

Optimal Bandwidth Selection for Robust Generalized Method of Moments Estimation

Daniel Wilhelm*

University College London

January 31, 2013

Abstract

A two-step generalized method of moments estimation procedure can be made robust to heteroskedasticity and autocorrelation in the data by using a nonparametric estimator of the optimal weighting matrix. This paper addresses the issue of choosing the corresponding smoothing parameter (or bandwidth) so that the resulting point estimate is optimal in a certain sense. We derive an asymptotically optimal bandwidth that minimizes a higher-order approximation to the asymptotic mean-squared error of the estimator of interest. We show that the optimal bandwidth is of the same order as the one minimizing the mean-squared error of the nonparametric plugin estimator, but the constants of proportionality are significantly different. Finally, we develop a data-driven bandwidth selection rule and show, in a simulation experiment, that the particular bandwidth chosen may yield significant second-order gains.

JEL classification: C12; C13; C14; C22; C51

Keywords: GMM; higher-order expansion; optimal bandwidth; mean-squared error; long-run variance.

*Department of Economics, University College London, 30 Gordon St, London WC1H 0AX, United Kingdom; E-Mail address: d.wilhelm@ucl.ac.uk. I thank Christian Hansen, Alan Bester, the co-editor and two referees for helpful comments.

1 Introduction

Since the seminal paper by Hansen (1982) the generalized method of moments (GMM) has become a popular method for the estimation of partially specified models based on moment conditions. In time series applications, two-step GMM estimators can be made robust to heteroskedasticity and autocorrelation (HAC) by using a nonparametric plug-in estimator of the optimal weighting matrix. The goal of this paper is to develop a selection rule for the corresponding smoothing parameter of the nonparametric estimator such that the resulting point estimator minimizes a suitably defined mean-squared error (MSE) criterion.

Many instances of poor finite sample performance of GMM estimators have been reported in the literature. See for example Hansen, Heaton, and Yaron (1996) and references therein. As an attempt to improve the properties different extensions and new estimators have been proposed, e.g. the empirical likelihood estimator introduced by Owen (1988), Qin and Lawless (1994), the exponential tilting estimator of Kitamura and Stutzer (1997) and Imbens, Spady, and Johnson (1998) and the continuous updating estimator by Hansen, Heaton, and Yaron (1996). Newey and Smith (2004) show that all these estimators are members of a larger class of generalized empirical likelihood (GEL) estimators. A different approach termed “fixed-b” asymptotics is based on deriving more accurate approximations of estimators and test statistics based on an asymptotic sequence in which the HAC smoothing parameter tends to infinity at the same rate as the sample size. See for example Kiefer and Vogelsang (2002a,b, 2005). Instead of treating the smoothing parameter as proportional to the sample size, Sun and Phillips (2008) and Sun, Phillips, and Jin (2008) develop a higher-order asymptotic theory based on which they find the optimal rate at which the smoothing parameter (here a bandwidth) minimizes the coverage probability error or length of confidence intervals.

Similar in spirit, the present paper derives the optimal growth rate of the bandwidth to minimize an asymptotic mean-squared error (AMSE) criterion. We approximate the MSE of the second-step GMM estimator by the MSE of the first few terms in a stochastic expansion. Since the proposed semiparametric estimator is first-order equivalent to ordinary GMM estimators in the iid case, the optimal bandwidth derived in this paper will minimize its second-order effects on the estimator and lead to second-order efficiency gains. In an unpublished dissertation, Jun (2007) independently develops a similar expansion and arrives at the same MSE-optimal bandwidth as derived in this paper, however under a slightly different set of assumptions.¹

Other bandwidth choices for HAC-robust estimation have been suggested by Andrews

(1991), Newey and West (1994) and Andrews and Monahan (1992), for example, and are very popular in applied research. In this paper, we show that these are suboptimal choices if MSE-optimal point estimation is of main interest. In finite samples, the existing methods can select bandwidths that are significantly different from the MSE-optimal bandwidth even though they share the same asymptotic order relative to the sample size. The difference is due to the other methods minimizing the AMSE of the weighing matrix estimator instead of minimizing the AMSE of the GMM estimator itself; they guarantee accurate estimates of the optimal weighting matrix, but not necessarily of the parameter of interest. For example, in a linear instrumental variable model, the MSE-optimal bandwidth adapts to the amount of serial correlation in the regression errors, the degree of over-identification and to the amount of variation in the instruments. On the other hand, the existing bandwidth choices mentioned above, are independent of the latter two determinants.

In the linear regression framework with potential autocorrelation and heteroskedasticity, there are several papers (e.g. Robinson (1991), Xiao and Phillips (1998) and Tamaki (2007)) that derive higher-order expansions of the MSE of semiparametric frequency domain estimators to determine an optimal bandwidth that minimizes higher-order terms of such expansions. In the present paper, however, we allow for nonlinear models and over-identification which significantly complicate the problem and require a different set of tools to derive such expansions.

To approximate higher-order moments of the GMM estimator we develop a stochastic expansion of the estimator similar to the approach in Nagar (1959). See Rothenberg (1984) for an introduction to Nagar-type expansions and for further references. Several other authors have analyzed higher-order properties of GMM and GEL estimators using similar tools. Rilstone, Srivastava, and Ullah (1996) and Newey and Smith (2004) provide expressions for the higher-order bias and variance of GMM and GEL estimators when the data are iid. Anatolyev (2005) derives the higher-order bias in the presence of serial correlation.

Finally, Goldstein and Messer (1992) present general conditions under which functionals of nonparametric plug-in estimators achieve the optimal rate of convergence. Depending on the functional under-smoothing the plugin estimator relative to the smoothing parameter used to optimally estimate the nonparametric quantity itself may be necessary.

The paper is organized as follows. The first section introduces the econometric setup, derives a higher-order expansion of the two-step GMM estimator and the optimal bandwidth that minimizes an approximate MSE based on that expansion. The third section describes an approach to estimate the infeasible optimal bandwidths, followed by a simu-

lation experiment that demonstrates the procedure’s performance in finite samples. The paper concludes with an appendix containing all mathematical proofs.

Let $\text{vec}(\cdot)$ denote the column-by-column stacking operation and $\text{vech}(\cdot)$ the column-by-column stacking operation of entries on and above the diagonal of a symmetric matrix. By $K_{m,n}$ denote the $mn \times mn$ commutation matrix so that, for any $m \times n$ matrix M , $K_{m,n}\text{vec}(M) = \text{vec}(M')$. Let \otimes be the Kronecker product and $\nabla^r F(\beta)$, with $r \in \mathbb{Z}$ and $F(\beta)$ a matrix being r times differentiable in β , denote the matrix of r -th order partial derivatives with respect to β , recursively defined as in [Rilstone, Srivastava, and Ullah \(1996\)](#). $\nabla^0 F(\beta) := F(\beta)$. This notation for derivatives will sometimes be used to save space and simplify notation. $\|\cdot\|$ denotes the Euclidean (matrix) norm, M' the transpose of a matrix M , “with probability approaching one” is abbreviated “w.p.a. 1” and “with probability one” by “w.p. 1”. The notation $x_T = O_p(1)$ means that the sequence $\{x_T\}_{T=1}^\infty$ is uniformly tight.

2 Optimal Bandwidth

In this section, we introduce the basic framework, define an appropriate MSE criterion and find the optimal bandwidth that minimizes it. The idea is to derive a higher-order approximation of the second-step estimator and to the MSE from the moments of this approximation. A higher-order analysis is required in this setup because first-order asymptotics do not depend on the smoothing parameter.

Consider estimation of a parameter $\beta_0 \in \mathcal{B}$ from the moment equation $Eg(X_t, \beta_0) = 0$ given a data sample $\{x_t\}_{t=1}^T$. If the dimension of the range of g is at least as large as the dimension of the parameter β_0 , then a popular estimator of β_0 is the two-step GMM estimator defined as follows. First, estimate β_0 by some \sqrt{T} -consistent estimator, say $\tilde{\beta}$, that is then used to construct a consistent estimator $\hat{\Omega}(\tilde{\beta})$ of the long-run variance $\Omega_0 := \sum_{s=-\infty}^\infty \Gamma(s)$, $\Gamma(s) := E[g(X_{t+s}, \beta_0)g(X_t, \beta_0)']$. In a second step, compute the GMM estimator $\hat{\beta}$ of β_0 with weighting matrix $\hat{\Omega}(\tilde{\beta})^{-1}$, viz.

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B}} \hat{g}(\beta)' \hat{\Omega}_T(\tilde{\beta})^{-1} \hat{g}(\beta), \quad (2.1)$$

where $\hat{g}(\beta) := T^{-1} \sum_{t=1}^T g(x_t, \beta)$. The second step improves the first-step estimator in terms of efficiency. In fact, $\hat{\beta}$ is optimal in the sense that it achieves the lowest asymptotic variance among all estimators of the form $\hat{\beta}_W := \arg \min_{\beta \in \mathcal{B}} \hat{g}(\beta)' W \hat{g}(\beta)$ for some positive definite weighting matrix W (see [Hansen \(1982\)](#)).

In the special case of an iid process $\{X_t\}$, Ω_0 collapses to $\Omega_0 = E[g(X_t, \beta_0)g(X_t, \beta_0)']$ and can simply be estimated by its sample analog $\hat{\Omega}_T(\tilde{\beta}) := T^{-1} \sum_{t=1}^T g(x_t, \tilde{\beta})g(x_t, \tilde{\beta})'$.

When the iid assumption is not justified, one can perform inference robust to autocorrelated and/or heteroskedastic X_t processes. Robustness here means that potential dependence and heteroskedasticity are treated nonparametrically and one does not have to be explicit about the data generating process of the X_t 's. To that end, one needs to “smooth” the observations $g(x_{t+s}, \tilde{\beta})g(x_t, \tilde{\beta})'$ over s to ensure that $\hat{\Omega}_T(\tilde{\beta})$ is consistent. In this paper, we use a nonparametric kernel estimator of the form

$$\hat{\Omega}_T(\tilde{\beta}) := \frac{1}{T} \sum_{s=1-T}^{T-1} \sum_{t=\max\{1, 1-s\}}^{\min\{T, T-s\}} k\left(\frac{s}{S_T}\right) g_{t+s}(\tilde{\beta})g_t(\tilde{\beta})'$$

with $g_t(\beta) := g(x_t, \beta)$ and k a kernel function. S_T , with $S_T \rightarrow \infty$ as $T \rightarrow \infty$, is a so-called bandwidth parameter that governs the degree of smoothing. Andrews (1991) derives a range of rates (in terms of T) at which S_T is allowed to diverge in order to guarantee consistency of $\hat{\Omega}_T(\tilde{\beta})$. These conditions, however, do not suggest rules for choosing S_T for a fixed sample size T . Small values of S_T imply averaging over only few observations which decreases the variability of the estimator $\hat{\Omega}_T(\tilde{\beta})$, but increases its bias. On the other hand, large bandwidths yield inclusion of more distant lags in the above sum, thereby increasing the variance, but decreasing the bias of the estimator. Below we show that the choice of S_T affects the bias and variance of the second-step estimator $\hat{\beta}$ in a similar way. This trade-off can be used to derive decision rules on how to pick S_T in finite samples. For example, Andrews (1991) derives the optimal bandwidth minimizing a truncated asymptotic mean-square error (AMSE) criterion that balances bias and variance of $\hat{\Omega}_T(\tilde{\beta})$, thereby guaranteeing “good” properties of the estimator of the optimal weighting matrix. However, in the GMM estimation framework, the second-step estimator $\hat{\beta}$ is the quantity of interest and, thus, the bandwidth should be chosen so as to take into account the bias and variance trade-off of $\hat{\beta}$, rather than that of $\hat{\Omega}_T(\tilde{\beta})$. To that end the subsequent analysis develops a higher-order expansion of the MSE of the second-step estimator and then minimize the leading terms with respect to the bandwidth.

Assumption 2.1. (a) The process $\{X_t\}_{t=-\infty}^{\infty}$ taking values in $\mathcal{X} \subset \mathbb{R}^m$ is fourth-order stationary and α -mixing with mixing coefficients $\alpha(j)$ satisfying $\sum_{j=1}^{\infty} j^2 \alpha(j)^{(\nu-1)/\nu} < \infty$ for some $\nu > 1$. (b) $\{x_t\}_{t=1}^T$ is an observed sample of $\{X_t\}_{t=-\infty}^{\infty}$. (c) $h(\cdot, \beta) := (g(\cdot, \beta)', \text{vec}(\nabla g(\cdot, \beta) - E\nabla g(\cdot, \beta))', \text{vec}(\nabla^2 g(\cdot, \beta) - E\nabla^2 g(\cdot, \beta))')'$ is a measurable function for every $\beta \in \mathcal{B}$. (d) $\sup_{t \geq 1} E[\|h(x_t, \beta_0)\|^{4\nu}] < \infty$. (e) $\sup_{t \geq 1} E\|g(x_t, \beta_0)\|^2 < \infty$. (f) $\sup_{t \geq 1} E[\sup_{\beta \in \mathcal{B}} \|\nabla^k g(x_t, \beta)\|^2] < \infty$ for $k = 1, 2, 3$. (g) There is a first-step estimator $\tilde{\beta}$ satisfying $\tilde{\beta} - \beta_0 = O_p(T^{-1/2})$.

As a slight abuse of notation, in the remainder, x_t represents the random variable X_t as well as the observation x_t , but the distinction should be clear from the context.

Furthermore, dropping β as an argument of a function means that the function is evaluated at β_0 , e.g. $\hat{\Omega}_T := \hat{\Omega}_T(\beta_0)$ or $g_t := g_t(\beta_0)$. Let $G_0 := EG(x_t)$, $G(x_t, \beta) := \partial g_t(\beta)/\partial \beta'$, $G_t(\beta) := G(x_t, \beta)$, and the sample counterpart $G_T(\beta) := T^{-1} \sum_{t=1}^T \partial g_t(\beta)/\partial \beta'$.

Following Parzen (1957), let q be the characteristic exponent that characterizes the smoothness of the kernel k at zero: $q := \max\{\alpha \in [0, \infty) : g_\alpha \text{ exists and } 0 < |g_\alpha| < \infty\}$ with $g_\alpha := \lim_{z \rightarrow 0} \frac{1-k(z)}{|z|^\alpha}$. For example, for the Bartlett, Parzen and Tukey-Hanning kernel the values of q are 1, 2 and 2, respectively.

Assumption 2.2. Let the kernel k satisfy the following conditions: **(a)** $k : \mathbb{R} \rightarrow [-1, 1]$ satisfies $k(0) = 1$, $k(x) = k(-x) \forall x \in \mathbb{R}$, $\int_{-\infty}^{\infty} k^2(x)dx < \infty$, $\int_{-\infty}^{\infty} |k(x)|dx < \infty$, $k(\cdot)$ is continuous at 0 and at all but a finite number of other points, and $S_T \rightarrow \infty$, $S_T^2/T \rightarrow 0$, $S_T^q/T \rightarrow 0$ for some $q \in [0, \infty)$ for which $g_q, \|f^{(q)}\| \in [0, \infty)$ where $f^{(q)} := \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^q \Gamma(j)$. **(b)** $\int_{-\infty}^{\infty} \bar{k}(x)dx < \infty$ with

$$\bar{k}(x) := \begin{cases} \sup_{y \geq x} |k(y)|, & x \geq 0 \\ \sup_{y \leq x} |k(y)|, & x < 0 \end{cases}.$$

Assumptions 2.1 and 2.2(a) imply Assumptions A, B and C in Andrews (1991) applied to $\{g_t\}$ and $\{G_t\}$, allowing us to use his consistency and rate of convergence results for HAC estimators. The necessity of Assumption 2.2(b) is explained in Jansson (2002).

Assumption 2.3. **(a)** $g : \mathcal{X} \times \mathcal{B} \rightarrow \mathbb{R}^l$, $l \geq p$, and $\beta_0 \in \text{int}(\mathcal{B})$ is the unique solution to $Eg(x_t, \beta_0) = 0$, $\mathcal{B} \subset \mathbb{R}^p$ is compact. **(b)** $\text{rank}(G_0) = p$. **(c)** For any $x \in \mathcal{X}$, $g(x, \cdot)$ is twice continuously differentiable in a neighborhood \mathcal{N} of β_0 . **(d)** There exists a function $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ with $\sum_{s=-\infty}^{\infty} Ed(x_{t+s}, x_t) < \infty$ so that g satisfies the Lipschitz condition $\|\nabla(g_{t+s}(\beta)g_t(\beta)') - \nabla(g_{t+s}(\beta_0)g_t(\beta_0)')\| \leq d(x_{t+s}, x_t)$ w.p. 1 for $\beta \in \mathcal{N}$. **(e)** There exists a function $b : \mathcal{X} \rightarrow \mathbb{R}$ with $Eb(x_t) < \infty$ such that $\|\nabla^k g(x, \beta) - \nabla^k g(x, \beta_0)\| \leq b(x)\|\beta - \beta_0\|$ for $k = 2, 3$. **(f)** Ω_0 is positive definite.

The following proposition is the first main result of the paper, presenting an expansion of the second-step GMM estimator $\hat{\beta}$ up to the lowest orders involving the bandwidth S_T . This approximation constitutes a crucial ingredient for the computation of the optimal bandwidth.

Proposition 2.1. *Under Assumptions 2.1–2.3, $\hat{\beta}$ satisfies the stochastic expansion*

$$\hat{\beta} = \beta_0 + \kappa_{1,T}T^{-1/2} + \kappa_{2,T}S_T^{1/2}T^{-1} + \kappa_{3,T}S_T^{-q}T^{-1/2} + o_p(\eta_T T^{-1/2}) \quad (2.2)$$

with $\eta_T := S_T^{1/2}T^{-1/2} + S_T^{-q}$, $\kappa_{i,T} = O_p(1)$ for $i = 1, 2, 3$,

$$\begin{aligned}\kappa_{1,T} &:= -H_0F_T \\ \kappa_{2,T} &:= H_0\sqrt{T/S_T} \left(\hat{\Omega}_T - \bar{\Omega}_T \right) P_0F_T \\ \kappa_{3,T} &:= H_0S_T^q \left(\bar{\Omega}_T - \Omega_0 \right) P_0F_T\end{aligned}$$

and $F_T(\beta) := \sqrt{T}\hat{g}(\beta)$, $F_T := F_T(\beta_0)$, $\bar{\Omega}_T := E\hat{\Omega}_T$, $\Sigma_0 := (G'_0\Omega_0^{-1}G_0)^{-1}$, $H_0 := \Sigma_0G'_0\Omega_0^{-1}$ and $P_0 := \Omega_0^{-1} - \Omega_0^{-1}G_0H_0$.

Since the lowest-order term, $-H_0F_TT^{-1/2}$, does not depend on the bandwidth, the expansion (2.2) illustrates the well-known fact that nonparametric estimation of the GMM weighting matrix does not affect first-order asymptotics as long as that nonparametric estimator is consistent. The other two terms in the expansion involve the bandwidth and arise from the bias ($S_T^q(\bar{\Omega}_T - \Omega_0)$) and variance ($\sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T)$) of the nonparametric estimator of the weighting matrix. In the iid case, these two components converge to zero as $T \rightarrow \infty$ whereas, in the presence of time series dependence, they converge to a non-degenerate random variable. Therefore, in the iid case, the next higher-order term after $\kappa_{1,T}T^{-1/2}$ is of order T^{-1} (see Newey and Smith (2004)) which, here is part of the remainder and plays no role in determining the optimal bandwidth derived below.

Anatolyev (2005) does not explicitly present a stochastic expansion such as (2.2), but computes the higher-order bias B_T of $\hat{\beta}$ which turns out to be of order T^{-1} , i.e. $E[\hat{\beta} - \beta_0] = B_TT^{-1} + o(T^{-1})$, and therefore does not depend on the bandwidth. Interestingly, one can show that the two higher-order terms in (2.2) do not contribute to that bias (Jun (2007)).

In the class of GMM estimators defined by (2.1) and indexed by the bandwidth S_T , we now characterize the most efficient one under quadratic loss. Specifically, we rank estimators according to the MSE of the approximation $\zeta_T := \beta_0 + \kappa_{1,T}T^{-1/2} + \kappa_{2,T}S_T^{1/2}T^{-1} + \kappa_{3,T}S_T^{-q}T^{-1/2}$.

Theorem 2.1. *Suppose Assumptions 2.1–2.3 hold. Let $\mathcal{W} \in \mathbb{R}^{p \times p}$ be a weighting matrix. Define the weighted MSE $MSE_T := E[(\zeta_T - \beta_0)' \mathcal{W} (\zeta_T - \beta_0)]$. Then*

$$MSE_T = \nu_1T^{-1} + \nu_2S_TT^{-2} + \nu_3S_T^{-2q}T^{-1} + o(\eta_T^2T^{-1})$$

with

$$\begin{aligned}\nu_1 &:= \sum_{s=-\infty}^{\infty} E[g'_tH'_0\mathcal{W}H_0g_{t+s}] \\ \nu_2 &:= \lim_{T \rightarrow \infty} \frac{T}{S_T} E \left[F'_T P'_0 (\hat{\Omega}_T - \bar{\Omega}_T) H'_0 \mathcal{W} H_0 (\hat{\Omega}_T - \bar{\Omega}_T) P_0 F_T \right]\end{aligned}$$

$$\begin{aligned}
& - 2 \lim_{T \rightarrow \infty} \frac{T}{S_T} E \left[F_T' P_0' (\hat{\Omega}_T - \bar{\Omega}_T) H_0' \mathcal{W} H_0 F_T \right] \\
\nu_3 := & \lim_{T \rightarrow \infty} S_T^{2q} E \left[F_T' P_0' (\bar{\Omega}_T - \Omega_0) H_0' \mathcal{W} H_0 (\bar{\Omega}_T - \Omega_0) P_0 F_T \right]
\end{aligned}$$

where all the limits exist, are finite and generally non-zero.

As explained above, the bias of the approximation ζ_T is zero so that the MSE expansion (2.1) represents only the variance of ζ_T . Despite the lack of a bias component of ζ_T , the expansion displays the first-order tradeoff that is relevant for choosing a bandwidth: $\nu_2 S_T T^{-2}$ increases in S_T and $\nu_3 S_T^{-2q} T^{-1}$ decreases in S_T . The terms have the standard order of squared bias and variance of HAC estimators as derived in Andrews (1991).

Under additional conditions, the moments of the approximation ζ_T correspond to the moments of a formal Edgeworth expansion of the cdf of the second-step estimator $\hat{\beta}$. The (finite) moments of such an Edgeworth expansion approximate the distribution of $\hat{\beta}$ up to the specified order even when the corresponding moments of $\hat{\beta}$ do not exist (Götze and Hipp (1978), Rothenberg (1984), Magdalinos (1992)). In this sense, we can regard (2.1) as an approximation of the MSE of $\hat{\beta}$ when $\hat{\beta}$ possesses second moments and as the MSE of an approximate estimator that shares the same Edgeworth expansion up to a specified order.

Remark 2.1. *For linear instrumental variable models with iid variables and normal errors, Kinal (1980) shows that $\hat{\beta}$ has finite moments up to order $l - p$. Similar results have been conjectured for GMM and generalized empirical likelihood estimators (e.g. Kunitomo and Matsushita (2003), Guggenberger (2008)). Therefore, one should be careful in interpreting the MSE approximation in cases when the degree of over-identification is less than two. Other loss functions may then be more appropriate (see Zaman (1981), for example).*

Having established Theorem 2.1, the calculation of an MSE-optimal bandwidth, S_T^* , becomes straightforward: for the second-order term to attain its fastest possible rate of convergence, the terms of order $S_T T^{-2}$ and $S_T^{-2q} T^{-1}$ have to be balanced which is the case for $S_T^* = c^*(q) T^{1/(1+2q)}$ and some constant $c^*(q)$. We refer to this bandwidth as the MSE($\hat{\beta}$)-optimal bandwidth. The bandwidth minimizing the MSE of $\hat{\Omega}_T(\tilde{\beta})$ as derived in Andrews (1991) we call the MSE($\hat{\Omega}$)-optimal bandwidth.

Corollary 2.1. *Under the assumptions of Theorem 2.1 and if $l > p$, minimizing the lowest order of MSE_T involving the bandwidth yields the optimal bandwidth growth rate $T^{1/(1+2q)}$. Moreover, $S_T^* = c^*(q) T^{1/(1+2q)}$ minimizes the higher-order AMSE defined as the*

limit of $HMSE_T := T^{(1+4q)/(1+2q)} \{MSE_T - \nu_1 T^{-1}\}$ with

$$c^*(q) := \left(\frac{c_0 \nu_3}{\nu_2} \right)^{1/(1+2q)}, \quad c_0 := \begin{cases} 2q, & \text{sign}(\nu_2) = \text{sign}(\nu_3) \\ -1, & \text{sign}(\nu_2) \neq \text{sign}(\nu_3) \end{cases}. \quad (2.3)$$

The expressions for ν_2 and ν_3 show that the optimal bandwidth growth rate is governed by the convergence rate of the covariances between the moment functions and the HAC estimator. The bias and variance of the HAC estimator itself only play an indirect role; in particular, the AMSE of $\hat{\beta}$ is not an increasing function of the AMSE of the HAC estimator. In consequence, none of the existing procedures minimizing the AMSE of the HAC estimator (Andrews (1991), Newey and West (1994) and Andrews and Monahan (1992) among others) are optimal in the above sense.

The convergence rate of the MSE of $\hat{\beta}$ is not affected by the bandwidth choice because it is of order $O(T^{-1})$, only the second-order terms of the MSE, converging at an optimal rate of $O(T^{-(1+4q)/(1+2q)})$, are. By choosing kernels of very high order q , this rate can be made arbitrarily close to $O(T^{-2})$. Nevertheless, kernels of order smaller than or equal to 2 are popular because, unlike kernels of higher order, they can produce positive definite covariance matrix estimates.

Remark 2.2. *Interestingly, the semiparametric estimator $\hat{\beta}$ converges at rate $T^{-1/2}$, but the optimal bandwidth minimizing $MSE(\hat{\beta})$ is of the same order as the optimal bandwidth minimizing $MSE(\hat{\Omega})$. This result contrasts the findings in other semiparametric settings such as those studied by Powell and Stoker (1996) and Goldstein and Messer (1992), for example, in which under-smoothing the nonparametric plugin estimator leads to $T^{-1/2}$ -convergence rates of smooth functionals of that nonparametric plugin estimator.*

To gain more insight into which features of the data generating process determine the value of the optimal bandwidth and to be able to directly estimate the quantities involved, the following proposition derives more explicit expressions for the constants ν_i .

Proposition 2.2. *Assume that $\{g_t\}$ follows a linear Gaussian process, viz.*

$$g_t = \sum_{s=0}^{\infty} \Psi_s e_{t-s}$$

for $t = 1, \dots, T$, $e_t \sim N(0, \Sigma_e)$ iid and Ψ_s satisfies $\sum_{s=0}^{\infty} s^4 \|\Psi_s\| < \infty$. Define $\mu_i := \int_{-\infty}^{\infty} k^i(x) dx$ for $i = 1, 2$ and $\Omega_0^{(q)} := 2\pi f^{(q)}$. Then

$$\begin{aligned} \nu_1 &= \text{tr}(\Omega_0 H_0' \mathcal{W} H_0), \\ \nu_2 &= (2\mu_1 + \mu_2)(l - p) \text{tr}(\Sigma_0 \mathcal{W}), \\ \nu_3 &= g_q^2 \text{tr} \left(\Omega_0^{(q)} H_0' \mathcal{W} H_0 \Omega_0^{(q)} P_0 \right). \end{aligned}$$

2.1 Linear IV Model

In this sub-section, we specialize the expressions in Proposition 2.2 to a stylized instrumental variable model that allows us to analyze the difference between the $\text{MSE}(\hat{\beta})$ -optimal and the $\text{MSE}(\hat{\Omega})$ -optimal bandwidths and subsequently serves as the data generating process for the Monte Carlo simulations. Let y_t and w_t be random variables satisfying

$$y_t = \beta_0 w_t + \varepsilon_t$$

and z_t an l -dimensional random vector of instruments such that

$$w_t = \gamma' z_t + v_t$$

where $\iota := (1, \dots, 1)' \in \mathbb{R}^l$ and $\gamma \in \mathbb{R}$. Define $x_t = (y_t, w_t, z_t')'$ so that $g(x_t, \beta) = (y_t - \beta w_t) z_t$. Further let $\{\varepsilon_t\}$ and $\{v_t\}$ be AR(1) processes with autocorrelation coefficient $|\rho| \in (0, 1)$, viz. $\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$ and $v_t = \rho v_{t-1} + u_t$, where

$$\begin{pmatrix} \eta_t \\ u_t \end{pmatrix} \sim iid N \left(0, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix} \right), \quad \sigma_{12} \neq 0,$$

and $z_t \sim iid N(0, \sigma_z^2 I_l)$ independent of $\{(\eta_t, u_t)\}$. Then, one can show that

$$\begin{aligned} c^*(1) &= \left(\frac{4c_0 g_1^2 \rho^2 (1-\rho)^2 [\sigma_z^4 l - (1-\rho)^4]}{(2\mu_1 + \mu_2)(1+\rho)^2 \sigma_z^8 l (l-1)} \right)^{1/(1+2q)}, \\ c^*(2) &= \left(\frac{4c_0 g_2^2 \rho^2 [\sigma_z^4 l - (1-\rho)^4]}{(2\mu_1 + \mu_2) \sigma_z^8 l (l-1)} \right)^{1/(1+2q)}. \end{aligned}$$

In this specific example, we can easily compare the $\text{MSE}(\hat{\beta})$ -optimal with the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth derived in Andrews (1991), $S_T = (qg_q^2/\mu_2 \alpha(q)T)^{1/(1+2q)}$ with $\alpha(q) = 2\text{vec}(\Omega_0^{(q)})' \mathcal{W}_A \text{vec}(\Omega_0^{(q)}) / \text{tr}(\mathcal{W}_A(I_{l^2} + K_u)) (\Omega_0 \otimes \Omega_0)$. For $\mathcal{W}_A = I_{l^2}$, the constants of proportionality become $\alpha(1) = 8\rho^2 / [(1-\rho)^2(1+\rho)^2(l+1)]$ and $\alpha(2) = 8\rho^2 / [(1-\rho)^4(l+1)]$.

Notice that the $\text{MSE}(\hat{\beta})$ -optimal bandwidth adapts to the persistence of the error processes (ρ), to the degree of over identification ($l-1$) and to the amount of variation in the instruments (σ_z^2). On the contrary, the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth depends only on ρ and decreases with l . Therefore, we expect there to be scenarios in which the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth is clearly not $\text{MSE}(\hat{\beta})$ -optimal and the two bandwidths may differ significantly. The simulation evidence in Section 4 confirms these findings.

3 Data-driven Bandwidth Choice

The optimal bandwidth S_T^* is infeasible because it depends on several unknown quantities. In the case in which $\{g_t\}$ is a linear Gaussian process, we require knowledge of Ω_0 , $\Omega_0^{(q)}$, and

G_0 . In this section, I describe a data-driven approach to select the optimal bandwidth by estimating the required quantities based on parametric approximating models for $\{g_t\}$, similarly as proposed in Andrews (1991). The idea is to first construct the first-step estimator $\tilde{\beta}$, then fit a parsimonious auto-regressive (AR) model to $\{g_t(\tilde{\beta})\}_{t=1}^T$ and, finally, to substitute its parameter estimates into analytical formulae for Ω_0 and $\Omega_0^{(q)}$ assuming that the AR is the true model. Together with the usual sample average estimator for G_0 , these estimates, are then substituted into the expressions of ν_2 and ν_3 in Proposition 2.2 to yield estimates of the optimal bandwidths.

We focus on estimating a univariate AR(1) model for each component of $\{g_t(\tilde{\beta})\}$, although other approximating models like vector autoregressions or models with more lags could be considered. Let $\hat{\rho}_i$ and $\hat{\sigma}_i$ be the estimated coefficient and the residual variance of the i -th estimated AR(1) process. We can construct estimators of Ω_0 and $\Omega_0^{(q)}$ as $\hat{\Omega}_0 := \text{diag}(\hat{\omega}_1, \dots, \hat{\omega}_l)$ and $\hat{\Omega}_0^{(q)} := \text{diag}(\hat{\omega}_1^{(q)}, \dots, \hat{\omega}_l^{(q)})$ with $\hat{\omega}_i := \hat{\sigma}_i^2 / (1 - \hat{\rho}_i)^2$, $\hat{\omega}_i^{(1)} := 2\hat{\sigma}_i^2 \hat{\rho}_i / [(1 - \hat{\rho}_i)^3 (1 + \hat{\rho}_i)]$ and $\hat{\omega}_i^{(2)} := 2\hat{\sigma}_i^2 \hat{\rho}_i / [(1 - \hat{\rho}_i)^4]$. Then, estimate H_0 , Σ_0 and P_0 by $\hat{H}_0 := \hat{\Sigma}_0 G_T(\tilde{\beta})' \hat{\Omega}_0^{-1}$, $\hat{\Sigma}_0 := (G_T(\tilde{\beta})' \hat{\Omega}_0^{-1} G_T(\tilde{\beta}))^{-1}$ and $\hat{P}_0 := \hat{\Omega}_0^{-1} - \hat{\Omega}_0^{-1} G_T(\tilde{\beta}) \hat{H}_0$. Finally, substitute all these expressions into the formulae of Proposition 2.2 to get estimates $\hat{\nu}_2$ and $\hat{\nu}_3$ of ν_2 and ν_3 , and the estimator of the optimal bandwidth,

$$\hat{S}_T := \left(\frac{c_0 \hat{\nu}_3}{\hat{\nu}_2} \right)^{1/(1+2q)} T^{1/(1+2q)}.$$

The difference in performance one incurs by using \hat{S}_T instead of the infeasible bandwidth minimizing the finite-sample MSE of $\hat{\beta}$ has four sources: the error made by replacing the MSE of $\hat{\beta}$ by the MSE of the first terms in the higher-order expansion (ζ_T), the error due to the large sample approximation of the MSE, the estimation error in $\hat{\Omega}_0$ and $\hat{\Omega}_0^{(q)}$, and the error made by potential misspecification of the approximating parametric model for $\{g_t\}$. In practice, one hopes that these errors are small. As mentioned in the discussion after Theorem 2.1, the first type of error vanishes with the sample size under additional assumptions. The second and third type also disappear as $T \rightarrow \infty$. The fourth type of error can typically be conjectured to be negligible because the MSE of $\hat{\beta}$ tends to be relatively flat around its minimum (as is the case in the Monte Carlo simulations, for example), so that misspecification in the approximating model is not expected to have a large impact on the properties of the resulting GMM estimator. Nevertheless, the applied researcher should bear in mind that the plugin procedure is not “automatic” and some thought has to go into selecting an appropriate approximating model and the potential impact of the aforementioned types of errors has to be considered.

Remark 3.1. *As in Andrews and Monahan (1992) one may want to consider pre-whitening*

the series $\{g_t(\tilde{\beta})\}_{t=1}^T$ before fitting the AR process. The reported increases in accuracy of test statistics in *Andrews and Monahan (1992)* and *Newey and West (1994)* are expected to occur with the procedure presented here as well.

4 Simulations

In this section, we discuss a small simulation experiment that illustrates the theoretical findings from the previous sections. We simulate the model from Section 2.1, denoted by “AR(1)”, for different degrees of serial correlation ($\rho \in \{0.1, 0.5, 0.9\}$), weak and strong instruments ($\gamma \in \{0.1, 1\}$) and increasing number of instruments ($l \in \{2, 3, 4, 5, 10, 15, 20\}$). We also consider a variant of the model in which the AR(1) error process is replaced by an MA(1), i.e. $\varepsilon_t = \rho\eta_{t-1} + \eta_t$ and $v_t = \rho u_{t-1} + u_t$. The variance of the instruments is chosen so as to balance signal and noise across different values of l and ρ : $\sigma_z^2 = 1/((1 - \rho^2)l^2)$ for the AR(1) and $\sigma_z^2 = (1 + \rho^2)/l^2$ for the MA(1). We simulate 1,000 samples and, to save space, present only results for sample size $T = 128$, $\sigma_{12} = 0.5$, $\beta_0 = 1$ and the Parzen kernel. Other parameter combinations yield similar results.

Table 1 reports four different bandwidths (“bw”) averaged over the simulation samples: “optimal”, “Andrews”, “naive” and “sim”, referring to the $\text{MSE}(\hat{\beta})$ -optimal, the $\text{MSE}(\hat{\Omega})$ -optimal, the naive choice $S_T = T^{1/(1+2q)}$ and to the (infeasible) bandwidth that minimizes the simulated $\text{MSE}(\hat{\beta})$ over a grid of bandwidths, respectively. The table also shows the bias, standard deviation (“SD”) and MSE of $\hat{\beta}$. In almost all cases considered here, the $\text{MSE}(\hat{\beta})$ -optimal bandwidth is closer to the one minimizing the simulated MSE than the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth. The $\text{MSE}(\hat{\beta})$ -optimal bandwidth tends to be around 20% or more smaller than the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth one. The results, together with unreported simulations at intermediate values of ρ , show that the $\text{MSE}(\hat{\beta})$ -optimal bandwidth leads to lower MSE values than the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth, except when $|\rho|$ is close to one.

Tables 2 and 3 show the ratios of MSE (“MSE ratio”) and higher-order MSE (“HMSE ratio”), as defined in Corollary 2.1, based on the $\text{MSE}(\hat{\beta})$ -optimal bandwidth divided by those based on the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth. The number of instruments is fixed at $l = 4$, but other numbers yield quantitatively the same results. μ^2/l denotes the standardized concentration parameter measuring the strength of the instruments (*Stock, Wright, and Yogo (2002)*). The MSE ratios demonstrate that the $\text{MSE}(\hat{\beta})$ -optimal bandwidth yields an HMSE up to 55% smaller than that of the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth. The MSE is smaller by up to 2%, again, except when $|\rho|$ is close to one.

Unlike the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth defined by *Andrews (1991)*, the $\text{MSE}(\hat{\beta})$ -

optimal bandwidth is formally not defined for the case $l = p$ in which the estimator $\hat{\beta}$ is independent of the weighting matrix. To study scenarios in which l/p is close to this boundary, we conclude this section by considering a robustness check in which l/p approaches one. Table 4 reports the same MSE and HMSE ratio as Tables 2 and 3, but for a sequence $l/p \in \{5/1, 4/2, 3/2, 4/3, 5/4, 8/7, 10/9\}$ that approaches one. ρ and γ are fixed at values 0.5 and 1, but other values yield similar results. The MSE based on the $\text{MSE}(\hat{\beta})$ -optimal bandwidth stays close to or slightly smaller than the one based on the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth for all values of l/p . Similarly, the HMSE is significantly smaller for the optimal bandwidth. In the case of the MA(1) model, the HMSE gains are even around 40%.

5 Conclusion

This paper develops a selection procedure for the bandwidth of a HAC estimator of the optimal GMM weighting matrix which minimizes the asymptotic MSE of the resulting two-step GMM estimator. We show that it is of the same order as the bandwidth minimizing the MSE of the nonparametric plugin estimator, but the constants of proportionality differ significantly. The simulation study suggests that the data-driven version of the selection procedure works well in finite samples and significantly reduces the second-order MSE of the GMM estimator.

Notes

¹I thank Michael Jansson for making me aware of this work.

References

- ANATOLYEV, S. (2005): “GMM, GEL, Serial Correlation, and Asymptotic Bias,” *Econometrica*, 73(3), 983–1002.
- ANDREWS, D. W. K. (1991): “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation,” *Econometrica*, 59(3), 817–58.
- ANDREWS, D. W. K., AND J. C. MONAHAN (1992): “An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator,” *Econometrica*, 60(4), 953–966.

- GOLDSTEIN, L., AND K. MESSER (1992): “Optimal Plug-in Estimators for Nonparametric Functional Estimation,” *The Annals of Statistics*, 20(3), 1306–1328.
- GÖTZE, F., AND C. HIPPE (1978): “Asymptotic Expansions in the Central Limit Theorem under Moment Conditions,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 42, 67–87.
- GUGGENBERGER, P. (2008): “Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator,” *Econometric Reviews*, 27(4), 526–541.
- HALL, P., AND C. C. HEYDE (1980): *Martingale Limit Theory and Its Applications*. Academic Press, New York.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.
- HANSEN, L. P., J. HEATON, AND A. YARON (1996): “Finite-Sample Properties of Some Alternative GMM Estimators,” *Journal of Business and Economic Statistics*, 14(3), 262–280.
- IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica*, 66(2), 333–357.
- JANSSON, M. (2002): “Consistent Covariance Matrix Estimation for Linear Processes,” *Econometric Theory*, 18, 1449–1459.
- JUN, B. H. (2007): “Essays in Econometrics,” Ph.D. thesis, University of California, Berkeley.
- KIEFER, N., AND T. VOGELSANG (2002a): “Heteroskedasticity-Autocorrelation Robust Standard Errors Using The Bartlett Kernel Without Truncation,” *Econometrica*, 70(5), 2093–2095.
- (2002b): “Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size,” *Econometric Theory*, 18, 1350–1366.
- (2005): “A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests,” *Econometric Theory*, 21, 1130–1164.
- KINAL, T. W. (1980): “The Existence of Moments of k-Class Estimators,” *Econometrica*, 48(1), pp. 241–249.

- KITAMURA, Y., AND M. STUTZER (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica*, 65(4), 861–874.
- KUNITOMO, N., AND Y. MATSUSHITA (2003): “Finite Sample Distributions of the Empirical Likelihood Estimator and the GMM Estimator,” Discussion Paper F-200, CIRJE.
- MAGDALINOS, M. A. (1992): “Stochastic Expansions and Asymptotic Approximations,” *Econometric Theory*, 8(3), 343–367.
- 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.
- NEWNEY, W. K., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2111–2245. Elsevier Science B.V.
- NEWNEY, W. K., AND R. J. SMITH (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72(1), 219–255.
- NEWNEY, W. K., AND K. WEST (1994): “Automatic Lag Selection in Covariance Matrix Estimation,” *The Review of Economic Studies*, 61(4), 631–653.
- OWEN, A. B. (1988): “Empirical Likelihood Ratio Confidence Intervals for a Single Functional,” *Biometrika*, 75(2), 237–249.
- PARZEN, E. (1957): “On Consistent Estimates of the Spectrum of a Stationary Time Series,” *The Annals of Mathematical Statistics*, 28(2), 329–348.
- POWELL, J. L., AND T. M. STOKER (1996): “Optimal Bandwidth Choice for Density-weighted Averages,” *Journal of Econometrics*, 75, 291–316.
- QIN, J., AND J. LAWLESS (1994): “Empirical Likelihood and General Estimating Equations,” *The Annals of Statistics*, 22(1), 300–325.
- RILSTONE, P., V. K. SRIVASTAVA, AND A. ULLAH (1996): “The Second-order Bias and Mean Squared Error of Nonlinear Estimators,” *Journal of Econometrics*, 75(2), 369–395.
- ROBINSON, P. M. (1991): “Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models,” *Econometrica*, 59(5), 1329–1363.

- ROTHENBERG, T. (1984): “Approximating the Distributions of Econometric Estimators and Test Statistics,” in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. II, pp. 881–935. Elsevier Science Publishers B.V.
- STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments,” *Journal of Business and Economic Statistics*, 20(4).
- SUN, Y., AND P. C. B. PHILLIPS (2008): “Optimal Bandwidth Choice for Interval Estimation in GMM Regression,” Discussion Paper 1661, Cowles Foundation, Yale University.
- SUN, Y., P. C. B. PHILLIPS, AND S. JIN (2008): “Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing,” *Econometrica*, 76(1), 175–194.
- TAMAKI, K. (2007): “Second Order Optimality for Estimators in Time Series Regression Models,” *Journal of Multivariate Analysis*, 98, 638–659.
- WHITE, H., AND I. DOMOWITZ (1984): “Nonlinear Regression with Dependent Observations,” *Econometrica*, 52(1), 143–162.
- XIAO, Z., AND P. C. B. PHILLIPS (1998): “Higher-order Approximations for Frequency Domain Time Series Regression,” *Journal of Econometrics*, 86, 297–336.
- ZAMAN, A. (1981): “Estimators Without Moments: The Case of the Reciprocal of a Normal Mean,” *Journal of Econometrics*, 15(2), 289–298.

A Proofs

Lemma A.1. *Under Assumptions 2.1–2.2 and 2.3(a)–(c), $\hat{\beta} - \beta_0 = \psi T^{-1/2} + o_p(T^{-1/2})$ with $\psi = -G_0^{-1} \sqrt{T} \hat{g} = O_p(1)$.*

Proof. We need to check the assumptions of Newey and McFadden (1994, Theorem 3.2) with \hat{W} replaced by $\hat{\Omega}_T(\tilde{\beta})$. First of all, by Assumptions 2.1–2.2 and Andrews (1991, Theorem 1(b)) $\hat{\Omega}_T(\tilde{\beta}) \xrightarrow{p} \Omega_0$. (i), (ii) and (v) hold by assumption. (iii) and (iv) hold by Assumption 2.1 and White and Domowitz (1984, Theorem 2.3, 2.4). Q.E.D.

Proof of Proposition 2.1. Step I: Expansion of the optimal weighting matrix. A Taylor expansion of $\hat{\Omega}_T(\tilde{\beta})$ around β_0 yields

$$\begin{aligned} \text{vec}\left(\hat{\Omega}_T(\tilde{\beta})\right) &= \text{vec}\left(\hat{\Omega}_T\right) + \nabla\hat{\Omega}_T(\bar{\beta})(\tilde{\beta} - \beta_0) \\ &= \text{vec}(\Omega_0) + \text{vec}\left(\hat{\Omega}_T - \Omega_0\right) + \nabla\Omega_0(\tilde{\beta} - \beta_0) + \left(\nabla\hat{\Omega}_T - \nabla\Omega_0\right)(\tilde{\beta} - \beta_0) \\ &\quad + \left(\nabla\hat{\Omega}_T(\bar{\beta}) - \nabla\hat{\Omega}_T\right)(\tilde{\beta} - \beta_0) \end{aligned} \quad (\text{A.1})$$

where $\bar{\beta}$ lies on the line segment joining $\tilde{\beta}$ and β_0 . By Assumptions 2.1–2.2 and Andrews (1991, Proposition 1(a),(b)), $\hat{\Omega}_T - \Omega_0 = \omega_{1,T}S_T^{1/2}T^{-1/2} + \omega_{2,T}S_T^{-q}$ with $\omega_{i,T} = O_p(1)$, $i = 1, 2$, and $\nabla\hat{\Omega}_T - \nabla\Omega_0 = O_p(\eta_T)$. Next, we show that $\nabla\hat{\Omega}_T(\bar{\beta}) - \nabla\hat{\Omega}_T = O_p(\|\tilde{\beta} - \beta_0\|)$. To this end, let $\bar{g}_t := g_t(\bar{\beta})$. Notice that, by Assumption 2.1(g), $\tilde{\beta} \in \mathcal{N}$ w.p.a. 1 and, thus, $\bar{\beta} \in \mathcal{N}$ w.p.a. 1. From the Lipschitz condition, Assumption 2.3(d), we get

$$\begin{aligned} \left\|\nabla\hat{\Omega}_T(\bar{\beta}) - \nabla\hat{\Omega}_T\right\| &\leq \frac{1}{T} \sum_{s=1-T}^{T-1} \sum_{t=\max\{1,1-s\}}^{\min\{T,T-s\}} \left\|k\left(\frac{s}{S_T}\right) [\nabla\bar{g}_{t+s}\bar{g}'_t - \nabla g_{t+s}g'_t]\right\| \\ &\leq \frac{C}{T} \sum_{s=1-T}^{T-1} \sum_{t=\max\{1,1-s\}}^{\min\{T,T-s\}} d(x_{t+s}, x_t) \|\tilde{\beta} - \beta_0\| \\ &= C \left(\sum_{s=-\infty}^{\infty} E d(x_{t+s}, x_t) + o_p(1) \right) \|\tilde{\beta} - \beta_0\| \end{aligned} \quad (\text{A.2})$$

which holds w.p.a. 1 and for some constant C . (A.1) together with (A.2) and the first-order asymptotics in Assumption 2.1(g) then imply

$$\hat{\Omega}_T(\tilde{\beta}) = \Omega_0 + \omega_{1,T}S_T^{1/2}T^{-1/2} + \omega_{2,T}S_T^{-q} + R_T \quad (\text{A.3})$$

with $R_T = O_p(\eta_T T^{-1/2})$, $\omega_{1,T} := \sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) = O_p(1)$ and $\omega_{2,T} := S_T^q(\bar{\Omega}_T - \Omega_0) = O(1)$.

Step II: Expansion of the second-step GMM estimator. Write the second-stage estimator $\hat{\theta} := (\hat{\beta}', \hat{\lambda}')'$ of $\theta_0 := (\beta'_0, 0)' \in \mathcal{B} \times \Lambda$, $\Lambda := [0, \infty)^l$, as the solution to

$$\begin{pmatrix} G_T(\hat{\beta})'\hat{\lambda} \\ \hat{g}(\hat{\beta}) + \hat{\Omega}_T(\tilde{\beta})\hat{\lambda} \end{pmatrix} = 0. \quad (\text{A.4})$$

Further, define $\hat{m}(\theta) := \frac{1}{T} \sum_{t=1}^T m_t(\theta)$ with

$$m_t(\theta) := - \begin{pmatrix} G(x_t, \beta)'\lambda \\ g_t(\beta) + \Omega_0\lambda \end{pmatrix}$$

for some $\theta := (\beta', \lambda')' \in \mathcal{B} \times \Lambda$. Then use the expansion in (A.3) to rewrite (A.4) as

$$0 = \hat{m}(\hat{\theta}) - \begin{pmatrix} 0 \\ (\omega_{1,T}S_T^{1/2}T^{-1/2} + \omega_{2,T}S_T^{-q} + R_T)' \hat{\lambda} \end{pmatrix}. \quad (\text{A.5})$$

Next, consider $\hat{\lambda} = -\hat{\Omega}_T(\tilde{\beta})^{-1}\hat{g}(\tilde{\beta})$. By Assumptions 2.1–2.2 and Andrews (1991, Proposition 1(a),(b), Theorem 1(b)), $\hat{\Omega}_T(\tilde{\beta}) - \Omega_0 = O_p(\eta_T)$. Also, by an expansion of $\hat{g}(\tilde{\beta})$ around β_0 , Lemma A.1, Assumption 2.1 and the CLT, $\hat{g}(\tilde{\beta}) = (I_l - G_0H_0)\hat{g} + o_p(T^{-1/2})$, and thus

$$\begin{aligned} \hat{\lambda} &= -[\Omega_0 + O_p(\eta_T)]^{-1} ((I_l - G_0H_0)F_T T^{-1/2} + o_p(T^{-1/2})) \\ &= -P_0F_T T^{-1/2} + O_p(\eta_T T^{-1/2}) + o_p(T^{-1/2}). \end{aligned} \quad (\text{A.6})$$

Consider the following expansion of $\hat{m}(\hat{\theta})$ around θ_0 :

$$\hat{m}(\hat{\theta}) = \hat{m}(\theta_0) + \nabla \hat{m}(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2} \nabla^2 \hat{m}(\bar{\theta}) \left[(\hat{\theta} - \theta_0) \otimes (\hat{\theta} - \theta_0) \right],$$

where $\bar{\theta}$ lies on the line segment joining $\hat{\theta}$ and θ_0 . By Assumptions 2.1(f) and 2.3(c)–(d), $\nabla^2 \hat{m}(\bar{\theta})$ is $O_p(1)$. Lemma A.1, Assumption 2.1 and the CLT for mixing sequences then imply

$$\hat{m}(\hat{\theta}) = \hat{m}(\theta_0) + M_1(\hat{\theta} - \theta_0) + O_p(T^{-1}). \quad (\text{A.7})$$

Substituting (A.6) and (A.7) into (A.5) and solving for $\hat{\theta} - \theta_0$ yields

$$\begin{aligned} \hat{\theta} - \theta_0 &= -M_1^{-1} \hat{m}(\theta_0) - M_1^{-1} \begin{pmatrix} 0 \\ (\omega'_{1,T}S_T^{1/2}T^{-1/2} + \omega'_{2,T}S_T^{-q})P_0F_T T^{-1/2} \end{pmatrix} \\ &\quad + O_p(T^{-1}) + O_p(\eta_T^2 T^{-1/2}) + o_p(\eta_T T^{-1/2}) \end{aligned}$$

where

$$M_1 = - \begin{pmatrix} 0 & G'_0 \\ G_0 & \Omega_0 \end{pmatrix}, \quad M_1^{-1} = - \begin{pmatrix} -\Sigma_0 & H_0 \\ H'_0 & P_0 \end{pmatrix}.$$

Therefore,

$$\sqrt{T}(\hat{\beta} - \beta_0) = -H_0F_T + H_0\omega'_{1,T}P_0F_T S_T^{1/2}T^{-1/2} + H_0\omega'_{2,T}P_0F_T S_T^{-q} + o_p(\eta_T).$$

Since F_T , $\omega_{1,T}$ and $\omega_{2,T}$ are $O_p(1)$, we also have that $\kappa_{i,T} = O_p(1)$ for $i = 1, 2, 3$. Q.E.D.

Proof of Theorem 2.1. We need to derive the order of $E[\kappa'_{i,T}\mathcal{W}\kappa_{j,T}]$, $i, j \in \{1, 2, 3\}$. Consider the case $i = j = 1$:

$$\begin{aligned} E[\kappa'_{1,T}\mathcal{W}\kappa_{1,T}] &= E[F'_T H'_0 \mathcal{W} H_0 F_T] = \frac{1}{T} \sum_{s,t=1}^T E[g'_t H'_0 \mathcal{W} H_0 g_s] \\ &= \sum_{s=-(T-1)}^{T-1} \left(1 - \frac{s}{T}\right) E[g'_t H'_0 \mathcal{W} H_0 g_{t-s}] \rightarrow \sum_{s=-\infty}^{\infty} E[g'_t H'_0 \mathcal{W} H_0 g_{t-s}]. \end{aligned}$$

The limiting sum can be shown to be finite using Assumption 2.1, the Hölder Inequality and the mixing inequality of Hall and Heyde (1980, Corollary A.2). Similarly, we can show that $E[\kappa'_{2,T}\mathcal{W}\kappa_{2,T}]$ and $E[\kappa'_{3,T}\mathcal{W}\kappa_{3,T}]$ are $O(1)$, $E[\kappa'_{1,T}\mathcal{W}\kappa_{2,T}] = O(S_T^{1/2}T^{-1/2})$, but $E[\kappa'_{1,T}\mathcal{W}\kappa_{3,T}] = o(\eta_T)$ and $E[\kappa'_{2,T}\mathcal{W}\kappa_{3,T}] = o(1)$. Q.E.D.

Proof of Proposition 2.2. From the proof of Theorem 2.1,

$$E[\kappa'_{1,T}\mathcal{W}\kappa_{1,T}] \rightarrow \sum_{s=-\infty}^{\infty} E[g'_t H'_0 \mathcal{W} H_0 g_{t-s}] = \text{vec}(\Omega_0)' \text{vec}(H'_0 \mathcal{W} H_0).$$

By Andrews (1991, Proposition 1(b)) we have $S_T^q(\bar{\Omega}_T - \Omega_0) \rightarrow -g_q \Omega_0^{(q)}$ and thus

$$\begin{aligned} E[\kappa'_{3,T}\mathcal{W}\kappa_{3,T}] &= E[F'_T P'_0 S_T^q(\bar{\Omega}_T - \Omega_0) H'_0 \mathcal{W} H_0 S_T^q(\bar{\Omega}_T - \Omega_0) P_0 F_T] \\ &\rightarrow g_q^2 \sum_{s=-\infty}^{\infty} E \left[g'_t P'_0 \Omega_0^{(q)} H'_0 \mathcal{W} H_0 \Omega_0^{(q)} P_0 g_{t-s} \right] \\ &= g_q^2 \text{vec}(\Omega_0)' \text{vec} \left(P'_0 \Omega_0^{(q)} H'_0 \mathcal{W} H_0 \Omega_0^{(q)} P_0 \right) \\ &= g_q^2 \text{tr} \left(\Omega_0^{(q)} H'_0 \mathcal{W} H_0 \Omega_0^{(q)} P_0 \right) \end{aligned}$$

because, for conformable matrices A, B , $\text{tr}(AB) = \text{tr}(BA)$, and $P_0 G_0 H_0 = 0$. Next, consider the terms

$$\begin{aligned} E[\kappa'_{1,T}\mathcal{W}\kappa_{2,T}] &= E \left[F'_T H'_0 \mathcal{W} H_0 \sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) P_0 F_T \right] \\ &= E \left[(F_T \otimes F_T)' \text{vec} \left(H'_0 \mathcal{W} H_0 \sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) P_0 \right) \right] \\ &= E \left[(F_T \otimes F_T)' (P'_0 \otimes H'_0 \mathcal{W} H_0) \text{vec} \left(\sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) \right) \right] \\ &= E \left[\left(F_T \otimes F_T \otimes \text{vec} \left(\sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) \right) \right)' \right] \text{vec} (P'_0 \otimes H'_0 \mathcal{W} H_0) \end{aligned}$$

and, similarly,

$$\begin{aligned} E[\kappa'_{2,T}\mathcal{W}\kappa_{2,T}] &= E \left[F'_T P'_0 \sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) H'_0 \mathcal{W} H_0 \sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T) P_0 F_T \right] \\ &= \frac{T}{S_T} \text{vec}(P'_0 \otimes P'_0)' E \left[F_T \otimes F_T \otimes (\hat{\Omega}_T - \bar{\Omega}_T) \otimes (\hat{\Omega}_T - \bar{\Omega}_T) \right] \text{vec}(H'_0 \mathcal{W} H_0). \end{aligned}$$

In order to find expressions for these two cross-products involving $\sqrt{T/S_T}(\hat{\Omega}_T - \bar{\Omega}_T)$, we make use of the BN-decomposition of the linear process $\{g_t\}$, viz.

$$g_t = \Psi e_t + \tilde{e}_{t-1} - \tilde{e}_t$$

where $\Psi := \sum_{j \geq 0} \Psi_j$, $\tilde{e}_t := \sum_{j \geq 0} \tilde{\Psi}_j e_{t-j}$ and the tail sums $\tilde{\Psi}_j := \sum_{k \geq j+1} \Psi_k$. With this representation of $\{g_t\}$ we can calculate limiting variances and covariances of g_t based only

on Ψe_t and disregard the transient part of the process, $\tilde{e}_{t-1} - \tilde{e}_t$. Since $e_t \sim N(0, \Sigma_e)$, third and fourth moments are zero. Therefore,

$$\begin{aligned}
& S_T^{1/2} T^{3/2} E \left[F_T \otimes F_T \otimes \text{vec} \left(\sqrt{T/S_T} (\hat{\Omega}_T - \bar{\Omega}_T) \right) \right] \\
&= E \left[\sum_{s,t=1}^T g_t \otimes g_s \otimes \left(\sum_{v=1-T}^{T-1} \sum_{u=\max\{1,1-v\}}^{\min\{T,T-v\}} k \left(\frac{v}{S_T} \right) \text{vec} (g_{u+v} g'_u - E[g_{u+v} g'_u]) \right) \right] \\
&= E \left[\sum_{t=1}^T \Psi e_t \otimes \Psi e_t \otimes \left(\sum_{\substack{u=1 \\ u \neq t}}^T \text{vec} (\Psi e_u \Psi e'_u - E[\Psi e_u \Psi e'_u]) \right) \right] + o(1) \\
&\quad + E \left[\sum_{s,t=1}^T \Psi e_t \otimes \Psi e_s \otimes \left(\sum_{\substack{v=1-T \\ v \neq 0}}^{T-1} \sum_{u=\max\{1,1-v\}}^{\min\{T,T-v\}} k \left(\frac{v}{S_T} \right) \text{vec} (\Psi e_{u+v} \Psi e'_u) \right) \right] \\
&= \sum_{s,t=1}^T \sum_{\substack{v=1-T \\ v \neq 0}}^{T-1} \sum_{u=\max\{1,1-v\}}^{\min\{T,T-v\}} k \left(\frac{v}{S_T} \right) E [\Psi e_t \otimes \Psi e_s \otimes \Psi e_u \otimes \Psi e_{u+v}] + o(1) \\
&= \sum_{s,t=1}^T k \left(\frac{t-s}{S_T} \right) E [\Psi e_t \otimes \Psi e_s \otimes \Psi e_s \otimes \Psi e_t] \\
&\quad + \sum_{s,t=1}^T k \left(\frac{t-s}{S_T} \right) E [\Psi e_t \otimes \Psi e_s \otimes \Psi e_t \otimes \Psi e_s] + o(1)
\end{aligned}$$

One can show that $E[\Psi e_t \otimes \Psi e_s \otimes \Psi e_s \otimes \Psi e_t] = E[\Psi e_t \otimes \Psi e_s \otimes \Psi e_t \otimes \Psi e_s] = \text{vec}(\Omega_0 \otimes \Omega_0)$ when $s \neq t$ and 0 otherwise. Noticing $\frac{1}{S_T T} \sum_{s,t=1}^T k((t-s)/S_T) \rightarrow \mu_1$, we then have

$$\begin{aligned}
\sqrt{\frac{T}{S_T}} E [\kappa'_{1,T} \mathcal{W} \kappa_{2,T}] &\rightarrow 2\mu_1 \text{vec}(\Omega_0 \otimes \Omega_0)' \text{vec} (P_0 \otimes H'_0 \mathcal{W} H_0) \\
&= 2\mu_1 \text{tr}((\Omega_0 \otimes \Omega_0)(P_0 \otimes H'_0 \mathcal{W} H_0)) \\
&= 2\mu_1 \text{tr}(\Omega_0 P_0 \otimes \Omega_0 H'_0 \mathcal{W} H_0) \\
&= 2\mu_1 \text{tr}(I_l - G_0 H_0) \text{tr}(\Omega_0 \Omega_0^{-1} G_0 \Sigma_0 \mathcal{W} \Sigma_0 G'_0 \Omega_0^{-1}) \\
&= 2\mu_1 (l - \text{tr}(H_0 G_0)) \text{tr}(\Sigma_0 \mathcal{W} \Sigma_0 G'_0 \Omega_0^{-1} G_0) \\
&= 2\mu_1 (l - p) \text{tr}(\Sigma_0 \mathcal{W})
\end{aligned}$$

which uses the fact that $H_0 G_0 = I_p$. By a similar derivation, $E[\kappa'_{2,T} \mathcal{W} \kappa_{2,T}] \rightarrow \mu_2 (l - p) \text{tr}(\Sigma_0 \mathcal{W})$ so that $\nu_2 = (2\mu_1 + \mu_2)(l - p) \text{tr}(\Sigma_0 \mathcal{W})$. Q.E.D.

		AR(1)							
		$\gamma = 0.1$				$\gamma = 1$			
ρ		optimal	Andrews	naive	sim	optimal	Andrews	naive	sim
0.1	bw	1.978	2.530	2.639	2.278	2.177	2.532	2.639	0.000
	bias	0.383	0.384	0.385	0.385	0.012	0.012	0.012	0.013
	SD	0.534	0.535	0.534	0.534	0.092	0.093	0.093	0.092
	MSE	55.217	55.406	55.473	55.417	1.113	1.118	1.116	1.105
0.5	bw	2.114	2.755	2.639	0.000	2.322	2.723	2.639	1.733
	bias	0.393	0.395	0.394	0.392	0.012	0.012	0.012	0.013
	SD	0.578	0.584	0.583	0.567	0.092	0.093	0.093	0.092
	MSE	62.517	63.544	63.398	60.752	1.112	1.118	1.115	1.111
0.9	bw	2.297	3.048	2.639	2.278	2.567	3.090	2.639	5.000
	bias	0.350	0.347	0.348	0.349	0.011	0.011	0.011	0.011
	SD	0.793	0.794	0.795	0.794	0.085	0.085	0.085	0.084
	MSE	96.049	95.977	96.304	96.136	0.942	0.933	0.944	0.924
		MA(1)							
		$\gamma = 0.1$				$\gamma = 1$			
ρ		optimal	Andrews	naive	sim	optimal	Andrews	naive	sim
0.1	bw	1.979	2.530	2.639	2.278	2.177	2.531	2.639	0.000
	bias	0.383	0.384	0.385	0.385	0.012	0.012	0.012	0.013
	SD	0.534	0.535	0.535	0.534	0.092	0.093	0.093	0.092
	MSE	55.248	55.447	55.522	55.463	1.111	1.116	1.115	1.104
0.5	bw	2.063	2.666	2.639	2.278	2.267	2.654	2.639	1.733
	bias	0.374	0.375	0.375	0.375	0.012	0.012	0.012	0.012
	SD	0.561	0.561	0.562	0.562	0.093	0.093	0.093	0.092
	MSE	58.152	58.235	58.347	58.345	1.116	1.119	1.118	1.112
0.9	bw	2.108	2.754	2.639	2.278	2.330	2.741	2.639	2.278
	bias	0.367	0.367	0.367	0.367	0.011	0.011	0.011	0.011
	SD	0.540	0.539	0.538	0.538	0.092	0.092	0.092	0.092
	MSE	54.558	54.395	54.224	54.193	1.088	1.091	1.089	1.088

Table 1: Bandwidths (“bw”), bias, standard deviation (“SD”) and MSE of $\hat{\beta}$ when computed based on the $\text{MSE}(\hat{\Omega})$ -optimal (“optimal”), the $\text{MSE}(\hat{\Omega})$ -optimal (“Andrews”), $S_T = T^{1/(1+2q)}$ (“naive”) or the simulated MSE-minimizing (“sim”) bandwidth.

		AR(1)					
ρ	l	$\gamma = 0.1$			$\gamma = 1$		
		MSE ratio	HMSE ratio	μ^2/l	MSE ratio	HMSE ratio	μ^2/l
0.1	2	0.996	0.590	0.724	0.999	0.898	64.697
	3	1.004	0.872	0.482	1.000	0.938	43.507
	4	0.997	0.891	0.373	0.996	0.944	32.729
	5	1.000	0.900	0.282	0.996	0.938	25.933
	10	0.997	0.901	0.149	0.990	0.930	13.092
	15	0.996	0.905	0.103	0.993	0.925	8.771
	20	0.999	0.905	0.092	0.986	0.912	6.515
0.5	2	1.008	0.454	0.732	0.999	0.901	65.719
	3	0.997	0.656	0.470	0.997	0.936	43.961
	4	0.984	0.919	0.375	0.995	0.941	33.321
	5	1.001	0.890	0.271	0.995	0.928	26.226
	10	0.992	0.889	0.149	0.990	0.908	13.206
	15	0.995	0.883	0.104	1.001	0.897	8.888
	20	1.002	0.883	0.091	1.012	0.887	6.602
0.9	2	0.988	0.748	0.777	1.004	0.895	75.421
	3	1.035	0.684	0.569	1.005	0.929	51.464
	4	1.001	0.993	0.394	1.009	0.926	38.605
	5	1.007	0.873	0.298	1.006	0.911	30.451
	10	1.008	0.863	0.157	1.016	0.876	15.011
	15	1.014	0.853	0.122	1.068	0.858	10.263
	20	1.015	0.846	0.097	1.102	0.853	7.636

Table 2: Ratios of MSE (“MSE ratio”) and higher-order MSE (“HMSE ratio”) based on the $\text{MSE}(\hat{\beta})$ -optimal bandwidth divided by those based on the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth. μ^2/l is the standardized concentration parameter measuring the strength of the instruments.

		MA(1)					
ρ	l	$\gamma = 0.1$			$\gamma = 1$		
		MSE ratio	HMSE ratio	μ^2/l	MSE ratio	HMSE ratio	μ^2/l
0.1	2	0.989	0.410	0.724	0.999	0.898	64.699
	3	1.005	0.867	0.482	1.000	0.938	43.508
	4	0.996	0.895	0.373	0.996	0.944	32.726
	5	1.000	0.899	0.282	0.996	0.939	25.935
	10	0.997	0.902	0.150	0.990	0.931	13.090
	15	0.996	0.905	0.103	0.993	0.925	8.770
	20	0.999	0.905	0.092	0.987	0.912	6.513
0.5	2	0.987	0.965	0.767	0.998	0.906	65.368
	3	1.009	0.922	0.468	0.998	0.938	43.610
	4	0.999	0.908	0.375	0.997	0.941	33.020
	5	0.999	0.904	0.270	0.998	0.932	26.071
	10	0.992	0.892	0.151	0.985	0.915	13.107
	15	0.998	0.895	0.100	0.996	0.909	8.813
	20	1.001	0.892	0.091	1.003	0.896	6.538
0.9	2	0.982	0.984	0.788	0.998	0.909	65.827
	3	1.001	0.894	0.451	0.996	0.936	43.661
	4	1.003	0.879	0.368	0.997	0.939	33.214
	5	1.003	0.889	0.260	0.997	0.928	26.170
	10	0.992	0.886	0.150	0.989	0.908	13.126
	15	1.001	0.886	0.100	0.999	0.899	8.848
	20	1.000	0.885	0.089	1.010	0.888	6.558

Table 3: Ratios of MSE (“MSE ratio”) and higher-order MSE (“HMSE ratio”) based on the $\text{MSE}(\hat{\beta})$ -optimal bandwidth divided by those based on the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth. μ^2/l is the standardized concentration parameter measuring the strength of the instruments.

model	l/p	MSE ratio	HMSE ratio
AR(1)	5	0.995	0.928
	2	0.999	0.946
	1.50	0.986	0.762
	1.33	0.997	0.993
	1.25	1.002	0.984
	1.14	0.978	0.738
	1.11	1.005	0.930
MA(1)	5	0.998	0.932
	2	1.000	0.974
	1.50	0.991	0.798
	1.33	0.998	0.855
	1.25	1.016	0.860
	1.14	0.976	0.643
	1.11	0.998	0.592

Table 4: Robustness check: ratios of MSE (“MSE ratio”) and higher-order MSE (“HMSE ratio”) based on the $\text{MSE}(\hat{\beta})$ -optimal bandwidth divided by those based on the $\text{MSE}(\hat{\Omega})$ -optimal bandwidth.