Optimal inference in dynamic models with conditional moment restrictions*

Bent Jesper Christensen†
University of Aarhus and CREATES

Michael Sørensen‡
University of Copenhagen and CREATES

July 4, 2008

Abstract

By an application of the theory of optimal estimating function, optimal instruments for dynamic models with conditional moment restrictions are derived. The general efficiency bound is provided, along with estimators attaining the bound. It is demonstrated that the optimal estimators are always at least as efficient as the traditional optimal generalized method of moments estimator, and usually more efficient. The form of our optimal instruments resembles that from Newey (1990), but involves conditioning on the history of the stochastic process. In the special case of i.i.d. observations, our optimal estimator reduces to Newey's. Specification and hypothesis testing in our framework are introduced. We derive the theory of optimal instruments and the associated asymptotic distribution theory for general cases including non-martingale estimating functions and general history dependence. Examples involving time-varying conditional volatility and stochastic volatility are offered.

Key words: optimal estimating function, generalized method of moments, conditional moment restrictions, dynamic models, optimal instruments, martingale estimating function, specification test.

JEL codes: C12, C13, C22, C32.

*We are grateful to Gary Chamberlain, Steven Heston, Yongmiao Hong, Rustem Ibragimov, Guido Imbens, Nick Kiefer, Whitney Newey, Jim Stock, and seminar participants at Cornell University and Harvard University for useful comments, and to Center for Research in Econometric Analysis of TimE Series (CREATES), funded by the Danish National Research Foundation, and the Danish Social Science Research Council for financial support. Some of this research was carried out when Christensen was visiting the Department of Economics, Harvard University, and the generosity and hospitality of the Department are gratefully acknowledged.

†Address: School of Economics and Management, University of Aarhus, 322 University Park, DK-8000 Århus C, Denmark. Email: bjchristensen@creates.au.dk.

‡Address: Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark. Email: michael@math.ku.dk.

1 Introduction

Ever since it was introduced by Hansen (1982), the generalized method of moments (GMM) has been an immensely popular method for estimation and inference in econometrics. The estimators are widely applicable and robust to model misspecification. The particular form of GMM, commonly referred to as optimal GMM, where the weight matrix or norm is chosen to minimize the asymptotic covariance matrix of the resulting estimator, is well known and frequently implemented in applications. However, we show that it is possible to modify the estimating equations and thereby improve efficiency without changing the underlying model assumptions. Our theory applies to general dynamic models with conditional moment restrictions. In effect, we consider a wider class of estimators that contains GMM as a special case, and we determine the optimal estimator within this generalized class. The optimal estimating equation is quite explicit and readily computable, and we show that the estimator achieves the general efficiency bound derived by Hansen (1985) and Hansen, Heaton & Ogaki (1988).

Our estimator utilizes the information in the conditional moment restrictions efficiently. In the case of optimal GMM, the conditional moment conditions are averaged across the sample, possibly mitigating the information loss entailed in unconditioning by expanding the moment conditions using instrumental variables. The question is how to choose instruments to utilize the conditional moment restrictions efficiently. Essentially two different approaches have been considered in the literature. One is to explicitly treat the moment conditions as conditional, and derive optimal instruments given the relevant conditioning variables, producing a number of estimating equations equal to the number of parameters. This is the approach followed by Newey (1990), who shows that nonparametric estimation of the optimal instruments can yield asymptotically efficient estimators in the i.i.d. case. Robinson (1991) allows for some restricted forms of serial dependence, but imposes conditional homoskedasticity. Newey (1993) considers the case of conditional heteroskedasticity depending on i.i.d. regressors. The alternative approach which has been pursued in the literature is to treat the moment conditions as unconditional, and let the number of instruments and hence moment conditions expand with sample size for efficiency purposes. This is the approach of Chamberlain (1987), who shows that the semiparametric efficiency bound for static models may be achieved this way in the limit.

In this paper, we solve the optimal inference problem for general dynamic conditionally heteroskedastic models. The form of our optimal weights (or instruments) resembles that from Newey (1990), but involves conditioning on the history of the stochastic process, rather than on i.i.d. regressors. In the special case of i.i.d. observations, our optimal estimator reduces to that of Newey. Our focus is on the form of the estimating equations in general dynamic cases, rather than on the computation of the instruments if they are not known analytically. In some cases, they may be determined arbitrarily precisely by simulation. In other cases, nonparametric estimation of instruments in dynamic models following Wefelmeyer (1996) may be pursued.

The modern statistical theory of optimal estimating functions dates back to the papers by Godambe (1960) and Durbin (1960). Indeed, the basic idea was in a sense

already used in Fisher (1935). The theory for dynamic models was developed in a series of papers by Godambe (1985), Godambe & Heyde (1987), Heyde (1988), and several others; see the references in Heyde (1997). Important particular instances are likelihood inference, the quasi-likelihood of Wedderburn (1974) and the closely related generalized estimating equations developed by Liang & Zeger (1986) to deal with problems of longitudinal data analysis, see also Prentice (1988) and Li (1997).

We use recent developments in this literature, reviewed in Bibby, Jacobsen & Sørensen (2004) and Sørensen (2008), to obtain improved estimators in the situation where the estimating function is a martingale and involves a finite lag length. After developing these estimators, we go on to show new results on optimal estimators in more general cases allowing for general history dependence and where the estimating functions need not be martingales. While the usual optimal GMM applies a common average weight matrix across all time periods, at least asymptotically, as we show, the idea behind the optimal estimator is to utilize the conditioning information in dynamic models optimally by applying time-varying weight matrices. By considering these time-varying weights as instruments, our optimal estimator may hence be regarded as GMM with optimal instruments.

The paper is laid out as follows. In Section 2, we study the martingale estimating function case with finite lag structure. To fix ideas, we briefly review the usual GMM approach, given a set of conditional moment conditions, including optimal GMM, where optimality is over choice of weight matrix (or norm). We then introduce the optimal estimator, which involves optimal choice of time-varying weights (or instruments), and show that it is strictly more efficient than optimal GMM. This result is obtained based only on the same stationarity and ergodicity conditions used in GMM, anyway, as well as ability to calculate conditional second moments. We do not rely on detailed distributional assumptions such as in maximum likelihood, and, when deriving our optimal estimator, we take the same conditional moment restrictions as in GMM for given. We also offer a pseudo-likelihood interpretation of the optimal estimator. In addition, we introduce specification and hypothesis testing in our framework, in the spirit of Newey & West (1987a). In Section 3, we generalize the theory to cover arbitrary history dependence, which requires new developments relative to both the econometrics and mathematical statistics literatures. In Section 4, we generalize to the non-martingale case, which requires the introduction of novel techniques, involving a suitable operator whose domain is the space of weight matrices (or instruments). We establish the form of the optimal estimating equation, including the case where conditional moment restrictions are generalized to conditions on prediction errors. The theory of optimal prediction-based estimating function of Sørensen (2000) is recovered as a particular case. Concluding remarks are made in Section 6, and proofs of results are in the Appendix. Throughout, the theory is illustrated by several examples, including GARCH, diffusion, and stochastic volatility models.

2 Martingale estimating functions

2.1 GMM and optimal inference

Suppose we model the observed time series $X_1, X_2, \ldots, X_T$ by a stochastic dynamic model indexed by a K-dimensional parameter θ that we wish to estimate. In general, we may consider multivariate data X_t, e.g. in cases where some of the coordinates are explanatory variables. We assume that the stochastic process $\{X_t\}$ is stationary and ergodic. Let $m_t(\theta) = m(X_t, \ldots, X_{t-L}; \theta)$ be a vector of functions of θ and the observations at the time points $t, t-1, \ldots, t-L$ for some given lag length L (we assume $t > L$) satisfying the conditional moment restrictions

$$E_\theta(m_{t+1}(\theta) \,|\, \mathcal{F}_t) = 0 \tag{2.1}$$

for all θ. Here, $\mathcal{F}_t$ is the σ-field generated by $X_1, \ldots, X_t$, and E_θ denotes expectation under the model with parameter value θ ($E_\theta(\cdot \,|\, \mathcal{F}_t)$ denotes conditional expectation given $\mathcal{F}_t$). The relations among observations given by $m_t(\theta)$ are preferably chosen on the basis of economic theory. The restriction to finite lag length L is relaxed to general history dependence ($L = \infty$) in Section 3 below. To interpret condition (2.1), note that it implies that $\{m_t(\theta)\}$ is a martingale difference sequence.

The GMM method makes use of a vector of instruments $z_t(\theta) = z(X_t, \ldots, X_{t-L+1}; \theta)$ dependent on data at time points $t, \ldots, t-L+1$ to define

$$h_{t+1}(\theta) = z_t(\theta) \otimes m_{t+1}(\theta), \tag{2.2}$$

which, by construction, satisfies the conditional moment restrictions

$$E_\theta(h_{t+1}(\theta) \,|\, \mathcal{F}_t) = 0. \tag{2.3}$$

We denote the dimension of the vector $h_{t+1}(\theta)$ by M and assume that it has finite second moments. A basic tool for estimating θ is the estimating function

$$H_T(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} h_t(\theta). \tag{2.4}$$

Since $\{m_t(\theta)\}$ and hence $\{h_t(\theta)\}$ are martingale difference sequences, $H_T(\theta)$ is a martingale estimating function (precisely, $(T-L)H_T(\theta) = \sum_{t=L+1}^{T} h_t(\theta)$ is a martingale since $E_\theta(|H_T(\theta)|) < \infty$ and $E_\theta(\sum_{t=L+1}^{T+1} h_t(\theta) \,|\, \mathcal{F}_t) = \sum_{t=L+1}^{T} h_t(\theta)$). When $M = K$, the exactly identified case, an estimator θ can be obtained by solving the estimating equation

$$H_T(\theta) = 0. \tag{2.5}$$

Frequently, $M > K$, so that we have more equations than parameters to be estimated, and, in effect, $M - K > 0$ overidentifying restrictions. In this case, a GMM estimator may be obtained by minimizing the quadratic form

$$M_T(\theta) = H_T(\theta)' W H_T(\theta), \tag{2.6}$$

for a suitably chosen $M \times M$-matrix W. Here and later, x' denotes the transpose of a vector or matrix x. Specifically, an initial estimator θ^I is obtained by minimizing (2.6)

for $W = I_M$ (the M-dimensional identity matrix). Then the optimal GMM estimator $\tilde{\theta}_T$ of Hansen (1982) is obtained by minimizing (2.6) with $W = \hat{V}(\theta^I)^{-1}$, where

$$\hat{V}(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} h_t(\theta) h_t(\theta)' \tag{2.7}$$

is an estimate of $V(\theta) = \mathrm{Var}_\theta(h_t(\theta))$, the covariance matrix of $h_t(\theta)$. It is assumed that $V(\theta)$ is non-singular, or equivalently that the coordinates of $h_t(\theta)$ are linearly independent functions of $X_t, \ldots, X_{t-L}$. Among all possible choices of W in (2.6), this is the one for which the asymptotic variance of the resulting estimator is minimal. In the exactly identified case $M = K$, optimal GMM is the solution to (2.5), so the estimator is defined for $M \geq K$.

Heteroskedasticity and autocorrelation consistent covariance estimators involving estimates of $\mathrm{Cov}_\theta(h_t(\theta), h_{t-j}(\theta))$ for $j \neq 0$ are often used instead of (2.7) (see Hansen (1982) and Newey & West (1987b)), but this is unnecessary here because the fundamental model assumption (2.3) implies that $\{h_t(\theta)\}$ is a martingale difference sequence, and hence, in particular, serially uncorrelated.

Under regularity conditions including ergodicity of the model and differentiability of $h_t(\theta)$ with respect to θ, Hansen (1982) shows that the asymptotic distribution of the optimal GMM estimator is given by

$$\sqrt{T}\left(\tilde{\theta}_T - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, \left(D(\theta_0)' V(\theta_0)^{-1} D(\theta_0)\right)^{-1}\right) \tag{2.8}$$

as $T \to \infty$, where θ_0 denotes the true parameter value and $D(\theta)$ is the $M \times K$-matrix

$$D(\theta) = E_\theta\left(\frac{\partial h_t(\theta)}{\partial \theta'}\right). \tag{2.9}$$

Asymptotic standard errors may be calculated by substituting $\hat{V}(\theta)$ from (2.7) for $V(\theta)$ in (2.8), and

$$\hat{D}(\theta) = (T-L)^{-1} \sum_{t=L+1}^{T} \frac{\partial h_t(\theta)}{\partial \theta'} \tag{2.10}$$

for $D(\theta)$, both evaluated at the estimate of θ_0, either θ^I or $\tilde{\theta}_T$. The asymptotic distribution result follows easily from considering the first order conditions for minimization of the quadratic form (2.6), given by

$$\frac{\partial H_T(\theta)'}{\partial \theta} W H_T(\theta) = 0. \tag{2.11}$$

The first order conditions simply amount to premultiplying the M-dimensional estimating equation (2.5) by a suitable $K \times M$ weight matrix $\partial H_T(\theta)'/\partial\theta \cdot W$, thus yielding K equations in the K unknown parameters. Hence, $\tilde{\theta}_T$, originally defined as the minimizer of (2.6), can equivalently be considered as the solution of the K equations (2.11). From this viewpoint, an asymptotically equivalent estimator is obtained by solving the simplified first order conditions, where the initial estimator θ^I obtained by minimizing (2.6) for $W = I_M$ is substituted not only in W, estimating $V(\theta)^{-1}$ by

$W = \hat{V}(\theta^I)^{-1}$, but also in $\partial H_T(\theta)'/\partial\theta$, estimating this by $\hat{D}(\theta^I)'$. Thus, the simplified estimating equation is

$$\hat{D}(\theta^I)'\hat{V}(\theta^I)^{-1}H_T(\theta) = 0, \tag{2.12}$$

of dimension K, to be solved for θ entering only through $H_T(\cdot)$. The resulting estimator is consistent and has the same asymptotic distribution (2.8) as the optimal GMM estimator. In effect, among estimators defined by estimating functions of the form $wH_T(\theta)$, with w a $K \times M$ weight matrix, that corresponding to $w = D(\theta)'V(\theta)^{-1}$ is optimal, and asymptotically equivalent to optimal GMM, also with consistent estimators in w, as in (2.12).

Since $H_T(\theta)$ is given as the raw average of the M-dimensional vector functions $h_t(\theta)$, it is natural to consider a generalized class of estimating equations, moving the $K \times M$ weight matrix ($\hat{D}(\theta^I)'\hat{V}(\theta^I)^{-1}$ in (2.12)) under the summation sign and allowing it to vary across time. Thus, the estimating function is generalized from the raw average type $wH_T(\theta) = w\sum_t h_t(\theta)/(T-L)$ to a weighted average of the moment conditions $h_t(\theta)$,

$$G_T(\theta) = \frac{1}{T-L}\sum_{t=L+1}^{T} w_t(\theta)h_t(\theta) = 0, \tag{2.13}$$

with $K \times M$ weight-matrices $w_t(\theta) = w(X_{t-1}, \ldots, X_{t-L}; \theta)$ that are arbitrary functions of $X_{t-1}, \ldots, X_{t-L}$ satisfying that $w_t(\theta)h_t(\theta)$ has finite second moments. This defines a very general class of estimators. Since $w_t(\theta)$ does not depend on X_t, the estimating function $G_T(\theta)$ is still a martingale. The K equations in (2.13) may be solved directly with respect to θ as it enters both $w_t(\cdot)$ and $h_t(\cdot)$, or an initial consistent estimator may be substituted in $w_t(\cdot)$, e.g. θ^I as in the GMM case, or the optimal GMM estimator, and the asymptotic properties of the final estimator are unaltered by this two-step procedure.

To be sure, estimators that are asymptotically equivalent to optimal GMM may be constructed in many different ways by imposing that the weight matrices in (2.13) be time-invariant,

$$w_t(\cdot) = \hat{D}(\cdot)'\hat{V}(\cdot)^{-1}. \tag{2.14}$$

Thus, solving $\hat{D}(\theta)'\hat{V}(\theta)^{-1}H_T(\theta) = 0$ with respect to θ yields another asymptotic equivalent, indeed, a one-step estimator, whereas that from (2.12) has θ^I in both $\hat{D}(\cdot)$ and $\hat{V}(\cdot)$ fixed when solving for θ, and the optimal GMM estimator itself is the third special case with weights not dependent on t produced by using θ^I in $\hat{V}(\cdot)$ but not in $\hat{D}(\theta)$ in (2.14). Note that neither of these three cases corresponds to minimizing the quadratic form (2.6) with $W = W(\theta)^{-1}$, i.e., allowing W to vary with the argument θ in the minimization. In particular, the first order conditions for such a problem would include derivatives of W with respect to θ, too, in contrast to (2.11). For most specifications $W(\cdot)$ this would lead to inconsistent estimators, but imposing $W = \hat{V}(\theta)^{-1}$ from (2.7) and minimizing (2.6) for this choice yields the continuous updating estimator of Hansen, Heaton & Yaron (1996) which is consistent. It is also another example of a one-step estimator, and a fourth case of an asymptotic equivalent to optimal GMM.

Replacing the time-invariant weights (2.14) with time-varying weights depending on both data and unknown parameters, $w_t(\theta)$ in (2.13), opens up for improved use

of conditioning information and more efficient estimators, relative to optimal GMM. This is the main point of the present paper. The improved estimators make use of the $M \times K$-matrix of conditional expectations

$$d_t(\theta) = E_\theta \left(\frac{\partial h_t(\theta)}{\partial \theta'} \,\middle|\, \mathcal{F}_{t-1}^L \right), \tag{2.15}$$

where $\mathcal{F}_t^L$ is the σ-field generated by $X_t, \ldots, X_{t-L+1}$, and of the $M \times M$ conditional covariance matrix of the moment conditions $h_t(\theta)$ given $\mathcal{F}_{t-1}^L$,

$$\Phi_t(\theta) = \mathrm{Var}_\theta \left(h_t(\theta) \,|\, \mathcal{F}_{t-1}^L \right) = E_\theta \left(h_t(\theta) h_t(\theta)' \,|\, \mathcal{F}_{t-1}^L \right). \tag{2.16}$$

Throughout, we write $\Phi_t(\theta)^{-1}$ both for the ordinary inverse of $\Phi_t(\theta)$, when it exists, and for the generalized inverse, which is relevant e.g. when instruments such as $z_t(\theta)$ in (2.2) are used to expand the set of moment conditions, since in this case the rank of $\Phi_t(\theta)$ is bounded by the dimension of $m_t(\theta)$.

The theory of optimal estimating functions gives criteria for an estimating function to be optimal within the class of all estimating functions of the form (2.13) with $h_t(\theta)$ given. A modern exposition of this theory can be found in Heyde (1997) (see also the review in Bibby, Jacobsen & Sørensen (2004)). The optimal estimating function is the one that is closest to the score function in an L^2-sense and which minimizes the asymptotic variance of the corresponding estimator. We denote the optimal choice of the weight matrix $w_t(\theta)$ by $w_t^*(\theta)$, and the optimal estimating function by

$$G_T^*(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} w_t^*(\theta) h_t(\theta). \tag{2.17}$$

Our first main result concerns the optimal choice of the weight matrix, fully exploiting the information contained in the conditional moment restrictions (2.3) in the dynamic case. The theorem is proved in the appendix.

Theorem 2.1 *The optimal estimating function in the class of all estimating functions of the form (2.13) is obtained when the weight-matrix is chosen as*

$$w_t^*(\theta) = d_t(\theta)' \Phi_t(\theta)^{-1}. \tag{2.18}$$

The asymptotic distribution as $T \to \infty$ of the estimator $\hat{\theta}_T$ obtained by solving the estimating equation $G_T^*(\theta) = 0$ is (under standard regularity conditions including invertibility of $\mathcal{J}(\theta)$)

$$\sqrt{T} \left(\hat{\theta}_T - \theta \right) \xrightarrow{\mathcal{D}} N \left(0, \mathcal{J}(\theta)^{-1} \right), \tag{2.19}$$

where

$$\mathcal{J}(\theta) = E_\theta \left(d_t(\theta)' \Phi_t(\theta)^{-1} d_t(\theta) \right). \tag{2.20}$$

The asymptotic information matrix $\mathcal{J}(\theta)$ may be consistently estimated as $\hat{\mathcal{J}}(\hat{\theta}_T)$ with

$$\hat{\mathcal{J}}(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} d_t(\theta)' \Phi_t(\theta)^{-1} d_t(\theta). \tag{2.21}$$

Theorem 2.1 presents the optimal choice of instruments or weights $w_t(\theta)$ in (2.13) in the present dynamic setting. It follows that $\mathcal{J}(\theta_0)$ is the efficiency bound in the class of estimators defined by different choices of instruments, i.e., $\mathcal{J}(\theta_0)^{-1}$ is the lower bound on the asymptotic covariance matrix of estimators of θ. Furthermore, the optimal estimator $\hat{\theta}_T$ from the Theorem attains this efficiency bound.

Several points are in order. Firstly, in the restricted class where $w_t(\theta) = w_t$, i.e., instruments do not depend on parameters, efficiency bounds were derived by Hansen (1985) (Lemma 4.3) and Hansen, Heaton & Ogaki (1988) (Theorem 4.2) for general settings not necessarily restricted to the martingale difference case. We return to the non-martingale case in Section 4 below. In the martingale difference case, their bound coincides with ours, $\mathcal{J}(\theta_0)^{-1}$, which is derived by different techniques. When seeking estimators achieving the efficiency bound in the restricted class, Hansen (1985) considered the simplified case where the conditional variance $\Phi_t(\theta) \equiv \Gamma(\theta)$ is constant across time t and deterministic. The estimator obtained using the weights

$$w_t = d_t(\theta_0)'\Gamma(\theta_0)^{-1}, \tag{2.22}$$

with θ_0 the true parameter value, was shown to achieve the efficiency bound in this situation. Note that dependence on true θ_0 was not a problem for deriving Hansen's efficiency bound, but for construction of a feasible estimator an initial consistent estimate must be plugged in.

The optimal estimator is well-defined even in the case $M < K$, which would be underidentified in the GMM case. The reason is that the optimal weights $w_t^*(\theta)$ by their definition ensure the right number of estimating equations, K, that in wide generality are different.

The optimal weights resemble those from GMM, which are defined in the exactly or overidentified case $M \geq K$. Thus, the identities

$$E_\theta(d_t(\theta)) = D(\theta) \quad \text{and} \quad E_\theta(\Phi_t(\theta)) = V(\theta) \tag{2.23}$$

imply that if we replace $d_t(\theta)$ and $\Phi_t(\theta)$ in (2.18) by their unconditional expectations, hence making them constant and deterministic, then the optimal estimating equation weights reduce to the optimal GMM weights (2.14) (specifically, if θ^I is substituted for θ in $\hat{V}(\theta)$). Similarly, if $\Phi_t(\theta)$ is made constant, but $d_t(\theta)$ is allowed to vary through time, weights like (2.22) from Hansen (1985) are obtained. Again, these are only optimal when the true conditional variance $\Phi_t(\theta_0)$ is constant and deterministic, ruling out e.g. ARCH-type effects. Theorem 2.1 covers the general situation with time-varying conditional second moments, which are important e.g. in many macroeconomic and financial applications.

The form of the optimal weights (or instruments) in (2.18) also resembles that from the cross-section case considered by Newey (1990), who studies the i.i.d. case with conditional moment restrictions, conditioning on i.i.d. regressors, whereas the conditioning in both $d_t(\theta)$ and $\Phi_t(\theta)$ in our optimal weights for the dynamic conditionally heteroskedastic case is on the history of the time series. The restriction to lag length L is relaxed in Section 3 below. Of course, if regressors are useful, they may readily be included as additional coordinates in X_t. This way the cross-section

regression model is a special case of ours, and in the i.i.d. case our optimal estimator reduces to that considered by Newey (1990).

Writing out the weighted moment conditions (2.13) for the case $h_t(\theta) = m_t(\theta)$ makes it clear that the standard use of vector-valued instruments (as opposed to matrices) such as $z_t(\theta)$ in (2.2), following Hansen (1982), is equivalent to choosing weight matrices $w_t(\theta) = z_{t-1}(\theta) \otimes I_p$, where $p = \dim(m_t)$, since $z_{t-1}(\theta) \otimes m_t(\theta)$ may be recast as $(z_{t-1}(\theta) \otimes I_p) m_t(\theta)$. These weight matrices (or instruments) obviously have very special structure and are of dimension $qp \times p$, where $q = \dim(z_t)$, and possibly $qp > K$, the dimension of the parameter θ. Theorem 2.1 then shows that it suffices to take weights of dimension $K \times p$, that is, the same number of moment conditions and parameters. In particular, the optimal weights from the Theorem may as well be based on the original set of moments $m_t(\theta)$, rather than the expanded set $h_t(\theta)$ obtained using the instruments in $z_{t-1}(\theta)$. In effect, the optimal instruments (weights) from the Theorem will, if necessary, undo the multiplication by $z_{t-1}(\theta)$. To see this, note that the two classes of estimating functions of the form (2.13) based on $h_t(\theta) = z_{t-1}(\theta) \otimes m_t(\theta)$ and $\tilde{h}_t(\theta) = m_t(\theta)$, respectively, coincide because $w_t(\theta)(z_{t-1}(\theta) \otimes m_t(\theta)) = w_t^\dagger(\theta) m_t(\theta)$ and $\tilde{w}_t(\theta) m_t(\theta) = \tilde{w}_t^\dagger(\theta)(z_{t-1}(\theta) \otimes m_t(\theta))$, where

$$w_t^\dagger(\theta) = w_t(\theta)(z_{t-1}(\theta) \otimes I_p) \quad \text{and} \quad \tilde{w}_t^\dagger(\theta) = \tilde{w}_t(\theta)(z_{t-1}(\theta)^- \otimes I_p)$$

with $z_{t-1}(\theta)^- = z_{t-1}(\theta)'/z_{t-1}(\theta)' z_{t-1}(\theta)$. This follows because $(z_{t-1}(\theta)^- \otimes I_p)$ is a generalized inverse of $(z_{t-1}(\theta) \otimes I_p)$, i.e. $(z_{t-1}(\theta)^- \otimes I_p)(z_{t-1}(\theta) \otimes I_p) = I_p$. In particular, the optimal estimating equations based on $m_t(\theta)$ and $h_t(\theta) = z_{t-1}(\theta) \otimes m_t(\theta)$ coincide. The original $m_t(\theta)$ may be recovered from the expanded $h_t(\theta)$, and the optimal estimator may as well be constructed by applying the optimal instruments $w_t^*(\theta)$ from the Theorem to the former. Of course, the very reason that instruments are introduced in (2.2) is that they capture relevant conditioning information, but in our approach this may be accommodated by including the relevant data series among the coordinates of X_t, which in turn enter the weights $w_t^*(\theta)$ in the optimal manner.

In many cases, e.g. when $X_1, X_2, \ldots, X_T$ is a discrete time sample from a continuous time process, there is no closed form expression for the conditional covariance matrix $\Phi_t(\theta)$ from (2.16), and it must be determined numerically, e.g. by simulation. An approach to circumvent the potentially considerable computational burden associated with recalculating $\Phi_t(\theta)$ at each trial value of the parameter θ without any loss of efficiency relative to the optimal estimator is to replace $\Phi_t(\theta)$ in (2.18) by $\Phi_t(\bar{\theta}_T)$, where $\bar{\theta}_T$ is a preliminary consistent estimator of θ, for instance the estimator θ^I obtained by minimizing (2.6) with $W = I_M$ or the optimal GMM estimator $\tilde{\theta}_T$. In this way, the simulation need only be carried out a single time, for one particular parameter value, $\bar{\theta}_T$. We refer to the estimating function

$$G_T^\diamond(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} w_t^\diamond(\theta) h_t(\theta) \tag{2.24}$$

with $w_t^\diamond(\theta) = d_t(\theta)' \Phi_t(\bar{\theta}_T)^{-1}$ as the modified optimal estimating function. The asymptotic distribution of the estimator obtained by solving the estimating equation $G_T^\diamond(\theta) = 0$ with respect to θ as it enters only $d_t(\theta)$ and $h_t(\theta)$ is the same as in (2.19). If the

function $d_t(\theta)$ from (2.15) is complicated to calculate, too, the weights $w_t^*(\bar{\theta}_T) = d_t(\bar{\theta}_T)'\Phi_t(\bar{\theta}_T)^{-1}$ can be used, and the estimating equation solved only with respect to θ in $h_t(\theta)$, again without any loss of relative efficiency. Again, if necessary, approximations may be inserted for Φ_t and d_t in the time-varying optimal weights, and, following Wefelmeyer (1996), if each of the consecutive approximations is based on data only through $t-1$, then the martingale property of the estimating function is retained. More generally, this property holds asymptotically for consistent estimators.

In some applications, $\Phi_t(\theta)$ or $d_t(\theta)$ may simply not be known, even up to simulation, e.g., if only a few moment conditions are available, as opposed to a full model specification. In such situations, they can be determined e.g. by flexible functional form regression, kernel smoothing, or similar. Often, asymptotic properties of the proposed estimators are unchanged if an estimator or approximation of these functions is used. Chamberlain (1992) discusses the cross-section case, suggesting parametric regression estimates of the conditional variances and emphasizing that the asymptotics of the estimator of θ do not require that the errors in approximating the conditional variances be negligible, and Wefelmeyer (1996) presents methods for the dynamic case.

It is illuminating to compare to the cross-section case, where the data are the i.i.d. vectors X_i, with some coordinates, say, z_i, the regressors or instruments used to form the conditional moment conditions, viz. $E_\theta(h_i(\theta) \,|\, z_i) = 0$. Newey (1990) considered this situation under the further condition of conditional homoskedasticity, $\Phi_i(\theta) = \Phi(\theta)$, and suggested that if $d_i(\theta) = d(\theta \,|\, z_i) = E_\theta(\partial h_i(\theta)/\partial\theta' \,|\, z_i)$ were unknown, then a series estimation may be used, approximating $d(\theta \,|\, \cdot)$ by a power (or trigonometric) series in z_i. With the order of the approximation increasing at a suitable rate, the resulting estimator reaches the efficiency bound, in this case $(E_\theta(d(\theta \,|\, z)'\Phi(\theta)^{-1}d(\theta \,|\, z)))^{-1}$. Note that in the efficiency bound $\mathcal{J}(\theta)^{-1}$ from (2.19) in our dynamic heteroskedastic case, the conditioning is on $\mathcal{F}_{t-1}$ rather than z_i, and occurs in both $d_t(\theta)$ and $\Phi_t(\theta)$. Robinson (1991) also considered the conditional homoskedasticity case $\Phi_i(\theta) = \Phi(\theta)$ and in the case of $d(\theta \,|\, z_i)$ unknown suggested using the sample average of $\partial h_i(\theta)/\partial\theta'$ across observations with common z_i, here assuming discrete regressors. Chamberlain (1987) suggested instead to expand the set of moment conditions, replacing $h_i(\theta)$ by $B(z_i)h_i(\theta)$ and carrying out optimal GMM based on the unconditional conditions $E_\theta(B(z_i)h_i(\theta)) = 0$. For $B(z)$ of type $(1, z, z^2, \ldots, z^q)$ and q increasing with sample size, he was able to construct sequences of estimators of θ with asymptotic variances converging to the lower bound. This result for the i.i.d. case is strengthened by Theorem 2.1, where we provide the efficiency bound for the dynamic heteroskedastic case, as well as an estimator reaching the bound in general, using only a fixed number (K) of estimating equations.

The difference between the asymptotic precisions (inverse variances) of the optimal estimator $\hat{\theta}_T$ and the optimal GMM estimator $\tilde{\theta}_T$ can be interpreted and further studied by means of the following lemma.

Lemma 2.2 *The minimum mean square error predictor of $h_t(\theta)$ given $w_t^*(\theta)h_t(\theta)$ is*

$$\hat{h}_t(\theta) = D(\theta)\mathcal{J}(\theta)^{-1}w_t^*(\theta)h_t(\theta) \tag{2.25}$$

and the prediction error covariance matrix is

$$\mathrm{Var}_\theta\left(h_t(\theta) - \hat{h}_t(\theta)\right) = V(\theta) - D(\theta)\mathcal{J}(\theta)^{-1}D(\theta)'. \tag{2.26}$$

The minimum mean square error predictor of $w_t^*(\theta)h_t(\theta)$ *given* $h_t(\theta)$ *is*

$$\widehat{w_t^* h}_t(\theta) = D(\theta)'V(\theta)^{-1}h_t(\theta) \tag{2.27}$$

and the prediction error covariance matrix is

$$\mathrm{Var}_\theta\left(w_t^*(\theta)h_t(\theta) - \widehat{w_t^* h}_t(\theta)\right) = \mathcal{J}(\theta) - D(\theta)'V(\theta)^{-1}D(\theta). \tag{2.28}$$

Note that (2.28) is simply the asymptotic precision of the optimal estimator $\hat{\theta}_T$ from (2.19) less the asymptotic precision of the optimal GMM estimator $\tilde{\theta}_T$ from (2.8). By Lemma 2.2, this difference in precision is itself a covariance matrix (in particular, of a certain prediction error), and hence positive semi-definite. This is consistent with the implication of Theorem 2.1, that $\hat{\theta}_T$ is asymptotically at least as efficient as the optimal GMM estimator $\tilde{\theta}_T$. Indeed, we have the following strengthening of this conclusion.

Theorem 2.3 *The optimal estimator* $\hat{\theta}_T$ *is strictly more efficient than the optimal GMM estimator* $\tilde{\theta}_T$*, i.e.,*

$$\mathcal{J}(\theta)^{-1} < (D(\theta)'V(\theta)^{-1}D(\theta))^{-1}, \tag{2.29}$$

except for the special case where the two estimators are identical.

In (2.29), the strict inequality indicates that the right-hand side minus the left-hand side is a strictly positive semi-definite matrix different from the zero matrix. It is illuminating to recast the inequality in the more picturesque form

$$E_\theta\left(d_t(\theta)'\Phi_t(\theta)^{-1}d_t(\theta)\right) > E_\theta\left(d_t(\theta)\right)' E_\theta\left(\Phi_t(\theta)\right)^{-1} E_\theta\left(d_t(\theta)\right), \tag{2.30}$$

using (2.20) and (2.23).

If, by divine inspiration, one happens to start with $\tilde{h}_t(\theta) = w_t^*(\theta)h_t(\theta)$ in the definition (2.4) of the estimating function $H_T(\theta)$, then there is obviously no scope for improvement, and the optimal weight matrix corresponding to $\tilde{h}_t(\theta)$ is the identity matrix. In this case, the estimators $\hat{\theta}_T$ and $\tilde{\theta}_T$ are identical. The point of the present paper, however, is that divine inspiration is not necessary, since a formula is given for the optimal instruments (the weight matrices $w_t^*(\theta)$ from (2.18)) corresponding to any M-dimensional $h_t(\theta)$, which may therefore as well be taken to be the original $m_t(\theta)$. Indeed, $\hat{\theta}_T$ and $\tilde{\theta}_T$ are quite different in many natural cases, as illustrated in the following example.

Example 2.4 Consider the model for the short rate of interest proposed by Cox, Ingersoll & Ross (1985), that is, the solution of the stochastic differential equation

$$dX_t = -\beta(X_t - \alpha)dt + \sigma\sqrt{X_t}dW_t, \tag{2.31}$$

where $\alpha, \sigma > 0$. We assume that the process is stationary and ergodic ($\beta > 0$ and $2\alpha\beta > \sigma^2$). The volatility parameter σ is easy to estimate precisely in the presence of high-frequency financial data, e.g. consistent estimation is possible with observations from a fixed time interval as long as the sampling frequency within the interval increases beyond bounds. In contrast, consistent estimation of the drift parameters requires an observation interval of increasing length, the rate of mean reversion β being the hardest parameter to estimate precisely. Specifically, suppose the data are observations at equidistant time points $X_0, X_\Delta, X_{2\Delta}, \ldots, X_{T\Delta}$. If the spacing between observations $\Delta \to 0$ while $T \to \infty$ so as to keep the length $T\Delta$ of the total observation interval fixed, then for purposes of asymptotics for volatility estimation the drift may be ignored, and by quadratic variation arguments $\hat{\sigma}^2 = \sum_i (X_{i\Delta} - X_{(i-1)\Delta})^2 / X_{(i-1)\Delta}$ is consistent for σ^2. Thus, for simplicity we now treat σ as known and take $\Delta = 1$ in the discussion of the remaining parameters α and β, where the asymptotics are for $T \to \infty$. The conditional expectation of X_{t+s} given X_t is given by

$$E_\theta(X_{t+s} \,|\, X_t) = \alpha + e^{-\beta s}(X_t - \alpha), \tag{2.32}$$

so $m_t(\alpha, \beta) = X_t - \alpha - e^{-\beta}(X_{t-1} - \alpha)$ seems a promising starting point for an estimating function for (α, β). Natural instruments are $z_{t-1} = (1, X_{t-1})'$, yielding

$$h_t(\alpha, \beta) = \begin{pmatrix} m_t(\alpha, \beta) \\ X_{t-1} m_t(\alpha, \beta) \end{pmatrix}.$$

In this case, $K = M = 2$ and $L = 1$, and the optimal GMM estimator of (α, β) is given by the estimating function

$$H_T(\alpha, \beta) = \begin{pmatrix} \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} \left[X_t - \alpha - e^{-\beta s}(X_{t-1} - \alpha) \right] \\ \\ \dfrac{1}{T} \displaystyle\sum_{t=1}^{T} X_{t-1} \left[X_t - \alpha - e^{-\beta s}(X_{t-1} - \alpha) \right] \end{pmatrix}.$$

The estimating equation $H_T(\alpha, \beta) = 0$ has an explicit solution. Writing $\bar{X} = \sum_{t=1}^{T} X_t / T$ and $\bar{X}_{-1} = \sum_{t=1}^{T} X_{t-1} / T$ for the sample average of X_t and its lagged value, respectively, we define

$$\tilde{\rho}_T = \frac{\sum_{t=1}^{T} (X_t - \bar{X})(X_{t-1} - \bar{X}_{-1})}{\sum_{t=1}^{T} (X_{t-1} - \bar{X}_{-1})^2},$$

a slightly modified version of the standard first order autocorrelation coefficient. With these definitions, the optimal GMM estimator is given by

$$\tilde{\alpha}_T = \frac{\bar{X} - \tilde{\rho}_T \bar{X}_{-1}}{1 - \tilde{\rho}_T}, \tag{2.33}$$

$$\tilde{\beta}_T = -\log(\tilde{\rho}_T). \tag{2.34}$$

A simple and asymptotically equivalent estimator may be obtained by replacing the shifted sample average $\bar{X}_{-1}$ by the ordinary average $\bar{X}$ throughout, in which case α

is estimated by $\bar{X}$ and β by $-\log(\rho_T)$, where ρ_T is the standard first order autocorrelation. On the other hand, the optimal estimating function in this model is

$$G_T^*(\alpha,\beta) = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\Psi_t(\alpha,\beta)} \left[X_t - \alpha - e^{-\beta s}(X_{t-1}-\alpha)\right] \\ \\ \frac{1}{T}\sum_{t=1}^{T} \frac{X_{t-1}}{\Psi_t(\alpha,\beta)} \left[X_t - \alpha - e^{-\beta s}(X_{t-1}-\alpha)\right] \end{pmatrix}, \tag{2.35}$$

where the conditional variances are given by

$$\Psi_t(\alpha,\beta) = \frac{\sigma^2}{\beta}\left((\tfrac{1}{2}\alpha - X_{t-1})e^{-2\beta} - (\alpha - X_{t-1})e^{-\beta} + \tfrac{1}{2}\alpha\right). \tag{2.36}$$

Since all moments of X_t are finite, (2.35) has finite variance, so the optimal estimating function is well-defined.

Because of the the non-linearity, the optimal estimating equation $G_T^*(\alpha,\beta) = 0$ must be solved numerically. However, an explicit estimator with the same asymptotic variance is obtained by substituting the optimal GMM estimator $(\tilde{\alpha}_T, \tilde{\beta}_T)$ obtained in closed form in (2.33)-(2.34) above, or simply $(\bar{X}, -\log(\rho_T))$, for (α,β) in the expression for $\Psi_t(\alpha,\beta)$ from (2.36) in the optimal estimating function (2.35). This allows analytical solution along the lines of the optimal GMM case. In particular, introduce the weights $\tilde{w}_t^{\Psi} = \tilde{\Psi}_t^{-1} / \sum_{s=1}^{T} \tilde{\Psi}_s^{-1}$, where $\tilde{\Psi}_t = \Psi_t(\tilde{\alpha}_T, \tilde{\beta}_T)$, and let $\bar{X}^{\Psi} = \sum_{t=1}^{T} \tilde{w}_t^{\Psi} X_t$ and $\bar{X}_{-1}^{\Psi} = \sum_{t=1}^{T} \tilde{w}_t^{\Psi} X_{t-1}$ be the appropriate conditional precision weighted sample averages of X_t and X_{t-1}, respectively, noting that in the latter, the weight applied to X_{t-1} is $\tilde{w}_t^{\Psi}$, not $\tilde{w}_{t-1}^{\Psi}$. Similarly, the suitably reweighted sample autocorrelation is

$$\tilde{\rho}_T^{\Psi} = \frac{\sum_{t=1}^{T} \tilde{w}_t^{\Psi}(X_t - \bar{X}^{\Psi})(X_{t-1} - \bar{X}_{-1}^{\Psi})}{\sum_{t=1}^{T} \tilde{w}_t^{\Psi}(X_{t-1} - \bar{X}_{-1}^{\Psi})^2}.$$

The optimal (up to asymptotic equivalence) estimator is then given by

$$\begin{aligned} \hat{\alpha}_T &= \frac{\bar{X}^{\Psi} - \tilde{\rho}_T^{\Psi}\bar{X}_{-1}^{\Psi}}{1 - \tilde{\rho}_T^{\Psi}}, \\ \hat{\beta}_T &= -\log\left(\tilde{\rho}_T^{\Psi}\right). \end{aligned}$$

Again, a slightly simpler and still asymptotically equivalent estimator may be obtained by substituting $\bar{X}^{\Psi}$ for $\bar{X}_{-1}^{\Psi}$ everywhere, in which case α is estimated by $\bar{X}^{\Psi}$, and β by $-\log\left(\hat{\rho}_T^{\Psi}\right)$, where $\hat{\rho}_T^{\Psi}$ is the precision weighted autocorrelation

$$\hat{\rho}_T^{\Psi} = \frac{\sum_{t=1}^{T} \tilde{w}_t^{\Psi}(X_t - \bar{X}^{\Psi})(X_{t-1} - \bar{X}^{\Psi})}{\sum_{t=1}^{T} \tilde{w}_t^{\Psi}(X_{t-1} - \bar{X}^{\Psi})^2}.$$

The idea is compelling: Optimal GMM leads to estimators given by the sample average for the long run mean parameter α and (minus the logarithm of) the first order autocorrelation coefficient for the rate of mean reversion β, i.e., the well-known estimators for a Gaussian Ornstein-Uhlenbeck process. In contrast, the optimal estimator

uses the weighted average and weighted autocorrelation coefficient, instead, reflecting time-varying volatility which is the key feature of the square-root process.

To illustrate the efficiency gain, we calculate the ratio between the asymptotic variance of the optimal GMM estimator and the optimal estimator of the rate of mean reversion β, the most problematic parameter. For realistic parameter values, we use those obtained from a time series of 1-month U.S. T-bill yields in Christensen, Poulsen & Sørensen (2001). The data are the same as those analyzed by Chan et al. (1992) and are obtained from the CRSP bond data file. The raw data are converted into continuously compounded annualized yields. The parameter values for the period October 1979 to September 1982 when yields are in per cent and the time unit is one year are $\alpha = 11$, $\beta = 2.4$ and $\sigma^2 = 3.2$. Since we have assumed that observations are at time points one time unit apart, we calculate the efficiency gain for parameter values corresponding to monthly, weekly and daily observations. For instance, the parameters values when the time unit is one month are obtained by dividing β by 12 and σ by $\sqrt{12}$, while α is unchanged. The efficiency gain is the percentage by which the asymptotic variance of the optimal GMM estimator exceeds the asymptotic variance of the optimal estimator. The results are given in Table 2.1. The first line corresponds to the parameter values from the data set. In the next two lines the volatility parameter σ has been multiplied by 2 resp. 3 to investigate the effect of increased volatility. It is noted that the benefits from using the optimal estimator are increasing in both sampling frequency and volatility. Efficiency gains of the same order of magnitude are found for the period October 1982 through December 1989 where the Fed pursued an interest rate targeting policy rather than the money supply rule of the 1979 through 1982 period; see Sanders & Unal (1988). At the actual parameter estimates, $\alpha = 7.7, \beta = 1.0, \sigma^2 = 0.7$, gains between 9 and 10 per cent were obtained for all three frequencies for the later period. Importantly, these gains relative to optimal GMM are achieved using simple, explicit estimators, whereas the likelihood function in this model involves Bessel functions, entering via the density function of the non-central χ^2-distribution.

	Sampling rate		
	monthly	weekly	daily
σ	11	12	13
2σ	40	46	52
3σ	71	90	108

Table 2.1: The percentage efficiency gain by using the optimal estimator of β.

□

2.2 A pseudo-likelihood interpretation

In order to further interpret our estimators, we introduce the pseudo-log-likelihood function

$$\log L_T(\theta) = -\frac{1}{2} G_T^*(\theta)' \mathcal{J}(\theta^I)^{-1} G_T^*(\theta), \tag{2.37}$$

where θ^I again is the initial GMM estimator with identity weighting (or some other initial consistent estimator, such as optimal GMM). By differentiation, we get the associated pseudo-score function

$$s_T(\theta) = -\partial_\theta G_T^*(\theta)' \mathcal{J}(\theta^I)^{-1} G_T^*(\theta). \tag{2.38}$$

We may now define the maximum pseudo-likelihood estimator as the solution to the equation $s_T(\theta) = 0$. With these definitions, it is possible to give a pseudo-likelihood interpretation of the optimal estimator. This complements the previous interpretation as GMM with optimal instruments. The key result is given in the following Theorem.

Theorem 2.5

1. *The maximum pseudo-likelihood estimator coincides with the optimal estimator* $\hat{\theta}_T$.

2. *The pseudo-score function is asymptotically equivalent to minus the optimal estimating function* G_T^*.

The notion that the optimality property may be related to a pseudo-likelihood property is pursued further in Section 2.4 on hypothesis testing below.

The pseudo-likelihood function is not in general a proper likelihood. Of course, it is not unique, either, and perhaps a more natural pseudo-log-likelihood function is

$$\log \tilde{L}_T(\theta) = -\frac{1}{2(T-L)} \sum_{t=L+1}^{T} h_t(\theta)' \Phi_t(\theta^I)^{-1} h_t(\theta). \tag{2.39}$$

Because of the functional form of this pseudo-log-likelihood function, $(T-L)\log \tilde{L}_T(\theta)$ could actually be a proper log-likelihood function in special cases, provided $\Phi_t(\theta^I)$ is replaced by $\Phi_t(\theta)$. The associated pseudo-score is now

$$\tilde{s}_T(\theta) = -\frac{1}{T-L} \sum_{t=L+1}^{T} \frac{\partial h_t(\theta)'}{\partial \theta} \Phi_t(\theta^I)^{-1} h_t(\theta). \tag{2.40}$$

In this case, solution of the pseudo-likelihood equation $\tilde{s}_T(\theta) = 0$ produces an estimator $\hat{\tilde{\theta}}_T$ that is consistent with respect to $\tilde{\theta}_0 = \arg\min_\theta E_{\theta_0}(h_t(\theta)'\Phi_t(\theta_0)^{-1}h_t(\theta))$ (see p. 106-7 in Amemiya (1985)), and even though $\tilde{s}_T(\theta)$ need not be a martingale, $\tilde{\theta}_0$ actually coincides with the true θ_0. To establish the latter point, define $u_t(\theta) = \Phi_t(\theta_0)^{-\frac{1}{2}} h_t(\theta)$ and note that we have

$$\begin{aligned} &E_{\theta_0}(h_t(\theta)'\Phi_t(\theta_0)^{-1}h_t(\theta) \,|\, \mathcal{F}_{t-1}^L) = E_{\theta_0}(u_t(\theta)'u_t(\theta) \,|\, \mathcal{F}_{t-1}^L) \\ &= \sum_{i=1}^{M} \mathrm{Var}_{\theta_0}(u_t(\theta)_i \,|\, \mathcal{F}_{t-1}^L) + \sum_{i=1}^{M} E_{\theta_0}(u_t(\theta)_i \,|\, \mathcal{F}_{t-1}^L)^2 = M + \sum_{i=1}^{M} E_{\theta_0}(u_t(\theta)_i \,|\, \mathcal{F}_{t-1}^L)^2, \end{aligned}$$

which is minimized for $\theta = \theta_0$ because $E_{\theta_0}(u_t(\theta_0)_i | \mathcal{F}_{t-1}^L) = 0$.

In the common class considered in the following example, $\tilde{s}_T(\theta)$ is a martingale estimating function, just like the estimating functions leading to the optimal GMM estimator and the optimal estimator.

Example 2.6 In many applications, the estimating functions $h_t(\theta)$ are given by the deviation from the conditional mean of a basic function of interest $f(X_t, \ldots, X_{t-R})$, motivated by economic theory. Thus, we have

$$h_t(\theta) = f(X_t, \ldots X_{t-L}) - E_\theta(f(X_t, \ldots, X_{t-L}) \,|\, \mathcal{F}_{t-1}^L). \tag{2.41}$$

In this class of estimating functions, we have that $d_t(\theta) = \partial h_t(\theta)/\partial\theta'$. The reason is that while in general $d_t(\theta) = E_\theta(\partial h_t(\theta)/\partial\theta' \,|\, \mathcal{F}_{t-1}^L)$ by (2.15), we have here that $\partial h_t(\theta)/\partial\theta' = -\frac{\partial}{\partial\theta'} E_\theta(f(X_t, \ldots, X_{t-L}) \,|\, \mathcal{F}_{t-1}^L)$ only depends on the information in $\mathcal{F}_{t-1}^L$. Thus, we now get conclusions in a sense stronger than those in Theorem 2.5. In particular, we have:

1. The maximum pseudo-likelihood estimator $\hat{\tilde{\theta}}_T$ is equal to the modified optimal estimator obtained from $G_T^\diamond$ in (2.24).

2. The pseudo-score function (2.40) is equal to the modified optimal estimating function $G_T^\diamond$ given by (2.24).

Thus, while the second of the two conclusions is only valid asymptotically in Theorem 2.5, both hold in finite samples in the present class. Also, since the modified optimal estimator is asymptotically equivalent to the optimal estimator in general, the maximum pseudo-likelihood estimator is asymptotically equivalent to the optimal estimator within the present class. The fact that $f(\cdot)$ may be chosen almost freely (up to mild restrictions such as finite second moments) and an unbiased estimating function is constructed by subtracting the conditional expectation, if necessary computed by simulation, makes the approach quite generally applicable.

In a similar situation, Duffie & Singleton (1993) proposed a simulated moment estimator, replacing the conditional expectation by the corresponding unconditional expectation and computing the latter as the average for a long simulated path for given parameters, under a Markov assumption on X_t. In their illustrative example, X_t was the capital stock, technology shock and taste shock in a production based dynamic asset pricing model. The proposed estimator was chosen to minimize the GMM criterion (2.6), with the expectation in (2.41) computed as $\sum_{s=1}^{\mathcal{T}(T)} f(X_s^\theta, \ldots, X_{s-L}^\theta)/\mathcal{T}(T)$, with $\{X_s^\theta\}$ the simulated process given parameter θ, and $\mathcal{T}(T)$ the length of the simulated path. If $T/\mathcal{T}(T) \to \tau$ then the asymptotic variance in (2.8) is multiplied by the factor $1 + \tau$ also known from McFadden (1989) and Pakes & Pollard (1989).

The differences between these standard simulated moment methods and our approach are that we, in case the conditional expectations in (2.41) are not known analytically, would simulate each of them, using $\sum_{s=1}^{\mathcal{T}(T)} f(X_{t,s}^\theta, X_{t-1}, \ldots, X_{t-L})/\mathcal{T}(T)$ for the t'th, with $X_{t,s}^\theta$ drawn from the conditional distribution given the observed conditioning arguments $X_{t-1}, \ldots, X_{t-L}$ (this is the Lth order Markov case), and $\mathcal{T}(T) \to \infty$, and we use the optimally weighted average (2.17) of the functions $h_t(\theta)$, rather than the unweighted average $H_T(\theta)$. The increase in efficiency of our method relative to standard simulated moments is then the same as the general efficiency improvement of the optimal estimator compared to optimal GMM.

□

2.3 Diagnostic testing

A noted virtue of GMM is that upon calculation of the estimator $\tilde{\theta}_T$, a test of model specification may be based on the omnibus statistic

$$Q = (T-L)H_T(\tilde{\theta}_T)'\hat{V}(\tilde{\theta}_T)^{-1}H_T(\tilde{\theta}_T), \tag{2.42}$$

which is asymptotically χ^2-distributed on $M-K$ degrees of freedom. Here, $H_T(\theta)$ and $\hat{V}(\theta)$ are computed using (2.4) and (2.7), respectively.

An alternative test can be made based on the optimal estimator $\hat{\theta}_T$. In the martingale estimating function case we have the result in the following theorem.

Theorem 2.7 *Consider the test statistic*

$$Q_2 = (T-L)H_T(\hat{\theta}_T)'\hat{V}_2(\hat{\theta}_T)^{-}H_T(\hat{\theta}_T), \tag{2.43}$$

where $\hat{V}_2(\theta)^-$ is a generalized inverse of

$$\hat{V}_2(\theta) = \hat{V}(\theta) - \hat{D}(\theta)\hat{\mathcal{J}}(\theta)^{-1}\hat{D}(\theta)', \tag{2.44}$$

with $\hat{D}(\theta)$ and $\hat{\mathcal{J}}(\theta)$ from (2.10) and (2.21), respectively. As $T \to \infty$, Q_2 is asymptotically χ^2-distributed with degrees of freedom f equal to the rank of the matrix

$$V_2(\theta) = V(\theta) - D(\theta)\mathcal{J}(\theta)^{-1}D(\theta)'. \tag{2.45}$$

In particular, $M-K \leq f \leq M$. If d denotes the dimension of the intersection of the subspace generated by the column vectors of $V_2(\theta)$ and that generated by the columns of $D(\theta)\mathcal{J}(\theta)^{-1}D(\theta)'$ and if the rank of $D(\theta)$ is K, then $f = M-K+d$.

Note that since (2.45) by Lemma 2.2 is the covariance matrix of the prediction error of the minimum mean square error predictor of $h_t(\theta)$ given $w_t^*(\theta)h_t(\theta)$, the degrees of freedom f equal the dimension of the support of the random variable $(I_M - D(\theta)\mathcal{J}(\theta)^{-1}w_t^*(\theta))h_t(\theta)$, which is equal to the dimension of the linear space

$$\text{span}\{(I_M - D(\theta)\mathcal{J}(\theta)^{-1}w^*(\theta, x_1, \ldots, x_L))h(\theta, x_1, \ldots, x_L) \,|\, (x_1, \ldots, x_L) \in \mathcal{S}^L\},$$

where $\mathcal{S}$ denotes the state-space of the observed process. In the special case of optimal GMM, the weight matrix w^* is constant and from (2.14) given by $D(\theta)'V(\theta)^{-1}$, while $\mathcal{J}(\theta) = D(\theta)'V(\theta)^{-1}D(\theta)$, the inverse covariance matrix from (2.8). Hence, $D\mathcal{J}^{-1}w^* = D(D'V^{-1}D)^{-1}D'V^{-1}$ is a projection matrix of rank K, and the standard result $f = M-K$ follows in this case.

In general, the degrees of freedom f of the new test (2.43) are higher or at least as high as in the GMM case, since the optimal estimator is not defined to set as many combinations as possible of the functions in the vector $H_T(\theta)$ equal to zero. A natural procedure in practical applications when d is unknown is to calculate the p-value under both $M-K$ and M degrees of freedom, which yields bounds on the true p-value, by Theorem 2.7.

2.4 Hypothesis testing

In this subsection we will briefly discuss tests based on the optimal estimating function and the optimal estimator that are analogous to the well-known tests considered by Newey & West (1987a) in case of usual GMM-estimators. We will consider a general hypothesis of the form

$$H_0 : a(\theta) = 0, \tag{2.46}$$

where a is a differentiable function from $\mathbb{R}^K$ into $\mathbb{R}^R$.

The following four test statistics (Wald, QLR, LM, MD) are all asymptotically equivalent, and asymptotically χ^2-distributed on R degrees of freedom, as can be seen by a standard proof based on Taylor expansions. The *Wald test statistic* here takes the form

$$W_T = T a(\hat{\theta}_T)' \left(A(\hat{\theta}_T) \mathcal{J}(\hat{\theta}_T)^{-1} A(\hat{\theta}_T)' \right)^{-1} a(\hat{\theta}_T), \tag{2.47}$$

where

$$A(\theta) = \partial_\theta a(\theta),$$

writing $\partial_\theta \cdot = \partial \cdot / \partial \theta'$ here and in the following.

The *quasi-likelihood-ratio* (QLR) *test statistic* is given by

$$-2 \log Q = -2T \log L_T(\hat{\theta}_T^Q), \tag{2.48}$$

where $\hat{\theta}_T^Q$ is the estimator of θ obtained by maximizing $\log L_T(\theta)$ given by (2.37) under the restriction (2.46).

The *Lagrange multiplier* (LM), *score*, or *Rao test statistic* based on $G_T^*(\theta)$ is

$$LM_T = T G_T^*(\hat{\theta}_T^R)' \mathcal{J}(\hat{\theta}_T^R)^{-1} G_T^*(\hat{\theta}_T^R), \tag{2.49}$$

where $\hat{\theta}_T^R$ is the estimator of θ obtained from the optimal estimating function (2.17) under the hypothesis (2.46). With this restriction, the model can be parametrized by $\beta \in B \subseteq \mathbb{R}^{K-R}$, i.e., there exists a function $\varphi : B \mapsto \mathbb{R}^K$, such that $\theta = \varphi(\beta)$. The optimal estimating function for the parameter β is

$$G_T^R(\beta) = \partial_\beta \varphi(\beta)' G_T^*(\varphi(\beta)).$$

If $\hat{\beta}_T^R$ denotes the estimator obtained by solving $G_T^R(\beta) = 0$, then $\hat{\theta}_T^R = \varphi(\hat{\beta}_T^R)$.

Unlike in the case of exact likelihood inference, the restricted estimator $\hat{\theta}_T^R$ used to define the LM statistic does not in general coincide with the restricted estimator $\hat{\theta}_T^Q$ in the QLR test. The latter is based not on $G_T^R(\beta) = 0$ but on $\partial_\beta \varphi(\beta)' s_T(\varphi(\beta)) = 0$, with $s_T(\cdot)$ the pseudo-score (2.38), in the sense that for $\hat{\beta}_T^Q$ solving this equation, $\hat{\theta}_T^Q = \varphi(\hat{\beta}_T^Q)$.

It is possible to slightly alter the analysis in such a way that the restricted estimators in QLR and LM do coincide. Thus, consider instead of (2.48) based on $\log L_T$ from (2.37) the alternative QLR statistic based on $\log \tilde{L}_T$ from (2.39), and instead of (2.49) based on G_T^* the alternative LM statistic based on the pseudo-score $\tilde{s}_T$ from (2.40). In this case, both restricted estimators are based on solution in B of $\partial_\beta \varphi(\beta)' \tilde{s}_T(\varphi(\beta)) = 0$, and $\hat{\theta}_T^Q = \hat{\theta}_T^R$. This holds even if $\tilde{s}_T(\theta)$ differs from the modified optimal estimating function $G_T^\diamond(\theta)$ from (2.24), i.e., outside Example 2.6, in which

case the pseudo-score need not be a martingale. The additional property of asymptotic equivalence to the other four tests is gained under the special (but common) structure (2.41) on $h_t(\cdot)$ from the Example.

The minimum distance or minimum chi-square estimator $\hat{\beta}_T^D$ of β that parametrizes the model under the restriction (2.46) is obtained by minimizing $(\hat{\theta}_T - \varphi(\beta))' \mathcal{J}(\hat{\theta}_T)(\hat{\theta}_T - \varphi(\beta))$, or equivalently from the estimating function

$$\partial_\beta \varphi(\beta)' \mathcal{J}(\hat{\theta}_T)(\hat{\theta}_T - \varphi(\beta)).$$

The minimum distance estimator of θ is $\hat{\theta}_T^D = \varphi(\hat{\beta}_T^D)$. The *minimum distance* (MD) or *minimum chi-square test statistic* is

$$MC_T = T(\hat{\theta}_T - \hat{\theta}_T^D)' \mathcal{J}(\hat{\theta}_T)(\hat{\theta}_T - \hat{\theta}_T^D). \tag{2.50}$$

Clearly, the restricted estimator $\hat{\beta}_T^D$, and consequently $\hat{\theta}_T^D$, in general differ from the restricted estimators of β, respectively θ, in either of the above forms of the QLR and LM statistics.

As noted in relation to Theorem 2.1, the optimal estimating function $G_T^*(\theta)$ forming the basis of the four test statistics is closer to the efficient score in an L^2-sense than other estimating functions given by alternative weights (or instruments), such as those from standard GMM, viz. (2.11). Hence, in this sense, the LM test (2.49) based on G_T^* is asymptotically optimal, as it is closest to likelihood inference, and by asymptotic equivalence, all of the four proposed tests enjoy the optimality property.

3 General history dependence

So far we have restricted attention to the case where $h_t(\theta)$ and $w_t(\theta)$ depend on at most L lags of X_t. We now consider the more general case where $h_t(\theta)$ is allowed to depend on all observations up to and including time t, while $w_t(\theta)$ may depend on all observations through $t-1$. The form of $h_t(\theta)$ and $w_t(\theta)$ will typically vary with t, e.g., the dependence may be of the type $h_t(\theta) = X_t - (\nu_1(\theta)X_{t-1} + \cdots + \nu_{t-1}(\theta)X_1 + \nu_t(\theta))$. Two concrete examples are given below. Condition (2.3) is maintained, so that the estimating functions are martingales, as in the previous section.

A major reason for the interest in general history dependence is that allowing unbounded lag length in $h_t(\cdot)$ and $w_t(\cdot)$ typically increases efficiency in cases where X_t is not Lth order Markov. A similar phenomenon occurs already in the simpler setting of the previous section, for suppose that X_t indeed is Lth order Markov, but $h_t(\cdot)$ only depends on a smaller number of lagged X-values, say $\tilde{L}$ lags, $\tilde{L} < L$. Then the optimal estimator is defined by (2.15), (2.16), and (2.18), with L lags in the weights, and generally this does not reduce to the smaller number $\tilde{L}$ of lags that happens to occur in $h_t(\cdot)$. If only $\tilde{L}$ lags are permitted in the weights for an Lth order Markov process, $\tilde{L} < L$, then efficiency is lost. More generally, if X_t is not Lth order Markov for any finite L, but $h_t(\cdot)$ depends on L lags, $L < \infty$, then Theorem 2.1 delivers the optimal estimator using only the same number of lags in $w_t^*(\cdot)$ as in $h_t(\cdot)$, and shifting to more lags in the former would typically improve efficiency in this case,

too. The most general situation is when neither lag order is restricted, i.e., general history dependence.

With general history dependence, the estimating function is no longer an average of stationary terms, although we will still assume that the process $\{X_t\}$ is stationary. Under wide regularity conditions the estimator obtained by minimizing (2.6) is also in this case consistent and asymptotically normal with limiting distribution of the form (2.8), where

$$V(\theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h_t(\theta) h_t(\theta)' \tag{3.51}$$

and

$$D(\theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \frac{\partial h_t(\theta)}{\partial \theta'}. \tag{3.52}$$

As before, an initial estimator θ^I may be obtained by minimizing (2.6) with $W = I_M$, whereas the final estimator uses $W = \hat{V}(\theta^I)^{-1}$, with $\hat{V}(\cdot)$ from (2.7). As in the previous section, we refer to this as the optimal GMM estimator, or $\tilde{\theta}_T$. Note that with general history dependence, summation can start at $t = 1$ in (2.4), (2.7), (3.51), (3.52), and similar sums. It is an assumption that these sums converge either almost surely or in probability to a deterministic limit, and that the observed process is stationary and ergodic. If $h_t = h(X_t, \ldots, X_{t-L}; \theta)$ (at least from a certain value of t), then the definitions (3.51) and (3.52) coincide with the definitions in the previous section, where $V(\theta)$ was the covariance matrix of $h_t(\theta)$ and $D(\theta)$ was given by (2.9).

In case of general history dependence, we consider generalized estimators obtained from estimating functions of the form (2.13) with $L = 0$, where the weights $w_t(\theta)$ are arbitrary $\mathcal{F}_{t-1}$-measurable $K \times M$ matrices.

Theorem 3.1 *The optimal estimating function in the class of all estimating functions of the form (2.13), where the weights $w_t(\theta)$ are arbitrary $\mathcal{F}_{t-1}$-measurable $K \times M$ matrices, is given by (2.18), with $d_t(\theta)$ and $\Phi_t(\theta)$ redefined as*

$$d_t(\theta) = E_\theta \left(\frac{\partial h_t(\theta)}{\partial \theta'} \,\middle|\, \mathcal{F}_{t-1} \right) \tag{3.53}$$

and

$$\Phi_t(\theta) = \mathrm{Var}_\theta \left(h_t(\theta) \,|\, \mathcal{F}_{t-1} \right) = E_\theta \left(h_t(\theta) h_t(\theta)' \,|\, \mathcal{F}_{t-1} \right). \tag{3.54}$$

The asymptotic distribution of the optimal estimator $\hat{\theta}_T$ obtained by solving the estimating equation $G_T^*(\theta) = 0$ is given by (2.19), with $\mathcal{J}(\theta)$ redefined as

$$\mathcal{J}(\theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} d_t(\theta)' \Phi_t(\theta)^{-1} d_t(\theta), \tag{3.55}$$

provided that the sum converges almost surely or in probability to a deterministic limit.

If $\{X_t\}$ is Lth order Markov or L-dependent, then the functions $d_t(\theta)$ and $\Phi_t(\theta)$, and hence the optimal weights, are of the form considered previously in the stationary

case (when $t > L$). Note also that with general history dependenc, the efficiency bound $\mathcal{J}(\theta)$ is given by a probability limit.

For comparison of the asymptotic variances of GMM and the optimal estimator, Lemma 2.2 on prediction of the terms h_t and $w_t^*(\theta)h_t(\theta)$ is in the case of general history dependence replaced by the following lemma on prediction of the summation of these terms.

Lemma 3.2 *The minimum mean square error predictor of $G_T^*(\theta)$ given $H_T(\theta)$ is*

$$\widehat{G}_T^*(\theta) = D_T(\theta)'V_T(\theta)^{-1}H_T(\theta) \tag{3.56}$$

and the prediction error covariance matrix is

$$\mathrm{Var}_\theta\left(G_T^*(\theta) - \widehat{G}_T^*(\theta)\right) = \mathcal{J}_T(\theta) - D_T(\theta)'V_T(\theta)^{-1}D_T(\theta), \tag{3.57}$$

where

$$D_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} E_\theta\left(\frac{\partial h_t(\theta)}{\partial \theta'}\right),$$

$$V_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} E_\theta\left(h_t(\theta)h_t(\theta)'\right),$$

and

$$\mathcal{J}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} E_\theta\left(d_t(\theta)'\Phi_t(\theta)^{-1}d_t(\theta)\right).$$

The minimum mean square error predictor of $H_T(\theta)$ given $G_T^*(\theta)$ is

$$\widehat{H}_T(\theta) = D_T(\theta)\mathcal{J}_T(\theta)^{-1}G_T^*(\theta)$$

with prediction error covariance matrix

$$\mathrm{Var}_\theta\left(H_T(\theta) - \widehat{H}_T(\theta)\right) = V_T(\theta) - D_T(\theta)\mathcal{J}_T(\theta)^{-1}D_T(\theta)'.$$

From (3.51) and (3.55), the difference in asymptotic precision between the optimal estimator and optimal GMM is the limit as $T \to \infty$ of the right hand side of (3.57), and the left hand side is seen to be the covariance matrix of the errors in predicting the optimal estimating function based on the moment conditions, i.e., in particular positive semi-definite. The following theorem strengthens this conclusion.

Theorem 3.3 *Assume that (3.51), (3.52), and (3.55) converge in L^1 to a deterministic limit. Then the optimal estimator $\hat{\theta}_T$ is strictly more efficient than the optimal GMM estimator $\tilde{\theta}_T$, i.e.,*

$$\mathcal{J}(\theta)^{-1} < (D(\theta)'V(\theta)^{-1}D(\theta))^{-1}, \tag{3.58}$$

except for the special case where the two estimating functions are asymptotically identical.

Thus, Theorem 3.3 generalizes Theorem 2.3 to the case of general history dependence.

Example 3.4 Consider the GARCH(1,1)-model of Bollerslev (1986), defined iteratively by

$$X_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2, \tag{3.59}$$

where the innovation sequence $\{\epsilon_t\}$ is i.i.d. with variance one. Identification and positive conditional variances, σ_t^2, require $\alpha > 0$, $\beta \geq 0$. Since this is a model of volatility, it is natural to specify h_t from the squared observations as

$$h_t(\theta) = X_t^2 - E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1}), \tag{3.60}$$

which in the GARCH(1,1) case is given by

$$h_t(\omega, \alpha, \beta) = X_t^2 - (\omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2). \tag{3.61}$$

Since σ_t^2 is defined recursively, h_t is of the form $h_t(\theta) = X_t^2 - (\nu_1(\theta) X_{t-1}^2 + \cdots + \nu_{t-1}(\theta) X_1^2 + \nu_t(\theta))$, where $\theta = (\omega, \alpha, \beta)$. Iterating on (3.59) shows immediately that $\nu_1(\theta) = \alpha$ and $\nu_k(\theta) = \alpha\beta^{k-1}$, for $k = 1, 2, \ldots, t-1$. Also, $\nu_t(\theta) = \omega(1-\beta^{t-1})/(1-\beta) + \beta^{t-1}\sigma_1^2$. The recursions are typically started at σ_1^2 given by the sample variance of X_t or the unconditional variance $\mathrm{Var}_\theta(X_1)$. Bollerslev (1986) shows that this is finite and given by $\omega/(1-\alpha-\beta)$ under the weak stationarity condition $\alpha + \beta < 1$. We need in addition that $h_t(\theta)$ has finite variance, which requires finite fourth moments of X_t. A necessary and sufficient condition for the latter is $\beta^2 + 2\alpha\beta + \alpha^2 E_\theta(\epsilon_t^4) < 1$. This was shown by Bollerslev (1986) for normal innovations ($E_\theta(\epsilon_t^4) = 3$) and by He & Teräsvirta (1999) in general.

The GARCH-model is a natural example where $h_t(\theta)$ is of unbounded lag length, so the methods of the previous section in general would not suffice. The exception is the special case $\beta = 0$, i.e., the ARCH(1)-model of Engle (1982), where $L = 1$.

As a second example, consider the GARCH-M (GARCH-in-mean) model of Engle, Lilien & Robins (1987). This is relevant in many financial applications where X_t is an asset return whose mean depends on conditional volatility. A typical specification is

$$X_t = \gamma_0 + \gamma_1 \sigma_t^2 + \sigma_t \epsilon_t, \qquad \sigma_t^2 = \omega + \alpha(X_{t-1} - \gamma_0 - \gamma_1 \sigma_{t-1}^2)^2 + \beta \sigma_{t-1}^2,$$

where the parameter of interest is γ_1, the market price of risk. The natural estimating function for the conditional mean return parameters would be based on

$$h_t(\gamma_0, \gamma_1) = X_t - \gamma_0 - \gamma_1 \sigma_t^2. \tag{3.62}$$

Suppose first that (γ_0, γ_1) are the only unknown parameters. In this case, $\Phi_t(\theta)$ is given by σ_t^2, which is clearly time-varying in the GARCH framework. Estimation methods ignoring this, such as OLS regression of X_t on σ_t^2, or simply estimating the market price of risk off the long run risk-return tradeoff $(\bar{X} - \gamma_0)/\sigma^2$, are inefficient. The latter approach is actually the GMM method that emerges when treating $\Phi_t(\theta)$ as constant. In general, with all parameters unknown, the estimating equation should be based on both (3.60) and (3.62). In this case, the lower right corner of the matrix $\Phi_t(\theta)$ is σ_t^2, i.e., still time-varying, and of unbounded lag length. □

Example 3.5 Consider the continuous time stochastic volatility model given by

$$\begin{aligned} dY_t &= \sigma_t dW_t^1, \\ d\sigma_t^2 &= a(\sigma_t^2, \theta)dt + b(\sigma_t^2, \theta)dW_t^2, \end{aligned} \tag{3.63}$$

where W^1 and W^2 are (possibly correlated) standard Wiener processes. Suppose the process $\{Y_t\}$ has been observed at time points $0, 1, \ldots, T$ and define $X_t = Y_t - Y_{t-1}$. If Y_t is the logarithm of the price of a financial asset, then X_t is the return over the time interval $[t-1, t]$. Again, a natural choice of h_t is (3.60), where the conditional expectation in general depends on all past observations $X_1, \ldots X_{t-1}$, so again h_t is of unbounded lag length. In the GARCH model, this was computed recursively, thus already giving rise to general history dependence in that case. In the stochastic volatility case, simulation is typically needed. Generally,

$$E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1}) = E_\theta\left(\int_{t-1}^t \sigma_s^2 ds \,|\, \mathcal{F}_{t-1}\right) = \int_{t-1}^t E_\theta(\sigma_s^2 \,|\, \mathcal{F}_{t-1}) ds,$$

where, as usual, $\mathcal{F}_{t-1}$ denotes the σ-field generated by $X_1, \ldots, X_{t-1}$. For mean reverting volatility with affine drift,

$$a(x, \theta) = -\beta(x - \alpha),$$

the conditional mean of σ_s^2 given σ_{t-1}^2 takes the same form as in (2.32), so

$$E_\theta(\sigma_s^2 \,|\, \mathcal{F}_{t-1}) = \alpha + e^{-\beta(s-t+1)}\left(E_\theta(\sigma_{t-1}^2 \,|\, \mathcal{F}_{t-1}) - \alpha\right).$$

It follows that in this case,

$$E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1}) = \alpha + \left(E_\theta(\sigma_{t-1}^2 \,|\, \mathcal{F}_{t-1}) - \alpha\right)\frac{1 - e^{-\beta}}{\beta}. \tag{3.64}$$

Thus, the stochastic volatility model is amenable to rather explicit treatment. Indeed, higher moments may be included to increase information on volatility of volatility, e.g., $h_t(\theta)$ may be expanded by $X_t^4 - E_\theta(X_t^4 \,|\, \mathcal{F}_{t-1})$, where

$$E_\theta(X_t^4 \,|\, \mathcal{F}_{t-1}) = 3E_\theta\left(\left(\int_{t-1}^t \sigma_s^2 ds\right)^2 \,\middle|\, \mathcal{F}_{t-1}\right) = 3\int_{t-1}^t \int_{t-1}^t E_\theta(\sigma_s^2 \sigma_u^2 \,|\, \mathcal{F}_{t-1}) ds du,$$

provided that W^1 and W^2 are independent. It follows from the Markov property of the volatility process σ_t^2 that

$$E_\theta(\sigma_s^2 \sigma_u^2 \,|\, \mathcal{F}_{t-1}) = E_\theta(\phi(s - t + 1, u - t + 1, \sigma_{t-1}^2) \,|\, \mathcal{F}_{t-1}) \tag{3.65}$$

for $s, u > t - 1$, where $\phi(t_1, t_2, z) = E_\theta(\sigma_{t_1}^2 \sigma_{t_2}^2 \,|\, \sigma_0^2 = z)$. Even though the function ϕ can be found explicitly for some models, there is no closed-form expression for (3.65). For the Heston (1993) model, where σ_t^2 solves the stochastic differential equation

$$d\sigma_t^2 = -\beta(\sigma_t^2 - \alpha)dt + \tau\sigma_t dW_t^2, \tag{3.66}$$

again with affine drift, the function ϕ is given by

$$\begin{aligned}\phi(t_1,t_2,z) &= z^2e^{-\beta(t_1+t_2)} + z\left(1-e^{-\beta t_1}\right)\left(\frac{\alpha\tau^2}{\beta}e^{-\beta t_1} - Ae^{-\beta t_2}\right) \\ &\quad + \alpha Ae^{-\beta t_2}\left(1-\cosh(\beta t_1)\right) - \alpha e^{-\beta t_1}\left(1-\tfrac{1}{2}e^{-\beta t_1}\right) - \tfrac{1}{2}\alpha\end{aligned}$$

for $t_2 \geq t_1 \geq 0$, where $A = \tau^2(\alpha-1)/\beta - 2\alpha$. To calculate (3.65), we thus need $E_\theta(\sigma_{t-1}^2 \,|\, \mathcal{F}_{t-1})$ as in (3.64), and in addition $E_\theta(\sigma_{t-1}^4 \,|\, \mathcal{F}_{t-1})$. These conditional expectations given $\mathcal{F}_{t-1} = \sigma(X_1,\ldots,X_{t-1})$ must be calculated numerically. Typically, a nested extended Kalman filter algorithm would be applied, running the filter forward once at each trial parameter value when solving the estimating equation. If σ_t^2 is stationary and ergodic with all moments finite ($\beta > 0$, $2\alpha\beta > \tau^2$ in the Heston model), then so is the observed process X_t, and the estimating functions we consider are of finite variance.

An approximation to $h_t(\theta)$ can be obtained by the following method which applies also outside the Heston model, but maintaining independence of the two driving Wiener processes. If the volatility process σ_t^2 is exponentially mixing, then so is the sequence of returns $\{X_t\}$, see Sørensen (2000). Concerning the concept of mixing, see e.g. Doukhan (1994). In this quite common case, the conditional expectation $E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1})$ in (3.60) depends only weakly on observations made much earlier than time t, so that it makes sense to approximate it by $E_\theta(X_t^2 \,|\, X_{t-1},\ldots,X_{t-L})$ for a suitable L. This conditional expectation can be calculated by simulation along the same lines as in H. Sørensen (2003). Specifically, define

$$S_t = \int_{t-1}^t \sigma_s^2 ds$$

for $t = 1, 2, \ldots$. Then conditionally on $S_1,\ldots,S_t$ the random variables $X_t,\ldots,X_{t-L}$ are independent and X_{t-j} is normally distributed with mean zero and variance S_{t-j}. Let $\varphi(x,\xi^2)$ denote the normal density function with mean zero and variance ξ^2. Then the unconditional density of $(X_t,\ldots,X_{t-L})$ is

$$p^{(L+1)}(x_t,\ldots,x_{t-L}) = E_\theta\left(\prod_{j=0}^L \varphi(x_{t-j}, S_{t-j})\right),$$

where the expectation is with respect to $S_t,\ldots,S_{t-L}$. It follows that

$$\begin{aligned}&E_\theta(X_t^2 \,|\, X_{t-1}=x_{t-1},\ldots,X_{t-L}=x_{t-L}) \\ &= \frac{\int_{-\infty}^\infty x_t^2 p^{(L+1)}(x_t,\ldots,x_{t-L})dx_t}{p^{(L)}(x_{t-1},\ldots,x_{t-L})} = \frac{E_\theta\left(\int_{-\infty}^\infty x_t^2\varphi(x_t,S_t)dx_t \prod_{j=1}^L \varphi(x_{t-j},S_{t-j})\right)}{p^{(L)}(x_{t-1},\ldots,x_{t-L})} \\ &= \frac{E_\theta\left(S_t\prod_{j=1}^L\varphi(x_{t-j},S_{t-j})\right)}{E_\theta\left(\prod_{j=1}^L\varphi(x_{t-j},S_{t-j})\right)},\end{aligned} \tag{3.67}$$

where the last equality uses that $\int_{-\infty}^{\infty} x_t^2 \varphi(x_t, \xi^2) dx_t = \xi^2$ by the properties of φ. If fourth moments are included, $E_\theta(X_t^4 \,|\, X_{t-1} = x_{t-1}, \ldots, X_{t-L} = x_{t-L})$ may be calculated similarly, as an approximation to $E_\theta(X_t^4 \,|\, \mathcal{F}_{t-1})$, simply substituting $3S_t^2$ for S_t (but of course not for S_{t-j} in $\varphi(\cdot)$) in the last expression in (3.67). Here, the expectations in the numerator and denumerator can be calculated by simulating $S_{t-L}, \ldots, S_t$ a large number of times and then using the law of large numbers. Note that since the volatility process is assumed stationary, then so is the sequence S_j. Therefore, the same simulations can be reused for all values of t.

If t is small, this method can be used to calculate $E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1}) = E_\theta(X_t^2 \,|\, X_{t-1}, \ldots, X_1)$, so that no approximation is needed (let $L = t-1$), but if t is large this would be computationally very demanding. Unfortunately, the estimating function (2.13) with $h_t(\theta)$ given by (3.60) is not a martingale when $E_\theta(X_t^2 \,|\, \mathcal{F}_{t-1})$ is replaced by the approximation $E_\theta(X_t^2 \,|\, X_{t-1}, \ldots, X_{t-L})$. Non-martingale estimating functions are treated in the next section, where we also return to the stochastic volatility example.

□

4 Non-martingale estimating functions

For a class of non-martingale estimating functions it is in general much more difficult to find an optimal estimating function, except if the class of estimating functions is finite dimensional. Suppose the observed process is non-Markovian and stationary, and consider the class of unbiased estimating functions of the form

$$G_T(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} w(X_{t-1}, \ldots, X_{t-L}; \theta) \left[f(X_t) - \Pi(X_{t-1}, \ldots, X_{t-L}; \theta) \right], \quad (4.1)$$

where f and Π are given M-dimensional functions and w is a $K \times M$ matrix of weights that should be chosen optimally. As earlier, K denotes the dimension of the parameter $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^K$ is the parameter set. To ensure consistency of the resulting estimator, the functions f, Π and w should satisfy that $G_T(\theta)$ is unbiased, i.e. $E_\theta(w(X_{t-1}, \ldots, X_{t-L}; \theta) \left[f(X_t) - \Pi(X_{t-1}, \ldots, X_{t-L}; \theta) \right]) = 0$. An obvious choice of Π is $\Pi(X_{t-1}, \ldots, X_{t-L}; \theta) = E_\theta(f(X_t) \,|\, X_{t-1}, \ldots, X_{t-L})$, in which case the function $w(\cdot)$ can be chosen freely (except that the expectation must exist and for the optimality theory $G_T(\theta)$ must have finite variance). Note that even with this Π, the estimating function $G_T(\theta)$ is not in general a martingale, although it is if $\{X_t\}$ is an Lth order Markov process. A more general choice of Π is a predictor of $f(X_t)$, yielding a prediction-based estimating function in the sense of Sørensen (2000). If Π is the minimum mean square error predictor in a certain class of predictors, then w can be chosen as any other predictor in the class (except that $G_T(\theta)$ must have finite variance). The conditional expectation $E_\theta(f(X_t) \,|\, X_{t-1}, \ldots, X_{t-L})$ above is a special case: It is the minimum mean square error predictor in the class of all predictors with finite variance.

In the general non-Markov, non-martingale case, we introduce a general class of

estimating functions with $(i,j)'$th entry of the weight matrix w given by

$$w_{ij}(x,\theta) = \sum_{k=1}^{\infty} a_k^{ij}(\theta)\xi_k(x), \tag{4.2}$$

where x is L-dimensional (or, precisely, of dimension L times that of X_t, see (4.1)). The real functions ξ_k are given, but must satisfy that $E_\theta(\xi_k(X_{t-1}, \ldots, X_{t-L})[f(X_t) - \Pi(X_{t-1}, \ldots, X_{t-L}; \theta)]) = 0$, while we are free to choose the $K \times M$ matrices $a_k(\theta) = \{a_k^{ij}(\theta)\}$ in an optimal way. An example (for $L = 1$) is when w_{ij} is assumed to belong to the class of all square integrable functions (of x), which is spanned by, for instance, the Hermite functions. In order to find the optimal weight function, i.e. the optimal choice of the matrices $a_k(\theta)$, we need the $M \times K$ matrices

$$s_k(\theta) = E_\theta\left(\xi_k(X_L, \ldots, X_1)\frac{\partial}{\partial\theta'}\Pi(X_L, \ldots, X_1; \theta)\right), \tag{4.3}$$

$k \in \mathbb{N}$, and the $M \times M$ matrices

$$m_{kl}^T(\theta) = E_\theta\left(H_{T,k}(\theta)H_{T,l}(\theta)'\right), \tag{4.4}$$

$k, l \in \mathbb{N}$, where

$$H_{T,k}(\theta) = \frac{1}{\sqrt{T-L}}\sum_{t=L+1}^{T} \xi_k(X_{t-1}, \ldots, X_{t-L})\left[f(X_t) - \Pi(X_{t-1}, \ldots, X_{t-L}; \theta)\right]. \tag{4.5}$$

Define for every $\theta \in \Theta$ an operator $M_T(\theta)$ that maps a sequence of $M \times K$ matrices $b = \{b_k \,|\, k \in \mathbb{N}\}$ into another sequence of $M \times K$-matrices $M_T(\theta)b = \{(M_T(\theta)b)_k \,|\, k \in \mathbb{N}\}$ given by

$$(M_T(\theta)b)_k = \sum_{l=1}^{\infty} m_{kl}^T(\theta)b_l. \tag{4.6}$$

The sum is defined by summation within each entry. The domain of the operator $M_T(\theta)$ is the set of sequences b for which the sum (4.6) converges. Under weak moment conditions on the functions ξ_k, the domain consists of the sequences a for which (4.2) converges. Note that M_T can be thought of as an operator on the space of weight matrices w. Specifically, a weight matrix given by (4.2) is mapped into the weight matrix $\sum_{k=1}^{\infty} b_k\xi_k(x)$, where $b_k = \sum_{\ell=1}^{\infty} m_{k\ell}^T(\theta)a_\ell$.

The following theorem provides the lower bound $\mathcal{J}(\theta)^{-1}$ on the asymptotic variance of estimators in the case of non-martingale estimating functions of the form (4.1), as well as a general procedure for constructing an estimator reaching the bound.

Theorem 4.1 *Any estimating function with weights $w^*(x,\theta)$ given by (4.2) with $a_k^{ij}(\theta)$ given by a sequence $a^*(T,\theta) = \{a_k^*(T,\theta) \,|\, k \in \mathbb{N}\}$ of $K \times M$-matrices such that*

$$M_T(\theta)a^*(T,\theta)' = s(\theta)k(\theta), \tag{4.7}$$

with $s(\theta) = \{s_l(\theta) \,|\, l \in \mathbb{N}\}$ from (4.3), $s(\theta)k(\theta) = \{s_l(\theta)k(\theta) \,|\, l \in \mathbb{N}\}$, and $k(\theta)$ some invertible $K \times K$ matrix, is optimal in the class of all estimating functions of the

form (4.1) with weights $w(x,\theta)$ given by (4.2). Suppose the operator $M_T(\theta)$ has an inverse $M_T(\theta)^{-1}$ and that the sequence $s(\theta)$ belongs to the domain of $M_T(\theta)^{-1}$. Then an optimal estimating function $G_T^*(\theta)$ is defined by substituting

$$a^*(T,\theta) = (M_T(\theta)^{-1}s(\theta))'$$

in (4.2) and the resulting weights in (4.1). Suppose that the observed process X_t is sufficiently mixing that a central limit theorem holds for $\sqrt{T}G_T^*(\theta)$ and that $a_k^*(T,\theta) \to a_k^*(\theta)$ and

$$\sum_{k=1}^{\infty} a_k^*(T,\theta)s_k(\theta) \to \mathcal{J}(\theta),$$

where

$$\mathcal{J}(\theta) = \sum_{k=1}^{\infty} a_k^*(\theta)s_k(\theta). \tag{4.8}$$

Then (under regularity conditions) the optimal estimator $\hat{\theta}_T$ solving $G_T^*(\theta) = 0$ satisfies

$$\sqrt{T}(\hat{\theta}_T - \theta) \xrightarrow{\mathcal{D}} N(0, \mathcal{J}(\theta)^{-1}).$$

The theorem is proved in the appendix. The weight matrix $\frac{\partial}{\partial\theta}\Pi(X_{t-1},\ldots,X_{t-L};\theta)'$ $\Phi(X_{t-1},\ldots,X_{t-L})^{-1}$ where $\Phi(x_L,\ldots,x_1) = \mathrm{Var}_\theta\left(f(X_{L+1}) \,|\, X_L = x_l,\ldots,X_1 = x_1\right)$ is usually not optimal, as it does not satisfy (5.1). Sufficient conditions that a central limit theorem holds for $G_T^*(\theta)$ are that X_t is geometrically strong mixing and that the $(2+\varepsilon)'$th moment of G_T^* is finite. Strong mixing is also referred to as α-mixing. The α mixing coefficient indexed by t is a measure of the dependence of observations that are made at time points that are at least t time units apart. The value zero indicates no dependence. If the α mixing coefficients tend to zero, the process is called strongly mixing, and if they go exponentially fast to zero the process is called geometrically mixing. Geometrical strong mixing together with a moment condition implies a central limit theorem. For more details on the concept of mixing and on central limit theorems for mixing processes, see Doukhan (1994). The expression (4.8) for the asymptotic variance $\mathcal{J}(\theta)$ of $G_T^*(\theta)$ follows from the central limit theorem for strongly mixing processes.

The problem in the non-martingale case is that the matrices (4.4), and hence the operator $M_T(\theta)$, are very complicated when the estimating function is not an average of uncorrelated terms. In the martingale case ($\{X_t\}$ Lth order Markov and $\Pi(X_{t-1},\ldots,X_{t-L};\theta) = E_\theta(f(X_t) \,|\, X_{t-1},\ldots,X_{t-L})$) we have

$$m_{kl}^T(\theta) = E_\theta(\xi_k(X_L,\ldots,X_1)\xi_l(X_L,\ldots,X_1)\Phi(X_L,\ldots,X_1)),$$

and it is not difficult to see that the general optimality condition $M_T(\theta)a^*(\theta)' = s(\theta)$ from Theorem 4.1 is satisfied if and only if

$$\begin{aligned} &E_\theta(\xi_k(X_L,\ldots,X_1)\Phi(X_L,\ldots,X_1)w^*(X_L,\ldots,X_1)') \\ &\quad = E_\theta\left(\xi_k(X_L,\ldots,X_1)\frac{\partial}{\partial\theta'}E_\theta(f(X_{L+1}) \,|\, X_L,\ldots,X_1)\right), \end{aligned}$$

for all $k \in \mathbb{N}$. This condition holds if

$$w^*(x_L, \ldots, x_1) = \frac{\partial}{\partial\theta} E_\theta(f(X_{L+1})' \mid X_L = x_L, \ldots, X_1 = x_1)\Phi(x_L, \ldots, x_1)^{-1},$$

which is exactly (2.18) in the Markov case. Thus, Theorem 4.1 represents a generalization, and the martingale special case comes out as an example where the operator $M_T(\theta)$ is invertible.

In the general non-Markov, non-martingale case, we still have $m_{kl}^T(\theta) \to m_{kl}(\theta)$ as $T \to \infty$, provided $\{X_t\}$ is sufficiently mixing, e.g., geometrically strong mixing. An estimator with the same efficiency as the optimal estimator in the general case from Theorem 4.1 can be obtained from the limiting operator $M(\theta)$ by substituting the $m_{kl}(\theta)$s for the $m_{kl}^T(\theta)$s in $M_T(\theta)$ in the Theorem.

In practice, for computational purposes, the infinite sum (4.2) is truncated at some finite number of terms, say N. In this case, since in effect we are restricting the weight matrix to belong to a finite dimensional space, we can find the optimal weight matrix explicitly. We can think of this as a tractable approximation to the general weight functions (4.2) belonging to the infinite dimensional space spanned by $\{\xi_k(\cdot)\}$. Specifically, consider the class of estimating functions of the form (4.1) with the ijth entry of the weight matrix given by

$$w_{ij}(x, \theta) = \sum_{k=1}^{N} a_k^{ij}(\theta)\xi_k(x). \tag{4.9}$$

Define

$$H_T(\theta) = \begin{pmatrix} H_{T,1}(\theta) \\ \vdots \\ H_{T,N}(\theta) \end{pmatrix}, \quad s(\theta) = \begin{pmatrix} s_1(\theta) \\ \vdots \\ s_N(\theta) \end{pmatrix} \quad \text{and} \quad a(\theta) = \begin{pmatrix} a_1(\theta) & \cdots & a_N(\theta) \end{pmatrix},$$

where $H_{T,k}(\theta)$ is given by (4.5) and $s_k(\theta)$ by (4.3), while $a_k(\theta) = \{a_k^{ij}(\theta)\}$. Note that $H_T(\theta)$ is an NM-dimensional random vector, $s(\theta)$ is an $NM \times K$-matrix, and $a(\theta)$ a $K \times NM$-matrix. Let $V_T(\theta)$ be the covariance matrix of the stochastic vector $H_T(\theta)$ and let $f_1, \ldots, f_M$ denote the coordinate functions of f. The following corollary is proved in the appendix.

Corollary 4.2 *Suppose that $\xi_1, \ldots, \xi_N$ are linearly independent, and that $1, f_1, \ldots, f_M$ are linearly independent on the support of the conditional distribution of X_T given $X_1, \ldots, X_{T-1}$. Then the matrix $V_T(\theta)$ is invertible, and in the class of estimating functions of the form (4.1) with weights (4.9), that with weight function given by*

$$a^*(\theta) = s(\theta)' V_T(\theta)^{-1} \tag{4.10}$$

is optimal.

A similar theory covering the more general estimating functions

$$G_T(\theta) = \tag{4.11}$$
$$\frac{1}{T-L} \sum_{t=L+1}^{T} w(X_{t-1}, \ldots, X_{t-L}; \theta) \left[f(X_t, \ldots, X_{t-L}) - \Pi(X_{t-1}, \ldots, X_{t-L}; \theta) \right]$$

can be derived along the same lines as indicated here for the slightly simpler case (4.1). When Π is the minimum mean square predictor in a finite-dimensional class of predictors, and when the entries of the weight matrices are taken to belong to the same class, the theory of optimal prediction-based estimating functions in Sørensen (2000) is recovered as a particular case.

Example 4.3 Consider again the stochastic volatility model (3.63). As discussed in Example 3.5, a simpler and more tractable estimating function than the one with conditional expectations dependent on the entire history can be obtained by conditioning only on the L previous observations. To obtain an unbiased estimating function (and hence consistent estimators) the weight matrix can be taken to depend on the same lagged observations. Thus, we obtain

$$G_T(\theta) = \frac{1}{T-L} \sum_{t=L+1}^{T} w(X_{t-1}, \ldots, X_{t-L}; \theta) \left[X_t^2 - E_\theta(X_t^2 \mid X_{t-1}, \ldots, X_{t-L}) \right].$$

This is not a martingale estimating function, but it is a prediction-based estimating function, and an unbiased estimating function of the type (4.1). The conditional expectation easily can be calculated by simulation as explained in Example 3.5. However, even if we restrict the space of weight functions to a finite dimensional space, the optimal estimating function from Theorem 4.1 cannot be easily found, because the partial derivative with respect to θ makes it computationally demanding to calculate $s_k(\theta)$ numerically from (4.3). If the volatility of volatility coefficient b from (3.63) does not depend on θ, the situation is better, and $s_k(\theta)$ can be calculated by simulating many independent copies of the volatility process using the ideas of Example 3.5 and utilizing that for differentiation of (3.67),

$$\begin{aligned} &\frac{\partial}{\partial\theta} E_\theta \left(S_t \prod_{j=1}^{L} \varphi(x_{t-j}, S_{t-j}) \right) \qquad (4.12) \\ &= E_\theta \left(S_t \prod_{j=1}^{L} \varphi(x_{t-j}, S_{t-j}) \left[\frac{\partial}{\partial\theta} \log f_\theta(\sigma^2_{t-L-1}) + \int_{t-L-1}^{t} \frac{\partial_\theta a(\sigma_s^2, \theta)}{b(\sigma_s^2)^2} d\sigma_s^2 \right. \right. \\ &\qquad \left. \left. - \int_{t-L-1}^{t} \frac{a(\sigma_s^2, \theta) \partial_\theta a(\sigma_s^2, \theta)}{b(\sigma_s^2)^2} ds \right] \right) \\ &= E_\theta \left(S_t \prod_{j=1}^{L} \varphi(x_{t-j}, S_{t-j}) \left[\frac{\partial}{\partial\theta} \log f_\theta(\sigma^2_{t-L-1}) + \int_{t-L-1}^{t} \frac{\partial_\theta a(\sigma_s^2, \theta)}{b(\sigma_s^2)} dW_s^2 \right] \right), \end{aligned}$$

where $f_\theta(x)$ denotes the stationary distribution of σ_t^2, e.g., in the Heston (1993) model from (3.66) f_θ is a gamma density. Formula (4.12) can be obtained by differentiating with respect to θ the Radon-Nikodym derivative of P_θ with respect to some fixed P_{θ_0}, see the appendix for details. When the coefficient b depends on θ, this method cannot be applied because in that case the necessary Radon-Nikodym derivative of P_θ with respect to a fixed P_{θ_0} does not exist. Note that because the volatility process is stationary, the expression in the square parenthesis need only be calculated once and can then be used for all values of t, as discussed in Example 3.5. The derivative of

$E_\theta\left(\prod_{j=1}^L \varphi(x_{t-j}, S_{t-j})\right)$ in (3.67) can be calculated similarly. The stochastic integral with respect to σ_s^2 can be calculated by means of usual integrals, since by Ito's formula

$$\int_{t-L-1}^t \frac{\partial_\theta a(\sigma_s^2,\theta)}{b(\sigma_s^2)^2} d\sigma_s^2$$
$$= A(\sigma_t^2) - A(\sigma_{t-L-1}^2) - \tfrac{1}{2}\int_{t-L-1}^t \frac{\partial^2}{\partial\sigma^2\partial\theta} a(\sigma_s^2,\theta) ds + \int_{t-L-1}^t \frac{\partial_\theta a(\sigma_s^2,\theta) b'(\sigma_s^2)}{b(\sigma_s^2)} ds,$$

where

$$A(x) = \int_1^x \frac{\partial_\theta a(y,\theta)}{b(y)^2} dy. \tag{4.13}$$

Note also that if the stationary density is not explicitly known, then $\frac{\partial}{\partial\theta}\log f_\theta(x)$ can be calculated as

$$\frac{\partial}{\partial\theta}\log f_\theta(x) = 2\left(A(x) - \int_0^\infty A(x) f_\theta(x) dx\right) = 2\left(A(x) - E_\theta(A(\sigma_1^2))\right), \tag{4.14}$$

see the appendix. Thus, the non-martingale estimating function approach does allow analysis of stochastic volatility models.

A simplified approach is to approximate the conditional expectation in $G_T(\theta)$ by the minimum mean square predictor $\Pi(X_{t-1},\ldots,X_{t-L};\theta)$ in a suitable finite-dimensional class of predictors. To ensure unbiasedness of the estimating functions, the weights w must be vectors of predictors from the same class if predictors. In this way, a much more tractable class of estimating functions is obtained, and the general theory of the present section can be easily applied. In particular, the estimating function takes the form (4.1), and the optimality theory is given in Corollary 4.2. This is a prediction-based estimating function of the type treated in detail in Sørensen (2000). For a concrete example, let

$$\Pi(X_{t-1},\ldots,X_{t-L};\theta) = \bar\nu_0(\theta) + \bar\nu_1(\theta)X_{t-1}^2 + \cdots + \bar\nu_L(\theta)X_{t-L}^2$$

be the minimum mean square predictor of X_t^2 in the class of predictors of the form $\nu_0 + \nu_1 X_{t-1}^2 + \cdots + \nu_L X_{t-L}^2$. The quantities $\bar\nu_j(\theta)$ are given by

$$\begin{pmatrix} c_1(\theta) \\ c_2(\theta) \\ \vdots \\ c_L(\theta) \end{pmatrix} = \begin{pmatrix} c_0(\theta) & c_1(\theta) & \cdots & c_{L-1}(\theta) \\ c_1(\theta) & c_0(\theta) & \cdots & c_{L-2}(\theta) \\ \vdots & \vdots & & \vdots \\ c_{L-1}(\theta) & c_{L-2}(\theta) & \cdots & c_0(\theta) \end{pmatrix} \begin{pmatrix} \bar\nu_1(\theta) \\ \bar\nu_2(\theta) \\ \vdots \\ \bar\nu_L(\theta) \end{pmatrix}$$

and

$$\bar\nu_0(\theta) = E_\theta\left(X_1^2\right)[1 - \bar\nu_1(\theta) - \cdots - \bar\nu_L(\theta)],$$

where $c_0(\theta) = \mathrm{Var}_\theta\left(X_1^2\right)$ and $c_i(\theta) = \mathrm{Cov}_\theta\left(X_1^2, X_{i+1}^2\right)$, $i = 1,\ldots,L$. The covariance matrix is invertible. For the stochastic volatility models considered here, the covariances $c_i(\theta)$ and the mean and variance of X_i^2 can be found when the mean, variance and autocorrelation function of the volatility process are known, see Sørensen (2000). When the drift of the volatility model is linear, there are tractable expressions for

these quantities, see e.g. Bibby, Skovgaard & Sørensen (2005) where a large number of examples can be found. In order to determine the optimal weights by Corollary 4.2, we need $s(\theta)$ and $V_T(\theta)$, the covariance matrix of the stochastic vector $H_T(\theta)$. Since the functions ξ_k are all of the form X_{t-j}^2, $j = 1, \ldots, L$, the vector $s(\theta)$ can be found from $c_i(\theta)$, $i = 0, \ldots, L$. To find $V_T(\theta)$ we must also calculate mixed moments of the type $E_\theta(X_i^2 X_j^2 X_k^2)$ and $E_\theta(X_i^2 X_j^2 X_k^2 X_\ell^2)$, assumed finite. These mixed moments can be determined by simulation. They can also be calculated from the mixed moments $E_\theta(\sigma_s^2 \sigma_t^2 \sigma_u^2)$ and $E_\theta(\sigma_s^2 \sigma_t^2 \sigma_u^2 \sigma_v^2)$ of the volatility process when these are available, as they are, e.g., in the Heston (1993) model (3.66). The calculation is rather complicated and is discussed in detail in Sørensen (2000), where it is also explained how to find the mixed moments for the volatility process in certain other situations. Due to the computational burden involved in calculating the optimal weights, it is usually best just to determine them once and for all at θ equal to a consistent estimator, e.g., the one obtained with weights $1, X_{t-1}^2, \ldots, X_{t-K+1}^2$.

□

5 Conclusion

Our theory provides estimators for parameters in dynamic models with conditional moment restrictions that are quite explicit and are asymptotically optimal in wide generality. The estimators may be seen as generalized method of moments estimators with optimal choice of time-varying instruments that depend on both data and parameters.

Our results complement and contribute to both the econometrics and mathematical statistics literatures. Thus, our method is strictly more efficient than the usual GMM estimator with optimal weight matrix (or norm) of Hansen (1982), which is commonly used in econometrics and obtains as the special case where our time-varying instruments are made constant by replacing them by their unconditional mean. Our approach provides the means of computing estimators that reach the lower variance bound of Hansen (1985) and Hansen, Heaton & Ogaki (1988) in general, including in dynamic heteroskedastic models. The optimal instruments resemble those known from the i.i.d. case with conditional moment restrictions obtained by Newey (1990), but the conditioning in our case is on the history of the time series. Although tied in with the GMM literature, our results on the martingale estimating function case with finite lag length follow the mathematical statistics literature reviewed in Bibby, Jacobsen & Sørensen (2004), while our generalizations to the general history dependence and non-martingale cases are novel and contribute to both the econometrics and mathematical statistics literatures.

Our work points to a number of fruitful possibilities for future research. For example, it would be interesting to relate our approach to the semiparametric statistics literature. Here, optimality results in the dynamic case are sparse, but we expect that our estimators reach the relevant semiparametric efficiency bounds for dynamic models. In addition, our methodology lends itself to empirical application, and the efficiency gains relative to existing methods obtained in practice are naturally of

interest. Finally, our theory leads to new testing procedures, and their size and power properties in finite samples are natural objects of study. Ongoing research explores both these and several related ideas.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Bibby, B. M.; Jacobsen, M. & Sørensen, M. (2004). "Estimating Functions for Discretely Sampled Diffusion-Type Models". In Aït-Sahalia, Y. & Hansen, L. P., editors, *Handbook of Financial Econometrics*. Amsterdam: North-Holland. Forthcoming.

Bibby, B. M.; Skovgaard, I. M. & Sørensen, M. (2005). "Diffusion-type models with given marginals and autocorrelation function". *Bernoulli*, 11:191–220.

Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics*, 31:307–327.

Chamberlain, G. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions". *Journal of Econometrics*, 34:305–34.

Chamberlain, G. (1992). "Efficiency bounds for semiparametric regression". *Econometrica*, 60:567–596.

Chan, K. C.; Karolyi, G. A.; Longstaff, F. A. & Sanders, A. B. (1992). "An Empirical Comparison of Alternative Models of the Short-term Interest Rate". *Journal of Finance*, 47:1209–1227.

Christensen, B. J.; Poulsen, R. & Sørensen, M. (2001). "Optimal inference for diffusion processes with applications to the short rate of interest". Working Paper No. 102, Centre for Analytical Finance, University of Aarhus.

Cox, J. C.; Ingersoll, J. E. & Ross, S. A. (1985). "A Theory of the Term Structure of Interest Rates". *Econometrica*, 53(2):385–407.

Doukhan, P. (1994). *Mixing, Properties and Examples*. Springer, New York. Lecture Notes in Statistics 85.

Duffie, D. & Singleton, K. J. (1993). "Simulated moments estimation of Markov models of asset prices". *Econometrica*, 61:929–952.

Durbin, J. (1960). "Estimation of parameters in time-series regression models". *J. Roy. Statist. Soc. Ser. B*, 22:139–153.

Engle, R. F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation". *Econometrica*, 50:987–1008.

Engle, R. F.; Lilien, D. & Robins, R. (1987). "Estimating time-varying risk premia in the term structure: the ARCH-M model". *Econometrica*, 55:391–407.

Fisher, R. A. (1935). "The logic of inductive inference". *J. Roy. Statist. Soc.*, 98:39–54.

Godambe, V. P. (1960). "An optimum property of regular maximum likelihood estimation". *Annals of Mathematical Statistics*, 31:1208–1212.

Godambe, V. P. (1985). "The foundations of finite sample estimation in stochastic processes". *Biometrika*, 72:419–428.

Godambe, V. P. & Heyde, C. C. (1987). "Quasi likelihood and optimal estimation". *International Statistical Review*, 55:231–244.

Hall, P. & Heyde, C. C. (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*, 50:1029–1054.

Hansen, L. P. (1985). "A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators". *Journal of Econometrics*, 30:203–238.

Hansen, L. P.; Heaton, J. & Yaron, A. (1996). "Finite-sample properties of some alternative GMM estimators". *Journal of Business and Economic Statistics*, 14:262–280.

Hansen, L. P.; Heaton, J. C. & Ogaki, M. (1988). "Efficiency bounds implied by multiperiod conditional restrictions". *Journal of the American Statistical Association*, 83:863–871.

He, C. & Teräsvirta, T. (1999). "Fourth moment structure of the GARCH(p,q) process". *Econometric Theory*, 15:824–846.

Heston, S. L. (1993). "A Closed-form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options". *Review of Financial Studies*, 6:327–343.

Heyde, C. C. (1988). "Fixed sample and asymptotic optimality for classes of estimating functions". *Contemporary Mathematics*, 80:241–247.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer-Verlag, New York.

Karlin, S. & Taylor, H. M. (1981). *A Second Course in Stochastic Processes*. Academic Press, Orlando.

Li, B. (1997). "On the consistency of generalized estimating equations". In Basawa, I. V.; Godambe, V. P. & Taylor, R. L., editors, *Selected Proceedings of the Symposium on Estimating Functions*, pages 115–136. Hayward: Institute of Mathematical Statistics. IMS Lecture Notes – Monograph Series, Vol. 32.

Liang, K.-Y. & Zeger, S. L. (1986). "Longitudinal data analysis using generalized linear model". *Biometrika*, 73:13–22.

McFadden, D. (1989). "A method of simulated moments for estimation of discrete response models without numerical integration". *Econometrica*, 57:995–1026.

Newey, W. K. (1990). "Efficient instrumental variables estimation of nonlinear models". *Econometrica*, 58:809–837.

Newey, W. K. (1993). "Efficient estimation of models with conditional moment restrictions". In Maddala, G. S.; Rao, C. R. & Vinod, H. D., editors, *Handbook of Statistics, Vol. 11*. Elsevier Science Publishers.

Newey, W. K. & West, K. D. (1987a). "Hypothesis testing with efficient method of moments estimation". *International Economic Review*, 28:777–787.

Newey, W. K. & West, K. D. (1987b). "A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix". *Econometrica*, 55:703–708.

Pakes, A. & Pollard, D. (1989). "The asymptotics of simulation estimators". *Econometrica*, 57:1027–1058.

Prentice, R. L. (1988). "Correlated binary regression with covariates specific to each binary observation". *Biometrics*, 44:1033–1048.

Robinson, P. M. (1991). "Best nonlinear three-stage least squares estimation of certain econometric models". *Econometrica*, 59:755–786.

Sanders, A. & Unal, H. (1988). "On the intertemporal stability of the short term rate of interest". *Journal of Financial and Quantitative Analysis*, 23:417–423.

Sørensen, H. (2003). "Simulated Likelihood Approximations for Stochastic Volatility Models". *Scand. J. Statist.*, 30:257–276.

Sørensen, M. (2000). "Prediction-based Estimating Functions". *Econometrics Journal*, 3:123–147.

Sørensen, M. (2008). "Parametric inference for discretely sampled stochastic differential equations". In Andersen, T. G.; Davis, R. A.; Kreiss, J.-P. & Mikosch, T., editors, *Handbook of Financial Time Series*. Springer. Forthcoming.

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method". *Biometrika*, 61:439–447.

Wefelmeyer, W. (1996). "Quasi-likelihood models and optimal inference". *Ann. Statist.*, 24:405–422.

Appendix

Here we give proofs of the main results of the article.

Proof of Theorem 2.1: By Theorem 2.1 in Heyde (1997), the estimating function $G_T^*(\theta)$ with weight matrix $w_t^*(\theta)$ is optimal if and only if the matrix

$$E_\theta\left(\frac{\partial}{\partial\theta'}G_T(\theta)\right)^{-1} E_\theta\left(G_T(\theta)G_T^*(\theta)'\right) \tag{5.1}$$

is the same for all estimating functions $G_T(\theta)$ of the form (2.13). Since

$$\begin{aligned} E_\theta\left(\frac{\partial}{\partial\theta'}G_T(\theta)\right) &= -\frac{1}{T-L}\sum_{t=L+1}^{T} E_\theta\left(w_t(\theta)\frac{\partial h_t(\theta)}{\partial\theta'}\right) \\ &= E_\theta\left(w_T(\theta)\frac{\partial h_T(\theta)}{\partial\theta'}\right) \\ &= E_\theta\left(w_T(\theta)E_\theta\left(\frac{\partial h_T(\theta)}{\partial\theta'}\middle|\mathcal{F}_{T-1}^L\right)\right) \\ &= E_\theta\left(w_T(\theta)d_T(\theta)\right), \end{aligned} \tag{5.2}$$

where we have used (2.3), stationarity, iterated expectations, $w_t(\theta) \in \mathcal{F}_{t-1}^L$, and (2.15) (obtaining a non-essential simplification using the stationarity), and

$$\begin{aligned} E_\theta\left(G_T(\theta)G_T^*(\theta)'\right) &= \frac{1}{(T-L)^2}\sum_{t=L+1}^{T} E_\theta\left(w_t(\theta)h_t(\theta)h_t(\theta)'w_t^*(\theta)'\right) \\ &= \frac{1}{T-L}E_\theta\left(w_T(\theta)E_\theta\left(h_T(\theta)h_T(\theta)'|\mathcal{F}_{T-1}^L\right)w_T^*(\theta)'\right) \\ &= \frac{1}{T-L}E_\theta\left(w_T(\theta)\Phi_T(\theta)w_T^*(\theta)'\right), \end{aligned} \tag{5.3}$$

using stationarity, iterated expectations, $w_t(\theta)$ and $w_t^*(\theta) \in \mathcal{F}_{t-1}^L$, and (2.16), we see that if $w_t^*(\theta)$ is given by (2.18), then (5.1) equals $I_K/(T-L)$ for all $G_T(\theta)$, so $G_T^*(\theta)$ is optimal.

The asymptotic distribution of the estimator $\hat{\theta}_T$ follows in the usual way from the mean value theorem:

$$0 = G_T^*(\hat{\theta}_T) = G_T^*(\theta) + S_T(\hat{\theta}_T - \theta). \tag{5.4}$$

Here, the (i,j)'th entry of the $K \times K$ matrix S_T is

$$\frac{\partial}{\partial\theta_j}G_T^*(\theta_T^{(j)})_i,$$

where $\theta_T^{(j)}$ is a parameter value on the straight line connecting $\hat{\theta}_T$ and θ. Since the observed process $\{X_t\}$ is ergodic,

$$\frac{\partial}{\partial\theta'}G_T^*(\theta) \xrightarrow{P_\theta} \mathcal{J}(\theta) \tag{5.5}$$

as $T \to \infty$, and by the martingale central limit theorem, see Hall & Heyde (1980),

$$\sqrt{T}G_T^*(\theta) \xrightarrow{\mathcal{D}} N\left(0, \mathcal{J}(\theta)\right) \ , \tag{5.6}$$

with $\mathcal{J}(\theta)$ from (2.20). That the asymptotic covariance matrix of $\sqrt{T}G_T^*(\theta)$ and the expectation of $\frac{\partial}{\partial\theta'}G_T^*(\theta)$ are both equal to $\mathcal{J}(\theta)$ follows from the calculations in the first part of the proof by inserting $w_t^*(\theta)$ for $w_t(\theta)$ in (5.2) and (5.3). We also need that

$$S_T \xrightarrow{P_\theta} \mathcal{J}(\theta), \tag{5.7}$$

which follows from (5.5) under regularity conditions that ensure that the convergence here is uniform in a $\sqrt{T}$-shrinking neighbourhood of θ. Now (2.19) follows by combining (5.4), (5.6), and (5.7).

□

Proof of Lemma 2.2: The covariance matrices of the stochastic vectors $h_t(\theta)$ and $w_t^*(\theta)h_t(\theta)$, respectively, are $V(\theta)$ and

$$\begin{aligned}
\mathrm{Var}_\theta(w_t^*(\theta)h_t(\theta)) &= E_\theta\left(w_t^*(\theta)h_t(\theta)h_t(\theta)'w_t^*(\theta)'\right) \\
&= E_\theta\left(w_t^*(\theta)E_\theta(h_t(\theta)h_t(\theta)'|\mathcal{F}_{t-1}^L)w_t^*(\theta)'\right) = E_\theta\left(w_t^*(\theta)\Phi_t(\theta)w_t^*(\theta)'\right) \\
&= E_\theta\left(d_t(\theta)'\Phi_t(\theta)^{-1}\Phi_t(\theta)\Phi_t(\theta)^{-1}d_t(\theta)\right) = \mathcal{J}(\theta),
\end{aligned}$$

using iterated expectations, (2.16), (2.18), and (2.20). Moreover,

$$\begin{aligned}
\mathrm{Cov}_\theta(w_t^*(\theta)h_t(\theta), h_t(\theta)) &= E_\theta\left(w_t^*(\theta)h_t(\theta)h_t(\theta)'\right) \\
&= E_\theta\left(w_t^*(\theta)E_\theta(h_t(\theta)h_t(\theta)'|\mathcal{F}_{t-1}^L)\right) = E_\theta\left(w_t^*(\theta)\Phi_t(\theta)\right) \\
&= E_\theta\left(d_t(\theta)'\Phi_t(\theta)^{-1}\Phi_t(\theta)\right) = E_\theta\left(d_t(\theta)'\right) = D(\theta)'.
\end{aligned}$$

Finally, we use that for zero mean variables x and y with finite variances, the mean square error prediction of y given x is $\mathrm{Cov}(y,x)\mathrm{Var}(x)^{-1}x$ with prediction error covariance matrix given by $\mathrm{Var}(y) - \mathrm{Cov}(y,x)\mathrm{Var}(x)^{-1}\mathrm{Cov}(x,y)$.

□

Proof of Theorem 2.3: Theorem 2.1 yields the corresponding weak inequality, in the partial ordering of positive semi-definite matrices. We must show that equality only obtains when $\hat{\theta}_T = \tilde{\theta}_T$. From Lemma 2.2, since (2.28) is a covariance matrix, equality requires that the optimal estimating function takes the form given by $w_t^*(\theta)h_t(\theta) = \widehat{w_t^*h_t}(\theta)$. By (2.27), $\widehat{w_t^*h_t}(\theta) = w_t(\theta)h_t(\theta)$, where $w_t(\theta) = D(\theta)'V(\theta)^{-1}$. From (2.14), this is exactly the choice of weights leading to the optimal GMM estimator, so $\hat{\theta}_T = \tilde{\theta}_T$.

□

Proof of Theorem 2.7: By Taylor expansion, (5.4), and (5.7), $\sqrt{T-L}H_T(\hat{\theta}_T)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{T-L}}\sum_{t=L+1}^{T}\left[I_M - D(\theta)\mathcal{J}(\theta)^{-1}w_t^*(\theta)\right]h_t(\theta). \tag{5.8}$$

From Lemma 2.2, $\left[I_M - D(\theta)\mathcal{J}(\theta)^{-1}w_t^*(\theta)\right]h_t(\theta)$ is the prediction error of the minimum mean square error predictor of $h_t(\theta)$ given $w_t^*(\theta)h_t(\theta)$, which has covariance

matrix given by (2.45). Hence, (5.8) is asymptotically normal in $\mathbb{R}^M$ with mean zero and covariance (2.45). This is a singular multivariate normal distribution concentrated on a subspace with dimension equal to the rank of (2.45). Thus, the result holds if (2.45) is used in place of $\hat{V}_2(\hat{\theta}_T)$ in Q_2, and hence under standard regularity conditions also when consistent estimators are inserted. If d is the dimension of the intersection of the subspace generated by the column vectors of V_2 and the subspace generated by the columns of $D\mathcal{J}^{-1}D'$, then by Grassman's formula,

$$f = \text{rank}(V_2) = M - \text{rank}(D\mathcal{J}^{-1}D') + d \geq M - \text{rank}(D\mathcal{J}^{-1}D') \geq M - K.$$

□

Proof of Theorem 3.1: The optimality result is proved in essentially the same way as in the proof of Theorem 2.1. The only difference is that we cannot use stationarity of $w_t(\theta)$, $d_t(\theta)$, and $\Phi_t(\theta)$ to make a simplification that is not essential to the proof. The asymptotic distribution is found in a way similar to the proof of Theorem 2.1. The necessary regularity conditions must ensure that the martingale central limit result (5.6) hold with $\mathcal{J}(\theta)$ defined by (3.55), and that the convergence (5.5) is sufficiently uniform, e.g., uniform on compact sets containing the true parameter value θ_0.

□

Proof of Theorem 3.3: Theorem 3.1 yields the corresponding weak inequality in the partial ordering of positive semi-definite matrices. Since (3.57) is a covariance matrix that converges, by the assumptions of L^1-convergence, to $\mathcal{J}(\theta) - D(\theta)'V(\theta)^{-1}D(\theta)$, equality requires, by Lemma 3.2, that $\text{Var}_\theta(G_t^*(\theta) - \widehat{G}_t^*(\theta)) \to 0$. From (2.14) and (3.56), this happens only when the optimal estimating function is asymptotically equal to the optimal GMM estimating function.

□

Proof of Theorem 4.1: According to Theorem 2.1 in Heyde (1997), the estimating function G_T^* with weight matrix w^* is optimal if and only if the matrix (5.1) is the same for all estimating functions G_T of the form (4.1). Now,

$$\begin{aligned}
&\frac{\partial}{\partial\theta_i} w(X_{t-1},\theta)\,[f(X_t) - \Pi(X_{t-1};\theta)] \\
&\quad = \quad \frac{\partial}{\partial\theta_i} w(X_{t-1},\theta)\,[f(X_t) - \Pi(X_{t-1};\theta)] - w(X_{t-1},\theta)\frac{\partial}{\partial\theta_i}\Pi(X_{t-1};\theta),
\end{aligned}$$

and since $\frac{\partial}{\partial\theta_i} w(X_{t-1},\theta) = \sum_{k=1}^{\infty} \frac{\partial}{\partial\theta_i} a_k(\theta)\xi_k(x)$ belongs to the class of weight functions (we restrict the class to those for which the interchange of differentiation and summation is valid), it follows that

$$\begin{aligned}
E_\theta\left(\frac{\partial}{\partial\theta'} G_T(\theta)\right) &= -\frac{1}{T-L}\sum_{t=L+1}^{T} E_\theta\left(w(X_{t-1},\theta)\frac{\partial}{\partial\theta'}\Pi(X_{t-1};\theta)\right) \\
&= -E_\theta\left(w(X_0,\theta)\frac{\partial}{\partial\theta'}\Pi(X_0;\theta)\right) = -\sum_{k=1}^{\infty} a_k(\theta)s_k(\theta).
\end{aligned}$$

Because of this and since

$$E_\theta\left(G_T(\theta)G_T^*(\theta)'\right) =$$
$$\frac{1}{(T-L)^2}\sum_{t,s=L+1}^{T} E_\theta\left(w(X_{t-1},\theta)\left[f(X_t)-\Pi(X_{t-1};\theta)\right]\left[f(X_s)-\Pi(X_{s-1};\theta)\right]' w^*(X_{s-1},\theta)'\right)$$
$$= \frac{1}{T-L}\sum_{k=1}^{\infty} a_k(\theta)\sum_{\ell=1}^{\infty} m_{k\ell}^T(\theta)a_\ell^*(T,\theta)' = \frac{1}{T-L}\sum_{k=1}^{\infty} a_k(\theta)(M_T(\theta)a^*(T,\theta)')_k,$$

we see that (4.7) implies (5.1).

The asymptotic normality of the optimal estimator follows in a way analogous to the proof of Theorem 2.1, using that

$$\sqrt{T}G_T^*(\theta) \xrightarrow{\mathcal{D}} N\left(0,\mathcal{J}(\theta)\right),$$

with $\mathcal{J}(\theta)$ given by (4.8), based on a suitable general central limit theorem for mixing sequences, accounting for the fact that $(T-L)G_T^*(\theta)$ is not the sum of a stationary sequence because the weights depend on T. We have used the expression for the covariance matrix of $G_T^*(\theta)$ above and (4.7) (assuming that $k(\theta)$ is the identity matrix, which can always be achieved by a linear transformation of G^*). The expression for the expectation of $\frac{\partial}{\partial\theta'}G_T^*(\theta)$ above converges to $-\mathcal{J}(\theta)$.

□

Proof of Corollary 4.2: In this case the operator $M_T(\theta)$ is given by $M_T(\theta)x = V_T(\theta)x$ (with obvious notation). Therefore the corollary follows from Theorem 4.1 if it can be established that the matrix $V_T(\theta)$ is invertible. If the covariance matrix $V_T(\theta)$ is not strictly positive definite, the random vector $H_T(\theta)$ is concentrated on a subspace of $\mathbb{R}^{NM}$, i.e. in this case there exists a non-trivial linear combination of its coordinates which is identically equal to zero. This again implies that there exists function $c_0,\ldots,c_M$ that are not all equal to zero such that

$$\sum_{j=1}^{M} c_j(X_{T-1})f_j(X_T) + c_0(X_0,\cdots,X_{T-1}),$$

which contradicts the assumption on $1, f_1,\ldots,f_M$.

□

Proof of (4.12): Let P_θ denote the probability measure that determine the process between time $t-L-1$ and time t when the parameter value is θ and let θ_0 be some fixed parameter value. Then the Radon-Nikodym derivative of P_θ with respect to P_{θ_0} is

$$L(\theta) =$$
$$\frac{f_\theta(\sigma^2_{t-L-1})}{f_{\theta_0}(\sigma^2_{t-L-1})}\exp\left(\int_{t-L-1}^{t}\frac{a(\sigma_s^2,\theta)-a(\sigma_s^2,\theta_0)}{b(\sigma_s^2)^2}d\sigma_s^2 - \tfrac{1}{2}\int_{t-L-1}^{t}\left[\frac{a(\sigma_s^2,\theta)^2}{b(\sigma_s^2)^2}-\frac{a(\sigma_s^2,\theta_0)^2}{b(\sigma_s^2)^2}\right]ds\right)$$

It is here assumed that the probability measures are equivalent. Hence

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}E_\theta\left(S_t\prod_{j=1}^{L}\varphi(x_{t-j},S_{t-j})\right)\\
&= \frac{\partial}{\partial\theta}E_{\theta_0}\left(S_t\prod_{j=1}^{L}\varphi(x_{t-j},S_{t-j})L(\theta)\right) = E_{\theta_0}\left(S_t\prod_{j=1}^{L}\varphi(x_{t-j},S_{t-j})\frac{\partial}{\partial\theta}L(\theta)\right)\\
&= E_{\theta_0}\left(S_t\prod_{j=1}^{L}\varphi(x_{t-j},S_{t-j})\left[\frac{\partial}{\partial\theta}\log f_\theta(\sigma^2_{t-L-1})+\int_{t-L-1}^{t}\frac{\partial_\theta a(\sigma_s^2,\theta)}{b(\sigma_s^2)^2}d\sigma_s^2\right.\right.\\
&\qquad\left.\left.-\int_{t-L-1}^{t}\frac{a(\sigma_s^2,\theta)\partial_\theta a(\sigma_s^2,\theta)}{b(\sigma_s^2)^2}ds\right]L(\theta)\right)\\
&= E_{\theta}\left(S_t\prod_{j=1}^{L}\varphi(x_{t-j},S_{t-j})\left[\frac{\partial}{\partial\theta}\log f_\theta(\sigma^2_{t-L-1})+\int_{t-L-1}^{t}\frac{\partial_\theta a(\sigma_s^2,\theta)}{b(\sigma_s^2)^2}d\sigma_s^2\right.\right.\\
&\qquad\left.\left.-\int_{t-L-1}^{t}\frac{a(\sigma_s^2,\theta)\partial_\theta a(\sigma_s^2,\theta)}{b(\sigma_s^2)^2}ds\right]\right).
\end{aligned}
$$

□

Proof of (4.14): It is well known that the stationary density equals the speed measure, see Karlin & Taylor (1981),

$$f_\theta(x)=\frac{\kappa(\theta)}{b(x)^2}\exp\left(2\int_1^x\frac{a(y,\theta)}{b(y)^2}dy\right),$$

where

$$\kappa(\theta)=\frac{1}{\displaystyle\int_0^\infty\frac{1}{b(x)^2}\exp\left(2\int_1^x\frac{a(y,\theta)}{b(y)^2}dy\right)dx}.$$

From this it follows that

$$\kappa'(\theta)=-\kappa(\theta)2\int_0^\infty H(x)f_\theta(x)dx,$$

with H given by (4.13), and that

$$\frac{\partial}{\partial\theta}f_\theta(x)=\kappa'(\theta)f_\theta(x)/\kappa(\theta)+2H(x)f_\theta(x),$$

from which (4.14) follows.

□