

**Efficient Estimation of Nonparametric
Regression with Autocorrelated Errors**

Oliver Linton Enno Mammen
LSE Mannheim

Yale University, October, 2004

Introduction

- We discuss the efficient estimation of the model

$$B(L)Y_t = A(L)m(X_t) + \varepsilon_t,$$

where

- ε_t is an i.i.d. sequence with finite variance σ_ε^2 and independent of the regressors X_t
- the regressors X_t are assumed to follow some stationary process
- $A(L) = \sum_{j=0}^{\infty} a_j L^j$ and $B(L) = \sum_{j=0}^{\infty} b_j L^j$
- The function $m(\cdot)$ is assumed to be unknown but smooth.
- It is of interest to estimate m , and $A(L), B(L)$.

- Leading case of interest is the nonparametric regression model

$$Y_t = m(X_t) + u_t,$$

$t = 1, \dots, T$, where the residual process u_t satisfies

$$A(L)u_t = \varepsilon_t = \sum_{j=0}^{\infty} a_j u_{t-j}.$$

- In this case

$$A(L)Y_t = A(L)m(X_t) + \varepsilon_t,$$

which is in the above form with $A(L) = B(L)$.

- This is the model considered by Xiao, Linton, Carroll, and Mammen (JASA, 2003) XLCM.

Parametric Case

- Suppose that

$$m(x) = \beta^\top x$$

for some unknown parameters. This classic model has been extensively treated in the econometrics literature.

- The variance of OLS is proportional to the long run variance of the process $\{X_t u_t\}$ and least squares standard errors that ignore this fact are inconsistent and need to be modified in a non-trivial way.
- Also, one can generally improve efficiency of least squares estimators by using a GLS weighting scheme that reflects the error autocorrelation function.
- Cochrane-Orcutt, Hildreth-Lu, Prais-Winsten, and Durbin-Watson et als.
- Some cases where OLS=GLS. Specifically, when the regressors are polynomials in time, see for example Andersen

(1971, p581).

- Compare this with the case where $m(\cdot)$ is nonparametric, which has been analyzed in Robinson (1983), Masry (1996ab) for example.
 - In this case, standard kernel regression smoothers do not take account of the correlation structure in X_t or u_t and estimate the regression function in the same way as if these processes were independent.
 - Furthermore, the variance of such estimators is proportional to the short run variance of u_t , $\sigma_u^2 = \text{var}(u_t)$ and does not depend on the regressor or error covariance functions $\gamma_X(j) = \text{cov}(X_t, X_{t-j})$, $\gamma_u(j) = \text{cov}(u_t, u_{t-j})$, $j \neq 0$.
- Practitioners accustomed to correcting standard errors for dependence believe that the standard errors in nonparametric regression are therefore suspect.

- XLCM introduced an alternative non-parametric estimator of m that was more efficient than the usual estimators and took account of the autocorrelation structure.
- The estimation method of XLCM is unattractive because it requires several degrees of smoothing.
- We propose an alternative estimator of the function m based on solving a type 2 linear integral equation.
- We show that it has attractive theoretical and finite sample properties.
 - In particular, it has smaller asymptotic variance than the main method of XLCM
 - It also works in the case where u_t is a unit root process, whereas standard kernel regression and XLCM procedures do not.

Estimation Method

- First suppose that $A(L), B(L)$ are known.
- Letting $Z_t = B(L)Y_t$ we have

$$Z_t = \sum_{j=1}^{\infty} a_j m(X_{t-j}) + \varepsilon_t,$$

which is an additive autoregression with i.i.d. errors and with a restriction on the additive components.

- Define m as the minimizer of the criterion

$$E \left[\left\{ Z_t - \sum_{j=1}^{\infty} a_j m(X_{t-j}) \right\}^2 \right].$$

- It follows that m satisfies the first order condition

$$E \left[\left\{ Z_t - \sum_{j=1}^{\infty} a_j m(X_{t-j}) \right\} \sum_{k=1}^{\infty} a_k h(X_{t-k}) \right] = 0$$

for any measurable function h .

- This implies that, taking $h(\cdot)$ to be the

Dirac delta function,

$$\begin{aligned} \sum_{j=1}^{\infty} a_j E[Z_t | X_{t-j} = x] &= \sum_{j=1}^{\infty} a_j^2 m(x) \\ &+ \sum_{j \neq k} \sum a_j a_k E[m(X_{t-j}) | X_{t-k} = x]. \end{aligned}$$

- This can be re-expressed as

$$\begin{aligned} m(x) &= m^*(x) + \int \mathcal{H}(x, y) m(y) f_0(y) dy \\ m &= m^* + \mathcal{H}m, \end{aligned}$$

which is a linear type 2 integral equation in $L_2(f_0)$ with

$$m^*(x) = \sum_{j=1}^{\infty} a_j^* E[Z_t | X_{t-j} = x]$$

$$\mathcal{H}(x, y) = - \sum_{j=\pm 1}^{\pm \infty} a_j^+ \frac{f_{0,j}(y, x)}{f_0(y) f_0(x)},$$

where $a_j^* = a_j / \sum_{j=1}^{\infty} a_j^2$ and $a_j^+ = \sum_{k \neq 0} a_{j+k} a_j / \sum_{l=1}^{\infty} a_l^2$, while f_0 is the density of X_t and $f_{0,j}$ is the joint density of (X_t, X_{t-j}) .

Assumption A1. *The operator $\mathcal{H}(x, y)$ is Hilbert-Schmidt i.e.,*

$$\int \int \mathcal{H}(x, y)^2 f_0(x) f_0(y) dx dy < \infty.$$

- Under assumption A1, \mathcal{H} is a self-adjoint bounded compact linear operator on the Hilbert space of functions $L_2(p_0)$, and therefore has a countable number of eigenvalues:

$$\infty > |\lambda_1| \geq |\lambda_2| \geq \dots,$$

with $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$.

Assumption A2. *There exist no $m \in \mathcal{M}$ with $\|m\|_2 = 1$ such that $\sum_{j=1}^{\infty} a_j m(y_{t-j}) = 0$ with probability one.*

- This condition rules out a certain ‘concurvity’ in the stochastic process. That is, the data cannot be functionally related in this particular way.
- Under A1-A2 there exists a unique solution to the integral equation that satisfies

$$m = (I - \mathcal{H})^{-1} m^*.$$

Further Details for a special case

- Suppose that X_t is a scalar absolutely continuous random variable and that

$$A(L) = B(L) = 1 - \rho_0 L,$$

- i.e., this is nonparametric regression with AR(1) disturbances.

$$Y_t = m(X_t) + u_t,$$

$$u_t = \rho_0 u_{t-1} + \varepsilon_t$$

- Then we write

$$Z_t(\rho) = Y_t - \rho Y_{t-1}$$

for any ρ .

The estimation method is

1. For each ρ compute estimators of

$$\hat{m}_\rho^*, \hat{\mathcal{H}}_\rho$$

of $m_\rho^*, \mathcal{H}_\rho$

2. Solve an empirical version of the integral equation to obtain an estimator

$$\hat{m}_\rho$$

of m_ρ

3. Choose $\hat{\rho}$ to minimize the profiled negative log likelihood or least squares criterion with respect to ρ .

4. Let

$$\hat{m}(x) = \hat{m}_{\hat{\rho}}(x).$$

- For any $\{Z_t(\rho)\}$ and lag j define $\widehat{g}_j(x; \rho) = \widehat{a}_0$, where $(\widehat{a}_0, \widehat{a}_1)$ are the minimizers of

$$\sum_{t=j+1}^T \{Z_t(\rho) - a_0 - a_1(X_{t-j} - x)\}^2 K_h(X_{t-j} - x)$$
 with respect to (a_0, a_1) , where K is a symmetric probability density function, h is a positive bandwidth, and $K_h(\cdot) = K(\cdot/h)/h$.

- Then define

$$\widehat{m}_\rho^*(x) = \frac{1}{1 + \rho^2} \{\widehat{g}_0(x; \rho) - \rho \widehat{g}_1(x; \rho)\}$$

$$\widehat{\mathcal{H}}_\rho(x, y) = \frac{-\rho}{1 + \rho^2} \frac{\widehat{f}_{0,1}(x, y) + \widehat{f}_{0,1}(y, x)}{\widehat{f}_0(x) \widehat{f}_0(y)},$$

where

$$\widehat{f}_{0,1}(y, x) = \frac{1}{T-1} \sum_{t=2}^T K_h(y - X_t) K_h(x - X_{t-1}),$$

$$\widehat{f}_0(x) = \frac{1}{T} \sum_{t=1}^T K_h(x - X_t).$$

- Define \widehat{m}_ρ as any solution to the equation

$$m = \widehat{m}_\rho^* + \widehat{\mathcal{H}}_\rho m,$$

in $L_2(\widehat{f}_0)$

- Let

$$\widehat{\rho} = \arg \min_{\rho \in [-1, 1]} \widehat{S}_T(\rho),$$

where

$$\widehat{S}_T(\rho) = \sum_{t=2}^T \{Y_t - \rho Y_{t-1} - \widehat{m}_\rho(X_t) + \rho \widehat{m}_\rho(X_{t-1})\}^2.$$

- Finally, let

$$\widehat{m}(x) = \widehat{m}_{\widehat{\rho}}(x).$$

- How we solve the integral equation in practice. Note that one can rewrite as an integral equation on $[0, 1]^2$ as

$$m_\rho^\dagger(s) = m_\rho^{*\dagger}(s) + \int_0^1 \mathcal{H}_\rho^\dagger(s, t) m_\rho(t) dt,$$

where $\mathcal{H}_\rho^\dagger(s, t) = \mathcal{H}_\rho^\dagger(F_0^{-1}(s), F_0^{-1}(t))$ with $y = F_0^{-1}(s)$, $x = F_0^{-1}(t)$ and $m_\rho^\dagger(t) = m_\rho(F_0^{-1}(t))$ and $m_\rho^{*\dagger}(t) = m_\rho^*(F_0^{-1}(t))$ and F_0 is the c.d.f. of X_t . For simplicity we drop the superfluous \dagger superscript in the sequel.

- Let $\{t_{j,n}, j = 1, \dots, n\}$ be some equally spaced grid of points in $[0, 1]$, and let $q_{j,n} = \widehat{F}_0^{-1}(t_{j,n})$ be the empirical $t_{j,n}$ quantile of X_t . Now approximate the original equation by

$$\widehat{m}_\rho(q_{i,n}) = \widehat{m}_\rho^*(q_{i,n}) + \sum_{j=1}^n \widehat{\mathcal{H}}_\rho(q_{i,n}, q_{j,n}) \widehat{m}_\rho(q_{j,n}),$$

$$i = 1, \dots, n.$$

- The linear system can be written in matrix notation

$$(I_n - \widehat{\mathbf{H}}_\rho) \widehat{\mathbf{m}}_\rho = \widehat{\mathbf{m}}_\rho^*,$$

where I_n is the $n \times n$ identity,

$$\widehat{\mathbf{m}}_\rho = (\widehat{m}_\rho(q_{1,n}), \dots, \widehat{m}_\rho(q_{n,n}))^\top$$

$$\widehat{\mathbf{m}}_\rho^* = (\widehat{m}_\rho^*(q_{1,n}), \dots, \widehat{m}_\rho^*(q_{n,n}))^\top$$

$$\widehat{\mathbf{H}}_\rho = - \left[\frac{\rho}{1 + \rho^2} \frac{\widehat{f}_{0,1}(q_{i,n}, q_{j,n}) + \widehat{f}_{0,1}(q_{j,n}, q_{i,n})}{\widehat{f}_0(q_{i,n}) \widehat{f}_0(q_{j,n})} \right]_{i,j=1}^n$$

is an $n \times n$ matrix.

- We then find the solution values

$$\widehat{\mathbf{m}}_\rho = (\widehat{m}_\rho(q_{1,n}), \dots, \widehat{m}_\rho(q_{n,n}))^\top$$

to this system by direct inversion when n is less than say 2000.

Main Result

- We shall assume that the error process $\{u_t\}$ is independent of the process $\{X_t\}$ and we assume that $\{X_t\}$ is a α -mixing process.
- Suppose that Assumptions 1 to 7 hold.

Then,

$$\begin{aligned} & \sqrt{Th} (\widehat{m}(x) - m(x) - h^2 b(x)) \\ \implies & N \left(0, \frac{\sigma_\varepsilon^2}{f_0(x)(1 + \rho_0^2)} \|K\|_2^2 \right). \end{aligned}$$

- Compare this with
 - the usual kernel estimator, which has variance

$$\|K\|_2^2 \frac{\sigma_u^2}{f_0(x)} = \|K\|_2^2 \frac{\sigma_\varepsilon^2}{f_0(x)(1 - \rho_0^2)},$$

- the estimator of XLCM, which has variance

$$\|K\|_2^2 \frac{\sigma_\varepsilon^2}{f_0(x)}.$$

- Note that the standard nonparametric regression estimator is not consistent when $\rho_0 = 1$, whereas our procedure is.
- The XLCM procedure also does not work in the unit root case because it relies on the initial standard nonparametric regression estimator.

The parametric part

- When $|\rho_0| < 1$

$$\sqrt{T}(\hat{\rho} - \rho_0) \implies N(0, V),$$

- Consistency is shown by

$$\sup_{|\rho - \rho_0| \leq \epsilon} \left| T^{-1} \hat{S}_T(\rho) - S(\rho) \right| = o_p(1)$$

for some $\epsilon > 0$; and unique minimization of limit $S(\rho)$.

- Distribution theory follows from Taylor expansion.
- When $\rho_0 = 1$ we conjecture that

$$T(\hat{\rho} - \rho_0) \implies D$$

for some unit root distribution.

- Consistency is shown by

$$T^{-1} \hat{S}_T(1) \xrightarrow{P} \sigma_\epsilon^2$$

and for all $\epsilon > 0$

$$\liminf_{T \rightarrow \infty} \inf_{1 - \rho \geq \epsilon} \left| T^{-1} \left(\hat{S}_T(\rho) - \hat{S}_T(1) \right) \right| > 0.$$

Heuristic Argument

- For each ρ

$$\begin{aligned}
 \widehat{m}_\rho - m_\rho &= (I - \widehat{\mathcal{H}}_\rho)^{-1} \widehat{m}_\rho^* - (I - \mathcal{H}_\rho)^{-1} m_\rho^* \\
 &= (I - \widehat{\mathcal{H}}_\rho)^{-1} (\widehat{m}_\rho^* - m_\rho^*) \\
 &\quad + \left[(I - \widehat{\mathcal{H}}_\rho)^{-1} - (I - \mathcal{H}_\rho)^{-1} \right] m_\rho^* \\
 &= (I - \widehat{\mathcal{H}}_\rho)^{-1} \\
 &\quad \times \left[(\widehat{m}_\rho^* - m_\rho^*) + (\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho) (I - \mathcal{H}_\rho)^{-1} m_\rho^* \right] \\
 &= (I - \widehat{\mathcal{H}}_\rho)^{-1} \left[(\widehat{m}_\rho^* - m_\rho^*) + (\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho) m_\rho^* \right].
 \end{aligned}$$

- Therefore, provided

$$\widehat{m}_\rho^* - m_\rho^* \text{ and } \widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho$$

are small in an appropriate sense

$$\widehat{m}_\rho - m_\rho \simeq (I - \mathcal{H}_\rho)^{-1} \left[(\widehat{m}_\rho^* - m_\rho^*) + (\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho) m_\rho^* \right].$$

- It can be further shown that

$$\widehat{m}_\rho - m_\rho \simeq (\widehat{m}_\rho^* - m_\rho^*) + (\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho)m_\rho^*.$$

- When $\|\mathcal{H}_\rho\| < 1$ this can be seen as follows since

$$(I - \mathcal{H}_\rho)^{-1} = I + \mathcal{H}_\rho + \mathcal{H}_\rho^2 + \dots$$

and

$$\mathcal{H}_\rho(\widehat{m}_\rho^* - m_\rho^*) = o_p(\widehat{m}_\rho^* - m_\rho^*)$$

because applying the operator \mathcal{H}_ρ involves averaging.

- The full argument allowing also for the case that $\|\mathcal{H}_\rho\| \geq 1$ is given in Linton and Mammen (2003, Proposition 1).

- We proceed in the case that $\rho = \rho_0$. We have

$$\widehat{m}_{\rho_0}^*(x) - m_{\rho_0}^*(x) = \widehat{m}_{\rho_0}^{*,A}(x) + \widehat{m}_{\rho_0}^{*,B}(x) + R_T(x),$$

where

$$\sup_x |R_T(x)| = o_p(T^{-2/5}),$$

while

$$\begin{aligned} \widehat{m}_{\rho_0}^{*,A}(x) &= \frac{1}{1 + \rho_0^2} \frac{1}{Thf_0(x)} \times \\ &\sum_{t=1}^T K\left(\frac{x - X_t}{h}\right) \{Z_t(\rho_0) - E[Z_t(\rho_0)|X_t]\} \\ &- \rho_0 \sum_{t=2}^T K\left(\frac{x - X_{t-1}}{h}\right) \{Z_t(\rho_0) - E[Z_t(\rho_0)|X_{t-1}]\}, \end{aligned}$$

where

$$\begin{aligned} E[Z_t(\rho_0)|X_t] &= m(X_t) - \rho_0 E[m(X_{t-1})|X_t] \\ E[Z_t(\rho_0)|X_{t-1}] &= E[m(X_t)|X_{t-1}] - \rho_0 m(X_{t-1}). \end{aligned}$$

- Likewise

$$(\widehat{\mathcal{H}}_\rho - \mathcal{H}_\rho)m_\rho^*(x) = \widehat{m}_{\rho_0}^{*,C}(x) + \widehat{m}_{\rho_0}^{*,D}(x) + R_{HT}(x),$$

where

$$\sup_x |R_{HT}(x)| = o_p(T^{-2/5}).$$

- It follows that

$$\widehat{m}_{\rho_0}^{*,A}(x) + \widehat{m}_{\rho_0}^{*,C}(x) \simeq \frac{1}{1 + \rho_0^2 Th f_0(x)} \times \left[\sum_{t=1}^T K\left(\frac{x - X_t}{h}\right) \varepsilon_t - \rho_0 \sum_{t=2}^T K\left(\frac{x - X_{t-1}}{h}\right) \varepsilon_t \right].$$

- This term is asymptotically normal with zero mean and variance given in the theorem.

Numerical Results

- We suppose that

$$m(x) = \beta_0 x^2 / 2,$$

where

$$X_t \sim N(0, 1) \text{ and } \varepsilon_t \sim N(0, \sigma).$$

- We examine the cases $T \in \{800, 400, 200\}$ and $\rho_0 \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, and use $ns = 1000$ replications.
- We compute our estimator \hat{m} using 200 grid points and assuming in the first instance that ρ_0 is known.
- We also compute the standard local linear estimator \tilde{m} , in both cases the Gaussian kernel was used.

- We chose bandwidth to be optimal according to (asymptotic) weighted mean squared error

$$P_\infty(\hat{m}) = \text{plim}_{T \rightarrow \infty} T^{4/5} \int_{-c}^c [\hat{m}(x) - m(x)]^2 f_0(x) dx,$$

- This gives

$$h_{opt} = c_K c_M T^{-1/5},$$

where

$$c_K = (2c \|K\|_2^2 / \mu_2^2(K))^{1/5}$$

is to do with the kernel and

$$c_M = (\sigma_\varepsilon^2 / (1 + \rho_0^2) \beta_0^2 (F_0(c) - F_0(-c)))^{1/5},$$

where $F_0(x)$ is the c.d.f. of the covariate, is to do with the model.

- We have taken $c = 2$, which corresponds to an interval containing almost 95% of the covariate distribution.

- For the standard local linear estimator the optimal bandwidth is

$$c_K c_M^* T^{-1/5}$$

with

$$c_M^* = (\sigma_\varepsilon^2 / (1 - \rho_0^2)) \beta_0^2 (F_0(c) - F_0(-c))^{1/5}$$

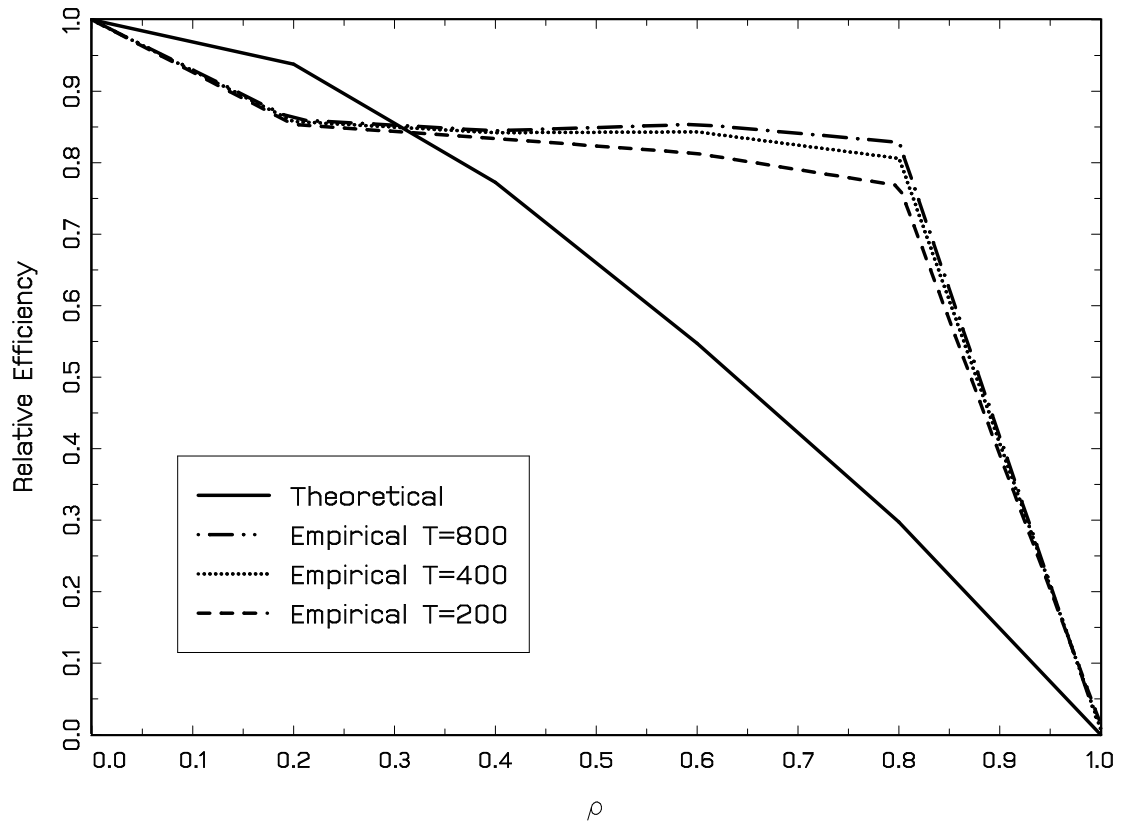
provided $\rho_0 \neq 1$ (when $\rho_0 = 1$ we set ρ_0 in the formula arbitrarily to 0.9).

- In Figure 1 below we report the relative value of the performance measure

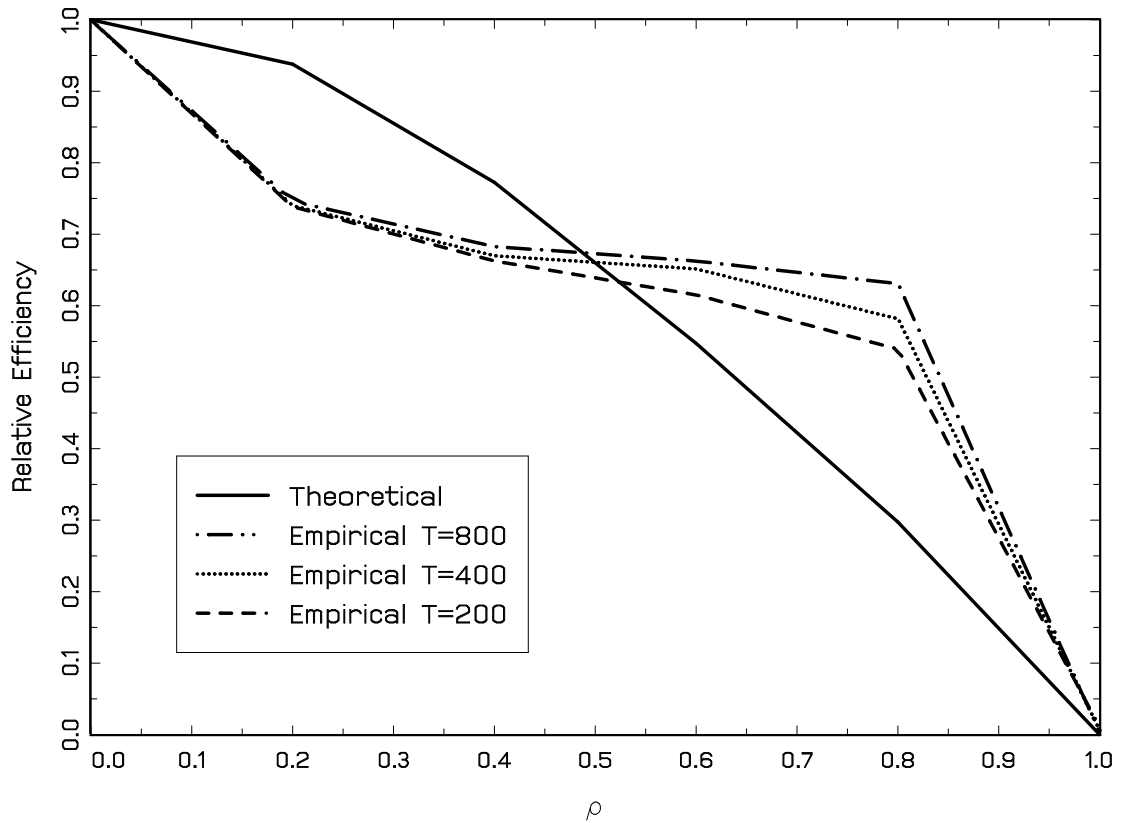
$$P_T(\hat{m}) = E \int_{-c}^c [\hat{m}(x) - m(x)]^2 f_0(x) dx$$

to $P_T(\tilde{m})$, where E is computed by the average over Monte Carlo simulations. Both estimators use their optimal bandwidths, and consequently their theoretical relative efficiency is

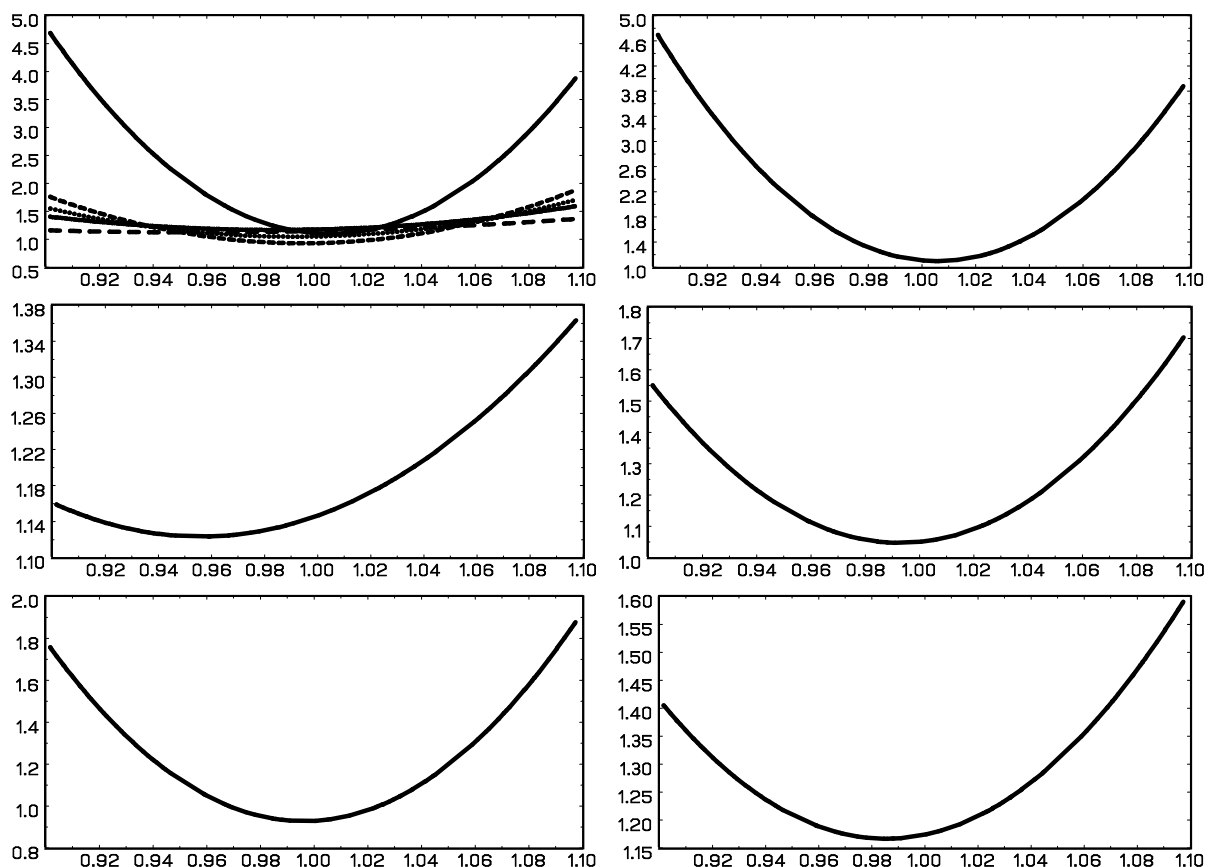
$$\left(\frac{1 - \rho_0^2}{1 + \rho_0^2} \right)^{4/5} .$$



Shows the empirical performance ratio $P_T(\hat{m})/P_T(\tilde{m})$ for different sample sizes along with the asymptotic value $P_\infty(\hat{m})/P_\infty(\tilde{m})$ predicted from the asymptotic theory. X_t iid $N(0, 1)$.



Shows the empirical performance ratio $P_T(\hat{m})/P_T(\tilde{m})$ for different sample sizes along with the asymptotic value $P_\infty(\hat{m})/P_\infty(\tilde{m})$ predicted from the asymptotic theory. X_t iid $N(0, 1)$.
 $X_t = 0.95X_{t-1} + u_t$ with $X_t \sim N(0, 1)$.



Shows Average Sum of Square Residuals
 against parameter value in the unit root case
 from five simulations

Concluding Remarks

- The more general model with $A(L) \neq B(L)$ can be treated similarly.
- Estimation works fine when A, B are low order. Otherwise need to parameterize in some way, so that

$$A(L) = \sum_{j=0}^{\infty} a_j(\theta) L^j$$

and $\theta \in \mathbb{R}^p$

- Can allow for heteroskedasticity with not much change in methods and results.
- For example suppose

$$Y_t = m(X_t) + \varepsilon_t \sigma_t$$

with σ_t^2 some volatility process where ε_t is independent of X_t . We could have

$$\sigma_t^2 = f(X_t, X_{t-1}, \dots, Y_{t-1}, \dots).$$

- Suppose that σ_t^2 is known. Then find m to minimize

$$E \left[\left(\frac{Y_t - m(X_t)}{\sigma_t} \right)^2 + \ln \sigma_t^2 \right].$$

The solution is

$$m(x) = \frac{E \left[\frac{Y_t}{\sigma_t^2} | X_t \right]}{E \left[\frac{1}{\sigma_t^2} | X_t \right]}.$$

In practice σ_t^2 has to be estimated from residuals.

- If for some smooth unknown function f

$$\sigma_t^2 = f(X_t)$$

then there is no gain, but otherwise there is.