_M = [\bar{z}_{\max(1,r-m+1),M}, ..., \bar{z}_{n-m,M}]'$ and $X = [x_{\max(m+1,r+1)} - \bar{x}, ..., x_n - \bar{x}]'$ the matrix of regressors. Also, Y is the stacked vector of the first demeaned element in y_t. Then define the $d \times Mp$ matrix $\hat{P}'_M = n^{-1}X'Z_M$ as well as the $Mp \times 1$ vector $\hat{P}^y_M = n^{-1}Z'_MY$. Let $\hat{\Xi}_M = K_M\hat{\Omega}^{-1}_M K_M$. Assuming that M is such that $M \geq d/p$, where d is the dimension of the parameter space, the estimator $\hat{\beta}_{n,M}$ can now be written as

$$\hat{\beta}_{n,M} = \left(\hat{P}'_M\hat{\Xi}_M\hat{P}_M\right)^{-1}\hat{P}'_M\hat{\Xi}_M\hat{P}^y_M. \tag{3.2}$$

For the truncated kernel with $K_M = I_{Mp}$, (3.2) is the standard GMM formula. The effects of using kernel weighted moments can be inferred from (3.2). The kernel matrix K_M distorts efficiency by using $\hat{\Xi}_M$ instead of the optimal $\hat{\Omega}_M^{-1}$ as weight matrix. As is shown below, these effects are second order for suitable choices of the kernel function $k(j, M)$ and bandwidth M. It is also shown that the second order loss of efficiency is more than compensated by a reduction of the second order bias for suitably chosen kernel functions.

We now turn to the formal requirements the kernel weight function $k(.,.)$ has to satisfy. We first define the constant s which plays a role in determining the rate of convergence of $\beta_{n,M}$ to $\beta_{n,\infty}$.

Definition 3.1. *Let λ_1 and $\tilde{\lambda}_1$ be as defined in Assumption (B). Let s_1 be the multiplicity of λ_1 and $\tilde{s}_1$ the multiplicity of $\tilde{\lambda}_1$. Define $s = 2s_1 - 1/2$ if $\lambda_1 > \tilde{\lambda}_1$, $s = 2s_1 + 3/2\tilde{s}_1 - 1/2$ if $\lambda_1 = \tilde{\lambda}_1$ and $s = 3/2\tilde{s}_1 - 1/2$ if $\lambda_1 < \tilde{\lambda}_1$.*

Assumption C. *The kernel function $k(j, M)$ is regular if $k(j, M) = k(j/M)$. Then $k(.)$ satisfies $k : \mathbb{R} \mapsto [-1, 1]$, $k(0) = 1$, $k(x) = k(-x) \forall x \in \mathbb{R}$, $k(x) = 0$ for $|x| > 1$. $k(.)$ is continuous at 0 and at all but a finite number of points. For $q \in (0, \infty)$ there exists a constant k_q such that $k_q = \lim_{x \to 0}(1 - k(x))/|x|^q$. We distinguish: i) $k_q = 0$ for all $q \in (0, \infty)$, ii) $k_q \neq 0$ for some $q \in (0, \infty)$.*

Assumption D. *The kernel function $k(.,.)$ is rate-adaptive if it satisfies $|k(x, y)| \leq c < \infty \; \forall (x, y) \in \mathbb{R} \times \mathbb{R}_+$, $k(-x, y) = k(x, y)$, $k(0, y) = 1$ for all $y \in \mathbb{R}_+$ and $k(x, y) = 0$ for $|x| > 1$. Furthermore, for λ defined in Assumption (B) and $c_0 = -\log(\lambda)$, $\lim_{y \to \infty} (1 - k(x/y, y^s \lambda^y)) / y^s \lambda^y = c_0^q c_1 |x|$ for all $x \in \mathbb{R}$ and some constant c_1.*

Assumption (C) corresponds to the assumptions made in Andrews (1991) except that we also require $k(x) = 0$ for $|x| > 1$. This assumption ensures that only a finite number of moment conditions, controlled by the bandwidth parameter, are used in estimation. The assumption could be relaxed at the cost of having to introduce additional bandwidth parameters to estimate the optimal weight matrix. This seems unattractive from a practical point of view and is not pursued here.

Assumption (C) rules out certain parametric kernel functions such as the Quadratic Spectral kernel but is satisfied by a number of well known kernels such as the Truncated, Bartlett, Parzen and Tukey-Hanning kernels.

In the case of regular kernels the constant k_q measures the higher order loss in efficiency induced by the kernel function which is proportional to $1 - k(i/M) = k_q M^{-q} |i|^q$ for large M and some q such that $k_q \neq 0$. For rate-adaptive kernels we obtain an efficiency loss of $1 - k(i/M, M^s \lambda^M) = M^s \lambda^M c_0^q |i|^q k_q$ which, as will be shown, is of the same order of magnitude as the efficiency loss due to truncating the number of instruments. The kernels are called rate-adaptive because their smoothness locally at zero adapts to the smoothness λ of the model. In Section 4 it is shown how the argument $M^s\lambda^M$ in $k(.,.)$ can be replaced by an estimate.

Kernel functions that satisfy Assumption (D) can be generated by exploiting the chain rule of differentiation. Consider functions of the form

$$\phi(v,z) = (2 - z(-\log(z))^q)\, v + (z(-\log(z))^q - 1)\, v^2 \tag{3.3}$$

for $v \in [-1,1]$ and some non-negative integer q. Then $\phi(1,z) = 1, \phi(0,z) = 0, \partial\phi(v,z)/\partial v|_{v=1} = z(-\log(z))^q$ for all z. It thus follows that z parametrizes the partial derivative of ϕ with respect to v. The constant q is chosen in accordance with a kernel $k(x)$ to which $\phi(v,z)$ is applied and for which $k_q \neq 0$. The rate adaptive kernel $k(j,M)$ is then obtained as

$$k(j,M) = \phi(k(j/M), M^s\lambda^M) \tag{3.4}$$

for any kernel $k(j/M)$ satisfying Assumption (C). It is shown in the proof of Proposition (3.4) that $\lim_{M\to\infty}\left(1 - \phi(k(i/M), M^s\lambda^M)\right)/M^s\lambda^M = |i|^q\, k_q c_0^q$. Also note that $\int \phi(k(x), M^s\lambda^M)^2 dx < \infty$ uniformly in $M \in (0,\infty]$ as long as $\lambda \in (0,1)$ and $\int k(x)^4 dx$ exists. The latter is the case for all kernels satisfying Assumption (C). We use the short hand notation $\int \phi(x)^2 dx = \lim_{M\to\infty}\int \phi(k(x), M^s\lambda^M)^2 dx$.

The bandwidth parameter M is chosen such that the approximate MSE of a weighted sum of the elements of $\hat\beta_{n,M}$ is minimized. We use the Nagar (1959) type approximation to the MSE used in Donald and Newey (2001). Let $\hat\beta_{n,M}$ be stochastically approximated by $b_{n,M}$ such that $n^{1/2}(\hat\beta_{n,M} - \beta) = b_{n,M} + r_{n,M}$ where $r_{n,M}$ is an error term. Define the approximate mean squared error $\varphi_n(M,\ell,k(.))$ of $\hat\beta_{n,M}$ as in Donald and Newey (2001) such that for $\ell \in \mathbb{R}^d$ with $\ell'\ell = 1$,

$$\ell' D^{1/2} E\left(b_{n,M} b_{n,M}'\right) D^{1/2\prime}\ell = 1 + \varphi_n(M,\ell,k(.)) + R_{n,M}$$

and require that the error terms $r_{n,M}$ and $R_{n,M}$ satisfy

$$\frac{\|r_{n,M}\|^2 + R_{n,M}}{\varphi_n(M,\ell,k(.))} = o_p(1) \text{ as } M\to\infty, n\to\infty, \sqrt{M}/n \to 0. \tag{3.5}$$

The only difference to Donald and Newey (2001) is that in our case $\varphi_n(M,\ell,k(.))$ is an unconditional expectation. As noted by Donald and Newey, the approximation is only valid for $M\to\infty$. Given the efficiency of $\beta_{n,\infty}$ this is the case of interest in the context of this paper.

Proposition 3.2. *Suppose Assumptions (A) and (B) hold and $\ell \in \mathbb{R}^d$ with $\ell'\ell = 1$. Let $\Gamma_{t-s}^{\varepsilon x} = E\varepsilon_t x_s$ and define $f_{\varepsilon x}(\lambda) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\Gamma_j^{\varepsilon x} e^{-i\lambda j}$. Define*

$$A_1 = p\,(4\pi)^{-1}\int_{-\pi}^{\pi} f_{\varepsilon x}(\lambda)' f_\varepsilon^{-1}(\lambda) d\lambda. \tag{3.6}$$

Let $\mathcal{A} = \ell' D^{-1/2} \mathcal{A}_1' \mathcal{A}_1 D^{-1/2} \ell$, *and* $\mathcal{B}_1 = \lim_{M \to \infty} (1 - \sigma_{1M}) / \left(M^{2s} \lambda^{2M} \right)$ *with*

$$\sigma_{1M} = \ell' D^{-1/2} P_M' \Omega_M^{-1} P_M D^{-1/2} \ell.$$

Also, $\mathcal{B}^{(q)} = \lim_{M \to \infty} \sigma_{2M} / \left(M^{2s} \lambda^{2M} \right)$ *where* $\sigma_{2M} = \ell' D^{-1/2} b_M' \Omega_M b_M D^{-1/2} \ell$, $b_M' = Q_M (I_{Mp} - P_M \left(P_M' \Omega_M^{-1} P_M \right)^{-1} P_M' \Omega_M^{-1})$ *and* $Q_M = P_M' \left(I_{Mp} - K_M \right) \Omega_M^{-1} + P_M' \Omega_M^{-1} (I_{Mp} - K_M)$. *Then,*
i) for $k(.)$ *such that Assumption (Cii) is satisfied it follows that*

$$\varphi_n(M, \ell, k(.)) = O(M^2/n) + O(M^{-2q}).$$

ii) for $k(.,.)$ *defined in (3.4) such that Assumption (D) holds,* $n, M \to \infty$ *and* $M^{2s-2} \lambda^{2M} n \to 1/\kappa$ *with* $0 < \kappa < \infty$,

$$\lim_n n/M^2 \varphi_n(M, \ell, k(.)) = \mathcal{A} \left(\int_{-\infty}^{\infty} \phi(x)^2 dx \right)^2 + c_0^{2q} k_q^2 \mathcal{B}^{(q)} / \kappa + \mathcal{B}_1 / \kappa.$$

iii) for $k(x) = \{|x| \leq 1\}$, $n, M \to \infty$ *and* $M^{2s-2} \lambda^{2M} n \to 1/\kappa$ *with* $0 < \kappa < \infty$,

$$\lim_n n/M^2 \varphi_n(M, \ell, k(.)) = 4\mathcal{A} + \mathcal{B}_1 / \kappa.$$

This result shows that using standard kernels introduces variance terms of order $O(M^{-2q})$ that are larger than for the truncated kernel. Using the rate-adaptive kernels overcomes this problem, leading to variance terms of the same order as in the standard GMM case. Nevertheless, kernel weighting introduces additional variance terms of order $M^{2s} \lambda^{2M}$. Intuitively, the kernel function distorts the optimal weight matrix resulting in an increased variance of higher order terms in the expansion. As will be shown, this increased variance can be traded off against a reduction in the bias by an appropriate choice of the kernel function. Since any kernel other than the rate-adaptive kernels lead to slower rates of convergence, only rate-adaptive kernels will be considered from now on.

An immediate corollary resulting from Lemma (3.2) is that the feasible estimator has the same asymptotic distribution as the optimal infeasible estimator as long as $M/\sqrt{n} \to 0$.

Corollary 3.3. *Assume that the assumptions of Lemma (3.2) hold. If* $n, M \to \infty$ *and* $M/\sqrt{n} \to 0$ *as* $n \to \infty$ *then* $\sqrt{n} \left(\hat{\beta}_{n.M} - \beta \right) - D^{-1} d_0 = o_p(1)$.

Because $\mathcal{B}^{(q)}$ and $\mathcal{B}_1$ are difficult to estimate we do not minimize $\varphi_n(M, \ell, k(.))$ directly. Instead we propose the following criterion.

Proposition 3.4. *Let* $\bar{M}^*$ *minimize* $\varphi_n(M, \ell, k(.))$. *Then, as* $n \to \infty$, $\bar{M}^* / M^* \to 1$ *where*

(3.7) $$M^* = \arg\min_{M \in I} \mathrm{MIC}(M) = \arg\min_{M \in I} \frac{M^2}{n} \mathcal{A} \left(\int_{-\infty}^{\infty} \phi(x)^2 dx \right)^2 - \log \sigma_M$$

with $I = \{[d/p]+1, [d/p]+2, ...\}$ and $[a]$ denotes the largest integer smaller than a. Here $\sigma_M = \sigma_{1M} - \sigma_{2M}$ with σ_{1M}, σ_{2M} defined in Proposition (3.2). If in addition, for some constant c_1, $\log \sigma_{1M} = c_1 M^{2s} \lambda^{2M} + o(\lambda_r^{2M})$ where λ_r satisfies $0 < \lambda_r < \lambda^{3/2}$, then $\bar{M}^*/M^* - 1 = O(n^{-1/2} (\log n)^{1/2})$.

Remark 2. *By construction, $\hat{\beta}_{n,M^*}$ is higher order efficient under quadratic risk in the class of all GMM estimators $\hat{\beta}_{n,M}$, $M \in I$.*

Remark 3. *For the truncated kernel $\int_{-\infty}^{\infty} \phi(x)^2 dx = 2$ and $\sigma_{2M} = 0$. Note that σ_{1M} measures the second order loss of efficiency caused by a finite number of instruments. This result follows from the fact that $\left(P_M' \Omega_M^{-1} P_M\right)^{-1}$ is the asymptotic variance matrix of the estimator based on M instruments. Then σ_{1M} has the interpretation of a generalized measure of relative efficiency of $\hat{\beta}_M$. It corresponds to $N^{-1}\sigma_\varepsilon^2 H^{-1} f'(I - P^k) f H^{-1}$ of Donald and Newey (2001, their notation). The term σ_{2M} measures the additional loss in efficiency due to the kernel. The constant $\mathcal{A}$ measures the simultaneity bias caused by estimates of the optimal instruments. When $m = 1$ such that ε_t is serially uncorrelated, $\mathcal{A}_1 = \frac{p}{2}\sigma_\varepsilon^{-2} \ell' D^{-1/2} \sigma_{\varepsilon x}$. Noting that here Mp is the total number of instruments, it can be seen easily that when $m = 1$ the penalty term essentially is the same as in Donald and Newey (2001).*

Remark 4. *It is shown in the proof of Proposition (4.3) that $M^* = \log n/(2c_0) + o(\log n)$ and asymptotically does not depend on $\mathcal{A}$, $\mathcal{B}^{(q)}$, $\mathcal{B}_1$ or k up to order $o(\log n)$. If the result $\bar{M}^*/M^* - 1 = o(1)$ is sufficient then it is possible to replace $\mathrm{MIC}(M)$ by the simpler criterion $n^{-1}(Mp)^2 - \log \sigma_{1M}$ where Mp is the total number of instruments. Note that this simplification is allowed only for models where $\log \sigma_{1M}$ decays at an exponential rate. This is the case for the ARMA class.*

Remark 5. *A further simplification is available if the definition of $\varphi_n(M, \ell, k(.))$ is based on*

$$\operatorname{tr} D^{1/2} E b_{n,M} b_{n,M}' D^{1/2\prime}$$

instead of $\ell' D^{1/2} E(b_{n,M} b_{n,M}') D^{1/2\prime} \ell$. Then the approximate MSE depends on $\operatorname{tr} \tilde{\sigma}_{1M} = \log |\tilde{\sigma}_{1M}| = \log |P_M' \Omega_M^{-1} P_M| - \log |D|$ where $\tilde{\sigma}_{1M} = D^{-1/2} P_M' \Omega_M^{-1} P_M D^{-1/2}$. The simplified criterion for this case is $n^{-1}(Mp)^2 + \log \left|P_M' \Omega_M^{-1} P_M\right|^{-1}$ and knowledge of the variance lowerbound D^{-1} is no longer required. Note that this formulation of $\mathrm{MIC}(M)$ is quite similar to Hall and Inoue (2000) except for the penalty term $(Mp)^2/n$.

4. Fully Feasible GMM

In this section we derive the missing results that are needed to obtain a fully feasible procedure. In particular one needs to replace the unknown optimal bandwidth parameter M_n^* by an estimate $\hat{M}^*$. In order to have a fully feasible procedure we need consistent estimates of the constants $\mathcal{A}_1, D$ and σ_M, converging at sufficiently fast rates.

The following analysis shows that estimation of $\mathcal{A}_1$ can be done nuisance parameter free in the sense that consistent estimates of $\mathcal{A}_1$ do not depend on additional unknown parameters. Unfortunately the same is not true for D and σ_M. We use an approximating parametric model for $C(L)$ to estimate D and σ_M.

We first consider the simpler estimation problem for the constant $\mathcal{A}_1$ where

$$\mathcal{A}_1 = \frac{p}{4\pi}\int f_{\varepsilon x}(\lambda)' f_{\varepsilon}^{-1}(\lambda) d\lambda = \frac{p}{2}\sum_{j=-\infty}^{\infty} \zeta_j \Gamma_j^{\varepsilon x'}.$$

Note that ζ_j are the coefficients in the series expansion of $f_{\varepsilon}^{-1}(\lambda)$.

Consistent estimates of the MA(m-1) representation of ε_t can be obtained by using consistent estimates of the parameter β to obtain estimates $\hat{\varepsilon}_t$. An MA(m-1) model is then estimated for $\hat{\varepsilon}_t$. This can be done by using a nonlinear least squares or pseudo maximum likelihood procedure as described in chapter 8 of Brockwell and Davis (1991). This procedure is outlined in the proof of Lemma (4.1). Because of the exponential decay of ζ_j and the fact that m is finite, $\Gamma_j^{\varepsilon x}$ can be replaced by a simple sample average based on estimated residuals $\hat{\Gamma}_j^{\varepsilon x} = n^{-1}\sum_{t=\min(j+1.1)}^{\min(n-m,n-j)} \hat{\varepsilon}_{t+m} x_{t-j}$. Using these estimates one forms $\widehat{\mathcal{A}}_1$ by

(4.1) $$\widehat{\mathcal{A}}_1 = \frac{p}{2}\sum_{j=-n+1}^{n} \hat{\zeta}_j \hat{\Gamma}_{j-m}^{\varepsilon x}.$$

To summarize, we state the following proposition.

Proposition 4.1. *Let Assumptions (A) and (B) be satisfied. Let $\widehat{\mathcal{A}}_1$ be defined in (4.1). Then* $\sqrt{n}(\widehat{\mathcal{A}}_1 - \mathcal{A}_1) = O_p(1)$.

We use a finite order VAR(h) approximation to $C(L)^{-1} \equiv \breve{C}(L)$ with $\breve{C}(L) = \sum_{j=0}^{\infty} \breve{C}_j L^j$ to estimate the parameters D and σ_M. It follows that $y_t = \breve{C}_0^{-1}\breve{C}(1)\mu_y + \sum_{j=1}^{\infty}\pi_j y_{t-j} + v_t$ where $\pi_j = \breve{C}_0^{-1}\breve{C}_j$ and $v_t = \breve{C}_0^{-1}u_t$. Let $\pi(L) = I - \sum_{j=1}^{\infty}\pi_j L^j$ and $Ev_t v_t' = \Sigma_v$. The approximate model with VAR coefficient matrices $\pi_{1,h}, ...\pi_{h,h}$ is then given by

(4.2) $$y_t = \mu_{y,h} + \pi_{1,h} y_{t-1} + ... + \pi_{h,h} y_{t-h} + v_{t,h}$$

where $\Sigma_{v.h} = Ev_{t,h}v_{t,h}'$ is the mean squared prediction error of the approximating model. VAR approximations in a bandwidth selection context was proposed by Andrews (1991). There however, h is kept fixed such that the resulting bandwidth choice is asymptotically suboptimal. We let $h \to \infty$ at a data-dependent rate, to be described below, leading to the approximating model being asymptotically equivalent to $\pi(L)$.

It was shown by Berk (1974) and Lewis and Reinsel (1985) that the parameters $(\pi_{1,h},...,\pi_{h,h})$ are root-n consistent and asymptotically normal for $\pi(h) = (\pi_1,...,\pi_h)$ if h does not increase too quickly, i.e. if h is chosen such that $h^3/n \to 0$. At the same time h must not increase too slowly to avoid asymptotic biases. Berk (1974) shows that h needs to increase such that $n^{1/2}\sum_{j=h+1}\pi_j \to 0$ as $h, n \to \infty$. Ng and Perron (1995) argue that information criteria such as the Akaike criterion do not satisfy these conditions and can therefore not be used to choose h. Moreover, the results of Hannan and Kavalieris (1984, 1986) imply that if $\hat{h}$ is selected by AIC or BIC then $\hat{h} - h = o_p(\sqrt{\log n})$ and $\hat{h}$ fails to be adaptive in the sense of Ng and Perron (1995).

To avoid the problems that arise from using information criteria to select the order of the approximating model we use the sequential testing procedure analyzed in Ng and Perron (1995). Let $\pi(h) = (\pi_1',...,\pi_h')'$, $Y_{t,h} = (y_t' - \bar{y}',...,y_{t-h+1}' - \bar{y}')'$ and $M_h = \sum_{t=h+1}^n Y_{t-1,h}Y_{t-1,h}'$ and define $M_h^{-1}(1)$ to be the lower-right $p \times p$ block of M_h^{-1}. Let Γ_k be the $kp \times kp$ matrix whose (m,n)th block is Γ_{n-m}^{yy} and $\Gamma_{1,h}' = [\Gamma_{-1}^{yy},...,\Gamma_{-h}^{yy}]$ where $\Gamma_{j-i}^{yy} = \mathrm{Cov}(y_{t-i}, y_{t-j}')$. The coefficients of the approximate model satisfy the Yule-Walker equations $(\pi_{1,h},...,\pi_{h,h}) = \Gamma_{1,h}\Gamma_h^{-1}$. Let $\hat{\Gamma}_{1,h} = (n-h)^{-1}\sum_{t=h}^{n-1} Y_{t,h}(y_{t+1} - \bar{y})'$ and $\hat{\Gamma}_h = (n-h)^{-1}\sum_{t=h}^{n-1} Y_{t,h}Y_{t,h}'$. The estimated error covariance matrix is $\hat{\Sigma}_{v,h} = n^{-1}\sum_{t=h+1}^n \hat{v}_{t,h}\hat{v}_{t,h}'$ where $\hat{v}_{t,h} = y_t - \hat{\pi}_{1,h}y_{t-1} - ... - \hat{\pi}_{h,h}y_{t-h}$ with coefficients $\hat{\pi}(h)' = \hat{\Gamma}_{1,h}'\hat{\Gamma}_h^{-1}$. A Wald test for the null hypothesis that the coefficients of the last lag h are jointly 0 is then, in Ng and Perron's notation,

$$J(h,h) = n\,(\mathrm{vec}\,\hat{\pi}_{h,h})'\left(\hat{\Sigma}_{v,h} \otimes M_h^{-1}(1)\right)^{-1}(\mathrm{vec}\,\hat{\pi}_{h,h}).$$

We adopt the following lag order selection procedure from Ng and Perron (1995).

Definition 4.2. *The general-to-specific procedure chooses i)* $\hat{h}_n = h$ *if, at significance level* α, $J(h,h)$ *is the first statistic in the sequence* $J(i,i), \{i = h_{\max},...,1\}$, *which is significantly different from zero or ii)* $\hat{h}_n = 0$ *if* $J(i,i)$ *is not significantly different from zero for all* $i = h_{\max},...,1$ *where* $h_{\max}$ *is such that* $h_{\max}^3/n \to 0$ *and* $n^{1/2}\sum_{j=h_{\max}+1}\|\pi_j\| \to 0$ *as* $n \to \infty$.

In order to calculate the impulse response coefficients associated with (4.2) define the matrix

$$A_h = \begin{bmatrix} \pi_{1,h} & \pi_{2,h} & \cdots & \pi_{h,h} \\ I & 0 & \cdots & 0 \\ & \ddots & & \vdots \\ & & I & 0 \end{bmatrix}$$

with dimensions $hp \times hp$. The j-th impulse coefficient of the approximating model is given by $C_{j,h} = E_h' A_h^j E_h$ with $E_h' = [I_p, 0, ..., 0]$. The autocovariance function Γ_j^{yy} is then approximated by $\Gamma_{j,h}^{yy} = \sum_{l=0}^{\infty} C_{l+j,h} \Sigma_{v,h} C_{l,h}'$ for all $j = 1, 2, ...$ Likewise we approximate the optimal weight matrix by the infinite dimensional matrix $\Omega_h = \sum_{j=-m+1}^{m-1} \gamma^{\varepsilon}(j) \Omega_h(j)$ where the infinite dimensional matrix $\Omega_h(j)$ has typical k, l-th block $\Gamma_{l-k-j,h}^{yy}$. We denote the k, l-th block of Ω_h^{-1} by $\vartheta_{kj,h}$. We define D_h by letting $D_h = \sum_{k=1,j=1}^{\infty} \Gamma_{k,h}^{xy} \vartheta_{kj,h} \Gamma_{-j,h}^{yx}$.

Estimates $\hat{C}_{j,h} = E_h' \hat{A}_h^j E_h$ of $C_{j,h}$ are obtained by substituting $\hat{\pi}(h)$ for $\pi(h)$ in A_h such that $\hat{A}_h$ is defined in the obvious way. Substituting estimates $\hat{C}_{j,h}$ for $C_{j,h}$ in D_h leads to an estimate $\hat{D}_h$. A fully feasible version $\hat{D}_{\hat{h}}$ is obtained by replacing h with $\hat{h}$ as defined in Definition (4.2). In the proof of Proposition (4.3) it is shown that $\sqrt{n}$-consistency of $\hat{\pi}(\hat{h})$ implies $\hat{D}_{\hat{h}} - D = O_p(n^{-1/2})$. Next, σ_M is approximated by $\sigma_{M,h}$ in a similar way. We use the approximate autocovariance matrices $\Gamma_{j,h}^{yy}$ to form the possibly infinite dimensional matrices $P_{M,h}' = \left[\Gamma_{1,h}^{xy}, ...\Gamma_{M,h}^{xy}\right]$ where $P_{\infty,h}'$ is defined in the obvious way. We then form the matrix $Q_{M,h} = P_{M,h}' (I_{Mp} - K_M) \Omega_{M,h}^{-1} + P_{M,h}' \Omega_{M,h}^{-1} (I_{Mp} - K_M)$ and $b_{M,h}^{(q)} = Q_{M,h}(I_{Mp} - P_{M,h}' \left(P_{M,h}' \Omega_{M,h}^{-1} P_{M,h}\right)^{-1} P_{M,h}' \Omega_{M,h}^{-1})$. The parameter $\sigma_{M,h}$ is obtained from

$$\sigma_{M,h} = \sigma_{1M,h} - \sigma_{2M,h} \text{ with } \sigma_{2M,h} = \ell' D_h^{-1/2} b_{M,h} \Omega_{M,h} b_{M,h}' D_h^{-1/2} \ell \tag{4.3}$$

where $D_h = P_{\infty,h}' \Omega_h^{-1} P_{\infty,h}$ and $\sigma_{1M,h} = \ell' D_h^{-1/2} P_{M,h}' \Omega_{M,h}^{-1} P_{M,h} D_h^{-1/2} \ell$. An estimate of $\sigma_{1M,h}$ is based on $\hat{\pi}(h)$ in the same way as described above and $\sqrt{n}$-consistency uniformly in M is implied by the result for $\hat{D}_{\hat{h}}$. In practice, infinite dimensional matrices such as $P_{\infty,h}'$ and Ω_h have to be replaced by finite dimensional matrices. The resulting approximation can be made arbitrarily accurate by an appropriate choice of the relevant dimensions without affecting the convergence results.

We now define the estimate $\hat{\sigma}_{M,\hat{h}}$ where we replace K_M in $b_{M,h}$ by $\hat{K}_M$ with typical element $\phi(k(i/M), \sqrt{\log \hat{\sigma}_{1M,\hat{h}}})$. Moreover, we replace all elements in (4.3) with corresponding estimates based on $\hat{\pi}(\hat{h})$. In Proposition (4.3) we establish that $\left(\hat{\sigma}_{M,\hat{h}} - \sigma_M\right) / \left(M^{2s} \lambda^{2M}\right) = O_p(n^{-1/2} (\log n)^{1/2+s'})$ uniformly in M where s' is defined in Proposition (4.3). Establishing this result requires a stronger form of uniform convergence of the approximating parameters $\hat{\pi}(\hat{h})$ than was needed to show $\hat{D}_{\hat{h}} - D = O_p(n^{-1/2})$, thus explaining the slower rate of convergence. Also let $\widehat{\mathcal{A}} = \ell' \hat{D}_{\hat{h}}^{-1/2} \widehat{\mathcal{A}}_1' \widehat{\mathcal{A}}_1 \hat{D}_{\hat{h}}^{-1/2} \ell$ where $\widehat{\mathcal{A}}_1$ is defined in (4.1). Define $\hat{M}^*$ as

$$\hat{M}^* = \arg\min_{M \in I} \frac{M^2}{n} \widehat{\mathcal{A}} \left(\int_{-\infty}^{\infty} \phi(x)^2 dx\right)^2 - \log \hat{\sigma}_{M,\hat{h}}. \tag{4.4}$$

The following result can be established.

Proposition 4.3. *Let $\hat{M}^*$ be defined in (4.4) and M^* as in (3.7). Then* $\left(\hat{M}^*/M^* - 1\right) = o_p(1)$. *If in addition* $\log \sigma_{1M} = c_1 M^{2s}\lambda^{2M} + o(\lambda_r^{2M})$ *with* $0 < \lambda_r < \lambda^{3/2}$ *then* $\left(\hat{M}^*/M^* - 1\right) = O_p(n^{-1/2}(\log n)^{1/2+s'})$ *with* $s' = s_1\left\{\lambda_1 \geq \tilde{\lambda}_1\right\} + \tilde{s}_1\left\{\lambda_1 \leq \tilde{\lambda}_1\right\}$.

Remark 6. *It is always the case that* $\log \sigma_{1M} = c_1 M^{2s}\lambda^{2M} + o(\lambda_r^{2M})$ *for some* λ_r *such that* $0 < \lambda_r < \lambda$. *Here we require that* λ_r *is not too close to* λ, *ie. that the remainder term disappears sufficiently fast.*

Ultimately, one is interested in the properties of a fully automated estimator $\hat{\beta}_{n,\hat{M}^*}$ where the data determined optimal bandwidth $\hat{M}^*$ is plugged into the kernel function and the data-dependent kernel $\phi(k(i/\hat{M}^*), \sqrt{\log \hat{\sigma}_{1\hat{M}^*,\hat{h}}})$ is used. In order to analyze this estimator we need an additional Lipschitz condition for the class of permitted kernels.

Assumption E. *The kernel* $k(j/M)$ *satisfies* $|k(x) - k(y)| \leq c_1 |x^q - y^q| \; \forall x, y \in [0,1]$ *for some* $c_1 < \infty$ *and* $q \geq 1$.

Assumption (E) corresponds to the assumptions made in Andrews (1991). Using the previous results we are now in a position to state one of the main results of this paper which establishes that an automated bandwidth selection procedure can be used to pick the number of instruments based on sample information alone.

Theorem 4.4. *Suppose Assumptions (A) and (B) hold and ether i)* $k(.,.)$ *is defined in 3.4 and satisfies Assumptions (D) and* $k(.)$ *used as an argument in* $\phi(.,.)$ *satisfies Assumptions (Cii) and (E) where* $\phi(.,.)$ *is as defined in (3.3) or ii)* $k(.,.) = \{|x| \leq 1\}$. *Assume that* $\log \sigma_{1M} = c_1 M^{2s}\lambda^{2M} + o(\lambda_r^{2M})$ *with* $0 < \lambda_r < \lambda^{3/2}$. *Let* $\hat{M}^*$ *be defined in (4.4) then for case i)* $n/\sqrt{M^*}(\hat{\beta}_{n,\hat{M}^*} - \hat{\beta}_{n,M^*}) = O_p((\log n)^{\max(s'-q,-1)})$ *and for case ii)* $n/\sqrt{M^*}(\hat{\beta}_{n,\hat{M}^*} - \hat{\beta}_{n,M^*}) = o_p(1)$. *Also, if i) and* $s' - q < 0$ *or ii) holds then*

$$\lim_{n\to\infty}\left(\varphi_n(\hat{M}^*, \ell, k(.,.), b_{n,\hat{M}^*}) - \varphi_n(M^*, \ell, k(.,.), b_{n,M^*})\right) = 0.$$

Remark 7. *Under the conditions of the Theorem,* M^* *can be replaced by* $\bar{M}^*$ *in the last display as well as in Proposition (4.3).*

Remark 8. *Although the constant* s' *is typically unknown it is often reasonable to assume that* $s' = 1$ *which requires that* $\lambda_1 \neq \tilde{\lambda}_1$ *and there are no multiplicities in the largest roots.*

Theorem (4.4) shows that using the feasible bandwidth estimator $\hat{M}^*$ results in estimates $\hat{\beta}_{n,\hat{M}^*}$ that have asymptotic mean squared errors which are equivalent to asymptotic mean squared errors of estimators where a nonrandom optimal bandwidth sequence M^* is used. An immediate consequence of the Theorem is also that $\hat{\beta}_{n,\hat{M}^*}$ is first order asymptotically equivalent to the infeasible estimator $D^{-1}d_0$. The theorem however shows in addition higher order equivalence of $\hat{\beta}_{n,\hat{M}^*}$ and $\hat{\beta}_{n,M^*}$. In this respect Theorem (4.4) is stronger than Proposition 4 of Donald and Newey (2001).

5. Bias Reduction and Kernel Selection

In this section we analyze the asymptotic bias of $\hat{\beta}_{n,M}$ as a function of the sample size n and the bandwidth parameter M.

Theorem 5.1. *Suppose Assumptions (A) and (B) hold and $k(.,.)$ satisfies Assumption (D). If $M \to \infty$ and $M/n^{1/2} \to 0$ then*

$$\lim_{n\to\infty} \sqrt{n}/ME(b_{n,M} - \beta) = D^{-1}\mathcal{A}_1' \int \phi^2(x)dx.$$

Remark 9. *This result also holds for kernels satisfying Assumption (C) if $\int \phi^2(x)dx$ is replaced by $\int k^2(x)dx$.*

A simple consequence of this result is that for many standard kernels the asymptotic bias of the kernel weighted GMM estimator is lower than the bias for the standard GMM estimator based on the truncated kernel.

Corollary 5.2. *Suppose Assumptions (A) and (B) hold and $k(.,.)$ satisfies Assumption (D) with $\int \phi^2(x)dx \leq 2$ or Assumption (C) with $\int k^2(x)dx \leq 2$. If $n, M \to \infty$, $M/n^{1/2} \to 0$ then*

$$\lim_{n\to\infty} \left\|\sqrt{n}/ME(b_{n,M} - \beta)\right\| \leq \lim_{n\to\infty} \left\|\sqrt{n}/ME(b_{n,M}^T - \beta)\right\|$$

where $b_{n,M}^T$ is the stochastic approximation to the GMM estimator based on the truncated kernel.

It can be shown easily that substituting well known kernels such as the Bartlett, Parzen or Tukey-Hanning in $\phi(v,z)$ leads to $\int \phi^2(x)dx$ being equal to 16/15, 67/64 and 1.34 respectively.

We now turn to the question of optimality in the higher order MSE sense of the choice of kernel function. Let $\kappa^* = \lim 1/\left(\lambda^{2M_T^*} M_T^{*2s-2}\right)$ where M_T^* is optimal for the truncated kernel. Note that by optimality of M_T^*, $0 < \kappa^* < \infty$. From Proposition (3.2) it follows at once that any kernel for which $\mathcal{A}\left(\left(\int_{-\infty}^{\infty} \phi(x)^2 dx\right)^2 - 4\right) + c_0^2 k_q^2 \mathcal{B}^{(q)}/\kappa^* < 0$ dominates the truncated kernel. In Theorem (5.3) a simple variational argument is used to show that we can always find a kernel $k(.,.)$ such that this inequality is satisfied uniformly on a compact subset of the parameter space. This result raises the question of finding an optimal or at least dominating kernel. When $q = 2$ is fixed this problem has been solved for standard kernels by Priestley (see Priestley, 1981 p.569). In our context of rate-adaptive kernels with fully flexible q it is not known whether closed form solutions of the associated functional optimization problem exist. In any event, such solutions most likely depend on the constant $\mathcal{B}^{(q)}$ which is difficult to estimate.

To avoid these complications we propose the following data-dependent solution to the optimal kernel selection problem. Let $\phi(k(x), M^s\lambda^M)$ be as described before. Because $\phi(.,.)$ enforces adaptiveness of the kernel we only choose $k(x)$, which by the Weierstrass Theorem can be approximated by polynomials. Let τ be a finite integer. Define $k(x) = 1+\sum_{i=1}^{\tau}\psi_i x^i$ for $x \in [0,1)$, $k(x) = k(|x|)$ for $x < 0$ and $k(x) = 0$ if $|x| \geq 1$. Let $\psi = (\psi_1, ..., \psi_\tau)$. Then for $j = 1, ..., \tau$ define $U_\psi^j \subset \mathbb{R}^\tau$, such that U_ψ^j is compact and for $\psi^j = (\psi_1, .., \psi_{j-1}, 0, \psi_{j+1}, .., \psi_\tau)$, it follows $\psi^j \notin U_\psi^j$. Also let $\mathcal{K}^0 = \{|x| \geq 1\}$. The permissible class of kernels is $\mathcal{K}_j = \mathcal{K}^0 \cup \mathcal{K}^j$ where for $j \geq 1$, $\mathcal{K}^j = \left\{k(x)|k(x) = 1 + \sum_{i=j}^{\kappa}\psi_i x^i, k(x) \in [-1,1], \psi \in U_\psi^j\right\}$ and it is understood that $k(x)$ also satisfies the restrictions outlined before. The optimal kernel $k^*(x)$ with $\phi(k(x)) = 2k(x) - k(x)^2$ satisfies

$$\text{(5.1)} \quad \mathcal{A}\left(\int_{-\infty}^{\infty}\phi(k^*(x))^2dx\right)^2 + k_{q^*}^{*2}\mathcal{B}^{(q^*)}/\kappa^* \leq \mathcal{A}\left(\int_{-\infty}^{\infty}\phi(k(x))^2dx\right)^2 + k_q^2\mathcal{B}^{(q)}/\kappa^*, \text{ all } \underset{\text{for all } q\geq q', q'\geq 1}{k(.,.) \in \mathcal{K}_q}$$

Note that optimality is pointwise in $\mathcal{A}$ and $\mathcal{B}^{(q)}$ which means that in general k^* depends on $\mathcal{A}$ and $\mathcal{B}^{(q)}$. It will be shown that (5.1) is a reasonable optimality criterion because one of the main objectives in this section is to construct kernel functions that dominate the truncated kernel. To see why (5.1) implies dominance note that the particular choice of κ^* guarantees that k^* has the same variance term $\mathcal{B}_1/\kappa^*$ as the truncated kernel when evaluated along the sequence M_T^*. Once the kernel k^* is selected, its MSE, when evaluated under its own optimal M^* sequence, can be no worse than under the M_T^* sequence. A data-dependent optimal kernel is defined as $\hat{k}^*$ where

$$\text{(5.2)} \qquad \hat{k}^* = \underset{k\in\mathcal{K}_q, q=q',2,..,\tau; q'\geq 1}{\arg\min} \hat{\mathcal{A}}\left(\int_{-\infty}^{\infty}\phi(k(x))^2dx\right)^2 + \frac{n}{\hat{M}_T^{*2}}\left(2\frac{\hat{M}_T^*}{\log n}\right)^{-q}\hat{\sigma}^k_{2\hat{M}_T^*,\hat{h}}.$$

The notation $\hat{\sigma}^k_{2\hat{M}_T^*,\hat{h}}$ is used to emphasize that the $\hat{K}_M$-matrix used to construct σ_{2M} contains diagonal elements $\phi(k, \sqrt{\hat{\sigma}_{1,\hat{M}_T^*}})$ depending on k and $\hat{M}_T^*$ is the optimal bandwidth for the truncated kernel. We establish the following result.

Theorem 5.3. *Let $\hat{k}^*(x)$ be defined as in (5.2). Then for any $q' \in [1,\tau]$, $\tau < \infty$,*

$$\sup_x\left(\hat{k}^*(x) - k^*(x)\right) = O_p(n^{-1/2}(\log n)^{1/2+s'}).$$

Let $\hat{\beta}_{n,M^*(k^*),k^*}$ be the kernel weighted GMM estimator with kernel k^* and let $\hat{\beta}_{n,\hat{M}^*(\hat{k}^*),\hat{k}^*}$ be the GMM estimator based on $\hat{k}^*$. Here, $\hat{M}^*(k) = \arg\min \hat{\mathcal{A}}M^2n^{-1} - \log\hat{\sigma}^{k^*}_{M,\hat{h}}$ where for $\hat{M}(\hat{k}^*)$ we use $\hat{\sigma}^{\hat{k}^*}_M$ with

$\hat{\sigma}^{\hat{k}^*}_{2M,\hat{h}}$ *where* $\phi(\hat{k}^*, \sqrt{\log \hat{\sigma}_{1M,\hat{h}}})$ *is used as kernel weight. Then,*

$$n/\sqrt{M^*}\left(\hat{\beta}_{n,\hat{M}^*(\hat{k}^*),\hat{k}^*} - \hat{\beta}_{n,M^*(k^*),k^*}\right) = O_p((\log n)^{\max(s'-q',-1)}).$$

Furthermore, let $\Theta = \{A_1, ..., A_{p_a}, B_1, ..., B_{q_b}; p_a, q_b \text{ finite}\}$ *be the set of all reduced form models that satisfy Assumption (B). Let* Θ_0 *be a compact set* $\Theta_0 \subset \Theta$ *such that* $\sup_{\Theta_0} \mathcal{B}^{(q)} < \infty$ *and* $\inf_{\Theta_0} \mathcal{A} > 0$*. Then, for some collection of sets* U^j_ψ*, each sufficiently large, there exists a* $\tilde{k} \in \mathcal{K}_{q'}$ *with* $\tilde{k}_{q'} = \lim_{x\to 0}\left(1-\tilde{k}(x)\right)/|x|^{q'}$ *for any* $q' \in [1,\tau], \tau < \infty$ *such that* $\sup_{\Theta_0} \mathcal{A}\left(\left(\int_{-\infty}^{\infty}\phi(\tilde{k}(x))^2 dx\right)^2 - 4\right) + \tilde{k}^2_{q'}\mathcal{B}^{(q')}/\kappa^* < 0$.

The second part of the theorem implies that the truncated kernel is always dominated by k^*. This is the case because the truncated kernel $\{|x| \leq 1\} \in \mathcal{K}$ and there is an element in $\mathcal{K}$ that strictly dominates it.

6. Bias Correction

Another important issue is whether the bias term can be corrected for. The benefits of such a correction are analyzed first. It turns out that correcting for the bias term increases the optimal rate of expansion for the bandwidth parameter and consequently accelerates the speed of convergence to the asymptotic normal limit distribution.

Using the result in Theorem (5.1) the following bias corrected estimator is proposed

$$\hat{\beta}^c_{n,M} = \hat{\beta}_{n,M} - \frac{M}{n}\left(\hat{P}'_M\hat{\Xi}_M\hat{P}_M\right)^{-1}\widehat{\mathcal{A}}'_1\int\phi^2(x)dx. \tag{6.1}$$

Note that for standard GMM (truncated kernel) the bias correction term is simply $2\frac{M}{n}\left(\hat{P}'_M\hat{\Omega}^{-1}_M\hat{P}_M\right)^{-1}\widehat{\mathcal{A}}'_1$. The bias term $\mathcal{A}_1$ can be estimated by the methods described in the previous section. The quality of the estimator $\mathcal{A}_1$ determines the impact of the correction on the higher order convergence rate of the estimator. If $\widehat{\mathcal{A}}_1 - \mathcal{A}_1$ is only $o_p(1)$ then the convergence rate of $\hat{\beta}^c_{n,M}$ is essentially the same as the one for $\hat{\beta}_{n,M}$. If $\widehat{\mathcal{A}}_1 - \mathcal{A}_1 = O_p(n^{-\eta})$ for $\eta \in (0,1/2]$ then the convergence rate of the estimator is improved. The mean squared error of the bias corrected estimator is defined as before by

$$nD^{1/2}\ell' E b^c_{n,M} b^c_{n,M}{}'\ell D^{1/2} = 1 + \varphi^c_n(M,\ell,k(.)) + R^c_{n,M}$$

where $\sqrt{n}\left(\hat{\beta}^c_{n,M} - \beta\right) = b^c_{n,M} + r^c_{n,M}$ with the same restrictions imposed on the remainder terms $R^c_{n,M}$ and $r^c_{n,M}$ as in (3.5). We obtain the following result.

Theorem 6.1. *Suppose Assumptions (A) and (B) hold and $k(.,.)$ satisfies Assumptions (Ci) or (D). Then for any $\ell \in \mathbb{R}^d$ with $\ell'\ell = 1$, $\varphi_n^c(M, \ell, k(.,.)) = O(M/n) - \log \sigma_M$. The optimal M_c^* can be chosen by*

$$M_c^* = \arg\min \frac{Mp}{n} - \log \sigma_M.$$

If $\hat{M}_c^* = \arg\min \frac{Mp}{n} - \log \hat{\sigma}_M$, then $\left(\hat{M}_c^*/M_c^* - 1\right) = O_p(n^{-1/2}(\log n)^{s'+1/2})$ and

$$n/\sqrt{M_c^*}\left(\hat{\beta}^c_{n,\hat{M}_c^*} - \hat{\beta}_{n,M_c^*} + \frac{M_c^*}{n} D^{-1} A_1' \int h^2(x)dx\right) = o_p(1).$$

Remark 10. *The result remains valid if $\hat{\beta}_{n,\hat{M}_c^*}$ is replaced with $\hat{\beta}_{n,\hat{M}_c^*(\hat{k}^*),\hat{k}^*}$ in (6.1) as long as $q' > s'$.*

It follows from Theorem (6.1) that for $\hat{\beta}^c_{n,\hat{M}_c^*}$ the higher order MSE is $O(\log n/n)$ compared to $O((\log n)^2/n)$ for the GMM estimator without bias correction.

7. Monte Carlo Simulations

A small Monte Carlo experiment is conducted in order to assess the performance of the proposed moment selection and bias correction methods. For the simulations we consider the following data generating process

$$(7.1) \qquad \begin{aligned} y_{t,1} &= \beta y_{t,2} + u_t - \theta u_{t-1} \\ y_{t,2} &= \phi y_{t-1,2} + v_t. \end{aligned}$$

with $[u_t, v_t]' \sim N(0, \Sigma)$ where Σ has elements $\sigma_1^2 = \sigma_2^2 = 1$ and σ_{12}. The parameter β is the parameter to be estimated and is set to $\beta = 1$ in all simulations. All remaining parameters are nuisance parameters not explicitly estimated. The parameter σ_{12} is one of the determinants of the small sample bias of both Ordinary Least Squares (OLS) and GMM estimators and is set to .5. The parameter ϕ controls the quality of lagged instruments and is chosen in $\{.1, .3, .5\}$. Low values of ϕ imply that the model is poorly identified. The parameter θ finally is set to $\{-.9, -.5, 0, .5, .9\}$.

We generate samples of size $n = \{128, 512\}$ from Model (7.1). Starting values are $y_0 = 0$ and $[u_0, v_0]' = 0$. In each sample the first 1,000 observations are discarded to eliminate dependence on initial conditions.

Standard GMM estimators are obtained from applying Formula (3.2) with $K_M = I_M$. In order to estimate Ω_M we first construct an inefficient but consistent estimate $\tilde{\beta}_{n,1}$ based on (3.2) setting $K_M = I_M$ and $\Omega_M = I_M$. We then construct residuals $\tilde{\varepsilon}_t = y_{1t} - \tilde{\beta}_{n,1} y_{2t}$ and estimate $\hat{\Omega}_M$ as described in (3.1). Kernel weighted GMM estimators (KGMM) are based on the same inefficient initial estimate such that the estimate for $\hat{\Omega}_M$ is identical to the weight matrix used for the standard GMM estimators. In the second stage we again apply (3.2) with $\hat{\Omega}_M$ and the matrix $\hat{K}^*_M$ based on the optimal data-dependent kernel $\hat{k}^*$ defined in Equation (5.2) with $\mathcal{K}_q = \mathcal{K}_1$.

The estimated optimal bandwidth $\hat{M}^*_n$ is computed according to the procedure laid out in Theorem (5.3).[1] For each simulation replication we obtain a consistent first stage estimate $\tilde{\beta}_{n,1}$ to generate residuals $\tilde{\varepsilon}_t$. We estimate θ by fitting an MA(1) model to $\tilde{\varepsilon}_t$ using the Matlab procedure `arimax`. We then estimate the sample autocovariances $\hat{\Gamma}^{\varepsilon x}_j$ for $j = 0, ..., n/2$ where n is the sample size and form an estimate of $\mathcal{A}_1$ based on Formula (4.1). Next we use the procedure of Definition (4.2) to determine the optimal specification of the approximating VAR for $y_t = [y_{1t}, y_{2t}]'$ allowing for a maximum of $2 * [n^{1/3}]$ of lags. Based on the optimal lag length specification we compute the impulse coefficients of the VAR which are then used to estimate the remaining parameters D and σ_M needed for the criterion MIC(M) as well as for optimal kernel selection.

In Tables 1-3 we compare[2] the performance of feasible kernel weighted GMM with optimally chosen kernel, KGMM-Opt, to feasible standard GMM, GMM-Opt, with truncated kernel. In addition to automatic selection of $\hat{M}^*$ we consider both estimators, KGMM-X (with optimally chosen kernel) and GMM-X, with a fixed number of $X = 1, 20$ lagged instruments. We also report the performance of the corresponding bias corrected versions BGMM-Opt and BKGMM-Opt where $\hat{M}^*_c$ is selected according to the procedure described in Theorem (6.1).

In order to separate the effects of selecting M from the properties of using weights for the moment conditions we first consider GMM and KGMM with a fixed number of instruments. Tables 1-3 show that for ϕ small relative to the sample size there is little difference between the two estimators. They are also not very different from OLS. As the identification of the model improves, KGMM starts to dominate GMM both in terms of (median) bias as well as MSE and mean absolute error (MAE). This effect becomes more pronounced as more and more instruments are being used which can be explained by the predominance of bias terms in this case and the bias reducing property of the kernel weighted estimator.

Turning now to the fully feasible versions we see that the same results remain to hold. For poorly identified models the choice of M does not affect bias that much and all the estimators considered have roughly the same bias properties. Especially for poorly identified parametrizations optimal GMM is much more disperse than optimal KGMM. The reason for this lies in the fact that optimal GMM tends to be based on fewer instruments which results in somewhat lower bias but comes at the cost of increased variability. As identification, parameterized by ϕ, and/or sample size improve, optimally

[1] The Matlab code is available on request.

[2] Results for $\theta = \{-.9, .9\}$ are available on request.

chosen kernel weighted GMM starts to dominate standard GMM for most parameter combinations.

The bias corrected versions of both estimators attain further improvements both as far as bias as well as MSE and MAE are concerned when the model is well identified. In these circumstances the kernel weighted and bias corrected GMM tends to have a somewhat larger MSE than the non-weighted version. On the other hand the non-weighted version tends to overcorrect bias in some cases. A clear ranking is thus not possible. The bias corrected estimators tend to perform relatively poorly compared to GMM-Opt and KGMM-Opt when identification is weak or when $|\theta|$ is large. Overall, their performance is more sensitive to the underlying data-generating process.

In the theoretical development of the paper we have maintained the assumption of conditional homoskedasticity of the innovations. While the strongest results concerning higher order adaptiveness and optimality of bias correction in Theorems (4.4) and (6.1) are not expected to go through without the homoskedasticity assumption it is still expected that the optimal $M^* = \log n/(2\log\lambda)$ asymptotically. In this sense it is plausible that the criterion MIC(M) performs reasonably well even with heteroskedastic errors. We investigate this question by changing the first equation in Model (7.1) to $y_{1t} = \beta y_{2t} + \varepsilon_t - \theta\varepsilon_{t-1}$ where ε_t follows the IGARCH(1,1) process $\varepsilon_t = u_t h_t^{1/2}$ with $h_t = a_0 + a_1\varepsilon_{t-1}^2 + b_1 h_{t-1}$. We set $b_1 = .2$, $a_0 = .1$ and $a_1 = .8$. The innovations $[u_t, v_t]$ are defined as before. Since heteroskedasticity of this form is easy to detect in the data we assume that GMM estimators are now implemented with heteroskedasticity consistent covariance matrix estimators $\hat{\Omega}_M$. For simplicity we use the procedure of Newey and West (1987) with a fixed number of lags.

The results are reported in Table 4 for the case of $\theta = .5$. Results for other parametrizations are available on request. For a fixed number of moment conditions KGMM still dominates GMM in many cases for the larger sample size $n = 512$. The estimator GMM-Opt continues to perform well while KGMM-Opt now does worse when identification is weak but performs at least as well when identification is strong and/or sample sizes are large. As expected, bias correction is no longer effective in reducing bias. Moreover, when combined with kernel weighting, it produces severe outliers resulting in inflated dispersion measures. For this reason only inter decile ranges (IDR) are reported.

8. Conclusions

We have analyzed the higher order asymptotic properties of GMM estimators for time series models. Using expressions for the asymptotic Mean Squared Error a selection rule for the optimal number of lagged instruments is derived. It is shown that plugging an estimated version of the optimal rule into the estimator leads to a fully feasible GMM procedure.

A new version of the GMM estimator for linear time series models is proposed where the moment conditions are weighted by a kernel function. It is shown that optimally chosen kernel weights of the moment restrictions reduce the asymptotic bias and MSE. Correcting the estimator for the highest order bias term leads to an overall increase in the optimal rate at which higher order terms vanish asymptotically. A fully automatic procedure to chose both the number of instruments as well as the optimal kernel is proposed.

A. Proofs

Auxiliary Lemmas are collected in Appendix B which is available upon request. They are referred to in the text as Lemma (B.XX). Before stating the proofs a few commonly used terms are defined.

Definition A.1. *Let* $\mu_x = Ex_t$. *Define* $w_{t,i} = (x_{t+m} - \mu_x)(y_{t-i+1} - \mu_y)'$, $\Gamma_i^{xy} = Ew_{t,i}$ *and* $\Gamma_{-i}^{yx} = Ew'_{t,i}$ *and let* $\tilde{w}_{t,i} = w_{t,i} - \Gamma_i^{xy}$. *Next define* $w_{t,j-i}^y = (y_{t-i} - \mu_y)(y_{t-j} - \mu_y)'$ *with* $Ew_{t,j-i}^y = \Gamma_{j-i}^{yy}$. *Define* $v_{t,i} = \varepsilon_{t+m}(y_{t-i+1} - \mu_y)$ *and* $E\varepsilon_{t+m}y_s = \Gamma_{t-s}^{\varepsilon y}$. *Let the* j,k*-th element of* Ω, Ω^{-1} *and* Ω_M^{-1} *be* $\omega_{j,k}, \vartheta_{j,k}$ *and* $\vartheta_{j,k}^M$ *respectively. For a matrix* A, $\|A\|^2 = \operatorname{tr} AA'$.

Proof of Proposition (3.2). Consider a second order Taylor approximation of $\hat{D}_M^{-1}$ around D^{-1} such that for $\hat{d}_M = \hat{d}_M = \hat{P}_M' K_M \hat{\Omega}_M^{-1} n^{-1/2} \sum_{t=1}^{n-m} \varepsilon_{t+m} \tilde{z}_{t,M} K_M$,

$$\sqrt{n}(\beta_{n,M} - \beta) = D^{-1}[I - (\hat{D}_M - D)D^{-1} + (\hat{D}_M - D)D^{-1}(\hat{D}_M - D)D^{-1}]\hat{d}_M + o_p(M/\sqrt{n})$$

where for $M/n^{1/2} \to 0$ the error term is $o_p(M/\sqrt{n})$ by the Taylor theorem, and the fact that $\det D \neq 0$, $\hat{D}_M - D = O_p(M/n^{1/2})$ as shown in Lemmas (B.14)-(B.23) and $\hat{d}_M = O_p(1)$ by Lemmas (B.24) to (B.33). We decompose the expansion into $\hat{D}_M - D = H_1 + ... + H_4$ where $H_1 = P_M' K_M \Omega_M^{-1} K_M P_M - P'\Omega^{-1}P$, $H_2 = \hat{P}_M' K_M \Omega_M^{-1} K_M \hat{P}_M - P_M' K_M \Omega_M^{-1} K_M P_M$, $H_3 = -\hat{P}_M' K_M \Omega_M^{-1}(\hat{\Omega}_M - \Omega_M)\Omega_M^{-1} K_M \hat{P}_M$ and H_4 is defined in (A.14). Also, $\hat{d}_M = d_0 + d_1 + ... + d_9$ with d_i defined in (A.15)-(A.24) such that $\sqrt{n}(\beta_{n,M} - \beta) = b_{n,M} + o_p(M/\sqrt{n})$ with

$$b_{n,M} = D^{-1}\sum_{i=0}^{9} d_i - D^{-1}\sum_{i=1}^{4}\sum_{j=0}^{9} H_i D^{-1} d_j$$

The terms H_3 and H_4 contain a Taylor series expansion of $\hat{\Omega}_M^{-1}$ around Ω_M^{-1} given by

$$\hat{\Omega}_M^{-1} = \Omega_M^{-1} - \Omega_M^{-1}(\hat{\Omega}_M - \Omega_M)\Omega_M^{-1} + B + o_p(\left\|\hat{\Omega}_M - \Omega_M\right\|^2) \tag{A.1}$$

where B has typical element k, l given by $\operatorname{vec}(\hat{\Omega}_M - \Omega_M)' \frac{\partial^2 \vartheta_{kl}}{\partial \operatorname{vec}\Omega \partial \operatorname{vec}\Omega'} \operatorname{vec}(\hat{\Omega}_M - \Omega_M)$. The term $o_p(\left\|\hat{\Omega}_M - \Omega_M\right\|^2) = o_p(1)$ by Lemma (B.9). Let $g_k(M) = M^{-q}$ for regular kernels $k(j/M)$ and $g_k(M) = M^s \lambda^M$ for rate-adaptive kernels $k(j, M)$. The notation g_k indicates the dependence of the rate on the kernel used. Define the constant $c_k = 1$ for regular kernels and $c_k = -\log\lambda$ for rate adaptive kernels. In Lemmas (B.14) to (B.16) it is shown that $H_1 = H_{11} + H_{12} + H_{13} + H_{14}$ is

$$H_{11} \equiv P_M' \Omega_M^{-1} P_M - P'\Omega^{-1}P = O(M^{2s}\lambda^{2M}) \tag{A.2}$$

$$H_{12} \equiv P_M'(I - K_M)\Omega_M^{-1}(I - K_M)P_M = O(g_k(M)^2) \tag{A.3}$$

$$H_{13} \equiv -P_M' \Omega_M^{-1}(I - K_M)P_M = O(g_k(M)) \tag{A.4}$$

$$H_{14} \equiv -P_M'(I - K_M)\Omega_M^{-1}P_M = O(g_k(M)) \tag{A.5}$$

where $\equiv$ means 'equal by definition'. In Lemmas (B.17) to (B.20) the term $H_2 = H_{211} + H_{212} + H_{221} + H_{222}$ is analyzed to be

$$\text{(A.6)} \quad H_{211} \equiv -\left(\hat{P}_M - \check{P}_M\right)' K_M \Omega_M^{-1} K_M (\hat{P}_M - \check{P}_M) = O_p(M/n)$$

$$\text{(A.7)} \quad H_{212} \equiv \hat{P}_M' K_M \Omega_M^{-1} K_M (\hat{P}_M - \check{P}_M) + (\hat{P}_M - \check{P}_M)' K_M \Omega_M^{-1} K_M \hat{P}_M = O_p(n^{-1/2})$$

$$\text{(A.8)} \quad H_{221} \equiv -\left(\check{P}_M - P_M\right)' K_M \Omega_M^{-1} K_M' \left(\check{P}_M - P_M\right) = O_p(M/n)$$

$$\text{(A.9)} \quad H_{222} \equiv \check{P}_M' K_M \Omega_M^{-1} K_M (\check{P}_M - P_M) + (\check{P}_M - P_M)' K_M \Omega_M^{-1} K_M \check{P}_M = O_p(M/n^{1/2}).$$

where $\hat{P}_M$ is defined in Section 3 and $\check{P}_M' = \left[\check{\Gamma}_1^{xy}, ..., \check{\Gamma}_M^{xy}\right]$ with $\check{\Gamma}_j^{xy} = n^{-1} \sum_{t=\max(j+1,r-m)+1}^{n} \check{w}_{t,j}$.

Lemmas (B.21) and (B.22) show that $H_3 = H_{31} + H_{32} + H_{33} + H_{34}$ is

$$\text{(A.10)} \quad H_{31} \equiv (\hat{P}_M - P_M)' K_M \Omega_M^{-1} (\hat{\Omega}_M - \Omega_M) \Omega_M^{-1} K_M (\hat{P}_M - P_M) = o_p(M/n)$$

$$\text{(A.11)} \quad H_{32} \equiv -\hat{P}_M' K_M \Omega_M^{-1} (\hat{\Omega}_M - \Omega_M) \Omega_M^{-1} K_M (\hat{P}_M - P_M) = o_p(M/n)$$

$$\text{(A.12)} \quad H_{33} \equiv -(\hat{P}_M - P_M)' K_M \Omega_M^{-1} (\hat{\Omega}_M - \Omega_M) \Omega_M^{-1} K_M \hat{P}_M = o_p(M/n)$$

$$\text{(A.13)} \quad H_{34} \equiv -P_M' K_M \Omega_M^{-1} (\hat{\Omega}_M - \Omega_M) \Omega_M^{-1} K_M P_M = O_p(n^{-1/2})$$

and H_4 which is a remainder term defined as

$$\text{(A.14)} \quad H_4 \equiv \hat{P}_M' K_M (\hat{\Omega}_M^{-1} - \Omega_M^{-1} + \Omega_M^{-1} (\hat{\Omega}_M - \Omega_M) \Omega_M^{-1}) K_M \hat{P}_M = o_p(M/n)$$

where the last equality follows from Lemma (B.23).

Next we turn to the analysis of $\hat{d}_M$ which is decomposed as $\hat{d}_k = \sum_j^9 d_j$. Define

$$V_M = \left[n^{-1/2} \sum_t v_{t,1}', ..., n^{-1/2} \sum_t v_{t.M}'\right]'$$

with $V \equiv V_\infty$ such that it follows from Lemmas (B.24) to (B.33) that

$$\text{(A.15)} \quad d_0 \equiv P' \Omega^{-1} V = O_p(1)$$

$$\text{(A.16)} \quad d_1 \equiv P_M' \Omega_M^{-1} V_M - P' \Omega^{-1} V = O_p(M^s \lambda^M)$$

$$\text{(A.17)} \quad d_2 \equiv P_M' (I - K_M) \Omega_M^{-1} (I - K_M) V_M = O_p(g_k(M)^2)$$

$$\text{(A.18)} \quad d_3 \equiv -P_M' (I - K_M) \Omega_M^{-1} V_M - P_M' \Omega_M^{-1} (I - K_M) V_M = O_p(g_k(M))$$

$$\text{(A.19)} \quad d_4 \equiv \left(\hat{P}_M - \check{P}_M\right)' K_M \Omega_M^{-1} K_M V_M = O_p(M/n)$$

$$\text{(A.20)} \quad d_5 \equiv \left(\check{P}_M - P_M\right)' K_M \Omega_M^{-1} K_M V_M = O_p(M/n^{1/2})$$

$$\text{(A.21)} \quad d_6 \equiv \left(\hat{P}_M - P_M\right)' K_M \Omega_M^{-1} (\Omega_M - \hat{\Omega}_M) \Omega_M^{-1} K_M V_M = O_p(M/n)$$

$$\text{(A.22)} \quad d_7 \equiv P_M' K_M \Omega_M^{-1} (\Omega_M - \hat{\Omega}_M) \Omega_M^{-1} K_M V_M = O_p(n^{-1/2})$$

$$\text{(A.23)} \quad d_8 \equiv \hat{P}_M' K_M B K_M V_M + o_p(M/n) = O_p(M/n)$$

$$\text{(A.24)} \quad d_9 \equiv n^{-1/2} \sum_t \varepsilon_t \hat{P}_M' K_M \hat{\Omega}_M^{-1} K_M \left[\mathbf{1}_M \otimes (\bar{y} - \mu_y)\right] = O_p(M/n^{3/2}).$$

We first focus on regular kernels where $g_k(M) = M^{-q}$. We consider the terms in the expansion $D^{-1}\sum_{i=0}^{9} d_i - D^{-1}\sum_{i=1}^{4}\sum_{j=0}^{9} H_i D^{-1} d_j$ of the estimator which depend on M and n and are largest in probability. From the results in Equations (A.2) to (A.24) it follows that the largest such terms are H_{12}, H_{13}, H_{14}, H_{222}, d_0, d_2, d_3 and d_5. Of those terms we examine cross products of the form $Ed_i d_j'$, $Ed_i d_0' D^{-1} H_i$ and $EH_i D^{-1} d_0 d_0' D^{-1} H_j$. Letting $\mathcal{B}_1^{(q)} = c_0^{-q} k_q^{-1} \lim_{M\to\infty} (H_{13} + H_{14})/g_k(M)$, the largest terms vanishing at rate M^{-q} as $M \to \infty$ are $Ed_0 d_3' = -M^{-q} k_q \mathcal{B}_1^{(q)} + o(M^{-q})$ as shown in Lemma (B.39) and $-Ed_0 d_0' D^{-1}(H_{13} + H_{14}) = M^{-q} k_q \mathcal{B}_1^{(q)} + o(M^{-q})$ by Lemmas (B.24) and (B.42). The two terms cancel because they are of opposite sign.

Now define $\mathcal{B}_0^{(q)} = k_q^{-2} \lim_{M\to\infty} P_M'(I - K_M)\Omega_M^{-1}(I - K_M)P_M/g_k(M)^2$. Terms of order M^{-2q} include $Ed_0 d_2' = M^{-2q} k_q^2 \mathcal{B}_0^{(q)} + o(M^{-2q})$ by Lemma (B.38) and $-Ed_0 d_0' D^{-1} H_{12}' = -M^{-2q} k_q^2 \mathcal{B}_0^{(q)} + o(M^{-2q})$ by Lemma (B.35). Since $Ed_0 d_2'$ and $-Ed_0 d_0' D^{-1} H_{12}'$ are of opposite sign these terms cancel. We are left with $E(d_3 - (H_{13} + H_{14})D^{-1} d_0)(d_3 - (H_{13} + H_{14})D^{-1} d_0)' = O(M^{-2q})$ by Lemmas (B.16), (B.24), (B.39) and (B.43).

Terms that grow with M and are largest in order are $H_{222} D^{-1} d_0$ and d_5. It follows by Lemma (B.41) that the cross product term $EH_{222} D^{-1} d_0 d_5'$ is of lower order. We are left with $EH_{222} D^{-1} d_0 d_0' D^{-1} H_{222}' = O(n^{-1})$ by Lemma (B.40) and $Ed_5 d_5' = O(M^2/n)$ by Lemma (B.44). Then $\varphi_n(M, \ell, k(.)) = O(M^2/n) + O(M^{-2q})$.

Next we turn to the case of the rate-adaptive kernel where $g_k(M) = M^s \lambda^M$. Now H_{11}, H_{12}, H_{13}, H_{14}, H_{222}, d_0, d_1, d_2, d_3 and d_5 are largest in probability. In Lemmas (B.34) and (B.36) we show that $Ed_0 d_0' D^{-1} H_{11} = Ed_1 d_0'$ such that these terms cancel out. Because $Ed_0 d_1' = H_{11} + o(n^{-1})$ by Lemma (B.36) it follows that $Ed_0 d_1' D^{-1}(H_{13} + H_{14})$ is of lower order. The largest terms remaining are therefore $Ed_1 d_1' = M^{2s} \lambda^{2M} \mathcal{B}_1 + o(g_k(M)^2)$ where $\mathcal{B}_1 = \lim_{M\to\infty} -H_{11}/g_k(M)^2$ and $E(d_3 - (H_{13} + H_{14})D^{-1} d_0)(d_3 - (H_{13} + H_{14})D^{-1} d_0)' = O(M^{2s}\lambda^{2M})$ by Lemmas (B.16), (B.24), (B.25), (B.39) and (B.43). The largest term growing with M is not affected by the kernel choice and is therefore $Ed_5 d_5' = O(M^2/n)$ as before.

For part ii) and iii) we only need to consider the terms $A_n = Ed_5 d_5'$ and $B_n = E(d_3 - (H_{13} + H_{14})D^{-1} d_0)(d_3 - (H_{13} + H_{14})D^{-1} d_0)' + Ed_1 d_1'$. Since for all $n \geq 1$ we have $A_n \geq 0$ and $B_n \geq 0$ it follows that $\liminf_n A_n \geq 0$ and $\liminf_n B_n \geq 0$ such that $\mathcal{A}$, $\mathcal{B}^{(q)}$ and $\mathcal{B}_1$ are nonnegative.

From Lemma (B.44) it follows that

$$E\ell' D^{-1/2} d_5 d_5' D^{-1/2} \ell = M^2/n \left(\int \phi^2(x) dx \right)^2 \ell' D^{-1/2} \mathcal{A}_1' \mathcal{A}_1 D^{-1/2} \ell + o(M^2/n).$$

From Lemma (B.39) it follows that $M^s\lambda^M Ed_0d_3 = -k_q\mathcal{B}_1^{(q)} + o(1)$ and from Lemma (B.24) it follows that $Ed_0d_0' = D + o(1)$ such that

$$M^{2s}\lambda^{2M}E(H_{13}+H_{14})D^{-1}d_0d_0'D^{-1}(H_{13}+H_{14})' = c_0^{2q}k_q^2\mathcal{B}_1^{(q)}D^{-1}\mathcal{B}_1^{(q)\prime} + o(1).$$

This implies that

$$E(H_{13}+H_{14})D^{-1}d_0d_3' - E(H_{13}+H_{14})D^{-1}d_0d_0'D^{-1}(H_{13}+H_{14})' = o(M^{2s}\lambda^{2M})$$

or in other words $B_n = Ed_3d_3' - E(H_{13}+H_{14})D^{-1}d_0d_0'D^{-1}(H_{13}+H_{14}) + o(M^{2s}\lambda^{2M})$. Here $Ed_3d_3' = M^{2s}\lambda^{2M}\mathcal{B}_2^{(q)} + o(M^{2s}\lambda^{2M})$ as shown in Lemma (B.43) where $\mathcal{B}_2^{(q)}$ is defined in (B.19). ■

Proof of Proposition (3.4): First note that by Lemma (B.25)

$$\begin{aligned} 1-\sigma_{1M} &= \ell'D^{-1/2}(D - P_M'\Omega_M^{-1}P_M)D^{-1/2}\ell = \ell'D^{-1/2}Ed_1d_1'D^{-1/2}\ell + o(\max(n^{-1}, M^{2s}\lambda^{2M})) \\ &= M^{2s}\lambda^{2M}\mathcal{B}_1 + o(\max(n^{-1}, M^{2s}\lambda^{2M})). \end{aligned}$$

Since $D - P_M'\Omega_M^{-1}P_M \geq 0$ by standard arguments it follows that $1-\sigma_{1M} \downarrow 0$ as $M \to \infty$. Also, $Q_M P_M = H_{13} + H_{14}$ such that

$$\begin{aligned} b_M'\Omega_M b_M &= Q_M\Omega_M Q_M' - (H_{13}+H_{14})\left(P_M'\Omega_M^{-1}P_M\right)^{-1}(H_{13}+H_{14})' \\ &= M^{2s}\lambda^{2M}c_0^{2q}k_q^2(\mathcal{B}_2^{(q)} - \mathcal{B}_1^{(q)}D^{-1}\mathcal{B}_1^{(q)\prime}) + o(M^{2s}\lambda^{2M}) \end{aligned}$$

where the last line follows from $\mathcal{B}_1^{(q)} = c_0^{-q}k_q^{-1}\lim(H_{13}+H_{14})/g_k(M)$ and the fact that $\mathcal{B}_2^{(q)} = c_0^{-2q}k_q^{-2}\lim Q_M\Omega_M Q_M'/g_k(M)^2$ by similar arguments as in the proof of Lemma (B.43). More specifically, consider for example, $Q_M P_M = H_{13}+H_{14}$ with $H_{13} = M^s\lambda^M\sum_{j_1,j_2=1}^M \Gamma_{j_1}^{xy}|j_1|^q \frac{1-k(j_1,M)}{|j_1|^q M^{2s}\lambda^{2M}}\vartheta_{j_1,j_2}^M\Gamma_{-j_2}^{yx}$. We use the chain rule of differentiation to write

$$\lim_M \frac{1-k(i,M)}{|i|^q M^s\lambda^M} = \lim_M \frac{1-\phi(k(i/M), M^s\lambda^M)}{1-k(i/M)}\frac{1-k(i/M)}{(|i|/M)^q}\frac{1}{M^q M^s\lambda^M}.$$

Because $\partial\phi(v,z)/\partial v = 2 - z(-\log z)^q + 2(z(-\log z)^q - 1)v$ is continuous in v and $z \downarrow 0$, it follows that

$$\begin{aligned} &\lim_{M\to\infty}\frac{1-\phi(k(i/M), c_1M^s\lambda^M(1+(o_M(1)))}{M^s\lambda^M} \\ = &\lim_{M\to\infty}\left(\frac{-\log c_1}{M} + (k-1)\frac{-\log M}{M} + \log\lambda + \frac{-\log(1+o_M(1))}{M}\right)^q\frac{1-k(i/M)}{(i/M)^q} \\ = &\ k_q(-\log\lambda)^q\lim_{M\to\infty}\left(\frac{-\log c_1}{M\log\lambda} + (k-1)\frac{-\log M}{M\log\lambda} + 1 + \frac{-\log(1+o_M(1))}{M\log\lambda}\right)^q(1+o_M(1)) \\ = &\ k_q(-\log\lambda)^q \end{aligned}$$

Here, $o_M(1)$ is a term that goes to zero as $M \to \infty$. Define $\sigma_M = \sigma_{1M} - \sigma_{2M}$. It follows that $\sigma_M = 1 + O(M^{2s}\lambda^{2M})$. Let $x = 1 - \sigma_{1M} + \sigma_{2M}$ such that $-\log(1-x) = x + o(x)$. Then, $-\log\sigma_M = 1 - \sigma_{1M} + \sigma_{2M} + o(x) = M^{2s}\lambda^{2M}\left(k_q\mathcal{B}^{(q)} + \mathcal{B}_1\right) + o(M^{2s}\lambda^{2M})$. Since $\bar{M}^* \to \infty$ as $n \to \infty$ it follows that $\varphi_n(\bar{M}^*, \ell, k(.)) = \mathrm{MIC}(\bar{M}^*)\,(1 + o(1))$. The result then follows by the same arguments as in Hannan and Deistler (1988, p.333). If in addition $\log\sigma_{1M} = c_1M^{2s}\lambda^{2M} + o(\lambda_r^{2M})$ with λ_r such that $0 < \lambda_r < \lambda^{3/2}$ then $1 - \sigma_{1M} = c_1M^{2s}\lambda^{2M} + o(\lambda_r^{2M})$ and the same holds for σ_{2M} by construction of $k(.,.)$. Then, it follows that $\varphi_n(\bar{M}^*, \ell, k(.)) = \mathrm{MIC}(\bar{M}^*)\left(1 + O(n^{1/2}(\log n)^{1/2})\right)$ by the same arguments as in the proof of Proposition (4.3). .

Proof of Proposition (4.1): Since x_t contains elements of y_t it is enough to show without loss of generality that $\sum\hat{\zeta}_{j+m}\hat{\Gamma}^{\varepsilon y}_{j-k}$ is $\sqrt{n}$-consistent. Let $\tilde{\beta}$ be a $\sqrt{n}$-consistent first stage estimate. The estimated residuals $\hat{\varepsilon}_t = (y_t - \bar{y}) - \tilde{\beta}'(x_t - \bar{x})$ are used to estimate $\hat{\zeta}_j$. Let $g(\lambda,\theta) = \left|\theta(e^{i\lambda})\right|^2$ with $\theta(z) = 1 - \theta_1 z - \ldots\theta_{m-1}z^{m-1}$. Define the parameter space $\Theta_1 \subset \mathbb{R}^{m-1}$ such that $\theta = (\theta_1, \ldots, \theta_{m-1}) \in \Theta_1$ if $\theta(z) \neq 0$ for $|z| \leq 1$. By Assumption (B) $\exists\ \Theta_2 \subset \operatorname{int}\Theta_1, \Theta_2$ compact such that $\theta_0 \in \Theta_2$.

The periodogram of $\hat{\varepsilon}_t$ is $\hat{I}^{\varepsilon}_n(\lambda) = n^{-1}\sum_{t,s}\hat{\varepsilon}_t\hat{\varepsilon}_s e^{i\lambda(t-s)}$. The maximum likelihood estimator for θ is asymptotically equivalent to

$$\tilde{\theta} = \arg\min_{\theta}\Lambda^{\hat{\varepsilon}}_n(\theta) \tag{A.25}$$

with $\Lambda^{\hat{\varepsilon}}_n(\theta) = n^{-1}\sum_j \hat{I}^{\varepsilon}_n(\lambda_j)/g(\lambda_j,\theta)$ for $\lambda_j = 2\pi j/n$, $j = -n+1, \ldots, 0, \ldots, n-1$. Define $I^{\varepsilon}_n(\lambda) = n^{-1}\sum_{t,s}\varepsilon_t\varepsilon_s e^{i\lambda(t-s)}$, $I^{\varepsilon x}_n(\lambda) = n^{-1}\sum_{t,s}\varepsilon_t(x_s - \mu_x)e^{i\lambda(t-s)}$, $I^{x\alpha}_n(\lambda) = n^{-1}(\hat{\alpha}_0 - \alpha_0)\sum_{t,s}(x_t - \mu_x)e^{i\lambda(t-s)}$, $I^{\varepsilon\alpha}_n(\lambda) = n^{-1}(\hat{\alpha}_0 - \alpha_0)\sum_{t,s}\varepsilon_t e^{i\lambda(t-s)}$ and $I^{\alpha}_n(\lambda) = n^{-1}(\hat{\alpha}_0 - \alpha_0)^2\sum_{t,s}e^{i\lambda(t-s)}$ for $\hat{\alpha}_0 - \alpha_0 = \bar{y} - \mu_y - \tilde{\beta}'(\bar{x} - \mu_x)$. It follows that

$$\begin{aligned}\hat{I}^{\varepsilon}_n(\lambda) &= I^{\varepsilon}_n(\lambda) + (\tilde{\beta} - \beta)'I^{x}_n(\lambda)(\tilde{\beta} - \beta) + I^{\alpha}_n(\lambda)\\ &\quad + 2\left(\tilde{\beta} - \beta\right)' I^{\varepsilon x}_n(\lambda) + 2I^{\varepsilon\alpha}_n(\lambda) + 2(\hat{\alpha}_0 - \alpha_0)(\tilde{\beta} - \beta)'I^{\alpha x}_n(\lambda).\end{aligned}$$

Note that $I^{\alpha}_n(\lambda_j) = I^{x\alpha}_n(\lambda_j) = I^{\varepsilon\alpha}_n(\lambda_j) = 0$ for $j \neq 0$ and $I^{\alpha}_n(\lambda_j) = n(\hat{\alpha}_0 - \alpha_0)^2$, $I^{\varepsilon\alpha}_n(\lambda_j) = (\hat{\alpha}_0 - \alpha_0)\sum_t\varepsilon_t$ for $j = 0$. We now have

$$\begin{aligned}\Lambda^{\hat{\varepsilon}}_n(\theta) &= \Lambda^{\varepsilon}_n(\theta) + 2(\tilde{\beta} - \beta)'\Lambda^{\varepsilon x}_n(\theta) + (\tilde{\beta} - \beta)'\Lambda^{x}_n(\theta)(\tilde{\beta} - \beta)\\ &\quad + \left[2(\hat{\alpha}_0 - \alpha_0)n^{-1}\sum_t(\varepsilon_t + (x_t - \mu_x)) + (\hat{\alpha}_0 - \alpha_0)^2\right]/g(0,\theta).\end{aligned}$$

From standard arguments (see Brockwell and Davis 1991, ch. 10) it follows that $\Lambda^{ab}_n(\theta) \xrightarrow{a.s.} \Lambda^{ab}(\theta)$ with $\Lambda^{ab}(\theta) = 2\pi\int f_{ab}(\lambda)/g(\lambda,\theta)d\lambda$ and $\partial^k\Lambda^{ab}_n(\theta)/\partial\theta \xrightarrow{a.s.} \partial^k\Lambda^{ab}(\theta)/\partial\theta$ for $k < \infty$ such that $\Lambda^{\hat{\varepsilon}}_n(\theta) \to 2\pi\int f_{\varepsilon\varepsilon}(\lambda)/g(\lambda,\theta)d\lambda$ uniformly in $\theta \in \Theta_2$. Consistency of $\tilde{\theta}$ follows from standard arguments.

To establish $\sqrt{n}$-consistency note that $\sqrt{n}\partial\Lambda_n^{\varepsilon}(\theta_0)/\partial\theta = O_p(1)$, $n^{-1/2}\sum_t \varepsilon_t = O_p(1)$ and

$$n^{-1/2}\sum_t (x_t - \mu_x) = O_p(1).$$

Therefore

$$\sqrt{n}\partial\Lambda_n^{\hat{\varepsilon}}(\theta)/\partial\theta = \sqrt{n}\partial\Lambda_n^{\varepsilon}(\theta)/\partial\theta + \sqrt{n}2(\tilde{\beta}-\beta)'\partial\Lambda^{\varepsilon x}(\theta)/\partial\theta + o_p(1)\,. \tag{A.26}$$

We also define $\partial\Lambda^{\varepsilon}(\theta)/\partial\theta = 2\pi\int f_{\varepsilon}(\lambda)\partial g^{-1}(\lambda,\theta)/\partial\theta d\lambda$ such that

$$\left\|\frac{\partial\Lambda^{\varepsilon}(\tilde{\theta})}{\partial\theta}\right\| \leq \left\|\frac{\partial\Lambda^{\varepsilon}(\tilde{\theta})}{\partial\theta} - \frac{\partial\Lambda_n^{\hat{\varepsilon}}(\tilde{\theta})}{\partial\theta}\right\| + \left\|\frac{\partial\Lambda_n^{\hat{\varepsilon}}(\tilde{\theta})}{\partial\theta} - \frac{\partial\Lambda_n^{\hat{\varepsilon}}(\theta_0)}{\partial\theta}\right\| + \left\|\frac{\partial\Lambda_n^{\hat{\varepsilon}}(\theta_0)}{\partial\theta}\right\| \tag{A.27}$$

where $\|\partial\Lambda_n^{\hat{\varepsilon}}(\theta_0)/\partial\theta\| = O_p(n^{-1/2})$ by (A.26). Definition (A.25) for $\tilde{\theta}$ implies that

$$\left\|\frac{\partial\Lambda_n^{\hat{\varepsilon}}(\tilde{\theta})}{\partial\theta} - \frac{\partial\Lambda_n^{\hat{\varepsilon}}(\theta_0)}{\partial\theta}\right\| \leq 2\left\|\frac{\partial\Lambda_n^{\hat{\varepsilon}}(\theta_0)}{\partial\theta}\right\| = O_p(n^{-1/2}).$$

Finally

$$\begin{aligned}\frac{\partial\Lambda_n^{\hat{\varepsilon}}(\tilde{\theta})}{\partial\theta} - \frac{\partial\Lambda^{\varepsilon}(\tilde{\theta})}{\partial\theta} &= \partial\Lambda_n^{\varepsilon}(\tilde{\theta})/\partial\theta - \int 2\pi f_{\varepsilon}(\lambda)\partial g^{-1}(\lambda,\tilde{\theta})/\partial\theta d\lambda \\ &\quad +2(\tilde{\beta}-\beta)'(\partial(\Lambda^{\varepsilon x}(\theta_0)))/\partial\theta + o_p(1)\end{aligned}$$

where the second term is $O_p(n^{-1/2})$ since $(\tilde{\beta}-\beta) = O_p(n^{-1/2})$. The first term can be written as

$$\begin{aligned}\partial\Lambda_n^{\varepsilon}(\tilde{\theta})/\partial\theta &- \int 2\pi f_{\varepsilon}(\lambda)\partial g^{-1}(\lambda,\tilde{\theta})/\partial\theta d\lambda \\ &= n^{-1}\sum_j\left[I_n^{\varepsilon}(\lambda_j) - 2\pi f_{\varepsilon}(\lambda_j)\right]\partial g^{-1}(\lambda_j,\tilde{\theta})/\partial\theta \\ &\quad + n^{-1}\sum_j 2\pi f_{\varepsilon}(\lambda_j)\partial g^{-1}(\lambda_j,\tilde{\theta})/\partial\theta - \int 2\pi f_{\varepsilon}(\lambda)\partial g^{-1}(\lambda,\tilde{\theta})/\partial\theta d\lambda\end{aligned}$$

where the second term is $O(n^{-1})$. Now define $\xi_j(\theta) = (2\pi)^{-1}\int \partial g^{-1}(\lambda,\theta)/\partial\theta e^{i\lambda j}d\lambda$ such that $\partial g^{-1}(\lambda,\theta)/\partial\theta = \sum_j \xi_j(\theta)e^{-i\lambda j}$ and

$$\begin{aligned}& n^{-1}\sum_j\left[I_n^{\varepsilon}(\lambda_j) - 2\pi f_{\varepsilon}(\lambda_j)\right]\partial g^{-1}(\lambda_j,\tilde{\theta})/\partial\theta \\ =\;& n^{-2}\sum_j\sum_{t,s=1}^{n}\sum_{l=-\infty}^{\infty}\left(\varepsilon_t\varepsilon_s - E\varepsilon_t\varepsilon_s\right)\xi_l(\tilde{\theta})e^{i\lambda_j(t-s-l)} + O_p(n^{-1}) \\ =\;& n^{-1}\sum_{l=-n}^{n}\sum_{t=\max(l,1)}^{n-|\min(l,0)|}\left(\varepsilon_t\varepsilon_{t-l} - E\varepsilon_t\varepsilon_{t-l}\right)\xi_l(\tilde{\theta}) + O_p(n^{-1}) \\ \leq\;& \left(\sum_{\substack{l=-n\\ l\neq 0}}^{n}|l|^{-2}\left(n^{-1}\sum_t\left(\varepsilon_t\varepsilon_{t-l} - E\varepsilon_t\varepsilon_{t-l}\right)\right)^2\right)^{1/2}\left(\sum_{l=-n}^{n}|l|_l^2\,\xi_l(\tilde{\theta})^2\right)^{1/2} + n^{-1}\sum_t\left(\varepsilon_t^2 - E\varepsilon_t^2\right)\end{aligned}$$

where the second equality follows from $n^{-1}\sum_j e^{i\lambda_j(t-s)} = 0$ for $t \neq s$ and the inequality follows from the Cauchy-Schwarz inequality. Then note that $n^{-1}\sum_t (\varepsilon_t^2 - E\varepsilon_t^2) = O_p(n^{-1/2})$,

$$E\sum_{\substack{l=-n\\l\neq 0}}^{n} |l|^{-2}\left(n^{-1}\sum\nolimits_t (\varepsilon_t\varepsilon_{t-l} - E\varepsilon_t\varepsilon_{t-l})\right)^2 = \sum_{\substack{l=-n\\l\neq 0}}^{n} |l|^{-2} n^{-2}\sum\nolimits_t \left[E\varepsilon_t^2\varepsilon_{t-l}^2 - (E\varepsilon_t\varepsilon_{t-l})^2\right] = O(n^{-1}),$$

and $\sum_{l=-n}^{n} |l|^2 c_l(\theta)^2$ is uniformly converging for θ with $|\theta - \theta_0| < \delta$ for some $\delta > 0$ such that $\theta(z)$ has no zeros on or inside the unit circle. Consistency of $\tilde{\theta}$ then implies $\sum_{l=-n}^{n} |l|^2 c_l(\tilde{\theta})^2 = O_p(1)$. These results establish that $\left\| \frac{\partial \Lambda_n^{\hat{\varepsilon}}(\tilde{\theta})}{\partial \theta} - \frac{\partial \Lambda^{\varepsilon}(\tilde{\theta})}{\partial \theta} \right\| = O_p(n^{-1/2})$. From (A.27) it then follows that $\left\| \frac{\partial \Lambda^{\varepsilon}(\tilde{\theta})}{\partial \theta} \right\| = O_p(n^{-1/2})$ such that by a continuity argument $\sqrt{n}(\tilde{\theta} - \theta) = O_p(1)$.

To show consistency of $\sum_j \hat{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{\hat{\varepsilon}x} = \sum_j \hat{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{yx} + \tilde{\beta}'\sum_j \hat{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{xx}$ we consider without loss of generality $\sum_j \hat{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{yy}$ since x_t is composed of elements of $y_t, y_{t-1}, ..., y_{t-r}$. We next show that $\sum_{j=-n+1}^{n-1} \tilde{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{yy} - \sum_{j=-n+1}^{n-1} \zeta_{j+m}\Gamma_{j-k}^{yy} = O_p(n^{-1/2})$. Write

$$\sum_{j=-n+1}^{n-1} \tilde{\zeta}_{j+m}\hat{\Gamma}_{j-k}^{yy} - \sum_j \zeta_{j+m}\Gamma_{j-k}^{yy} = \sum_{j=-n+1}^{n-1} \tilde{\zeta}_{j+m}\left(\hat{\Gamma}_{j-k}^{yy} - \Gamma_{j-k}^{yy}\right) - \sum_{j=-n+1}^{n-1}\left(\tilde{\zeta}_{j+m} - \zeta_{j+m}\right)\Gamma_{j-k}^{yy}.$$

First consider

$$n^{1/2}\sum_{j=-n+1}^{n-1} \left\|\tilde{\zeta}_j - \zeta_j\right\| \left\|\Gamma_{j-k}^{xy}\right\| \leq n^{1/2}\sup_j \left\|\tilde{\zeta}_j - \zeta_j\right\| \sum_{j=-n+1}^{n-1} \left\|\Gamma_{j-k}^{yy}\right\|$$

where $P(\sup_j \left\|\tilde{\zeta}_j - \zeta_j\right\| > Cn^{-1/2})$ goes to zero for some C large enough by the previous result.

For any δ such that $|\theta - \theta_0| < \delta$ implies $\theta(z)$ has no zeros on or inside the unit circle consider

$$P\left(n^{1/2}\sum\nolimits_{j=-n+1}^{n-1} \left|\tilde{\zeta}_j\right| \left\|\hat{\Gamma}_{j-k}^{yy} - \Gamma_{j-k}^{yy}\right\| > \eta\right) \leq P\left(n^{1/2}\sup_{|\theta-\theta_0|<\delta}\sum\nolimits_{j=-n+1}^{n-1} |\zeta_k(\theta)| \left\|\hat{\Gamma}_{j-k}^{yy} - \Gamma_{j-k}^{yy}\right\| > \eta\right)$$
$$+P\left(\left|\tilde{\theta} - \theta_0\right| \geq \delta\right)$$

We use the triangular inequality $\left\|\hat{\Gamma}_{j-k}^{yy} - \Gamma_{j-k}^{yy}\right\| \leq \left\|\hat{\Gamma}_{j-k}^{yy} - \check{\Gamma}_{j-k}^{yy}\right\| + \left\|\check{\Gamma}_{j-k}^{yy} - \check{\Gamma}_{j-k}^{yy}\right\|$ such that

$$n^{1/2}\sup_{|\theta-\theta_0|<\delta}\sum\nolimits_{j=-n+1}^{n-1} |\zeta_j(\theta)| \left\|\hat{\Gamma}_{j-k}^{yy} - \check{\Gamma}_{j-k}^{yy}\right\| = O_p(1)$$

by Equation (B.9) and the fact that $\sup_{|\theta-\theta_0|<\delta}\sum_{j=-n+1}^{n-1} |\zeta_j(\theta)| = O(1)$ uniformly in n. In the same way it follows from Equation (B.10) that

$$n^{1/2}\sup_{|\theta-\theta_0|<\delta}\sum\nolimits_{j=-n+1}^{n-1} |\zeta_j(\theta)| E\left\|\check{\Gamma}_{j-k}^{yy} - \Gamma_{j-k}^{yy}\right\| = O(1).$$

This establishes that $\widehat{\mathcal{A}}_1 - \mathcal{A}_1 = O_p(n^{-1/2})$. The result then follows by Lemma (B.47). ■

Proof of Proposition (4.3): Let $\hat{L}_n(M) = -\log \hat{\sigma}_{M,\hat{h}} + \frac{M^2}{n}\widehat{\mathcal{A}}\left(\int_{-\infty}^{\infty}\phi(x)^2dx\right)^2$ and $L_n(M) = -\log \sigma_M + \frac{M^2}{n}\mathcal{A}\left(\int_{-\infty}^{\infty}\phi(x)^2dx\right)^2$. We first show that uniformly in M,

$$\hat{L}_n(M) = L_n(M)\left(1 + O_p(n^{-1/2}(\log n)^{1/2+s'}\right) \tag{A.28}$$

with $s' = s_1\left\{\lambda_1 \geq \tilde{\lambda}_1\right\} + \tilde{s}_1\left\{\lambda_1 \leq \tilde{\lambda}_1\right\}$. Consider

$$\left|\frac{\hat{L}_n(M) - L_n(M)}{L_n(M)}\right| \leq \left|\frac{\log \hat{\sigma}^2_{M,\hat{h}} - \log \sigma^2_M}{\log \sigma^2_M}\right| + \frac{\left(\mathcal{A} - \widehat{\mathcal{A}}\right)}{\mathcal{A}}$$

because $L_n(M) \geq M^2/n\mathcal{A}\left(\int_{-\infty}^{\infty}\phi(x)^2dx\right)^2$ and $\sigma^2_M \geq 1$. By Proposition (4.1) and Lemma (B.47) it follows that $\widehat{\mathcal{A}} - \mathcal{A} = O_p(n^{-1/2})$. Next, note that $\log \sigma^2_M = c_1 M^{2s}\lambda^{2M} + o\left(M^{2s}\lambda^{2M}\right)$ and $\log \hat{\sigma}^2_{M,\hat{h}} = \hat{\sigma}^2_{M,\hat{h}} - 1 + o_p(\hat{\sigma}^2_{M,\hat{h}} - 1)$. Let $g(M) = |M|^s \lambda^{|M|}$, $g'(M) = |M|^{s+s'/2}\lambda^{|M|}$, $g_\Gamma(M) = |M|^{s_1-1}\lambda_1^{|M|}$ and $g_\vartheta(M) = |M|^{\tilde{s}_1-1}\tilde{\lambda}_1^{|M|}$. First show

$$\begin{aligned} g(M)^{-2}M^{-s'}\left(\hat{\sigma}^2_{1M,\hat{h}} - \sigma^2_{1M}\right) &= g(M)^{-2}M^{-s'}\left|\ell'\hat{D}_{\hat{h}}^{-1/2}\hat{H}_{11,\hat{h}}\hat{D}_{\hat{h}}^{-1/2}\ell - \ell' D^{-1/2}H_{11}D^{-1/2}\ell\right| \\ &= O_p(n^{-1/2}(\log n)^{1/2}) \end{aligned} \tag{A.29}$$

with $\hat{H}_{11,\hat{h}} = \hat{P}'_{M,\hat{h}}\hat{\Omega}^{-1}_{M,\hat{h}}\hat{P}_{M,\hat{h}} - \hat{D}_{\hat{h}}$. It is enough to show that $\hat{H}_{11,\hat{h}} - H_{11} = O_p(n^{-1/2}(\log n)^{1/2}g'(M)^2)$. First we analyze individual components of $\hat{H}_{11,\hat{h}} - H_{11}$. Note that $\left\|\hat{\Gamma}^{xy}_{j,\hat{h}} - \Gamma^{xy}_j\right\| = O_p(n^{-1/2}(\log n)^{1/2}jg_\Gamma(j))$ by Lemma (B.46) such that $\left\|\hat{\Gamma}^{yx}_{j,\hat{h}}\right\| \leq \left\|\Gamma^{yx}_j\right\|(1 + O_p(n^{-1/2}(\log n)^{1/2}j)$ because $\left\|\Gamma^{yx}_{-j}\right\| = O(g_\Gamma(j))$. To analyze terms involving ϑ_{j_1,j_2+M} we use the expansion $\hat{\Omega}^{-1}_{\hat{h}} - \Omega^{-1} = \Omega^{-1}\left(\Omega - \hat{\Omega}_{\hat{h}}\right)\Omega^{-1} + o_p\left(n^{-1/2}\sqrt{\log n}\right)$ such that

$$\hat{\vartheta}_{j_1,j_2+M,\hat{h}} - \vartheta_{j_1,j_2+M} = \textstyle\sum_{k,l=1}^{\infty}\vartheta_{j_1,k}\left(\hat{\omega}_{k,l,\hat{h}} - \omega_{k,l}\right)\vartheta_{l,j_2+M}. \tag{A.30}$$

Note that

$$\begin{aligned} \left\|\hat{\omega}_{k,i,\hat{h}} - \omega_{k,i}\right\| &\leq \sum_{l=-m+1}^{m-1}\left(|\hat{\gamma}^\varepsilon_l - \gamma^\varepsilon_l|\left\|\Gamma^{yy}_{i+l-k}\right\| + |\hat{\gamma}^\varepsilon_l|\left\|\hat{\Gamma}^{yy}_{i+l-k,\hat{h}} - \Gamma^{yy}_{i+l-k}\right\|\right) \\ &= O_p(n^{-1/2}|i-k|\,g_\Gamma(i-k)(\log n)^{1/2}) \end{aligned}$$

by Lemma (B.46). It then follows that

$$\left\|\hat{\vartheta}_{j_1,j_2+M,\hat{h}} - \vartheta_{j_1,j_2+M}\right\| = O_p(n^{-1/2}(\log n)^{1/2}M^{s_2}\lambda^M)$$

where $s_2 = 2\tilde{s}_1 - 1$ if $\lambda_1 < \tilde{\lambda}_1$, $s_2 = s_1 + 2\tilde{s}_1$ if $\lambda_1 = \tilde{\lambda}_1$ and $s_2 = s_1$ if $\lambda_1 > \tilde{\lambda}_1$. This can be seen by noting that $\sum_{k,l=1}^{\infty}\vartheta_{j_1,k}\left(\hat{\omega}_{k,l,\hat{h}} - \omega_{k,l}\right)\vartheta_{l,j_2+M}$ is of the same order as the M-th autocorrelation of a

AR process with lag polynomial $(1-\lambda_1 L)^{s_1+1}\left(1-\tilde{\lambda}_1 L\right)^{\tilde{s}_1}$. Consider now the largest order element of $\hat{H}_{11,\hat{h}}-H_{11}$, namely $\hat{P}'_{M,\hat{h}}\left(\hat{\Omega}^{-1}_{M,\hat{h}}-\left[\hat{\Omega}^{-1}_{\hat{h}}\right]_M\right)\hat{P}_{M,\hat{h}}-P'_M\left(\Omega^{-1}_M-\left[\Omega^{-1}\right]_M\right)P_M$. The largest order terms in this difference are of the form

$$\sum_{j_1,\dots,j_6=1}^{\infty}\Gamma^{xy}_{j_1}\left(\hat{\vartheta}_{j_1,j_2,\hat{h}}-\vartheta_{j_1,j_2}\right)(\omega_{j_2,j_3+M}\vartheta_{j_3,j_4}\omega_{j_4+M,j_5})\vartheta_{j_5,j_6}\Gamma^{yx}_{-j_6}$$

and taking into account (A.30) we need to consider

$$\sum_{j_1,\dots,j_8=1}^{\infty}\Gamma^{xy}_{j_1}\vartheta_{j_1,j_2}\left(\hat{\omega}_{j_2,j_3,\hat{h}}-\omega_{j_2,j_3}\right)\vartheta_{j_3,j_4}(\omega_{j_4,j_5+M}\vartheta_{j_5,j_6}\omega_{j_6+M,j_7})\vartheta_{j_7,j_8}\Gamma^{yx}_{-j_8}$$
$$=O_p(n^{-1/2}\sqrt{\log n}M^{s_3}\lambda^{2M})$$

where $s_3=4\tilde{s}_1-1$ if $\lambda_1<\tilde{\lambda}_1$, $s_2=5s_1+4\tilde{s}_1-1$ if $\lambda_1=\tilde{\lambda}_1$ and $s_2=5s_1-1$ if $\lambda_1>\tilde{\lambda}_1$ where $s_3=2s+s'$. The remaining terms of $\hat{H}_{11,\hat{h}}-H_{11}$ are of smaller order. This establishes (A.29).

Next, show that $\sqrt{n}\sup\left|\hat{\sigma}_{2M,\hat{h}}-\sigma_{2M}\right|/g'(M)^2=O_p(1)$. We focus on a typical term in $\hat{\sigma}_{2M,\hat{h}}$, the matrix

$$\hat{S}_M=\hat{P}'_{M,h}\left(I_M-\hat{K}_M\right)\hat{\Omega}^{-1}_{M,\hat{h}}\left(I_M-\hat{K}_M\right)\hat{P}_{M,h}$$

with population analogue S_M. Other terms of $\hat{\sigma}_{2M,\hat{h}}$ can be handled in a similar way. Let $s_M=\sqrt{\log\sigma_{1M}}$, $\hat{s}_M=\sqrt{\log\hat{\sigma}_{1M,\hat{h}}}$, $\phi_j=\phi(k(j/M),s_M)$ and $\hat{\phi}_j=\phi(k(j/M),\hat{s}_M)$. The matrix $\hat{K}_M$ contains diagonal elements $\hat{\phi}_j$. Note that $1-\phi(v,z)=(1-v)(1-v+vz\,(-\log z)^q)$ such that $\left|\hat{\phi}_j-\phi_j\right|\leq|1-k(j/M)|\,|\hat{s}_M\,(-\log\hat{s}_M)^q-s_M\,(-\log s_M)^q|$. Also let $A_{j_1,j_2}=\Gamma^{xy}_{j_1}\vartheta^{M-1}_{j_1j_2}\Gamma^{yx}_{-j_2}$ and $\hat{A}^M_{j_1,j_2}=\hat{\Gamma}^{xy}_{j_1,\hat{h}}\hat{\vartheta}^{M-1}_{j_1j_2,\hat{h}}\hat{\Gamma}^{yx}_{-j_2,\hat{h}}$. Now,

$$\sum_{j_1,j_2}\left(\left(1-\hat{\phi}_{j_1}\right)\left(1-\hat{\phi}_{j_2}\right)-\left(1-\phi_{j_1}\right)\left(1-\phi_{j_2}\right)\right)\hat{A}^M_{j_1,j_2}$$
$$=\sum_{j_1,j_2}\left\{\left(\hat{\phi}_{j_1}-\phi_{j_1}\right)\left(\hat{\phi}_{j_2}-\phi_{j_2}\right)+\left(\phi_{j_1}-\hat{\phi}_{j_1}\right)\left(1-\phi_{j_2}\right)+\left(1-\phi_{j_1}\right)\left(\phi_{j_2}-\hat{\phi}_{j_2}\right)\right\}\hat{A}^M_{j_1,j_2}$$

such that

$$\left\|\hat{S}_M-S_M\right\|/g'(M)^2$$
$$\leq\ 2\sup_M\left|\frac{\hat{s}_M}{g'(M)}\left(\frac{-\log\hat{s}_M}{M}\right)^q-\frac{s_M}{g'(M)}\left(\frac{-\log s_M}{M}\right)^q\right|^2\sum_{j_1j_2}^{M}\left|\frac{1-k(j_1/M)}{(j_1/M)^q}\right|\left|\frac{1-k(j_2/M)}{(j_2/M)^q}\right|\left\|\hat{A}_{j_1,j_2}\right\|$$
$$+2\sup_M\left|\frac{\hat{s}_M}{g'(M)}\left(\frac{-\log\hat{s}_M}{M}\right)^q-\frac{s_M}{g'(M)}\left(\frac{-\log s_M}{M}\right)^q\right|\sum_{j_1j_2}^{M}\left|\frac{1-k(j_1/M)}{(j_1/M)^q}\right|\left|\frac{1-\phi(k(j_2/M),s_M)}{g(M)}\right|\left\|\hat{A}_{j_1,j_2}\right\|$$
$$+\sum_{j_1j_2}^{M}\left|\frac{1-\phi(k(j_1/M),s_M)}{g'(M)}\right|\left|\frac{1-\phi(k(j_1/M),s_M)}{g'(M)}\right|\left\|\hat{A}^M_{j_1,j_2}-A^M_{j_1,j_2}\right\|.$$

where

$$\sup_M \left| \frac{\hat{s}_M}{g'(M)} \left(\frac{-\log \hat{s}_M}{M} \right)^q - \frac{s_M}{g'(M)} \left(\frac{-\log s_M}{M} \right)^q \right|$$
$$\leq c_1 \sup \left| (\hat{s}_M - s_M) / g'(M) \right| + c_2 \sup \left| \left(\frac{-\log \hat{s}_M}{M} \right)^q - \left(\frac{-\log s_M}{M} \right)^q \right| = O_p(n^{-1/2} (\log n)^{1/2})$$

where $c_1 = \sup_M \left(-\log \left(M^s \lambda^M\right) / M\right)^q$ and $c_2 = 2 \sup |\sigma_{1M}/g'(M)|$. The result then follows because $\left(\log \left(M^s \lambda^M\right) / M\right)$ is uniformly continuous in M and $\sup |(\hat{s}_M - s_M) / g'(M)| = O_p(n^{-1/2} (\log n)^{1/2})$ by (A.29). Also note that $\sum_{j_1 j_2}^M \left\| \hat{\Gamma}^{xy}_{j_1, \hat{h}} \right\| \left\| \hat{\vartheta}^{M-1}_{j_1 j_2, \hat{h}} \right\| \left\| \hat{\Gamma}^{yx}_{-j_2, \hat{h}} \right\| = O_p(1)$ with probability going to one. By the arguments in the proof of Proposition (3.4) it follows that $\left| \left(1 - \phi(k(j_1/M), \sqrt{\log \sigma_{1M}})\right) / g'(M) \right| \to 0$ as $M \to \infty$. By the same arguments as in the proof of (A.29) it can be shown that

$$\sum_{j_1 j_2}^M |j_1|^q |j_2|^q \left\| \hat{\Gamma}^{xy}_{j_1, \hat{h}} \hat{\vartheta}^{M-1}_{j_1 j_2, \hat{h}} \hat{\Gamma}^{yx}_{-j_2, \hat{h}} - \Gamma^{xy}_{j_1} \vartheta^{M-1}_{j_1 j_2} \Gamma^{yx}_{-j_2} \right\| = O_p(n^{-1/2} (\log n)^{1/2})$$

such that the third term of the bound for $\left\| \hat{S}_M - S_M \right\| / g'(M)^2$ is $o_p(n^{-1/2} (\log n)^{1/2})$. It thus follows that $\left\| \hat{S}_M - S_M \right\| / g'(M)^2 = O_p(n^{-1/2} (\log n)^{1/2})$ uniformly in M. We have therefore established (A.28).

Let $\tilde{L}_n(M) = L_n(M) + g(M)^2 + \log \sigma_M$ where $g(M)^2 + \log \sigma_M = O(\lambda_r^{2M})$ with $\lambda_r < \lambda$ by Hannan and Kavalieris (1986, p.47). Since $\lambda_r^{2M} / g(M)^2 = (\lambda_r / \lambda)^{2M} M^{-2s} \to 0$ as $M \to \infty$ it follows that $L_n(M) = \tilde{L}_n(M) \left(1 + O(g_r(M)^2)\right)$ with $g_r(M) = M^s (\lambda_r / \lambda)^M$. Let $\tilde{M}^*$ minimize $\tilde{L}_n(M)$. By the same arguments as in Hannan and Deistler (1988) it now follows that $\hat{M}^* / \tilde{M}^* - 1 = o_p(1)$ and $L_n(\hat{M}^*) / L_n(\tilde{M}^*) = 1 + o_p(1)$. This establishes the first part of the Proposition.

Moreover, optimality of M^* implies that $-\log \sigma_{M^*+1} + \log \sigma_{M^*} + \frac{2M^*+1}{n} c \geq 0$ for some constant c. This leads to

$$\log n + 2s \log (M^* + 1) + 2 (M^* + 1) \log \lambda \leq \log \left(2M^* + 1 + o(g(M^*)^2)\right)$$

or

$$\frac{\log n}{M^*} \leq \frac{\log \left(2M^* + 1 + o(g(M^*)^2)\right) - 2s \log (M^* + 1) - 2 \log \lambda}{M^*} - 2 \log \lambda \to -2 \log \lambda.$$

In a similar way we note that $-\log \sigma_{M^*} + \log \sigma_{M^*-1} + \frac{2M^*-1}{n} c \leq 0$ such that

$$\frac{\log n}{M^*} \geq \frac{\log \left(2M^* - 1 + o(g(M^*)^2)\right) - 2s \log (M^* + 1)}{M^*} - 2 \log \lambda \to -2 \log \lambda.$$

Thus, $\log n / M^* = O(1)$. This implies that $\hat{M}^* = -\log n / (2 \log \lambda) + o_p(\log n)$. Optimality of $\hat{M}^*$ then implies that $\hat{L}_n \left(\hat{M}^*\right) = O_p((\log n)^2 / n)$ and $\log \hat{\sigma}_{\hat{M}^*, \hat{h}} = O_p((\log n)^2 / n)$. We have seen before that $\log \hat{\sigma}_{\hat{M}^*, \hat{h}} = \log \sigma_{\hat{M}^*} (1 + O_p(n^{-1/2} (\log n)^{1/2+s'})$ such that $\log \sigma_{\hat{M}^*} = O_p((\log n)^2 / n)$ which in turn implies that $g(\hat{M}^*) = \lambda^{\hat{M}^*}\hat{M}^{*s} = O_p(\log n/n^{1/2})$. Substituting for $\hat{M}^* = -\log n/(2\log\lambda) + e_{\hat{M}^*}$ with $e_{\hat{M}^*} = o_p(\log n)$ in $\lambda^{\hat{M}^*}\hat{M}^{*s}$ shows that $\lambda^{e_{\hat{M}^*}} = O_p(\log n)$ if $s = 0$ and $O_p(1)$ otherwise. Since $\lambda_r/\lambda < \lambda^{1/2}$ by assumption it follows that $(\lambda_r/\lambda)^{e_{\hat{M}}^*} = O_p(\sqrt{\log n})$ if $s = 0$ and $O_p(1)$ otherwise. Then, consider

$$g_r\left(\hat{M}^*\right) = (\lambda_r/\lambda)^{e_{\hat{M}^*}}(\lambda_r/\lambda)^{-\log n/2\log\lambda}\left(\frac{\log n}{-2\log\lambda}\right)^s + o_p(g_r(M^*)).$$

Note that $(\lambda_r/\lambda)^{e_{\hat{M}^*}}(\log n)^s = O_p((\log n)^{s+1/2})$ for all s such that $g_r\left(\hat{M}^*\right)^2 = O_p\left((\lambda_r/\lambda)^{-\log n/\log\lambda}(\log n)^{2s+1}\right)$ where

$$(\lambda_r/\lambda)^{-\log n/\log\lambda} = \left((\lambda_r/\lambda)^{-\log n/\log(\lambda_r/\lambda)}\right)^{\log\lambda_r/\log\lambda - 1} = n^{-(\log\lambda_r/\log\lambda - 1)}.$$

But $(\log\lambda_r/\log\lambda - 1) > 1/2$ if $\lambda_r < \lambda^{3/2}$. Then $g_r\left(\hat{M}^*\right)^2 = o_p(n^{-1/2}(\log n)^{1/2})$ and

$$\tilde{L}_n(\tilde{M}^*) \leq \tilde{L}_n(\hat{M}^*) = L_n(\hat{M}^*)\left(1 + o(g_r(\hat{M}^*)^2)\right) = \hat{L}_n(\hat{M}^*)\left(1 + O_p(n^{-1/2}(\log n)^{1/2+s'})\right)$$

where the last equality follows from (A.28) and the fact that $\hat{M}^* = O_p(\log n)$. Also, $\hat{L}_n(\hat{M}^*) \leq \hat{L}_n(\tilde{M}^*) = \tilde{L}_n(\tilde{M}^*)(1 + O_p(n^{-1/2}(\log n)^{1/2+s'}))$ by similar arguments as before. It follows that

$$\tilde{L}_n(\hat{M}^*)/\tilde{L}_n(\tilde{M}^*) = 1 + O_p(n^{-1/2}(\log n)^{1/2+s'}). \tag{A.31}$$

A second order mean value expansion of $\tilde{L}_n\left(\hat{M}^*\right)$ around $\tilde{M}^*$ leads to

$$\tilde{L}_n\left(\hat{M}^*\right) = \tilde{L}_n\left(\tilde{M}^*\right) + \frac{1}{2}\frac{\partial^2\tilde{L}_n\left(\bar{M}^*\right)}{\partial M^2}\left(\hat{M}^* - \tilde{M}^*\right)^2$$

where $\left|\bar{M} - \tilde{M}^*\right| \leq \left|\hat{M}^* - \tilde{M}^*\right|$ and we have used the fact that $\partial\tilde{L}_n\left(\tilde{M}^*\right)/\partial M = 0$. Let $\theta(M) = \partial^2\left(\tilde{L}_n(M)\right)/(\partial M)^2$. Then,

$$\left(\frac{\hat{M}^*}{\tilde{M}^*} - 1\right)^2 = O_p\left(\frac{\left|\tilde{L}_n(\hat{M}^*) - \tilde{L}_n(\tilde{M}^*)\right|}{\tilde{L}_n(\tilde{M}^*)}\frac{\tilde{L}_n(\tilde{M}^*)}{2\tilde{M}^{*2}\theta(\bar{M}^*)}\right) = O_p(n^{-1/2}(\log n)^{1/2+s'})$$

follows from (A.31), $\tilde{L}_n(\tilde{M}^*) \leq \tilde{L}_n(\bar{M}^*)$ and

$$\tilde{L}_n(\bar{M}^*)/\left(2\tilde{M}^{*2}\theta(\bar{M}^*)\right) = O_p(1).$$

This result implies that $\hat{M}^* - \tilde{M}^* = O_p(n^{-1/4}\tilde{M}^*(\log n)^{1/4+s'/2}) = o_p(1)$. Similarly, we note that for M^* maximizing $L_n(M)$ we have

$$\left(\frac{M^*}{\tilde{M}^*} - 1\right)^2 = \frac{\left|\tilde{L}_n(M^*) - \tilde{L}_n(\tilde{M}^*)\right|}{\tilde{L}_n(\tilde{M}^*)}\frac{\tilde{L}_n(\bar{M}^*)}{2\bar{M}^{*2}\theta(\bar{M}^*)} = o(g_r(M^*)^2) \tag{A.32}$$

since $L_n(M^*) \leq L_n(\tilde{M}^*) = \tilde{L}_n(\tilde{M}^*)\left(1 + o(g_r(\tilde{M}^*)^2)\right)$ and $\tilde{L}_n(\tilde{M}^*) \leq \tilde{L}_n(M^*) = L(M^*)\left(1 + o(g_r(M^*)^2)\right)$ where $g_r(M^*)^2 = O(n^{-1/2}(\log n)^{1/2})$. We have thus shown that $\hat{M}^* - M^* = O_p(n^{-1/4}M^*(\log n)^{1/4})$. We use this result to sharpen the convergence result for $\hat{M}^*$. By the same arguments as in Hannan and Deistler (1988, p.333-334) optimality of $\hat{M}^*$ implies that

$$\tilde{L}_n(\tilde{M}^*) \leq \tilde{L}_n(\hat{M}^*) \leq \tilde{L}_n(\tilde{M}^*)(1 + O_p(n^{-1/2}(\log n)^{1/2+s'}))$$

or, by rewriting the inequalities as $0 \leq \tilde{L}_n(\hat{M}^*) - \tilde{L}_n(\tilde{M}^*) \leq \tilde{L}_n(\tilde{M}^*)O_p(n^{-1/2}(\log n)^{1/2+s'})$,

$$0 \leq \frac{\hat{M}^{*2}}{\tilde{M}^{*2}} - 1 + \frac{\left(g(\hat{M}^*)^2 - g(\tilde{M}^*)^2\right)n}{\tilde{M}^{*2}} \leq \left(\frac{g(\tilde{M}^*)^2 n}{\tilde{M}^{*2}} + 1\right)O_p(n^{-1/2}(\log n)^{1/2+s'}) \tag{A.33}$$

where by a mean value expansion

$$\frac{\left(g(\hat{M}^*)^2 - g(\tilde{M}^*)^2\right)n}{\tilde{M}^{*2}} = \frac{\partial g(\bar{M}^*)^2/\partial M}{\tilde{M}^*}n\left(\frac{\hat{M}^*}{\tilde{M}^*} - 1\right)$$

with $\left|\bar{M}^* - \tilde{M}^*\right| \leq \left|\hat{M}^* - \tilde{M}^*\right|$. Note that $g(\tilde{M}^*)^2 n/\tilde{M}^{*2} = O(1)$ by optimality of $\tilde{M}^*$. Furthermore, $\partial g(M)^2/\partial M = g(M)^2(2s/M + 2\log\lambda)$. By the first order condition for $\tilde{M}^*$ it follows that $\partial g(\tilde{M}^*)^2/\partial M + 2\tilde{M}^*/n = 0$, or $n/\tilde{M}^*\partial g(\tilde{M}^*)^2/\partial M = -1/2$. From

$$\frac{\partial g(\bar{M}^*)^2/\partial M}{\partial g(\tilde{M}^*)^2/\partial M} = \left(\frac{\tilde{M}^*}{\bar{M}^*}\right)^{2s}\lambda^{2(\bar{M}^* - \tilde{M}^*)}\frac{\left(2s/\bar{M}^* + 2\log\lambda\right)}{(2s/\tilde{M}^* + 2\log\lambda)} \xrightarrow{p} 1$$

since $\tilde{M}^* - \bar{M}^* = o_p(1)$ by previous arguments it follows that $n/\tilde{M}^*\partial g(\bar{M}^*)^2/\partial M \xrightarrow{p} 1/2$. We now rewrite (A.33) as

$$0 \leq \frac{\hat{M}^*}{\tilde{M}^*} - 1 \leq \frac{\left(\frac{g(\tilde{M}^*)^2 n}{\tilde{M}^{*2}} + 1\right)}{\left(\frac{\hat{M}^*}{\tilde{M}^*} + 1 + \frac{\partial g(\bar{M}^*)^2/\partial M}{\tilde{M}^*}n\right)}O_p(n^{-1/2}(\log n)^{1/2+s'}) = O_p(n^{-1/2}(\log n)^{1/2+s'})$$

such that the result follows. ■

Proof of Theorem (4.4) The decomposition $\sqrt{n}\left(\hat{\beta}_{n,\hat{M}^*} - \hat{\beta}_{n,M^*}\right) = \hat{D}_{M^*}^{-1}\left(\hat{D}_{M^*} - \hat{D}_{\hat{M}^*}\right)\hat{D}_{\hat{M}^*}^{-1}\hat{d}_{\hat{M}^*} - \hat{D}_{M^*}^{-1}(\hat{d}_{\hat{M}^*} - \hat{d}_{M^*})$ is used. Note that $\hat{D}_{M^*} = O_p(1)$ and $d_{M^*} = O_p(1)$. The following calculations also establish $\hat{D}_{\hat{M}^*} = O_p(1)$ and $d_{\hat{M}^*} = O_p(1)$. It is therefore enough to show that $\sqrt{n/M^*}(\hat{D}_{M^*} - \hat{D}_{\hat{M}^*}) = O_p((\log n)^{\max(s'-q,-1)})$ and $\sqrt{n/M^*}(\hat{d}_{\hat{M}^*} - \hat{d}_{M^*}) = O_p((\log n)^{\max(s'-q,-1)})$. Define $\tilde{k}_M = \text{diag}(\phi(k(0), M^s\lambda^M), ..., \phi(k((n-m-1)/M), M^s\lambda^M))'$ and $\tilde{K}_M = (\tilde{k}_M \otimes I_p)$. We write $\tilde{K}_{\hat{M}^*}$ for matrices with elements $\phi(k((j)/\hat{M}^*), \hat{s}_{\hat{M}^*})$ where $\hat{s}_{\hat{M}^*}$ is defined in the proof of Theorem (4.3) and note that for the truncated kernel $\tilde{K}_{\hat{M}^*}$ is a matrix where the first $p\hat{M}^*$ diagonal elements are one and all the other elements are zero. Let $\hat{P}'_{n-m} = n^{-1}X'Z_{n-m}$ and

$$\Omega^*_M = \begin{bmatrix} \Omega_M & 0 \\ 0 & I_{n-M-m} \end{bmatrix}, \Omega^{*-1}_M = \begin{bmatrix} \Omega^{-1}_M & 0 \\ 0 & I_{n-M-m} \end{bmatrix}$$

with $\hat{\Omega}^*_M$ and $\hat{\Omega}^{*-1}_M$ defined in the same way replacing Ω_M and Ω^{-1}_M by $\hat{\Omega}_M$ and $\hat{\Omega}^{-1}_M$. Using these definitions we can rewrite $\hat{d}_M = n^{-1/2}\hat{P}'_{n-m}\tilde{K}_M\hat{\Omega}^{*-1}_M\tilde{K}_M Z'_{n-m}\varepsilon$. First consider

$$\begin{aligned}
\sqrt{n/M^*}(\hat{d}_{\hat{M}^*} - \hat{d}_{M^*}) &= \sqrt{n/M^*}(\hat{P}'_{n-m}\tilde{K}_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\tilde{K}_{\hat{M}^*}\frac{Z'_{n-m}\varepsilon}{\sqrt{n}} - \hat{P}'_{n-m}\tilde{K}_{M^*}\hat{\Omega}^{*-1}_{M^*}\tilde{K}_{M^*}\frac{Z'_{n-m}\varepsilon}{\sqrt{n}}) \\
&= \sqrt{n/M^*}(\hat{P}'_{n-m}\left[\Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*} + \tilde{K}_{M^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*} + \Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\tilde{K}_{M^*}\right. \\
&\quad \left. + \tilde{K}_{M^*}\left(\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*}\right)\tilde{K}_{M^*}\right]\frac{Z'_{n-m}\varepsilon}{\sqrt{n}})
\end{aligned}$$

with $\Delta_{\hat{M}^*} = \tilde{K}_{\hat{M}^*} - \tilde{K}_{M^*}$. From Assumption (E) it follows that for some constant c_1, $\left|k(j/\hat{M}^*) - k(j/M^*)\right| \leq c_1 |j|^q \left(\left(1/\hat{M}^*\right)^q - (1/M^*)^q\right)$. Then,

$$\begin{aligned}
&\left|\phi(k(j/\hat{M}^*), \hat{s}_{\hat{M}^*}) - \phi(k(j/M^*), s_{M^*})\right| \\
\leq\ & 2c_2 \left|\hat{s}_{\hat{M}^*}(-\log \hat{s}_{\hat{M}^*})^q - s_{M^*}(-\log s_{M^*})^q\right| + \left|2 - s_{M^*}(-\log s_{M^*})^q\right| \left|k(j/\hat{M}^*) - k(j/M^*)\right| \\
&+ \left|s_{M^*}(-\log s_{M^*})^q - 1\right| \left|k(j/\hat{M}^*)^2 - k(j/M^*)^2\right| \\
\leq\ & 2c_2 \left|\hat{s}_{\hat{M}^*}(-\log \hat{s}_{\hat{M}^*})^q - s_{M^*}(-\log s_{M^*})^q\right| + c_3 |j|^q 1/M^{*q}\left(\left(M^*/\hat{M}^*\right)^q - 1\right)
\end{aligned}$$

for some constants c_2 and c_3 because

$$\left|k(j/\hat{M}^*)^2 - k(j/M^*)^2\right| \leq 2\left|k(j/\hat{M}^*) - k(j/M^*)\right| \leq 2c_1 |j|^q 1/M^{*q}\left(\left(M^*/\hat{M}^*\right)^q - 1\right)$$

and $s_{M^*}(-\log s_{M^*})^q \to 0$ for $M^* \to \infty$. Now note that

$$\begin{aligned}
&\left|\hat{s}_{\hat{M}^*}(-\log \hat{s}_{\hat{M}^*})^q - s_{M^*}(-\log s_{M^*})^q\right| \\
\leq\ & \sup_M \left|(\hat{s}_M(-\log \hat{s}_M)^q - s_M(-\log s_M)^q) / \left(g(M)M^{q+s'}\right)\right| \left|g(\hat{M}^*)\hat{M}^{*q+s'}\right| \\
&+ \left|s_{\hat{M}^*}(-\log s_{\hat{M}^*})^q - s_{M^*}(-\log s_{M^*})^q\right|
\end{aligned}$$

By the proof of Proposition (4.3) it follows that $\sup_M \left|(\hat{s}_M(-\log \hat{s}_M)^q - s_M(-\log s_M)^q) / \left(g(M)M^{q+s'/2}\right)\right|$ is $O_p(n^{-1/2}(\log n)^{1/2})$ and $g(\hat{M}^*) = O_p(n^{-1/2}\log n)$ such that the first term in the previous inequality is $O_p(n^{-1}(\log n)^{3/2+q+s'/2}) = o_p(n^{-1/2}\log n^{1/2}/M^*)$. For the second term we write

$$\left|s_{\hat{M}^*}(-\log s_{\hat{M}^*})^q - s_{M^*}(-\log s_{M^*})^q\right| = \left|\frac{s_{\hat{M}^*}(-\log s_{\hat{M}^*})^q}{s_{M^*}(-\log s_{M^*})^q} - 1\right| s_{M^*}(-\log s_{M^*})^q$$

where $s_M = M^s \lambda^M + o\left(M^s \lambda^M\right)$. Then $\left(-\log s_{\hat{M}^*}\right)^q / \left(-\log s_{M^*}\right)^q - 1 = O_p(n^{-1/2} (\log n)^{1/2+s'/2})$ and $s_{\hat{M}^*}/s_{M^*} - 1 = \left(\hat{M}^*/M^*\right)^s \lambda^{(\hat{M}^*-M^*)} - 1 = O_p(n^{-1/2} (\log n)^{1/2+s'/2})$ by Theorem (4.3) and the delta method. Since $s_{M^*} \left(-\log s_{M^*}\right)^q = O(n^{-1/2} (\log n)^{q+1})$ it follows that $\left|s_{\hat{M}^*} \left(-\log s_{\hat{M}^*}\right)^q - s_{M^*} \left(-\log s_{M^*}\right)^q\right| = o_p(n^{-1/2} \log n^{1/2}/M^*)$. From Theorem (4.3) we also have $1/M^{*q} \left(\left(M^*/\hat{M}^*\right)^q - 1\right) = O_p(n^{-1/2} (\log n)^{1/2+s'} / M^{*q})$ such that

$$\left|\phi(k(j/\hat{M}^*), \hat{s}_{\hat{M}^*}) - \phi(k(j/M^*), s_{M^*})\right| = O_p(|j| \, n^{-1/2} (\log n)^{1/2+\max(s'-q,-1)}).$$

For the truncated kernel $\Delta_{\hat{M}^*} = 0$ unless $\hat{M}^* \neq M^*$. But $P\left(\left|\hat{M}^* - M^*\right| \geq 1\right) = 0$ with probability tending to one. So the rest of the proof is trivial for the truncated kernel and we only consider the more general case. Denote the k,j-th element of $\Omega^{*-1}_{M^*}$ by $\vartheta^{*M}_{j_1,j_2}$. Then, letting $\hat{k}(j_1, \hat{M}^*) = \phi(k(j/\hat{M}^*), \hat{s}_{\hat{M}^*})$

$$\begin{aligned}
\sqrt{n/M^*} \hat{P}'_{n-m} \Delta_{\hat{M}^*} \hat{\Omega}^{*-1}_{\hat{M}^*} \tilde{K}_{M^*} \frac{Z'_{n-m}\varepsilon}{\sqrt{n}} &= \sqrt{n/M^*} \frac{C_2}{\sqrt{n}} \sum_{t=1}^{n} \sum_{j_1,j_2=1}^{n-m} \hat{\Gamma}^{xy}_{j_1} \left[\hat{k}(j_1, \hat{M}^*) - k(j_1, M^*)\right] \hat{\vartheta}^{*\hat{M}^*}_{j_1,j_2} k(j_2, M^*) v_{t,j_2} \\
&= \sqrt{n/M^*} C_2 \left[d_4^\Delta + d_5^\Delta + d_6^\Delta + d_7^\Delta + d_8^\Delta + d_9^\Delta + d^\Delta\right]
\end{aligned}$$

where $d^\Delta = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{j_1,j_2=1}^{n-m} \Gamma^{xy}_{j_1} \left[\hat{k}(j_1, \hat{M}^*) - k(j_1, M^*)\right] \vartheta^{*M}_{j_1,j_2} k(j_2/M^*) v_{t,j_2}$,

$$d_4^\Delta = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sum_{j_1,j_2=1}^{n-m} \left(\hat{\Gamma}^{xy}_{j_1} - \check{\Gamma}^{xy}_{j_1}\right) \left[\hat{k}(j_1, \hat{M}^*) - k(j_1, M^*)\right] \vartheta^{*\hat{M}^*}_{j_1,j_2} k(j_2, M^*) v_{t,j_2}$$

and similarly for $d_5^\Delta, ..., d_9^\Delta$ corresponding to Definitions (A.20-A.24) for $d_5, ..., d_9$ where we replace K_M by $\tilde{\Delta}_{\hat{M}^*}$ and $\hat{\Omega}_M$ by $\hat{\Omega}^*_M$ in the same way as in d_4^Δ. We consider the largest term d_5^Δ

$$\left\|\sqrt{n/M^*} d_5^\Delta\right\| \leq C_1 \frac{1}{M^{*1/2}} O_p(n^{-1/2} (\log n)^{1/2+\max(s'-q,-1)}) \sum_{t=1}^{n} \sum_{j_1,j_2=1}^{n-m} |j_1|^q \left\|\left(\check{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right) \vartheta^{*\hat{M}^*}_{j_1,j_2} k(j_2, M^*) v_{t,j_2}\right\|.$$

By the same arguments as in the proof of Lemma (B.29) it follows that

$$\text{(A.34)} \qquad \sum_{j_1,j_2=1}^{n-m} |j_1|^q \left\|\left(\check{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right) \vartheta^{*\hat{M}^*}_{j_1,j_2} k(j_2, M^*) \sum\nolimits_{t=1}^{n} v_{t,j_2}\right\| = O_p(M^*)$$

where we have used that $\sum_{j_1} |j_1|^q \left\|\vartheta^{*\hat{M}^*}_{j_1,j_2}\right\| < \infty$ uniformly in j_2 since $\vartheta^{*\hat{M}^*}_{j_1,j_2}$ has the same summability properties as ϑ_{j_1,j_2}. Also note that $k(j_2, M^*) = 0$ for $|j_2| > M^*$. The bound (A.34) implies that $\sqrt{n/M^*} d_5^\Delta = O_p(n^{-1/2} (\log n)^{1+\max(s'-q,-1)})$. Using similar arguments based on the proofs of Lemmas (B.28,B.30-B.33) it can be shown that the remaining terms $d_4^\Delta, d_6^\Delta ..., d_9^\Delta$ are of smaller order. For d^Δ note that $\sum_{j_1,j_2=1}^{n} |j_1|^q \left\|\Gamma^{xy}_{j_1} \vartheta^{*\hat{M}^*}_{j_1,j_2}\right\| = O(1)$ such that

$$\sum_{j_1,j_2=1}^{n} |k(j_2, M^*)| \, |j_1|^q \left\|\Gamma^{xy}_{j_1} \vartheta^{*\hat{M}^*}_{j_1,j_2}\right\| \left(E \left\|\sum\nolimits_{t=1}^{n} v_{t,j_2}/\sqrt{n}\right\|^2\right)^{1/2} = O(1)$$

and thus $\sqrt{n/M^*}d^{\Delta} = O_p((\log n)^{\max(s'-q,-1)})$.

For $\sqrt{n/M^*}\hat{P}'_{n-m}\Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*}\frac{Z'_{n-m}\varepsilon}{\sqrt{n}} = \sqrt{n/M^*}C_2\left[d_4^{\Delta\Delta} + ... + d_9^{\Delta\Delta} + d^{\Delta\Delta}\right]$ we define

$$d^{\Delta\Delta} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sum_{j_1,j_2=0}^{n-m}\Gamma^{xy}_{j_1}\left[\hat{k}(j_1,\hat{M}^*) - k(j_1,M^*)\right]\vartheta^{*\hat{M}^*}_{j_1,j_2}\left[\hat{k}(j_2,\hat{M}^*) - k(j_2,M^*)\right]v_{t,j_2}$$

and similarly for the other terms. From $\sum_{j_1,j_2=1}^{n-m}\left(|j_1|^q|j_2|^q\left\|\Gamma^{xy}_{j_1}\vartheta^{*\hat{M}^*}_{j_1,j_2}\right\|^2 E\left\|\sum_{t=1}^{n}v_{t,j_2}/\sqrt{n}\right\|^2\right)^{1/2} = O(1)$ it follows that $\sqrt{n/M^*}d^{\Delta\Delta} = O_p(n^{-1/2}(\log n)^{1+2\max(s'-q,-1)}M^{*-1/2}) = o_p(1)$. For $\sqrt{n/M^*}d_5^{\Delta\Delta}$ note that for some C_1

$$\left\|d_5^{\Delta\Delta}\right\| \leq C_1 O_p(n^{-1}(\log n)^{1+2\max(s'-q,-1)})\sum_{j_1,j_2=1}^{\max(\hat{M}^*,M^*)}|j_1|^q|j_2|^q\left\|\left(\tilde{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right)\vartheta^{*\hat{M}^*}_{j_1,j_2}\sum_{t=1}^{n}v_{t,j_2}/\sqrt{n}\right\|$$

such that for any finite $\epsilon > 0$ and some C consider

$$P\left(M^{*-2}\sum_{j_1,j_2=0}^{\lceil M^*+\epsilon\rceil}|j_1|^q|j_2|^q\left\|\left(\tilde{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right)\vartheta^{*\hat{M}^*}_{j_1,j_2}\sum_{t=1}^{n}v_{t,j_2}\right\| > C\right) + P(\hat{M}^* > M^* + \epsilon).$$

where $\lceil M^* + \epsilon\rceil$ denotes the smallest integer larger than $M^* + \epsilon$. Using the Markov inequality it follows that

$$M^{*-2}\sum_{j_1,j_2=0}^{\lceil M^*+\epsilon\rceil}|j_1|^q|j_2|^q\left(E\left\|\tilde{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right\|^2\right)^{1/2}\left\|\vartheta^{*\hat{M}^*}_{j_1,j_2}\right\|\left(E\left\|\sum_{t=1}^{n}v_{t,j_2}/\sqrt{n}\right\|^2\right)^{1/2} = O(n^{-1/2})$$

by similar arguments as in the proof of Lemma (B.29). Therefore $\sqrt{n/M^*}d_5^{\Delta\Delta} = O_p(n^{-1}(\log n)^{1+2\max(s'-q,-1)})$. The remaining terms are of smaller order by the same arguments as before.

Finally, for $\sqrt{n/M^*}\hat{P}'_{n-m}\tilde{K}_{M^*}\left(\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*}\right)\tilde{K}_{M^*}\frac{Z'_{n-m}\varepsilon}{\sqrt{n}}$ we expand $\hat{\Omega}^{*-1}_{\hat{M}^*}$ around $\hat{\Omega}^{*-1}_{M^*}$ and $\hat{\Omega}^{*-1}_{M^*}$ around $\Omega^{*-1}_{M^*}$ as in (A.1) leading to

$$\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*} = \hat{\Omega}^{*-1}_{M^*}(\hat{\Omega}^*_{M^*} - \hat{\Omega}^*_{\hat{M}^*})\hat{\Omega}^{*-1}_{M^*} + o_p(\left\|\hat{\Omega}^*_{\hat{M}^*} - \hat{\Omega}^*_{M^*}\right\|) \tag{A.35}$$

and

$$\hat{\Omega}^{*-1}_{M^*} = \Omega^{*-1}_{M^*} - \Omega^{*-1}_{M^*}(\Omega^*_{M^*} - \hat{\Omega}^*_{M^*})\Omega^{*-1}_{M^*} + o_p(\left\|\Omega^*_{M^*} - \hat{\Omega}^*_{M^*}\right\|) \tag{A.36}$$

Note that $\hat{\Omega}^*_{\hat{M}^*} - \hat{\Omega}^*_{M^*} = 0$ if $\hat{M}^* \neq M^*$. Because $\hat{M}^*$ and M^* are integer valued by definition it also follows that $\hat{M}^* \neq M^* \Rightarrow \left|\hat{M}^* - M^*\right| \geq 1$. We thus note that for any $\varepsilon > 0$, $\left|\hat{\Omega}^*_{\hat{M}^*} - \hat{\Omega}^*_{M^*}\right| > \varepsilon \Rightarrow \left|\hat{M}^* - M^*\right| \geq 1$. Using the fact that $P(\left|\hat{M}^* - M^*\right| \geq 1)$ tends to zero it follows that

$$P\left(n^{1/2}\left\|\hat{\Omega}^*_{\hat{M}^*} - \hat{\Omega}^*_{M^*}\right\| > \varepsilon\right) \leq P(\left|\hat{M}^* - M^*\right| \geq 1) \to 0,$$

in fact $\left\|\hat{\Omega}^*_{\hat{M}^*} - \hat{\Omega}^*_{M^*}\right\|$ converges to zero in probability at arbitrarily fast rates. Also note that $\left\|\Omega^*_{M^*} - \hat{\Omega}^*_{M^*}\right\| = O_p(n^{-1/2}(\log n)^{1/2})$ by Lemma (B.9). Combining (A.35) and (A.36) then leads to

$$\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*} = \mathcal{O}_{M^*}(\hat{\Omega}^*_{M^*} - \hat{\Omega}^*_{\hat{M}^*})\mathcal{O}_{M^*} + o_p(\left\|\Omega^*_{M^*} - \hat{\Omega}^*_{M^*}\right\|)$$

with $\mathcal{O}_{M^*} = \Omega^{*-1}_{M^*} - \Omega^{*-1}_{M^*}(\Omega^*_{M^*} - \hat{\Omega}^*_{M^*})\Omega^{*-1}_{M^*}$. It thus follows that

$$\sqrt{n/M^*}\hat{P}'_{n-m}\tilde{K}_{M^*}\left(\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*}\right)\tilde{K}_{M^*}\frac{Z'_{n-m}\varepsilon}{\sqrt{n}} = o_p(1).$$

Next consider

$$\begin{aligned}\sqrt{n/M^*}(\hat{D}_{\hat{M}^*} - \hat{D}_{M^*}) &= \sqrt{n/M^*}(\hat{P}'_{n-m}\Big[\Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*} + \tilde{K}_{M^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*} + \Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\tilde{K}_{M^*} \\ &\quad + \tilde{K}_{M^*}\left(\hat{\Omega}^{*-1}_{\hat{M}^*} - \hat{\Omega}^{*-1}_{M^*}\right)\tilde{K}_{M^*}\Big]\hat{P}_{n-m}).\end{aligned}$$

First, we analyze $\hat{P}'_{n-m}\Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}K'_{M^*}\hat{P}_{n-m} = H_2^{\Delta} + H_3^{\Delta} + H^{\Delta}$ where $H_2^{\Delta} = H_{211}^{\Delta} + H_{212}^{\Delta} + H_{221}^{\Delta} + H_{222}^{\Delta}$, $H_3^{\Delta} = H_{31}^{\Delta} + H_{32}^{\Delta} + H_{33}^{\Delta} + H_{34}^{\Delta}$ and the definitions follow from the definitions in (A.6)-(A.13) with the appropriate substitutions for $\hat{\Omega}^{*-1}_{\hat{M}^*}$ and $\Delta_{\hat{M}^*}$. Furthermore let $H^{\Delta} = \sum_{i=0}^{n}\sum_{j=0}^{n}\Gamma_i^{xy}(k(i/\hat{M}^*) - k(i/M^*))\vartheta^*_{i,j}k(j/M^*)\Gamma^{yx}_{-j}$. Then

$$\begin{aligned}\sqrt{n/M^*}\left\|H^{\Delta}\right\| &\leq \sqrt{n/M^*}O_p(n^{-1/2}(\log n)^{1/2+\max(s'-q,-1)})\sum_{j_1,j_2=1}^{n-m}|j_1|\left\|\Gamma^{xy}_{j_1}\vartheta^{*\hat{M}^*}_{j_1,j_2}k(j_2/M^*)\Gamma^{yx}_{-j_2}\right\| \\ &= O_p(M^{*\max(s'-q,-1)})O(1).\end{aligned}$$

Now consider H_{222}^{Δ}

$$\sqrt{n/M^*}\left\|H_{222}^{\Delta}\right\| \leq \frac{\sqrt{n}}{\sqrt{M^*}}O_p(n^{-1/2}(\log n)^{1/2+\max(s'-q,-1)})\sum_{j_1,j_2=1}^{n-m}|j_1|^q\left\|\left(\check{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right)\vartheta^{*\hat{M}^*}_{j_1,j_2}k(j_2/M^*)\check{\Gamma}^{yx}_{-j_2}\right\|$$

and by the proof of Lemma (B.20) it follows that $E\sum_{j_1,j_2=1}^{n-m}|j_1|^q\left\|\left(\check{\Gamma}^{xy}_{j_1} - \Gamma^{xy}_{j_1}\right)\vartheta^{*\hat{M}^*}_{j_1,j_2}k(j_2/M^*)\check{\Gamma}^{yx}_{-j_2}\right\| = O(n^{-1/2}M^*)$ such that $\sqrt{n/M^*}\left\|H_{222}^{\Delta}\right\| = O_p(n^{-1/2}(\log n)^{1+\max(s'-q,-1)})$. Using the results of Lemma (B.22) we can show in the same way that $\sqrt{n/M^*}\left\|H_{34}^{\Delta}\right\| = o_p(1)$. All the remaining terms are of lower order by Lemmas (B.17-B.21).

Next, we turn to $\hat{P}'_{n-m}\Delta_{\hat{M}^*}\hat{\Omega}^{*-1}_{\hat{M}^*}\Delta_{\hat{M}^*}\hat{P}_{n-m} = H_2^{\Delta\Delta} + H_3^{\Delta\Delta} + H^{\Delta\Delta}$ where $H_2^{\Delta\Delta}, H_3^{\Delta\Delta}, H^{\Delta\Delta}$ are defined in the obvious way. It follows immediately that

$$\begin{aligned}\sqrt{n/M^*}\left\|H^{\Delta\Delta}\right\| &\leq \sqrt{n/M^*}O_p(n^{-1}(\log n)^{1+2\max(s'-q,-1)})\sum_{j_1,j_2=1}^{n-m}|j_1||j_2|\left\|\Gamma^{xy}_{j_1}\vartheta^{*\hat{M}^*}_{j_1,j_2}\Gamma^{yx}_{-j_2}\right\| \\ &= O_p(n^{-1/2}(\log n)^{-1/2+2\max(s'-q,-1)}).\end{aligned}$$

For $H_{222}^{\Delta\Delta}$ we note that $E\sum_{j_1,j_2=1}^{n-m}|j_1|^q|j_2|^q\left\|\Gamma_{j_1}^{xy}\vartheta_{j_1,j_2}^{*\hat{M}^*}\Gamma_{-j_2}^{yx}\right\|=O(1)$ such that again $\sqrt{n/M^*}\left\|H_{222}^{\Delta\Delta}\right\|=o_p(1)$. The same type of arguments also establish $\sqrt{n/M^*}\left\|H_{34}^{\Delta\Delta}\right\|=o_p(1)$. All the other terms are of lower order.

Finally, we turn to $\sqrt{n/M^*}\hat{P}_{n-m}'\tilde{K}_{M^*}\left(\hat{\Omega}_{\hat{M}^*}^{*-1}-\hat{\Omega}_{M^*}^{*-1}\right)\tilde{K}_{M^*}\hat{P}_{n-m}=o_p(1)$ by the same arguments based on $(\hat{\Omega}_{M^*}^*-\hat{\Omega}_{\hat{M}^*}^*)$ as before. ■

Proof of Proposition (5.1): We consider Ed_i and EH_iDd_j. First, $Ed_i=0$ for $i\leq 3$. The terms $d_4,d_6,...,d_9$ are of lower order by Lemmas (B.28,B.30-B.33). The terms $EH_iD^{-1}d_j$ are all of lower order. The largest order term is therefore Ed_5. By the proof of Lemma (B.44) it follows that $Ed_5=M/\sqrt{n}\mathcal{A}_1\int\phi^2(x)dx+o(M/\sqrt{n})$. ■

Proof of Theorem (5.3): By Proposition (4.1) it follows that $\widehat{\mathcal{A}}_1-\mathcal{A}_1=O_p(n^{-1/2})$. By the same arguments as in the proof of Lemma (B.47), $\sup_M\left|\hat{\sigma}_{2M,\hat{h}}-\sigma_{2M}\right|=O_p(n^{-1/2})$. Note that σ_{2M} is uniformly continuous in $k\in\mathcal{K}^j$ for each $j=0,1,...,\tau$. It is not uniformly continuous in $k\in\mathcal{K}_q$, however. Note that $\psi\in U_\psi^j\Rightarrow\psi_j\neq 0$ such that for $k(x)\in\mathcal{K}_j$ it follows that $k_j\neq 0$ and $k_i=0$ for all $i<j$. We therefore analyze the problem of finding $\hat{k}^{j*}$ for j fixed, where $\hat{k}^{j*}=\underset{k\in\mathcal{K}^j}{\arg\min}Q_{n,j}(k)$ and

$$Q_{n,j}(k)=\widehat{\mathcal{A}}\left(\int_{-\infty}^{\infty}\phi(k(x))^2dx\right)^2+\frac{n}{\hat{M}_T^{*2}}\left(2\frac{\hat{M}_T^*}{\log n}\right)^{-j}\hat{\sigma}_{2\hat{M}_T^*,\hat{h}}^{\phi}.$$

Also define and $Q_j(k)=\mathcal{A}\left(\int_{-\infty}^{\infty}\phi(x)^2dx\right)^2+k_j^2\mathcal{B}^{(j)}/\kappa$. Then, uniformly for $k\in\mathcal{K}^j$, $\lim\sigma_{2M_T^*}n/M_T^{*2}=k_j^2\mathcal{B}^{(j)}/\kappa^*$ such that $\hat{\sigma}_{2\hat{M}_T^*,\hat{h}}^{\phi}n/\hat{M}_T^{*2}-k_j^2\mathcal{B}^{(j)}/\kappa^*=O_p(n^{-1/2}(\log n)^{1/2+s'})$. Next, note that

$$\frac{\hat{M}_T^*}{\log n}-\frac{M_T^*}{\log n}=\left(\frac{\hat{M}_T^*}{M_T^*}-1\right)\frac{M_T^*}{\log n}=O_p(n^{-1/2}(\log n)^{1/2+s'})$$

where $M^*/\log n=(-2\log\lambda)^{-1}+o(1)$.We have shown that $\sup_{k\in\mathcal{K}^j}|Q_{n,j}(k)-Q_j(k)|=O_p(n^{-1/2}(\log n)^{1/2+s'})$ because Q_j is uniformly continuous in k for each j. It then follows from standard arguments that $\sup_x\left|\hat{k}^{q*}-k^{q*}\right|=O_p(n^{-1/2}(\log n)^{1/2+s'})$. To find the optimum in $\mathcal{K}_q$ we now define $\hat{k}^*=\hat{k}^{\hat{q}*}$ with $\hat{q}=\arg\max_{q\geq q'}Q_{n,q}(\hat{k}^{q*})$. But now, q is countable and finite and for each q, $Q_{n,q}(\hat{k}^{q*})$ converges to a fixed value. Thus, $\hat{q}$ converges. For the second part note that $\hat{M}^*\left(\hat{k}^*\right)/M^*(k^*)-1=O_p\left(n^{-1/2}(\log n)^{1/2+s'}\right)$ because it can be checked easily that the proof of Proposition (4.3) still goes through when k is replaced with $\hat{k}^*$. Then, if $q'=1$,

$$\begin{aligned}&\phi\left(\hat{k}(i/\hat{M}^*\left(\hat{k}^*\right)),\sqrt{\log\sigma_{1\hat{M}^*(\hat{k}^*),\hat{h}}}\right)-\phi\left(k^*(i/\hat{M}^*\left(\hat{k}^*\right)),\sqrt{\log\sigma_{1\hat{M}^*(\hat{k}^*),\hat{h}}}\right)\\=&\left(\hat{\psi}_1-\psi_1\right)i/\hat{M}^*\left(\hat{k}^*\right)+o_p\left(1/\hat{M}^*\left(\hat{k}^*\right)\right)\\=&O_p\left(n^{-1/2}(\log n)^{1/2+\max(s'-q,-1)}\right).\end{aligned}$$

When $q' > 1$ the error is $o_p(n^{-1/2}(\log n)^{1/2+\max(s'-q,-1)})$ such that the same arguments as in the proof of Theorem (4.4) go through. For the last part of the theorem first consider $q' = 1$ and let $k(x) = 1-x^2$ such that $\int h^2(x)dx = 64/45 < 2$ with $k_2 = 1$. Now consider a perturbation to $k(.)$, say $k_\varepsilon(x) = 1-\varepsilon x - x^2$ such that $k_1 = \varepsilon$. In other words, for k_ε the approximate MSE is $\mathcal{A}\left(\int_{-\infty}^{\infty} h_\varepsilon(x)^2dx\right)^2 + \varepsilon^2\mathcal{B}^{(1)}/\kappa^* + \mathcal{B}_1/\kappa^*$. We need to show that by choosing ε this can be made smaller than the approximate MSE for the truncated kernel, which is $4\mathcal{A} + \mathcal{B}_1/\kappa^*$. Note that $\int h_\varepsilon^2(x)dx = \int h^2(x)dx + \delta(\varepsilon)$ where $\delta(\varepsilon) = -\varepsilon + 8/21\varepsilon^2 + 4/3\varepsilon^3 + 2/5\varepsilon^4$. Choose ε such that $\left(\varepsilon^2 \sup_{\Theta_0} \mathcal{B}^{(1)}/\kappa(k_\varepsilon) + \delta(\varepsilon)\right)/\mathcal{A} \leq 4 - (64/45)^2$ which is possible because of the properties of Θ_0. This shows that k_ε dominates the truncated kernel when M_T^* is used as a bandwidth choice. But then clearly, if M is chosen optimally for k_ε it can not do worse than with M_T^*. Similar arguments can be used to handle the case where $q' > 1$. ■

Proof of Theorem (6.1) We consider the expansion of $\sqrt{n}\left(\beta_{n,M}^c - \beta\right)$as before. The analysis of the MSE of $\sqrt{n}\left(\beta_{n,M}^c - \beta\right)$ is then the same as the analysis for $\sqrt{n}\left(\beta_{n,M} - \beta\right)$ where we replace d_5 by

$$\bar{d}_5 = d_5 - \frac{M}{\sqrt{n}}\mathcal{A}_1' \int \phi^2(x)dx$$

and the additional term $d_{13} = M/\sqrt{n}(\widehat{\mathcal{A}}_1' - \mathcal{A}_1')\int \phi^2(x)dx = O_p(M/n)$, where the order of magnitude follows from Proposition (4.1), needs to be considered. First note that $Ed_5 - \frac{M}{\sqrt{n}}\mathcal{A}_1' \int \phi^2(x)dx = o(1)$. Then $E\bar{d}_5\bar{d}_5' = E(d_5 - Ed_5)(d_5 - Ed_5)' + o(1)$. From the proof of Lemma (B.44) it follows that $E\bar{d}_5\bar{d}_5' = O(M/n)$. Also $E\ell' H_{222}D^{-1}d_0\bar{d}_5'D^{-1/2}\ell = o(M/n)$ by Lemma (B.41) and $d_{13} = O_p(M/n)$ together with Lemma (B.40) shows that all remaining terms are at most of order M/n. By Remark (4), to first order, M^* does not depend on the constants of the M/n terms. Thus it is possible to minimize $Mp/n - \log\sigma_M$. Then, $\hat{M}_c^*/M_c^* - 1 = O_p(n^{-1/2}(\log n)^{1/2+s'})$ follows by Proposition (4.3). Finally, note that $\sqrt{M_c^*}\left(\hat{M}_c^*/M_c^* - 1\right)\hat{D}_{\hat{M}_c^*}^{-1}\widehat{\mathcal{A}}_1 \int \phi^2(x)dx = O_p(n^{-1/2}(\log n)^{1/2+s'}\sqrt{M_c^*}) = o_p(1)$ by Proposition (4.3) and Lemma (B.47). In light of Theorem (4.4) this establishes the last part of the Theorem. ■

References

Andrews, D. W. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

——— (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, pp. 543–564.

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

Berk, K. N. (1974): "Consistent Autoregressive Spectral Estimates," *Annals of Statistics*, 2, 489–502.

Brillinger, D. R. (1981): *Time Series, Data Analysis and Theory, Expanded Edition.* Holden-Day, Inc.

Brockwell, P. J., and R. A. Davis (1991): *Time Series: Theory and Methods.* Springer Verlag-New York, Inc., second edn.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

E.J.Hannan, and M.Deistler (1988): *The Statistical Theory of Linear Systems.* John Wiley & Sons.

Gawronski, W. (1977): "Matrix Functions: Taylor Expansion, Sensitivity and Error Analysis," *Applicationes Mathematicae*, XVI, 73–89.

Hall, A. R., and A. Inoue (2001): "A Canonical Correlations Interpretation of Generalized Method of Moments Estimation with Applications to Moment Selection," *mimeo.*

Hannan, E., and L. Kavalieris (1984): "Multivariate Linear Time Series Models," *Adv.Appl.Prob*, 16, 492–561.

——— (1986): "Regression, Autoregression Models," *Journal of Time Series Analysis*, 7, 27–49.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1053.

——— (1985): "A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators," *Journal of Econometrics*, 30, 203–238.

Hansen, L. P., J. C. Heaton, and M. Ogaki (1988): "Efficiency Bounds Implied by Multiperiod Conditional Moment Restrictions," *Journal of the American Statistical Association*, 83, 863–871.

Hansen, L. P., and K. J. Singleton (1991): "Computing Semiparametric Efficiency Bounds for Linear Time Series Models," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 387–412.

——— (1996): "Efficient Estimation of Linear Asset-Pricing Models With Moving Average Errors," *Journal of Business and Economic Statistics*, pp. 53–68.

Hayashi, F., and C. Sims (1983): "Nearly Efficient Estimation of Time Series Models with Predetermined, but not Exogenous, Instruments," *Econometrica*, 51(3), 783–798.

Kuersteiner, G. M. (1997): "Efficient Inference in Time Series Models with Conditional Heterogeneity," Ph.D. thesis, Yale University.

——— (2001): "Optimal Instrumental Variables Estimation for ARMA Models," *Journal of Econometrics*, 104, 359–405.

——— (2002): "Efficient IV Estimation for Autoregressive Models with Conditional Heteroskedasticity," *Econometric Theory*, 18, 547–583.

Lewis, R., and G. Reinsel (1985): "Prediction of Multivariate Time Series by Autoregressive Model Fitting," *Journal of Multivariate Analysis*, pp. 393–411.

Linton, O. (1995): "Seocnd Order Approximation in the Partially Linear Regression Model.," *Econometrica*, pp. 1079–1112.

Linton, O. B. (1997): "Second Order Approximation for Semiparametric Intrumental Variable Estimators and Test Statistics," *mimeo Yale University.*

Magnus, J. R., and H. Neudecker (1979): "The Commutation Matrix: Some Properties and Applications," *Annals of Statistics*, 7(2), 381–394.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27(4), 575–595.

Newey, W. K. (1988): "Adaptive Estimation of Regression Models via Moment Restrictions," *Journal of Econometrics*, 38, 301–339.

——— (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, pp. 1349–1382.

Ng, S., and P. Perron (1995): "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90, 268–281.

Parzen, E. (1957): "On Consistent Estimates of the Spectrum of a Stationary Time Series," *Annals of Statistics*, 28, 329–348.

Priestley, M. (1981): *Spectral Analysis and Time Series*. Academic Press.

Stoica, P., T. Söderström, and B. Friedlander (1985): "Optimal Instrumental Variable Estimates of the AR Parameters of an ARMA Process," *IEEE Transactions on Automatic Control*, 30, 1066–1074.

Xiao, Z., and P. C. Phillips (1998): "Higher Order Approximations for Frequency Domain Time Series Regression," *Journal of Econometrics*, pp. 297–336.

Table 1: Performance of GMM Estimators with $\theta = -.5$

		$n = 128$			$n = 512$		
Estimator	ϕ	Bias	MAE	MSE	Bias	MAE	MSE
OLS	0.1	0.51652	0.51672	0.52458	0.5174	0.52002	0.52203
GMM-1		0.54524	1.3457	2.5025	0.5095	1.0594	1.653
GMM-20		0.4481	0.45146	0.47529	0.43876	0.44586	0.4726
KGMM-20		0.44917	0.47111	0.53162	0.44736	0.46461	0.52371
GMM-Opt		0.48439	1.0699	2.1788	0.47735	0.8655	1.4054
BGMM-Opt		0.39075	1.0655	3.6218	0.33739	0.69265	0.99113
KGMM-Opt		0.48483	0.75581	1.1466	0.46242	0.66349	0.92119
BKGMM-Opt		0.4259	0.85469	1.3366	0.38961	0.77809	1.0823
OLS	0.3	0.52227	0.52425	0.53233	0.52319	0.52238	0.52433
GMM-1		0.37319	0.92571	1.7589	0.19208	0.64946	2.8861
GMM-20		0.4699	0.47633	0.49813	0.4479	0.45124	0.47233
KGMM-20		0.45252	0.46546	0.51769	0.40358	0.41276	0.4601
GMM-Opt		0.41832	0.8571	1.6191	0.2193	0.60238	2.8213
BGMM-Opt		0.1965	0.84522	1.6398	-0.01157	0.48039	0.64576
KGMM-Opt		0.40261	0.65701	0.93193	0.26319	0.46613	0.68559
BKGMM-Opt		0.26391	0.71097	1.0397	0.063104	0.50435	0.80424
OLS	0.5	0.46675	0.46775	0.47616	0.47121	0.46856	0.47072
GMM-1		0.064923	0.36801	0.63065	0.012124	0.17127	0.22942
GMM-20		0.4193	0.42041	0.44244	0.2859	0.28971	0.31147
KGMM-20		0.33826	0.35308	0.40205	0.18586	0.19964	0.23394
GMM-Opt		0.076623	0.36358	0.6227	0.01198	0.17069	0.2287
BGMM-Opt		-0.20138	0.52444	0.70699	-0.17292	0.27818	0.37291
KGMM-Opt		0.085374	0.32911	0.52499	0.013643	0.16615	0.21996
BKGMM-Opt		0.021164	0.36326	0.55213	-0.02036	0.17825	0.24791

Table 2: Performance of GMM Estimators with $\theta = 0$

		$n = 128$			$n = 512$		
Estimator	ϕ	Bias	MAE	MSE	Bias	MAE	MSE
OLS	0.1	0.49705	0.49725	0.50385	0.49483	0.49372	0.49516
GMM-1		0.49461	0.98011	1.8786	0.48621	1.1214	5.2146
GMM-20		0.50373	0.50334	0.52432	0.49198	0.49328	0.51224
KGMM-20		0.49943	0.51901	0.58448	0.48642	0.49693	0.55721
GMM-Opt		0.51603	0.82588	1.5626	0.48211	0.95272	5.0932
BGMM-Opt		0.48989	1.1095	2.8016	0.45823	1.326	12.9356
KGMM-Opt		0.50207	0.69179	1.1779	0.48339	0.66513	0.89137
BKGMM-Opt		0.52045	0.76303	1.1456	0.46361	0.74174	1.0006
OLS	0.3	0.45384	0.45547	0.4613	0.45384	0.45403	0.45568
GMM-1		0.28972	0.66679	1.2017	0.1072	0.47244	0.90049
GMM-20		0.44322	0.44483	0.46525	0.40133	0.40931	0.4302
KGMM-20		0.43339	0.4501	0.51347	0.34142	0.35464	0.40636
GMM-Opt		0.36738	0.62256	1.1651	0.2135	0.48551	0.87465
BGMM-Opt		0.28444	0.66182	1.101	0.11745	0.35947	0.49946
KGMM-Opt		0.36569	0.52213	0.7151	0.20702	0.39769	0.57164
BKGMM-Opt		0.28996	0.57883	0.84856	0.11337	0.40762	0.66559
OLS	0.5	0.38046	0.37889	0.38487	0.37494	0.37593	0.3775
GMM-1		0.042111	0.29423	0.54636	0.015656	0.13132	0.1768
GMM-20		0.32936	0.32894	0.35015	0.21703	0.21913	0.23704
KGMM-20		0.2695	0.28151	0.32307	0.15066	0.1643	0.19171
GMM-Opt		0.08868	0.31714	0.58005	0.030298	0.14035	0.18463
BGMM-Opt		0.04296	0.27811	0.41247	0.016908	0.13559	0.18328
KGMM-Opt		0.090384	0.28488	0.44025	0.02436	0.13273	0.17765
BKGMM-Opt		0.054996	0.28023	0.42958	0.015666	0.13461	0.18303

Table 3: Performance of GMM Estimators with $\theta = .5$

		$n = 128$			$n = 512$		
Estimator	ϕ	Bias	MAE	MSE	Bias	MAE	MSE
OLS	0.1	0.4708	0.47223	0.4798	0.46821	0.47019	0.47208
GMM-1		0.52711	1.1213	2.0323	0.52264	1.1775	2.2054
GMM-20		0.39267	0.39842	0.42841	0.36723	0.37541	0.40458
KGMM-20		0.40082	0.42934	0.4976	0.37235	0.40024	0.45747
GMM-Opt		0.41931	0.80943	1.357	0.41969	0.86022	1.5581
BGMM-Opt		0.32466	1.2808	13.0744	0.30929	0.60111	0.88186
KGMM-Opt		0.43155	0.67594	0.9787	0.40402	0.67063	0.97716
BKGMM-Opt		0.35219	0.70706	0.98258	0.32216	0.76215	1.232
OLS	0.3	0.38604	0.38641	0.3945	0.38685	0.38754	0.38952
GMM-1		0.19769	0.82882	1.6984	0.087729	0.496	0.87353
GMM-20		0.27067	0.28668	0.31713	0.24542	0.25758	0.28589
KGMM-20		0.26555	0.30311	0.36346	0.22235	0.25503	0.30632
GMM-Opt		0.23217	0.62133	1.0808	0.13734	0.41268	0.70857
BGMM-Opt		0.098508	0.57105	1.1184	0.018366	0.29295	0.46272
KGMM-Opt		0.24356	0.50607	0.742	0.16227	0.35728	0.54694
BKGMM-Opt		0.13393	0.53538	0.84079	0.025493	0.35742	0.59875
OLS	0.5	0.27857	0.28049	0.28786	0.28095	0.2819	0.28394
GMM-1		0.011779	0.3225	0.67531	-0.00211	0.11246	0.15008
GMM-20		0.15784	0.16854	0.19507	0.1009	0.10759	0.12591
KGMM-20		0.13523	0.16429	0.20386	0.065555	0.09014	0.11271
GMM-Opt		0.058289	0.253	0.46363	0.01898	0.10532	0.14026
BGMM-Opt		-0.01254	0.21014	0.33764	-0.02236	0.095026	0.12984
KGMM-Opt		0.065003	0.22165	0.34951	0.012326	0.10667	0.13995
BKGMM-Opt		0.003854	0.22473	0.42831	-0.01114	0.094756	0.13116

Table 4: Performance of GMM Estimators with $\theta = .5$ and Heteroskedasticity

		$n = 128$		$n = 512$	
Estimator	ϕ	Bias	IDR	Bias	IDR
OLS	0.1	0.36203	0.39565	0.37789	0.22017
GMM-1		0.50136	5.22	0.51375	8.5451
GMM-20		0.36409	0.9079	0.43471	1.163
KGMM-20		0.36288	0.99398	0.43279	1.2035
GMM-Opt		0.37306	1.8038	0.44232	2.2475
BGMM-Opt		0.35293	1.0481	0.42576	1.1657
KGMM-Opt		0.41278	2.0442	0.49196	2.7351
BKGMM-Opt		0.38757	1.931	0.48469	2.158
OLS	0.3	0.28962	0.29852	0.31254	0.19993
GMM-1		0.19565	2.991	0.039369	2.5622
GMM-20		0.24842	0.65978	0.28085	0.9905
KGMM-20		0.248	0.78716	0.26538	0.93733
GMM-Opt		0.22376	1.2064	0.22999	1.1617
BGMM-Opt		0.24336	0.73583	0.27032	0.97506
KGMM-Opt		0.2297	1.4793	0.24377	1.62
BKGMM-Opt		0.23773	1.37	0.25496	1.5437
OLS	0.5	0.21107	0.26352	0.22779	0.15279
GMM-1		-0.00746	1.0207	-0.00346	0.60612
GMM-20		0.12121	0.50708	0.08675	0.5928
KGMM-20		0.11991	0.54298	0.076	0.47081
GMM-Opt		0.082524	0.57197	0.072163	0.4773
BGMM-Opt		0.1116	0.49428	0.084331	0.57169
KGMM-Opt		0.077325	0.67164	0.057138	0.46211
BKGMM-Opt		0.086783	0.65593	0.06762	0.50118