Adaptive estimation in multiple time series with independent component errors

Adaptive Estimation in Multiple Time Series with Independent Component Errors

P. M. Robinson* and L. Taylor

Department of Economics, London School of Economics, London WC2A 2AE, UK

June 20, 2016

Abstract

This paper develops statistical methodology for semiparametric models for multiple time series of possibly high dimension N. The objective is to obtain precise estimates of unknown parameters (which characterize autocorrelations and cross-autocorrelations) without fully parameterizing other distributional features, while imposing a degree of parsimony to mitigate a curse of dimensionality. The innovations vector is modelled as a linear transformation of independent but possibly non-identically distributed random variables, whose distributions are nonparametric. In such circumstances, Gaussian pseudo-maximum likelihood estimates of the parameters are typically $\sqrt{n}$-consistent, where n denotes series length, but asymptotically inefficient unless the innovations are in fact Gaussian. Our parameter estimates, which we call "adaptive", are asymptotically as first-order efficient as maximum likelihood estimates based on correctly-specified parametric innovations distributions. The adaptive estimates use nonparametric estimates of score functions (of the elements of the underlying vector of independent random varables) which involve truncated expansions in terms of basis functions; these have advantages over the kernel-based score function estimates used in most of the adaptive estimation literature. Our parameter estimates are also $\sqrt{n}$-consistent and asymptotically normal. A Monte Carlo study of finite sample performance of the adaptive estimates, employing a variety of parameterizations, distributions and choices of N, is reported.

Keywords and phrases. Multiple time series, independent component analysis, efficient semiparametric estimation, adaptive estimation, stationary processes, forecast error.

---

*Corresponding author. *E-mail address:* p.m.robinson@lse.ac.uk

1. INTRODUCTION

In many substantive fields, such as in the natural, engineering and social sciences, regularly-spaced time series observations are recorded on several related variables. For example macroeconomic data may consist of quarterly observations on GDP, unemployment and interest rates, though in many studies the number of variables of interest can be far reater than 3. It is generally desirable to treat such observations as a single, multiple time series, rather than as individual series, because one may expect there to be causal relations across the series or common effects, and forecasting of a given series to be improved by using others. Whereas cross-sectional observations are often assumed to be statistically independent, the likely temporal dependence in time series data raises the possibility of non-instantaneous correlations, along with the instantaneous correlations possible with multivariate cross-sectional data. The modelling of multiple time series typically entails features common across two or more of the individual series, including common parameters, which can be more precisely estimated if information from all the time series is combined. The estimated multivariate model can then be used in forecasting.

The modelling and statistical analysis of multiple time series faces difficulties that are significantly greater than ones encountered in a univariate setting. Denote a multiple time series by x_t, $t = 1$, $2, ..., n$, where x_t is an $N \times 1$ vector. Though we shall proceed as if the x_t are observable our methods can be readily extended to situations in which the basic time series modelling problem concerns unobservable errors in a location or more general linear or nonlinear regression model, whence observable proxiues for these errors would be inserted in place of x_t in our computing formulae. It will be supposed that n is large relative to N, which is treated as fixed, but as suggested above N can itself be large, and the larger it is the greater the impact on modelling and subsequent statistical inference. For stationary series, an important class of dynamic models is

$$A(B;\theta_0)\, x_t = e_t, \quad t = 0, \pm 1, ..., \tag{1}$$

where B is the backshift operator, $A(B;\theta)$ is a known $N \times N$ matrix function of B and the $K \times 1$ vector θ, θ_0 is an unknown $K \times 1$ parameter vector while θ denotes any admissible value, and e_t is a sequence of unobservable $N \times 1$ vector random variables, independent across t, such that

$$\begin{aligned} E(e_t) &= 0, \\ E\left(e_t e_t^T\right) &= \Omega_0, \end{aligned}$$

where Ω_0 is an unknown $N \times N$ positive definite matrix and T denotes transposition. In particular,

we suppose the existence of a possibly infinite autoregressive representation,

$$A(B;\theta) = I_N - \sum_{j=1}^{\infty} A_j(\theta) B^j, \tag{2}$$

where I_N is the $N \times N$ identity matrix and the $A_j(\theta)$ are given $N \times N$ matrix functions of θ. In the finite vector autoregression of order p, VAR(p), model we have $A_j(\theta) = 0$, $j > p$, so

$$A(B;\theta) = I_N - \sum_{j=1}^{p} A_j(\theta) B^j. \tag{3}$$

However, (2) covers also stationary and invertible vector moving averages and autoregressive moving averages, and indeed along with these short memory models it also covers ones with long memory and negative dependence, such as fractional models. The need for a finite parameterization explains the notational dependence of the $A_j(\theta)$ on θ in these latter models, where though the $A_j(\theta)$ decay as j diverges, they never actually vanish.

The modelling of all elements of the $A_j(\theta)$ in terms of θ is important even in the VAR(p) (3). Here, whereas in the univariate time series case $N = 1$, where unrestricted $A_j(\theta)$ (so there is identification of each $A_j(\theta)$ with an element of θ) entails only $K = p$ parameters describing temporal dependence, when on the other hand $N > 1$, unrestricted matrices $A_j(\theta)$ give rise to $K = N^2 p$ parameters. The parameter dimension thus increases rapidly with N, presenting a 'curse of dimensionality'. For multiple time series it is thus often important to consider parsimonious modelling of the $A_j(\theta)$, a possibility formally permitted by the notational dependence of the $A_j(\theta)$ on θ. For example, the $A_j(\theta)$ can be chosen to be relatively sparse, with many *a priori* zero elements, even diagonal, for example.

If e_t (and thus x_t) is Gaussian, the distribution of e_t is entirely characterized by Ω_0, and likewise the joint distribution of x_t, $t = 1, 2, ..., n$, is entirely characterized by θ_0 and Ω_0. Gaussian maximum likelihood estimates of the latter parameters have been studied, being asymptotically efficient under additional regularity conditions. Such estimates are also of interest when Gaussianity is relaxed to milder assumptions, on moments, when they are termed pseudo-maximum likelihood estimates. In standard parameterizations, where θ_0 does not overlap with Ω_0, the (multivariate normal) limit distribution of the estimate of θ_0 is desirably the same irrespective of whether or not e_t is Gaussian. However, asymptotic efficiency is lost in the absence of Gaussianity.

There is thus interest in developing estimates of θ_0 that are asymptotically efficient in the presence of vector e_t with possibly non-Gaussian distribution. Gaussian maximum likelihood asymptotic theory extends relatively straightforwardly to non-Gaussian parametric distributions, but though

obvious candidates for the latter, such as multivariate-t, present themselves, there is immense variety in the possible choices, even relative to the univariate case $N = 1$, and often little basis for singling out one, and moreover the consistency-robustness of Gaussian-based estimates to departures from Gaussianity generally does not extend to non-Gaussian-based estimates. We can achieve the same asymptotic efficiency by what we call "adaptive" estimates, which do not require full parametric distributional assumptions on e_t. Such a goal was achieved by Stone (1975) in the context of location estimation in the setting of independent scalar observations, which was then extended by Bickel (1982) to linear regression. Time series extensions were developed by Kreiss (1987), Drost *et al.* (1997), Koul and Schick (1997), Robinson (2005), for example, again for $N = 1$. A principal theme underlying these works is estimation of the score function of the independent innovations e_t, that is, the negative of the ratio of the derivative of the probability density function of e_t to the density itself. As is well known, estimation of such functions becomes problematic for vector random variables, and to a rapidly increasing extent with increasing dimension, with decreasing precision in the score function estimates, infecting the properties of the adaptive parameter estimates, even in large samples.

This issue has long been recognised by the literature on independent component analysis (ICA), see e.g. Hyvarinen, Karhunen and Oja (2001), Vlassis (2001), Bach and Jordan (2002), Hastie and Tibshirani (2003), Samarov and Tsybakov (2004), Nascimento and Dias (2005), Chen and Bickel (2005, 2006), Samworth and Yuan (2012). With respect to the independent vectors e_t, this assumes the structure

$$e_t = M_0 \varepsilon_t, \tag{4}$$

where M_0 is an $N \times N$ nonsingular mixing matrix and the elements of ε_t are mutually independent zero-mean random variables. Some, but by no means all, of the literature, focusses on parametric distributions for ε_t. Various estimation methods, algorithms, and theoretical results, appear in the literature.

There is also a time series ICA literature, see eg Aires and Chedin (2000), Cheung and Xu (2001), Lin et al (2007), Lu et al (2009), Chen et al (2011), Garcia-Ferrer et al (2011) . This focusses not on (4) with (1) but on the structure

$$x_t = M_0 u_t, \tag{5}$$

where the elements of the $N \times 1$ unobservable vector u_t are mutually independent autocorrelated time series. Here, the fundamental dynamics are modelled in a univariate way, and then instantaneously mixed by the matrix M_0.

The modelling and motivation with respect to (5) differ from those with respect to combining (4) with (1). It is the latter setup which suits our goal of obtaining efficient estimates which avoid a curse of dimensionality. Note that without further restrictions M_0 is not identified, in particular unless at most one element of ε_t is Gaussian, M_0 is not identified even up to order and scaling. To avoid identifiability problems we explicitly fix M_0 to be the unique positive definite square root of Ω_0, so $\Omega_0 = M_0^2$, entailing $E\varepsilon_t\varepsilon_t^T = I_N$, though there is no loss of generality in the Gaussian case. Note too that a curse of dimensionality resides also in the fact that Ω_0, and thus M_0, have potentially $N(N+1)/2$ distinct unknown elements, which quantity again increases rapidly with N. Thus, some *a priori* restrictions on Ω_0 might be imposed, either directly or indirectly via M_0.

Our adaptive parameter estimates, which employ nonparametric score function estimates using truncated expansions in terms of specified basis functions, are described in the following section. Section 3 imposes regularity conditions and describes the consequent asymptotic statistical properties of the parameter estimates, in particular asymptotic normality with $\sqrt{n}$ rate and asymptotic efficiency. Section 4 reports a Monte Carlo study of finite sample behaviour, using a variety of parameterizations, distributions and choices of N, and examining relative mean squared error, relative mean squared forecast error and interval estimation bases on our central limit theorem. Section 5 contains some final comments.

2. ADAPTIVE ESTIMATES

Our adaptive estimate of θ_0, with its asymptotic efficiemcy in the presence of unknown error distributional form, is an approximate Gauss-Newton step from an initial $\sqrt{n}-$consistent estimate of θ_0. It is typical in the adaptive estimation literature to develop theory for such an estimate, rather than for an implicitly-defined semiparametric maximum likelihood estimate, because, it has the advantage of avoiding the initial consistency proof required in estabkishing a central limit theorm for the latter. Consequently, the basic nonparametric building blocks are not densities but score functions. Thus we require the elements of ε_t introduced in (4) to have differentiable probability density functions. In general the score function of a vector random variable with differentiable probability density function $f(z)$ has column vector score function $\psi(z) = -f'(z)/f(z)$, where $f'(z) = (\partial/\partial z) f(z)$. However, since the elements of ε_t are independent its vector score function can be expressed in terms of scalar score functions. In particular, denoting by ε_{it} the ith element of ε_t and by f_i, f_i' respectively the probability density function of ε_{it} and its derivative for $i = 1, ..., N$,

the score function of ε_{it} is

$$\psi_i(s) = -f_i'(s)/f_i(s), \tag{6}$$

for $i = 1, ..., N$, and the vector score function of ε_t is

$$\psi(z) = -(\psi_1(z_1), ..., \psi_N(z_N))^T, \tag{7}$$

where z_i denotes the ith element of z. Thus, (4) enables us to deal with only univariate score functions. It may be helpful to describe our construction of an adaptive esttimate and efficient inference in a step-by-step fashion, with discussion,

Step 1: innovation proxies. Estimation of the $\psi_i(s)$ in (6) requires observable proxies for the ε_t. For any admissible θ and any positive definite $N \times N$ matrix M, define the $N \times 1$ vectors

$$\begin{aligned} \varepsilon_1(\theta, M) &= M^{-1}x_1, \\ \varepsilon_t(\theta, M) &= M^{-1}\left(x_t - \sum_{j=1}^{t-1} A_j(\theta)\, x_{t-j}\right),\ t = 2, ..., n. \end{aligned} \tag{8}$$

In general $\varepsilon_t(\theta_0, M_0)$ only approximates ε_t, due to the truncation of the infinite series in (2). In the $VAR(p)$ case (3) we have, however, $\varepsilon_t(\theta_0, M_0) = \varepsilon_t$, $t \geq p+1$, and here the practitioner might prefer to take $\varepsilon_t(\theta, M) = 0$, $t \leq p$. Though ε_t has zero mean, demeaning the $\varepsilon_t(\theta, M)$ has been found to improve finite sample properties, so we introduce

$$\digamma_t(\theta, M) = \varepsilon_t(\theta, M) - n^{-1}\sum_{t=1}^{n}\varepsilon_t(\theta, M),\ t = 1, ..., n. \tag{9}$$

Now denoting the ith element of $\digamma_t(\theta, M)$ by $\digamma_{it}(\theta, M)$, we define the $n \times 1$ vectors

$$\Gamma_i(\theta, M) = (\digamma_{i1}(\theta, M), ..., \digamma_{in}(\theta, M))^T,\ i = 1, ..., N.$$

Step 2: score function estimation. Most of the adaptive estimation literature has employed kernel estimation of the score function using the ratio of a derivative-of-density estimate to a density estimate. The consequent stochastic denominator causes technical difficulties, and typically entails one or more forms of trimming, sometimes sample-splitting and discretization of the initial estimate, and requires strong conditions on some aspects. In a scalar observation setting, these problems were avoided by Beran (1976), who proposed directly estimating the score function, after modelling it as a linear combination, with unknown coefficients, of finitely many given basis functions. This works due to an integration-by-parts argument, heuristically described below. Since the number of basis functions is finite, Beran's (1976) score function estimate was parametric. For a nonparametric

score function we need to assume an infinite expansion in terms of basis functions, approximating this by a truncated expansion contaiining L terms, and then in asymptotic theory allowing L to increase, at a suitably slow rate, with n. This was achieved by Newey (1988) in a cross-sectional regression model for scalar observables, and by Robinson (2005) in a scalar time series model with parametric trend and errors that can be fractionally integrated, and stationary or nonstationary. Though a number of modifications are necessary, we follow the latter's notation as much as possible. Our basis functions are denoted $\phi_\ell(s)$, $\ell = 1, 2, ...$, and chosen to be at least continuously differentiable, having derivatives $\phi'_\ell(s)$, $\ell = 1, 2, ...$ For $L \geq 1$, scalar h_t, $t = 1, ..., n$, and $h = (h_1, , , h_n)^T$, define

$$\begin{aligned} \phi^{(L)}(h_t) &= (\phi_1(h_t), ..., \phi_L(h_t))^T, \Phi^{(L)}(h_t) = \phi^{(L)}(h_t) - n^{-1} \sum_{s=1}^{n} \phi^{(L)}(h_s), \\ \phi^{'(L)}(h_t) &= (\phi'_1(h_t), ..., \phi'_L(h_t))^T, \end{aligned}$$

and

$$W^{(L)}(h) = n^{-1} \sum_{t=1}^{n} \Phi^{(L)}(h_t)\Phi^{(L)}(h_t)^T, w^{(L)}(h) = n^{-1} \sum_{t=1}^{n} \phi^{'(L)}(h_t),$$
$$\widehat{a}^{(L)}(h) = W^{(L)}(h)^{-1} w^{(L)}(h), \quad \psi^{(L)}(h_t; \widehat{a}^{(L)}(h)) = \widehat{a}^{(L)}(h)^T \Phi^{(L)}(h_t).$$

The quantity $w^{(L)}$ arises because $\widehat{a}^{(L)}\left(\Gamma_i\left(\widetilde{\theta}, \widetilde{M}\right)\right)$ employed below is essentially a least squares estimate of the unknown coefficients of the basis functions of the ith score function approximation, after using the integration-by-parts property $E\left(\phi_\ell(\varepsilon_{it})\psi_i\left(\varepsilon_{it}\right)\right) = -E\left(\phi'_\ell(\varepsilon_{it})\right)$ to justify replacing $n^{-1}\sum_{t=1}^{n} \phi^{(L)}(h_t)\psi_i\left(h_t\right)$, which involves the unknown function ψ_i, by $w^{(L)}(h)$, which involves the given ϕ'_ℓ functions. The same basis functions $\phi_\ell(s)$ and L are used across i, but with variation across i and t introduced by the score estimates

$$\widetilde{\psi}_{it}^{(L)}(\theta, M) = \psi^{(L)}\left(\boldsymbol{F}_{it}\left(\theta, M\right); \widehat{a}^{(L)}\left(\Gamma_i\left(\theta, M\right)\right)\right), i = 1, ..., N, t = 1, ..., n.$$

Step 3: adaptive point estimation. Now assuming the $A_j(\theta)$ are differentiable introduce the $K \times 1$ vectors

$$\boldsymbol{F}'_{it}(\theta, M) = \frac{\partial}{\partial\theta}\boldsymbol{F}_{it}(\theta, M), \ i = 1, ..., N, \ t = 1, ..., n,$$

which from (9) are given linear functions of derivatives of the elements of the $A_j(\theta)$. Then define

$$\begin{aligned} r_L(\theta, M) &= \sum_{i=1}^{N}\sum_{t=1}^{n} \widetilde{\psi}_{it}^{(L)}(\theta, M)\boldsymbol{F}'_{it}(\theta, M), \\ J_{iL}(\theta, M) &= n^{-1}\sum_{t=1}^{n} \widetilde{\psi}_{it}^{(L)}(\theta, M)^2, \ i = 1, ..., N, \\ S_L(\theta, M) &= \sum_{i=1}^{N} J_{iL}(\theta, M) \sum_{t=1}^{n} \boldsymbol{F}'_{it}(\theta, M)\boldsymbol{F}'_{it}(\theta, M)^T. \end{aligned}$$

Essentially, $r_L(\theta, M)$ and $S_L(\theta, M)$ are used to estimate respectively the first and second derivatives of the semiparametric log likelihood, with $J_{iL}(\theta, M)$ being used to estimate the information for ε_{it}. Now for given initial, $\sqrt{n}$ -consistent estimates $\widetilde{\theta}$, $\widetilde{M}$ define the adaptive estimate (essentially a Gauss-Newton iterative step from $\widetilde{\theta}$, $\widetilde{M}$)

$$\widehat{\theta} = \widetilde{\theta} - S_L\left(\widetilde{\theta}, \widetilde{M}\right)^{-1} r_L\left(\widetilde{\theta}, \widetilde{M}\right). \tag{10}$$

Step 4: efficient inference. In the following section we establish the useful large sample approximation

$$\widehat{\theta} \underset{\cdot\ d}{\sim} \mathcal{N}\left(\theta_0, S_L\left(\widetilde{\theta}, \widetilde{M}\right)^{-1}\right), \tag{11}$$

implying that $\widehat{\theta}$ is asymptotically efficient. We might thence expect forecasts on the basis of (1) that employ $\widehat{\theta}$ to be generally more accurate than ones using $\widetilde{\theta}$, say. If desired we can iterate, applying (10) with $\widetilde{\theta}$ replaced on the right hand side by $\widehat{\theta}$, and so on, or to improve convergence (to an approximate nonparametric maximum likelihood estimate) by shrinking the steps, multiplying the correction term in (10) by a positive scalar less than 1.

Notice that the structure (4) and the consequent simple score vector (7) has led to the simple summations across i in the formulae for $r_L(\theta, M)$ and $S_L(\theta, M)$. A general strategy for choosing the initial estimates $\widetilde{\theta}$, $\widetilde{M}$ is exact or approximate Gaussian pseudo-maximum likelihood estimation, possibly the conditional-sum-of squares estimate (as in Box and Jenkins (1971)), which also uses directly the residual functions (8), see also Robinson (2005) in a scalar time series setting.

Given its popularity and computational convenience, especially in forecasting, the implications for the VAR(p) process (3) are worth describing. As discussed in Section 1, we may wish to impose a parsimonious parameterization on $A_1(\theta), ..., A_p(\theta)$, especially when N is large. Many of these are covered by the linear restrictions $v(\theta) = vec(A_1(\theta), ..., A_p(\theta)) = Q\theta + q$ for given $pN^2 \times K$ rank K matrix Q and $pN^2 \times 1$ vector q (often $q = 0$). Thus $\left(\partial/\partial\theta^T\right) v(\theta) = Q$. As mentioned above, in the VAR(p) case we might modify (8) by taking $\varepsilon_t(\theta, M) = 0$, $t \leq p$, and correspondingly dropping summands for $t = 1, ..., p$ from calculations. Thus write for $t > p$, $X_t = \left(x_{t-1}^T, ..., x_{t-p}^T\right)^T$ and $x_t - \sum_{j=1}^{p} A_j(\theta) x_{t-j} = x_t - \left(X_t^T \otimes I_N\right)(Q\theta + q)$. We can take $\widetilde{\theta}$ to be the least squares estimate

$$\widetilde{\theta} = \left(\sum_{t=p+1}^{n} Q^T\left(X_t X_t^T \otimes I_N\right) Q\right)^{-1} \sum_{t=p+1}^{n} \left(Q^T\left(X_t \otimes I_N\right) x_t - Q^T\left(X_t X_t^T \otimes I_N\right) q\right) \tag{12}$$

and likewise

$$\widetilde{\Omega}(\theta) = (n-p)^{-1} \sum_{t=p+1}^{n} \left(x_t - \sum_{j=1}^{p} A_j(\theta) x_{t-j}\right)\left(x_t - \sum_{j=1}^{p} A_j(\theta) x_{t-j}\right)^T, \tag{13}$$

$$\widetilde{M} \text{ positive definite, } \widetilde{\Omega}\left(\widetilde{\theta}\right) = \widetilde{M}^2.$$

In connection with calculating the $F'_{it}(\theta, M)$ note that for $t > p$,

$$\frac{\partial}{\partial \theta^T}\varepsilon_t(\theta, M) = -\frac{\partial}{\partial \theta^T}M^{-1}\left(A_1(\theta), ..., A_p(\theta)\right)X_t = -\left(X'_t \otimes M^{-1}\right)Q,$$

whence the $F'_{it}(\theta, M)$ are constant across θ. The above formulae were employed in the computations in the Monte Carlo study of Section 4, below which focusses on various VAR(1) settings.

3. ASYMPTOTIC NORMALITY

This section presents regularity conditions for asymptotic properties of the adpative estimate $\widetilde{\theta}$.

Assumption 1 *The multiple time series x_t is generated by (1), (2) and (4), where the ε_t, $t = 0, \pm 1, ...,$ are independent and identically distributed with elements that are independent and have zero means and unit variances, and M_0 is the unique positive definite square root of the finite, positive definite matrix Ω_0.*

Assumption 2 *The elements ε_{i0} of ε_0 satisfy $E\varepsilon_{i0}^4 < \infty$, $i = 1, ..., N$.*

Assumption 3 *For $i = 1, ..., N$, ε_{i0} has probability density function, $f_i(s)$, that is absolutely continuous, and*

$$0 < \mathcal{J}_i < \infty,$$

where $\mathcal{J}_i = \int \psi_i(s)^2 f_i(s)\, ds$ is the information of ε_{i0}.

Assumption 4 *On a sufficiently small neighbourhood $\mathcal{N}$ of θ_0, $A(s;\theta)$ is thrice continuously differentiable in θ for $|s| = 1$, $B(s;\theta) = A(s;\theta)^{-1} = I_N + \sum_{j=1}^{\infty} B_j(\theta)s^j$ exists for $|s| = 1$, and denoting by γ_j the modulus of any element of $B_j(\theta)$ or the supremum over $\mathcal{N}$ of the modulus of any element of $A_j(\theta)$ or of its first, second or third derivatives, with respect to any element of θ, we have $\sum_{j=1}^{\infty} j^3\gamma_j < \infty$.*

Assumption 5 *Denoting by $\overline{\theta}$ the Gaussian pseudo likelihood estimate of θ_0, the limiting covariance matrix of $n^{1/2}\left(\overline{\theta} - \theta_0\right)$ is finite and positive definite.*

Assumption 6 *As $n \to \infty$,*

$$n^{\frac{1}{2}}\left(\tilde{\theta} - \theta_0\right) = O_p(1), \quad n^{\frac{1}{2}}(\widetilde{M} - M_0) = O_p(1).$$

Assumption 7 *For $\ell = 1, , 2, ...,$ $\phi_\ell(s)$ satisfies*

$$\phi_\ell(s) = \phi(s)^\ell, \tag{14}$$

where $\phi(s)$ is strictly increasing and thrice continuously differentiable and is such that, for some $\kappa \geq 0$, $C < \infty$,

$$|\phi(s)| \leq 1(|s| \leq 1) + |s|^{\kappa} 1(|s| > 1), \ \left|\phi'(s)\right| + \left|\phi''(s)\right| + \left|\phi'''(s)\right| \leq C(1 + |\phi(s)|^{C}),$$

with ϕ', ϕ'' and ϕ''' denoting the first, second and third derivatives of ϕ.

Assumption 8 *As $n \to \infty$, $L \to \infty$ such that*

$$\underline{\lim}_{n\to\infty} \left(\frac{\log n}{L}\right) > 8\left\{\log\left(1 + 2^{\frac{1}{2}}\right) + \max(\log\varphi, 0)\right\} \simeq 7.05 + 8\max(\log\varphi, 0);$$

where $\varphi = (1 + |\phi(s_1)|) / (\phi(s_2) - \phi(s_1))$, $[s_1, s_2]$ being an interval on which the $f_i(s)$ are bounded away from zero.

The details of Assumption 1 have already been introduced, Assumptions 3 and 6 are standard, and Assumption 4 includes mild smoothess conditions on $A(s;\theta)$ and weak dependence conditions on x_t. Trade-offs are possible between Assumptions 2 and 8, with the possibility of stronger moment conditions permitting milder restrictions on the rate of growth of L with n, as described in a scalar time series setting by Robinson (2005). Assumption 5 appears unprimitive but is designed to minimise introduction of additional notation, employing the fact that the limiting covariance matrix of an asymptotically efficient estimate is a scalar multiple of that of the Gaussian pseudo likelihood estimate, where primitive conditions for the finiteness and non-singularity of the latter are available. The polynomial structure of Assumption 7, in terms of a single basic function $\phi(s)$, could be relaxed but seems sufficiently flexible for practical purposes.

Theorem *Let Assumptions 1-8 hold. Then as $n \to \infty$, $S_L\left(\widetilde{\theta}, \widetilde{M}\right)^{1/2}(\hat{\theta} - \theta_0) \ \to_d \mathcal{N}(0, I_K)$.*

The lengthy proof of this theorem is omitted, because it relatively straightforwardly extends that of Robinson (2005) for the case of scalar series, $N = 1$, the structure (4) having led to summations across $i = 1, ...N$ of similar formulae to those in that reference. We present the Theorem for a studentized statistic for ease of application, but the matrix $S_L\left(\widetilde{\theta}, \widetilde{M}\right)^{-1}$ converges in probability to the limiting covariance matrix of the maximum likelihood estimate that would be obtained from correctly specified parametric distributions for the elements of ε_t in (4), and $\hat{\theta}$ is thus asymptotically efficient.

4. FINITE SAMPLE PERFORMANCE

It is desirable to investigate the finite-sample properties of our asymptotically-justified adaptive estimates by Monte Carlo simulations. The main features of interest in designing these are perhaps the impact of various choices of dimension N, the degree of mixing afforded by the matrix M_0, and heterogeneity in the elements of ε_t. We used M_0 of form

$$M_0 = (1-c)\, I_N + c 1_N 1_N^T,$$

where 1_N is the $N \times 1$ vector of 1's. Then M_0 is positive definite for $c < 1$, and Ω_0 has similar structure, $\Omega_0 = (1-c)^2 I_N + (Nc + 2c(1-c)) 1_N 1_N^T$. We took $c = 0.5$ and 0.9. We focussed on the VAR(1) case of (3), subjecting $A_1(\theta)$ to linear restrictions as discussed in Section 2, in particular

$$A_1(\theta) = diag(\theta_1, ..., \theta_N), \text{ so } K = N, \tag{15}$$

denoting by θ_i the ith element of θ, and

$$A_1(\theta) = \theta I_N, \quad \text{so } K = 1. \tag{16}$$

In (15) we took elements of θ_0 within the interval $[0.5, 0.9]$, for example $\theta_0 = (0.50, 0.57, 0.63, 0.7, 0.77, 0.83, 0.90)^T$ when $N = 7$, while in (16) we took θ_0 as 0.5 and 0.9. We chose $N = 2$, and 7, along with $n = 50$ and 100, and also a high-dimensional case $N = 56$, with $n = 560$. The candidate distributions for ε_t are listed in Table 1.

Table 1: Source distributions.

0	$\mathcal{N}(0,1)$
1	$0.5\mathcal{N}(-3,1) + 0.5\mathcal{N}(3,1)$
2	$0.05\mathcal{N}(0,1) + 0.95\mathcal{N}(0,1)$
3	Laplace
4	$t(5)$
5	Laplace+$\mathcal{N}(0,1)$
6	$t(5) + U[0,1]$

The methods were implemented using (12) and (13) for $\widetilde{\theta}$ and $\widetilde{M}$, and with either $\phi(s) = s$ or $\phi(s) = s(1+s^2)^{-\frac{1}{2}}$ (which is bounded) in (14), with $L = 1, 2, 3$ and 4. As well as computing the

one-step estimate (10), we went on to compute an iterative sequence of estimates, defined as

$$\widehat{\theta}_{j+1} = \widehat{\theta}_j - 0.2 S_L \left(\widehat{\theta}_j, \widetilde{M}\right)^{-1} r_L \left(\widehat{\theta}_j, \widetilde{M}\right), \quad j = 1, 2, ..., \tag{17}$$

where $\widehat{\theta}_1 = \widehat{\theta}$, stopping when $\left|\hat{\theta}_{j+1} - \hat{\theta}_j\right| < 0.001$.

The results are based on $R = 1000$ replications, except for $N = 56$, where $R = 100$. For the purpose of the immediately following definitions only, for convenience we take $\hat{\theta}$ either to denote (10) or the final iterative estimate obtained from (17). We report relative mean squared error, RMSE =MSE($\hat{\theta}$)/MSE($\tilde{\theta}$), where MSE(θ) $= R^{-1} \sum_{i=1}^{R} \left(\theta^{(i)} - \theta_0\right)^2$, $\theta^{(i)}$ referring in each case to the ith replicate. We also report the relative out-of-sample 5 steps ahead forecast MSE, RFMSE=FMSE($\hat{\theta}$)/FMSE($\tilde{\theta}$), where FMSE(θ) $= R^{-1} \sum_{i=1}^{R} (\hat{x}_{n+5}^{(i)}(\theta) - x_{n+5}^{(i)})^2$ with $\hat{x}_{n+5}(\theta) = \theta^5 x_n$, only when $\theta_0 = 0.5$ (results for other θ_0 were similar).

Finally, for the case (16) we computed coverage of nominal 95% and 99% confidence intervals based on (11), reporting only results for $\theta_0 = 0.5$. In cases where there was a substantial difference between the one-step estimate (10) and the final iterative one obtained from we report results for both. The first column in each of the following tables corresponds to the value of the mixing parameter c, the second indicates the value of n, and the third the value of θ_0.

[Tables 2-6 about here]

The models relating to Tables 2 and 5, $N = 2$, are regarded as the baseline cases, with $K = 2$ and $K = 1$ respectively, where we take ε_t to be Gaussian. For $A_1(\theta) = I_N \theta$, (16), there is little difference between the two estimates for $c = 0.5$, as is to be expected since there is relatively little mixing and least squares (12) is efficient under Gaussianity. For $c = 0.9$ we see a slight improvement of the adaptive estimates' relative performance, again as expected since we now have a more even mixture of Gaussian innovations.

Table 2											
$A_1(\theta) = diag(\theta_1, ..., \theta_N)$, $K = N = 2$ elements of ε_t each distributed according to 0 in Table 1.											
				$\phi(s) = s$				$\phi(s) = s(1+s^2)^{-\frac{1}{2}}$			
c	n	θ_0	L	1	2	3	4	1	2	3	4
One-step											
0.5	50	0.5	RMSE	0.96	0.92	0.99	1.05	1.05	1.00	1.01	1.05
		0.9	RMSE	1.75	1.69	1.56	1.40	1.71	1.67	1.80	1.58
		0.5	RFMSE	0.52	0.54	0.66	0.66	0.58	0.64	0.60	0.71
	100	0.5	RMSE	0.70	0.74	0.81	0.84	0.78	0.79	0.74	0.84
		0.9	RMSE	1.05	1.12	1.23	1.33	1.27	1.36	1.24	1.14
		0.5	RFMSE	0.75	0.76	0.78	0.78	0.73	0.75	0.79	0.78
0.9	50	0.5	RMSE	0.51	0.48	0.63	0.62	0.65	0.63	0.60	0.69
		0.9	RMSE	1.41	1.48	1.73	1.51	1.92	1.45	1.56	1.69
		0.5	RFMSE	0.62	0.65	0.73	0.75	0.81	0.67	0.72	0.85
	100	0.5	RMSE	0.66	0.73	0.57	0.57	0.81	0.78	0.58	0.67
		0.9	RMSE	3.30	3.37	2.66	2.53	3.71	3.39	2.87	2.96
		0.5	RFMSE	0.93	1.01	0.96	0.99	1.00	1.00	0.93	0.97
Iterative											
0.9	50	0.5	RMSE	0.23	0.23	0.32	0.33	0.26	0.25	0.31	0.33
		0.9	RMSE	0.34	0.34	0.53	0.54	0.38	0.36	0.43	0.45
		0.5	RFMSE	0.68	0.67	0.66	0.58	0.66	0.65	0.64	0.65
	100	0.5	RMSE	0.12	0.13	0.13	0.15	0.14	0.15	0.15	0.17
		0.9	RMSE	0.16	0.17	0.17	0.21	0.19	0.21	0.21	0.31
		0.5	RFMSE	0.81	0.89	0.85	0.83	0.83	0.86	0.86	0.79

Table 3											
$A_1(\theta) = diag(\theta_1, ..., \theta_N)$, $K = N = 7$ elements of ε_t distributed according to $(0-6)$ in Table 1 with each distribution used only once.											
				$\phi(s) = s$				$\phi(s) = s(1+s^2)^{-\frac{1}{2}}$			
c	n	θ_0	L	1	2	3	4	1	2	3	4
One-step											
0.5	50	0.5	RMSE	0.56	0.60	0.74	0.69	0.62	0.57	0.68	0.74
		0.9	RMSE	1.38	1.30	1.24	1.09	1.26	1.06	1.21	1.14
		0.5	RFMSE	0.41	0.51	0.52	0.56	0.45	0.48	0.55	0.50
	100	0.5	RMSE	0.38	0.37	0.40	0.41	0.40	0.42	0.36	0.43
		0.9	RMSE	0.94	0.96	1.10	0.88	1.00	1.06	0.94	0.84
		0.5	RFMSE	0.70	0.70	0.70	0.72	0.67	0.72	0.73	0.69
0.9	50	0.5	RMSE	0.55	0.52	0.56	0.57	0.66	0.64	0.58	0.57
		0.9	RMSE	1.59	1.46	1.64	1.45	2.01	1.72	1.62	1.42
		0.5	RFMSE	0.62	0.68	0.67	0.59	0.64	0.70	0.66	0.66
	100	0.5	RMSE	0.83	0.78	0.75	0.69	0.96	0.94	0.76	0.70
		0.9	RMSE	3.54	3.65	3.65	3.47	3.95	4.01	3.65	3.23
		0.5	RFMSE	0.98	0.97	1.09	1.06	1.33	1.09	1.05	1.00
Iterative											
0.9	50	0.5	RMSE	0.17	0.19	0.19	0.27	0.14	0.17	0.20	0.26
		0.9	RMSE	0.23	0.26	0.23	0.37	0.20	0.22	0.24	0.31
		0.5	RFMSE	0.65	0.65	0.64	0.63	0.69	0.66	0.71	0.68
	100	0.5	RMSE	0.08	0.08	0.08	0.08	0.07	0.08	0.08	0.08
		0.9	RMSE	0.12	0.09	0.10	0.10	0.10	0.09	0.11	0.11
		0.5	RFMSE	0.83	0.84	0.86	0.84	0.83	0.88	0.81	0.78

Table 4											
$A_1(\theta) = diag(\theta_1, ..., \theta_N)$, $K = N = 56$ elements of ε_t distributed according to $(0-6)$ in Table 1 with each distribution used eight times.											
				$\phi(s) = s$				$\phi(s) = s(1+s^2)^{-\frac{1}{2}}$			
c	n	θ_0	L	1	2	3	4	1	2	3	4
One-step											
0.5	560	0.5	RMSE	3.44	4.02	4.07	3.89	4.26	3.57	4.06	3.65
		0.9	RMSE	19.2	22.1	15.0	20.7	32.0	16.8	25.1	18.1
		0.5	RFMSE	2.03	1.28	1.24	1.50	1.73	1.36	1.26	1.43
Iterative											
0.5	560	0.5	RMSE	0.07	0.07	0.08	0.07	0.08	0.08	0.08	0.07
		0.9	RMSE	0.19	0.20	0.14	0.22	0.21	0.19	0.22	0.15
		0.5	RFMSE	0.95	0.98	0.98	0.92	0.98	0.99	0.97	0.98
One-step											
0.9	100	0.5	RMSE	12.8	12.5	12.5	14.6	14.4	11.1	14.3	10.1
		0.9	RMSE	55.5	60.2	43.0	64.9	87.3	45.7	75.1	44.8
		0.5	RFMSE	3.44	2.33	2.04	2.71	3.51	2.33	1.88	2.23
Iterative											
0.9	50	0.5	RMSE	0.09	0.11	0.06	0.06	0.08	0.09	0.09	0.11
		0.9	RMSE	0.53	0.61	0.40	0.30	0.55	0.56	0.61	0.62
		RFMSE		0.99	1.00	0.98	0.96	1.00	1.01	0.99	1.00

Table 5											
$A_1(\theta) = \theta I_N,\ K = 1,\ N = 2$ elements of ε_t each distributed according to 0 in Table 1.											
				$\phi(s) = s$				$\phi(s) = s(1+s^2)^{-\frac{1}{2}}$			
c	n	θ_0	L	1	2	3	4	1	2	3	4
One-step											
0.5	50	0.5	RMSE	0.93	0.90	0.93	0.94	0.96	0.95	0.95	0.96
		0.9	RMSE	1.33	1.33	1.26	1.24	1.33	1.33	1.29	1.18
		0.5	RFMSE	0.44	0.44	0.58	0.56	0.45	0.50	0.50	0.55
			95%	0.94	0.92	0.86	0.86	0.92	0.92	0.89	0.85
			99%	0.98	0.97	0.94	0.94	0.98	0.97	0.96	0.93
	100	0.5	RMSE	0.84	0.84	0.84	0.87	0.88	0.84	0.87	0.92
		0.9	RMSE	1.15	1.07	1.07	1.09	1.14	1.15	1.09	1.06
		0.5	RFMSE	0.54	0.55	0.60	0.67	0.71	0.60	0.63	0.67
			95%	0.94	0.94	0.93	0.90	0.94	0.93	0.92	0.90
			99%	0.99	0.99	0.98	0.97	0.99	0.99	0.98	0.97
0.9	50	0.5	RMSE	0.84	0.81	0.85	0.86	0.86	0.86	0.86	0.89
		0.9	RMSE	1.11	1.10	1.07	1.07	1.11	1.11	1.09	1.00
		0.5	RFMSE	0.36	0.35	0.48	0.48	0.35	0.38	0.41	0.46
			95%	0.94	0.91	0.86	0.86	0.92	0.92	0.88	0.84
			99%	0.98	0.97	0.94	0.93	0.98	0.97	0.96	0.93
	100	0.5	RMSE	0.75	0.75	0.76	0.79	0.80	0.75	0.79	0.83
		0.9	RMSE	0.99	0.90	0.92	0.95	0.99	0.99	0.91	0.90
		0.5	RFMSE	0.47	0.48	0.52	0.61	0.64	0.53	0.56	0.59
			95%	0.94	0.94	0.92	0.90	0.94	0.93	0.92	0.90
			99%	0.99	0.99	0.98	0.97	0.99	0.98	0.98	0.97

Table 6											
$A_1(\theta) = \theta I_N,\ K = 1,\ N = 7$ elements of ε_t each distributed according to $(0-6)$ in Table 1 with each distribution used only once.											
				$\phi(s) = s$				$\phi(s) = s(1+s^2)^{-\frac{1}{2}}$			
c	n	θ_0	L	1	2	3	4	1	2	3	4
One-step											
0.5	50	0.5	RMSE	0.55	0.57	0.60	0.63	0.54	0.54	0.60	0.63
		0.9	RMSE	0.98	0.90	0.91	0.86	0.90	0.89	0.82	0.83
		0.5	RFMSE	0.20	0.22	0.27	0.28	0.20	0.19	0.25	0.29
			95%	0.86	0.84	0.80	0.76	0.87	0.84	0.78	0.73
			99%	0.97	0.91	0.91	0.88	0.95	0.93	0.89	0.85
	100	0.5	RMSE	0.49	0.48	0.47	0.49	0.46	0.48	0.47	0.52
		0.9	RMSE	0.71	0.69	0.69	0.63	0.66	0.68	0.66	0.67
		0.5	RFMSE	0.34	0.35	0.38	0.38	0.40	0.44	0.40	0.44
			95%	0.91	0.91	0.91	0.89	0.91	0.91	0.88	0.83
			99%	0.98	0.97	0.98	0.96	0.98	0.97	0.96	0.94
0.9	50	0.5	RMSE	0.50	0.52	0.53	0.58	0.51	0.51	0.55	0.59
		0.9	RMSE	0.82	0.77	0.76	0.77	0.77	0.80	0.73	0.79
		0.5	RFMSE	0.15	0.16	0.13	0.20	0.15	0.16	0.23	0.30
			95%	0.85	0.84	0.79	0.74	0.86	0.84	0.78	0.73
			99%	0.96	0.94	0.90	0.87	0.95	0.95	0.90	0.83
	100	0.5	RMSE	0.45	0.45	0.45	0.45	0.41	0.42	0.44	0.49
		0.9	RMSE	0.63	0.61	0.57	0.60	0.55	0.58	0.57	0.61
		0.5	RFMSE	0.34	0.37	0.29	0.34	0.31	0.37	0.37	0.36
			95%	0.91	0.89	0.89	0.88	0.92	0.90	0.88	0.84
			99%	0.98	0.96	0.97	0.95	0.98	0.97	0.96	0.93

However, for $A_1(\theta) = \text{diag}(\theta_1, \ldots, \theta_N)$, (15), we find some strange results for $c = 0.9$. The performance of the one-step adaptive estimate (10) is worse than least squares (12), and, moreover the relative performance of the former falls as we increase sample size. From inspection of additional results (not presented for the sake of brevity) we found that both estimates improve significantly with increasing sample size, but least squares sees a far more dramatic gain. On the other hand the iterative adaptive estimates dominate least squares and we see an improvement in this relative superiority with increasing sample size. This pattern continues and becomes more evident as we increase the dimension N, in Tables 3, 4 and 6, where also non-Gaussian distributions are introduced. For $A_1(\theta) = I_n\theta$, with increasng N the relative performance of the adaptive estimates increases across all parameter values, reflecting the inefficiency of least squares as we move further from the Gaussian benchmark. It seems that when there is a high degree of mixing, $c = 0.9$, in order for the adaptive estimates to achieve efficiency improvements over least squares in small samples the iterative estimator is required.

Irrespective of the sample size and the value of the mixing parameter, c, the relative performance of the adaptive estimates is poorer for $\theta_0 = 0.9$ compared to $\theta_0 = 0.5$. Thus it appears that their relative superiority is mitigated somewhat near the unit root.

The choice of L does not seem to make a large difference in terms of RMSE. For $\theta_0 = 0.5$, a larger L tends to reduce relative performance of the adaptive estimates, whereas for $\theta_0 = 0.9$ a larger L improves it. There does not seem to be a clear pattern in the results for the different forms of $\phi(s)$.

The forecast performance of the adaptive estimates looks encouraging. In nearly all situations they outperform least squares; only in the high dimensional case, $N = 56$, is the one-step estimate inferior. The simplest form of estimate, taking $L = 1$ and $\phi(s) = s$, provides the best results.

It appears that for smaller sample sizes coverage rates are fairly anti-conservative, but as n increases to 100 these rates return fairly closely to the nominal level. For smaller values of L the coverage tracks the nominal level very closely, but becomes quite anti-conservative as L increases. Coverage rates are relatively insensitive to the different forms of $\phi(s)$ or to the use of one-step and iterative estimates.

5. FINAL COMMENTS

In a semiparametric model for stationary multiple time series of possibly high dimension, we have presented adaptive estimates of the parameters, and rules of large sample statistical inference,

avoiding a curse of dimensionality by modelling the innovations vector as a linear transformation of independent but possibly non-identically distributed random variables, having nonparametric distributions. Our setting, which covers vector autoregressive moving average processes, is widely applicable. In the vector autoregressive case, a Monte Carlo simulation study has found generally good finite sample performance of our estimates, with respect to accuracy and to their use in forecasting and interval estimation.

ACKNOWLEDGMENTS

This research was supported by ESRC Grant ES/J007242/1. The paper was revised following comments of a co-editor of the special issue and two reviewers.

REFERENCES

Aires, F. and Chedin, A. (2000) Independent component analysis of multivariate time series: Application to the tropical SST variability. *J. Geophys. Res.* 105, 17437-17455.

Bach, F.R. and Jordan, M. I. (2002) Kernel independent component analysis. *J. Mach. Learning Res.* 3, 1-48.

Beran, R. (1976) Adaptive estimates for autoregressive processes. *Ann. Inst. Statist. Math.* 26, 77-89.

Bickel, P. (1982) On adaptive estimation. *Ann. Statist.* 10, 647-671.

Box, G.E.P. and Jenkins, G.M. (1971) *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day.

Chen, A. and Bickel, P. (2005) Efficient independent component analysis. *Ann. Statist.* 34, 2825-2855.

Chen, A. and Bickel, P. (2006) Consistent independent component analysis and prewhitening. *IEEE Trans. Sig. Proc.* 53, 3625-3632.

Chen, J-P., Chen, Y. and Haerdle, W. (2011) TVICA - time varying independent component analysis and its application to financial data. SFB Discussion Paper 2011-054.

Cheung, Y-M. and Xu, L. (2001) Independent component analysis ordering in ICA time series analysis. *Neurocomputing* 41, 145-152.

Drost, F.L., Klassen, C.A.J. and Werker, B.J.M. (1997) Adaptive estimation in time series models. *Ann. Statist.* 25, 786-818.

Garcia-Ferre, A., Gonzalez-Prieto, E. and Pena D. (2011) Exploring ICA for time series decomposition. Working Paper 11-16 Univ. Carlos III de Madrid.

Hastie, T. and Tibshirani, R. (2003) Consistent independent components analysis through product density estimation. In *Advances in Neural Information Processing Systems 15* (S. Becker and K. Obermayer, eds.) pp. 649-656. Cambridge, MA: MIT Press.

Hyvarinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis.* New York: Wiley.

Koul, H.L. and Schick, A. (1997) Efficient estimation in nonlinear autoregressive models. *Bernoulli* 3, 247-277.

Kreiss, J-P. (1987) On adaptive estimation in stationary ARMA processes. *Ann. Statist.* 15, 112-133.

Lin, J-C., Li, Y-H. and Liu, C-H. (2007) Building time series forecasting by independent component analysis mechanism. *Proc. World Cong. Eng. Vol II.*

Lu, C-J., Lee, T-S. and Chiu, C-C. (2009) Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems* 47, 115-125.

Nascimento, J. M. P. and Dias, J. M. D. (2005) Does independent component analysis play a role in unmixing hyperspectral data? *IEEE Trans. Geoscience Rem. Sensing.* 43, 175-187.

Newey, W.K. (1988) Adaptive estimation of regression models via moment restrictions. *J. Econometrics* 38, 301-339.

Robinson, P.M. (2005) Efficiency improvements in inference on stationary and nonstationary fractional time series. *Ann. Statist.* 33, 1800-1842.

Samarov, A. and Tsybakov, A. (2004) Nonparametric independent component analysis. *Bernoulli* 10, 565-582.

Samworth, R.J. and Yuan, M. (2012) Independent component analysis via nonparametric maximum likelihood estimation. *Ann. Statist.* 40, 2973-3002.

Vlassis, N. (2001) Efficient source adaptivity in independent component analysis. *IEEE Trans. Neur. Net.* 12, 559-565.