Optimal Weights in Prediction Error and Weighted Least Squares Methods

Kim Nolsøe[*], Jan Nygaard Nielsen[†], and Henrik Madsen[‡]

May 22, 2000

Abstract

In this paper an expression for the bias implied by prediction error methods (weighted least squares methods) in both i.i.d. samples and time series models with heteroscedasticity is derived provided that explicit expression for the (conditional) mean and variance are available. It is shown that prediction error methods including weighted least squares methods fit within the general theory of estimating functions, which facilitates the derivation of optimal weights in the sense of Heyde (1997) such that the properties of estimators, in particular unbiasedness, optimality and efficiency, obtained by using these methods may be discussed. Four examples are provided.

KEY WORDS: Estimating theory; Optimal estimation; Maximum likelihood; Parameter estimation; Prediction error methods; Weighted least-squares.

[*] Andersen Consulting E-mail: `kim.nolsoe@ac.com`
[†] E-mail: `jnn@imm.dtu.dk`
[‡] E-mail: `hm@imm.dtu.dk`

Contents

1 Introduction

Parameter estimation is an inherent part of system identification and a huge literature is devoted to the subject. Prediction Error Methods (PEM), in particular the special case of Weighted Least Squares (WLS), are often used and, henceforth, implemented in many automated model construction tools. It is the objective of this paper to discuss the properties of the estimators provided by these methods when the problem is cast in the theory of *estimating functions*. The theory of estimating functions dates back to (Godambe, 1960) for the i.i.d. case, see also (McLeish and Small, 1988; Godambe, 1991; Heyde, 1997) for the general theory. The methodology encompasses Least Squares (LS), WLS, conditional least-squares, minimum chi-squared, M-estimation and Maximum Likelihood (ML) under minor regularity conditions. Estimating functions are closely related to likelihood methodology in the sense that the optimal estimating function is the one with the highest (vector) correlation with the score function from ML theory (Heyde, 1997), but it is essentially a method of moments. In many applications only explicit expressions for the mean and variance are required, i.e. in the i.i.d. case the unconditional mean and variance and in the time series case the conditional mean and variance.

The main advantage of estimating functions is that precise mathematical statements about the optimal choice of estimating functions and the statistical properties of the estimating functions can be made explicitly. Using the optimality criterion defined in (Heyde, 1997), which is related to the socalled *Godambe information*, an explicit expression for the bias provided by the PEM method is derived and estimating functions with optimal weights are provided. The PEM estimator may behave very badly in the presence of heteroscedasticity, i.e. non-constant variance (Heyde, 1997). In many cases the estimating functions approach is easier to implement, because, in the theory of estimating functions, the focus is on functions that have the value of the parameter as a root rather than the parameter itself. Thus the parameter is obtained by solving (estimating) equations rather than optimizing an appropriately chosen criteria function wrt. the paramaters.

The remainder of the paper is organized as follows: Section 2 considers estimating functions from the linear and quadratic family and provides an expression for the bias of the WLS estimator. Both the i.i.d. case and the time series case are considered. Section 3 contains a discussion of relations between estimating functions and PEM. Some examples are provided in Section 4, and Section 5 provides some concluding remarks.

2 Estimating functions and weighted least squares

Consider a sample of n independent stochastic variables X_i. Assume that the mean and variance are given by

$$\alpha_i(\theta) = \mathrm{E}[X_i;\theta] \tag{1a}$$
$$\sigma_i^2(\theta) = \mathrm{V}[X_i;\theta] \tag{1b}$$

respectively, where $\theta \in \Theta$, a subset of $\mathbb{R}$, is a parameter to be estimated.

Initially, it is assumed that only one parameter must be estimated to simplify the notation. The model specification (1) covers general nonlinear regression models and allows for non-constant variance without making any distributional assumptions. The time series case with autocorrelated observations are considered in Section 2.3.

The weighted least squares estimate (a special case of the PEM estimate) of θ is found by optimizing the following criteria function

$$S_{\text{WLS}}(\theta) = \sum_{i=1}^{n} \frac{(X_i - \alpha_i(\theta))^2}{\sigma_i^2(\theta)}, \tag{2}$$

i.e. by solving the *estimating equation*

$$S_{\text{WLS}}'(\hat{\theta}) = -2\sum_{i=1}^{n} \alpha_i'(\hat{\theta}) \frac{(X_i - \alpha_i(\hat{\theta}))}{\sigma_i^2(\hat{\theta})} - \sum_{i=1}^{n} \frac{\sigma_i^2(\hat{\theta})'}{\sigma_i^2(\hat{\theta})} \cdot \frac{(X_i - \alpha_i(\hat{\theta}))^2}{\sigma_i^2(\hat{\theta})} = 0, \tag{3}$$

where a prime ($'$) denotes the derivative with respect to θ.

An *estimating function* $G(X_1, \ldots, X_n; \theta)$ is a function of both the data and the parameter vector. An *unbiased* estimating function satisfies the estimating equation

$$\text{E}[G(X_1, \ldots, X_n; \theta)] = 0. \tag{4}$$

The following example provides one important reason for considering estimating functions.

EXAMPLE 2.1. Assuming that an explicit expression for the density $p(X_i; \theta)$ of X_i, $i = 1, \ldots, n$, is available, the likelihood function is given by

$$L_n(\theta) = L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} p(X_i; \theta) \tag{5}$$

and the *score function* follows immediately

$$S_n(\theta) = S(X_1, \ldots, X_n; \theta) = [\ln L_n(\theta)]' = \sum_{i=1}^{n} [\ln p(X_i; \theta)]' \tag{6}$$

The Maximum Likelihood (ML) estimate is given as the solution to $S_n(\theta) = 0$, which implies that $S_n(\theta)$ is an estimating function. ♦

In many cases an explicit expression for the density p does not exist which implies that a ML estimator is not available, yet it is convenient to obtain estimators that are as closely related to the score function $S(\theta)$ as possible, because the ML estimator is known to be optimal (unless θ is on the boundary of Θ).

Using (3), it is evident that the PEM (WLS) method corresponds to the estimating function $G_{\text{WLS}}(\theta) = S_{\text{WLS}}'(\theta)$.

THEOREM 2.1. The PEM (WLS) estimator will generally not be unbiased. Eq. (3) will not in general provide a consistent estimator.

Proof. It is easily seen that

$$\text{E}[G_{\text{WLS}}(\theta)] = \sum_{i=1}^{n} \frac{\sigma_i^2(\theta)'}{\sigma_i^2(\theta)} \cdot \frac{\text{E}[(X_i - \alpha_i(\theta))^2]}{\sigma_i^2(\theta)} = \sum_{i=1}^{n} \frac{\sigma_i^2(\theta)'}{\sigma_i^2(\theta)}, \tag{7}$$

which shows that the WLS estimator will generally not be unbiased. ■

REMARK 2.1. If the $\sigma_i(\theta)$'s do not depend on θ then the WLS estimator is consistent and unbiased, which is readily seen from (3) using the fact that $\sigma_i^2(\theta)' = 0$. ▼

The EF considered in Theorem 2.1 is a special case of the PEM method, i.e. a WLS method where the inverse of the variance is used as weights. The remarkable result in the theorem is that the expected value of the derivative of the PEM (WLS) criterion (2), i.e. (3), is not zero. It is shown later in Lemma 2.1 that it is precisely this result that in general leads to biased estimates and an arbitrarily large efficiency loss.

In order to introduce a feasible definition of optimality, consider the simple Least Squares (LS) method, i.e.

$$S_{\text{LS}}(\theta) = \sum_{i=1}^{n} (X_i - \alpha_i(\theta))^2. \tag{8}$$

The LS estimator obviously solves

$$\sum_{i=1}^{n} \alpha_i'(\hat{\theta})(X_i - \alpha_i(\hat{\theta})) = 0, \tag{9}$$

which leads to the estimating function $G_{\text{LS}}(X_1, \ldots, X_n; \theta) = S_{\text{LS}}'(\theta)$. Thus the LS estimator is characterized by the weights $\alpha_i'(\hat{\theta})$. As it will become evident the weights $\alpha_i'(\theta)$ are not generally optimal and thus they will most likely lead to inefficient estimators for θ.

2.1 Estimating functions from the linear family

Consider the class of Estimating Functions from the *Linear family* (EFL) given by

$$G(X_1, \ldots, X_n; \theta) = \sum_{i=1}^{n} b_i(\theta)(X_i - \alpha_i(\theta)), \tag{10}$$

where $b_i(\theta)$ denotes the weights applied to the martingale difference $X_i - \alpha_i(\theta)$. An expression for the optimal weights $b_i^*(\theta)$ will be determined shortly.

REMARK 2.2. Clearly the LS estimator is obtained for $b_i(\theta) = \alpha_i'(\theta)$. ▼

Assume that an estimating function $G(X_1, \ldots, X_n; \theta)$ is given. Define the *standardized estimating function* $G^{(s)}(X_1, \ldots, X_n; \theta)$ by

$$G^{(s)}(X_1, \ldots, X_n; \theta) = -\frac{\mathrm{E}[G'(X_1, \ldots, X_n; \theta)]}{\mathrm{E}[G^2(X_1, \ldots, X_n; \theta)]} G(X_1, \ldots, X_n; \theta). \tag{11}$$

Using a shorter notation the variance of the standardized estimating function is found to be

$$\begin{aligned} \mathrm{V}[G_n^{(s)}(\theta)] &= \mathrm{V}[-\frac{\mathrm{E}[G_n'(\theta)]}{\mathrm{E}[G_n^2(\theta)]} G_n(\theta)] = \frac{\mathrm{E}[G_n'(\theta)]^2}{\mathrm{E}[G_n^2(\theta)]^2} \mathrm{V}[G_n(\theta)] \\ &= \frac{\mathrm{E}[G_n'(\theta)]^2}{\mathrm{E}[G_n^2(\theta)]^2} (\mathrm{E}[G_n^2(\theta)] - \mathrm{E}[G_n(\theta)]^2), \end{aligned} \tag{12}$$

which may be reduced to

$$\mathrm{V}[G_n^{(s)}(\theta)] = \frac{\mathrm{E}[G_n'(\theta)]^2}{\mathrm{E}[G_n^2(\theta)]} \tag{13}$$

provided that an unbiased estimating function is used, i.e. $\mathrm{E}[G_n(\theta)] = 0$.

Heyde (1997) provides three good reasons for assessing the optimality of an estimating function in terms of $\mathrm{V}[G_n^{(s)}(\theta)]$:

- The asymptotic distribution of the estimator $\hat{\theta}$ is normal, i.e. $(\theta - \hat{\theta}) \sim \mathrm{N}(0, \mathrm{V}[G_n^{(s)}(\theta)]^{-1})$. This implies that the most efficient estimator is the one that maximizes $\mathrm{V}[G_n^{(s)}(\theta)]$.
- The correlation between the score function (from likelihood theory) and the estimating function is maximized. In likelihood terminology $\mathrm{V}[G_n^{(s)}(\theta)]$ is the *Fisher information*.
- The numerator in (13) is a measure of sensitivity such that a large value of $\mathrm{E}[G_n^{'}(\theta)]$ implies a high sensitivity against biasedness. Conversely, a small value of the denominator $\mathrm{E}[G_n^2(\theta)]$ yields a small variance of the estimator.

Using (13) as the optimality criterion also solves the bias-variance problem illustrated in Figure 1: Consider two estimating functions $G_{1,n}^{(s)}(\theta)$ and $G_{2,n}^{(s)}(\theta)$, where the first one provides an unbiased estimator $\hat{\theta}_1$ and the second an estimator $\hat{\theta}_2$, which may be biased. Clearly if $\mathrm{V}[\hat{\theta}_1] < \mathrm{V}[\hat{\theta}_2]$ then $\hat{\theta}_1$ is preferable, but if $\mathrm{V}[\hat{\theta}_1] > \mathrm{V}[\hat{\theta}_2]$ then it is not clear which estimating function to prefer. The optimality criterion in (13) provides a solution to the bias-variance problem in the sense that a small bias is allowed in order to get as close as possible to the score function.

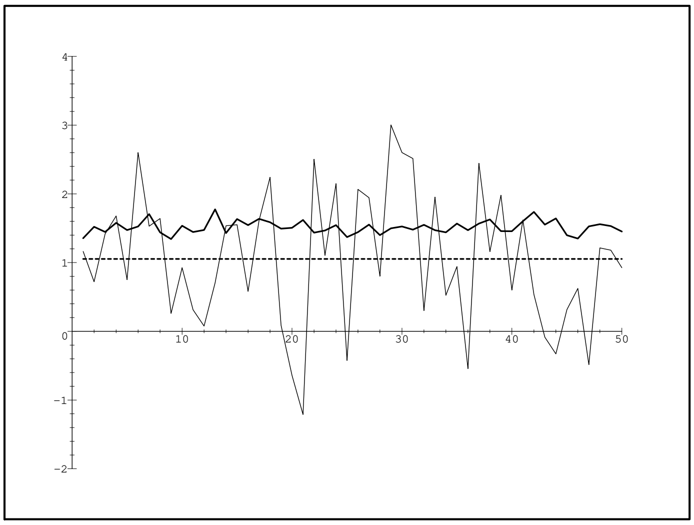

Figure 1: The bias-variance problem: The points on the thin line represent $\hat{\theta}_2$, the points on the thick line represent $\hat{\theta}_1$ and the dotted line represents the true value.

THEOREM 2.2. Given the optimality criterion (13) the optimal weights in (10) are given by

$$b_i(\theta) = \frac{\alpha_i^{'}(\theta)}{\sigma_i^2(\theta)}, \tag{14}$$

It holds that

$$\mathrm{V}[G_n^{(s)}(\theta)] = \sum_{i=1}^{n} \frac{(\alpha_i^{'}(\theta))^2}{\sigma_i^2(\theta)} \tag{15}$$

The optimal estimating function from the linear family is thus given by

$$G_n^*(\theta) = \sum_{i=1}^{n} \frac{\alpha_i^{'}(\theta)}{\sigma_i^2(\theta)}[X_i - \alpha_i(\theta)] \tag{16}$$

Proof. $\mathrm{V}[G_n^{(s)}(\theta)]$ is maximized by solving $\frac{\partial \mathrm{V}[G_n^{(s)}(\theta)]}{\partial b_j(\theta)} = 0$ for $j = 1, \ldots, n$, which results in the following n equations

$$\frac{2(\sum_{i=1}^n b_i \alpha_i'(\theta))\alpha_j'(\theta)(\sum_{i=1}^n b_i^2(\theta)\sigma_i^2(\theta)) - 2\sigma_j^2(\theta)b_j(\theta)(\sum_{i=1}^n b_i\alpha_i'(\theta))^2}{(\sum_{i=1}^n b_i^2(\theta)\sigma_i^2(\theta))^2} = 0.$$

Rewriting this system of equations yields

$$b_j(\theta) = \frac{\alpha_j'(\theta)}{\sigma_j^2(\theta)} \frac{(\sum_{i=1}^n b_i^2(\theta)\sigma_i^2(\theta))}{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))} = c\frac{\alpha_j'(\theta)}{\sigma_j^2(\theta)},$$

where c is an arbitrary, real constant indicating that the optimal weights are only unique up to the multiplicative constant c, which is, however, of no importance for the optimal estimating function. It is difficult to prove whether $\frac{\partial^2}{\partial b_i(\theta)\partial b_j(\theta)} \mathrm{V}[G_n^{(s)}(\theta)]$ is negative definite in order to ensure that the found optimum is a maximum. Instead the Cauchy-Schwarz inequality is used to verify that

$$\frac{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2}{\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta)} \leq \frac{\left(\sum_{i=1}^n \frac{(\alpha_i'(\theta))^2}{\sigma_i^2(\theta)}\right)^2}{\sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2}.$$

Consider

$$\begin{aligned}
\frac{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2}{\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta)} &= \frac{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2 \sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2}{\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta) \sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2} \\
&\leq \frac{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2 \sum_{i=1}^n \left(\frac{\alpha_i(\theta)'}{\sigma_i^2(\theta)}\right)^2}{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2} \\
&= \sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2.
\end{aligned}$$

Multiplying this expression by $\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta)$ on both sides of the equality sign yields

$$\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta) \sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2 \geq \left(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta)\right)^2.$$

The proof is now straightforward

$$\frac{(\sum_{i=1}^n b_i(\theta)\alpha_i'(\theta))^2}{\sum_{i=1}^n \sigma_i^2(\theta)b_i^2(\theta)} \leq \sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2 = \frac{\left[\sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2\right]^2}{\sum_{i=1}^n \left(\frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}\right)^2},$$

which shows that the optimum is a maximum. ■

REMARK 2.3. Notice that the first term in (3) is essentially the same as the optimal EFL (16). ▼

The result in Theorem 2.2 shows that it is possible to obtain an unbiased and efficient estimator as opposed to the PEM (WLS) estimator. Notice that this result is obtained without making any distributional assumptions as the EFL is based only on explicit expressions for the mean and the variance. The (W)LS estimator is only unbiased in the special case where $\sigma_i^2(\theta)$ does not depend on θ. Clearly the optimal estimating function given by (16) is just as easy to apply as the WLS method provided that explicit expressions for the mean and variance are available.

The next Theorem is important for assessing the efficiency of the PEM (WLS) estimator (3).

THEOREM 2.3. For the estimating function given by (3), it holds that

$$\begin{aligned} \mathrm{V}[G_{\mathrm{WLS}}^{(s)}(\theta)] &= \frac{\left[\sum_{i=1}^{n}\left(2\frac{(\alpha_i'(\theta))^2}{\sigma_i^2(\theta)} + 2\frac{(\sigma_i^2(\theta)')^2}{\sigma_i^4(\theta)} - \frac{\sigma_i^2(\theta)''}{\sigma_i^2(\theta)}\right)\right]^2}{\mathrm{E}[G_{\mathrm{WLS}}^2(\theta)]} \\ &\quad \times \left[\mathrm{E}[G_{\mathrm{WLS}}^2(\theta)] - \left(\sum_{i=1}^{n}\frac{\sigma_i^2(\theta)'}{\sigma_i^2(\theta)}\right)\right], \end{aligned} \tag{17}$$

where

$$\begin{aligned} \mathrm{E}[G_{\mathrm{WLS}}^2(\theta)] &= 4\sum_{i=1}^{n}\frac{(\alpha_i'(\theta))^2}{\sigma_i^2(\theta)} + \sum_{i=1}^{n}\frac{(\sigma_i^2(\theta)')^2}{\sigma_i^8(\theta)}\left(\mathrm{E}[(X_i-\alpha_i(\theta))^4] - \sigma_i^4(\theta)\right) \\ &\quad + \left(\sum_{i=1}^{n}\frac{\sigma_i^2(\theta)'}{\sigma_i^2(\theta)}\right)^2 + 4\sum_{i=1}^{n}\frac{\alpha_i'(\theta)\sigma_i^2(\theta)'}{\sigma_i^6(\theta)}\,\mathrm{E}[(X_i-\alpha_i(\theta))^3] \end{aligned} \tag{18}$$

Proof. See (Nolsøe, 1999, Section 2.4). ∎

LEMMA 2.1. It follows from Theorem 2.1 that $\mathrm{E}[G_{\mathrm{WLS}}(\theta)] \neq 0$. This implies that the simple optimality criterion given by (13) cannot be used. Instead the computationally more demanding criterion (12) must be used. In this case, it holds that

$$\theta - \hat{\theta} \sim \mathrm{N}\left(\frac{\sqrt{\mathrm{V}[\mathrm{G}_{\mathrm{WLS}}(\theta)]}\,\mathrm{E}[\mathrm{G}_{\mathrm{WLS}}(\theta)]}{\sqrt{\mathrm{V}[\mathrm{G}_{\mathrm{WLS}}^{(\mathrm{s})}(\theta)]}\,\mathrm{E}[\mathrm{G}_{\mathrm{WLS}}^2(\theta)]}, \frac{\mathrm{V}[\mathrm{G}_{\mathrm{WLS}}(\theta)]^2}{\mathrm{E}[\mathrm{G}_{\mathrm{WLS}}^2(\theta)]^2\,\mathrm{V}[\mathrm{G}_{\mathrm{WLS}}^{(\mathrm{s})}(\theta)]}\right) \tag{19}$$

where

$$\begin{aligned} \mathrm{V}[G_{\mathrm{WLS}}(\theta)] &= 4\sum_{i=1}^{n}\frac{(\alpha_i'(\theta))^2}{\sigma_i^2(\theta)} + \sum_{i=1}^{n}\frac{(\sigma_i^2(\theta)')^2}{\sigma_i^8(\theta)}\left(\mathrm{E}[(X_i-\alpha_i(\theta))^4] - \sigma_i^4(\theta)\right) \\ &\quad +4\sum_{i=1}^{n}\frac{\alpha_i'(\theta)\sigma_i^2(\theta)'}{\sigma_i^6(\theta)}\,\mathrm{E}[(X_i-\alpha_i(\theta))^3]. \end{aligned} \tag{20}$$

Proof. See (Nolsøe, 1999, Section 2.4). ∎

2.2 Estimating functions from the quadratic family

Next assume that a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a subset of $\mathbb{R}^p$, is to be estimated. Introduce the vector of p linearly independent functions $\mathbf{m}_i(\boldsymbol{\theta}) = (m_i(\boldsymbol{\theta})_1, \ldots, m_i(\boldsymbol{\theta})_p)^T$.

To introduce Estimating Functions from the Quadratic family (EFQ) consider the following moment restrictions

$$\mathbf{m}_i(\boldsymbol{\theta}) = \begin{pmatrix} X_i - \alpha_i(\boldsymbol{\theta}) \\ (X_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta}) \end{pmatrix}, \tag{21}$$

which are socalled martingale differences with expectation zero. One reason for increasing the number of moment restrictions is that this brings the resulting EF closer to the score function. Using EFL it is possible to obtain p equations that are linearly dependent such that it is not possible to estimate all p parameters. This problem may be eliminated by using a larger class of EFs, e.g. EFQ.

Consider the p-dimensional estimating function

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{A}_i(\boldsymbol{\theta})\mathbf{m}_i(\boldsymbol{\theta}), \tag{22}$$

where the optimal weights are given by, see (Heyde, 1997),

$$\mathbf{A}_i(\boldsymbol{\theta}) = \mathrm{E}[\partial_{\boldsymbol{\theta}^T}\mathbf{m}_i(\boldsymbol{\theta})]^T\mathbf{V}_i^{-1}(\boldsymbol{\theta}) \tag{23}$$

$$\mathbf{V}_i(\boldsymbol{\theta}) = \mathrm{V}[\mathbf{m}_i(\boldsymbol{\theta})] = \mathrm{E}[\mathbf{m}_i(\boldsymbol{\theta})\mathbf{m}_i(\boldsymbol{\theta})^T] \tag{24}$$

and the matrix of partial derivatives is

$$\partial_{\boldsymbol{\theta}^T}\mathbf{m}_i(\boldsymbol{\theta}) = \begin{pmatrix} \partial_{\theta_1} m_i(\boldsymbol{\theta})_1 & \dots & \partial_{\theta_p} m_i(\boldsymbol{\theta})_1 \\ \vdots & & \vdots \\ \partial_{\theta_1} m_i(\boldsymbol{\theta})_p & \dots & \partial_{\theta_p} m_i(\boldsymbol{\theta})_p \end{pmatrix}, \tag{25}$$

where, say, $\partial_{\theta_1} m_i(\boldsymbol{\theta})_1$ is short for the partial derivative $\frac{\partial}{\partial\theta_1} m_i(\boldsymbol{\theta})_1$.

Introduce the following notation for the moments

$$\alpha_i(\boldsymbol{\theta}) = \mathrm{E}[X_i;\boldsymbol{\theta}] \tag{26a}$$

$$\sigma_i^2(\boldsymbol{\theta}) = \mathrm{E}[(X_i - \alpha_i(\boldsymbol{\theta}))^2;\boldsymbol{\theta}] \tag{26b}$$

$$\eta_i(\boldsymbol{\theta}) = \mathrm{E}[(X_i - \alpha_i(\boldsymbol{\theta}))((X_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta}));\boldsymbol{\theta}] \tag{26c}$$

$$\psi_i(\boldsymbol{\theta}) = \mathrm{E}[((X_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta}))^2;\boldsymbol{\theta}] \tag{26d}$$

It follows immediately by applying (26) that

$$\mathbf{V}_i(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_i^2(\boldsymbol{\theta}) & \eta_i(\boldsymbol{\theta}) \\ \eta_i(\boldsymbol{\theta}) & \psi_i(\boldsymbol{\theta}) \end{pmatrix}$$

$$\mathbf{V}_i^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\psi_i(\boldsymbol{\theta})}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} & -\frac{\eta_i(\boldsymbol{\theta})}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} \\ -\frac{\eta_i(\boldsymbol{\theta})}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} & \frac{\sigma_i^2(\boldsymbol{\theta})}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} \end{pmatrix}$$

and, finally,

$$\mathrm{E}[\partial_{\boldsymbol{\theta}^T}\mathbf{m}_i(\boldsymbol{\theta})] = -\begin{pmatrix} \partial_{\theta_1}\alpha_i(\boldsymbol{\theta}) & \cdots & \partial_{\theta_p}\alpha_i(\boldsymbol{\theta}) \\ \partial_{\theta_1}\sigma_i^2(\boldsymbol{\theta}) & \cdots & \partial_{\theta_p}\sigma_i^2(\boldsymbol{\theta}) \end{pmatrix}$$

$$\mathrm{E}[\partial_{\boldsymbol{\theta}^T}\mathbf{m}_i(\boldsymbol{\theta})]^T\mathbf{V}_i^{-1}(\boldsymbol{\theta}) =$$

$$\begin{pmatrix} \frac{-\partial_{\theta_1}\alpha_i(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) + \eta_i(\boldsymbol{\theta})(\partial_{\theta_1}\sigma_i^2(\boldsymbol{\theta}))}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} & \frac{\partial_{\theta_1}\alpha_i(\boldsymbol{\theta})\eta_i(\boldsymbol{\theta}) - \sigma_i^2(\boldsymbol{\theta})(\partial_{\theta_1}\sigma_i^2(\boldsymbol{\theta}))}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} \\ \vdots & \vdots \\ \frac{-\partial_{\theta_p}\alpha_i(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) + \eta_i(\boldsymbol{\theta})(\partial_{\theta_1}\sigma_i^2(\boldsymbol{\theta}))}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} & \frac{\partial_{\theta_p}\alpha_i(\boldsymbol{\theta})\eta_i(\boldsymbol{\theta}) - \sigma_i^2(\boldsymbol{\theta})(\partial_{\theta_p}\sigma_i^2(\boldsymbol{\theta}))}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} \end{pmatrix}$$

This leads to the class of EFQ, i.e.

$$
\begin{aligned}
\mathbf{G}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\frac{-(\partial_{\boldsymbol{\theta}^T}\alpha_i(\boldsymbol{\theta}))^T\psi_i(\boldsymbol{\theta}) + \eta_i(\boldsymbol{\theta})(\partial_{\boldsymbol{\theta}^T}\sigma_i^2(\boldsymbol{\theta}))^T}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})}(X_i - \alpha_i(\boldsymbol{\theta})) \right. \\
&\quad \left. + \frac{(\partial_{\boldsymbol{\theta}^T}\alpha_i(\boldsymbol{\theta}))^T\eta_i(\boldsymbol{\theta}) - \sigma_i^2(\boldsymbol{\theta})(\partial_{\boldsymbol{\theta}^T}\sigma_i^2(\boldsymbol{\theta}))^T}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})}((X_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta})) \right), \qquad (27)
\end{aligned}
$$

where $\partial_{\boldsymbol{\theta}^T}\alpha_i(\boldsymbol{\theta}) = (\partial_{\theta_1}\alpha_i(\boldsymbol{\theta}), \ldots, \partial_{\theta_p}\alpha_i(\boldsymbol{\theta}))$ is introduced to allow a simpler notation.

REMARK 2.4. It is noticed that the optimal weights in front of the two martingale difference terms $X_i - \alpha_i(\boldsymbol{\theta})$ and $(X_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta})$ differ from those given by the WLS estimating equation (3). ▼

REMARK 2.5. To illustrate that the $\mathcal{O}_F$-optimal weights depend on the specification of the moment restrictions, consider the following alternative to (21), i.e.

$$
\mathbf{m}_i(\boldsymbol{\theta}) = \begin{pmatrix} X_i - \mathrm{E}[X_i;\boldsymbol{\theta}] \\ X_i^2 - \mathrm{E}[X_i^2;\boldsymbol{\theta}] \end{pmatrix} = \begin{pmatrix} X_i - \alpha_i(\boldsymbol{\theta}) \\ X_i^2 - \sigma_i^2(\boldsymbol{\theta}) - \alpha_i^2(\boldsymbol{\theta}) \end{pmatrix}. \qquad (28)
$$

Using these moment restrictions yield the following $\mathcal{O}_F$-optimal weights

$$
\begin{aligned}
\mathbf{G}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\frac{(\partial_{\boldsymbol{\theta}^T}\alpha_i(\boldsymbol{\theta}))^T(2\alpha_i(\boldsymbol{\theta})\eta_i(\boldsymbol{\theta}) + \psi_i(\boldsymbol{\theta})) - (\eta_i(\boldsymbol{\theta}) + 2\alpha_i(\boldsymbol{\theta}))(\partial_{\boldsymbol{\theta}^T}\sigma_i^2(\boldsymbol{\theta}))^T}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})} \right. \\
&\quad \times (X_i - \alpha_i(\boldsymbol{\theta})) \\
&\quad \left. - \frac{(\partial_{\boldsymbol{\theta}^T}\alpha_i(\boldsymbol{\theta}))^T\eta_i(\boldsymbol{\theta}) - \sigma_i^2(\boldsymbol{\theta})(\partial_{\boldsymbol{\theta}^T}\sigma_i^2(\boldsymbol{\theta}))^T}{\sigma_i^2(\boldsymbol{\theta})\psi_i(\boldsymbol{\theta}) - \eta_i^2(\boldsymbol{\theta})}(X_i^2 - \sigma_i^2(\boldsymbol{\theta}) - \alpha_i^2(\boldsymbol{\theta})) \right), \qquad (29)
\end{aligned}
$$

where the notation from (26) has been used. This result differs from (27), but there is no available theory to determine whether the moment restrictions given by (21) or (28) should be used. However, the EFQ given by either (27) or (29) is $\mathcal{O}_F$-optimal given the moment restrictions (21) or (28), respectively. Let EFQ′ denote the EFQ given by (29). ▼

2.3 Correlated observations

Now assume that the observations X_i for $i = 1, \ldots, n$ are correlated, i.e. the observations may be considered as a time series. Denote the conditional mean and conditional variance by

$$
\begin{aligned}
F(X_{i-1};\boldsymbol{\theta}) &= \mathrm{E}[X_i|X_{i-1};\boldsymbol{\theta}] \qquad (30a) \\
\Phi(X_{i-1};\boldsymbol{\theta}) &= \mathrm{V}[X_i|X_{i-1};\boldsymbol{\theta}] \qquad (30b)
\end{aligned}
$$

respectively.

The model specification (30) covers first order autoregressive processes and allows for heteroscedasticity e.g. the ARCH model (Engle, 1982), see also (Bollerslev, Chou and Kroner, 1992) for a review.

The earlier results for computing the optimal weights for estimating functions from the linear and quadratic family may immediately be generalized to the time series case.

LEMMA 2.2. The optimal EFL is given by

$$
G_n^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{F'(X_{i-1};\boldsymbol{\theta})}{\Phi(X_{i-1};\boldsymbol{\theta})}[X_i - F(X_{i-1};\boldsymbol{\theta})], \qquad (31)
$$

where $F(X_{i-1};\boldsymbol{\theta})$ and $\Phi(X_{i-1};\boldsymbol{\theta})$ are given by (30).

Proof. See (Heyde, 1997). ■

Clearly the martingale difference $X_i - F(X_{i-1};\boldsymbol{\theta})$ in (31) for $i = 1, \ldots, n$ may be interpreted as *one-step ahead prediction errors* for first order Markov processes.

REMARK 2.6. The optimal estimating functions from the quadratic family are obtained by a simple generalization of (26) and (27), see (Heyde, 1997; Nolsøe, 1999) for details. ▼

3 Estimating functions and prediction error methods

In this section some relations between the estimating function methodology and the prediction error method will be discussed.

Again, let X_i, $i = 1, \ldots, n$, denote a time series of correlated stochastic variables.

Let $\mathcal{F}_n = \sigma\{X_n, X_{n-1}, \ldots, X_0\}$ denote the information set (σ-algebra) up to and including time n. Finally, let

$$\varepsilon_i(\boldsymbol{\theta}) = X_i - \mathrm{E}[X_i|X_{i-1};\boldsymbol{\theta}] \tag{32}$$

denote the one-step ahead prediction errors for first order Markov processes, and introduce the criterion

$$V_n(\boldsymbol{\theta}, \mathcal{F}_n) = h\left(\frac{1}{n}\sum_{i=1}^{n} l(i, \boldsymbol{\theta}, \varepsilon_i(\boldsymbol{\theta}))\right) \tag{33}$$

where $l(\cdot,\cdot)$ is a function from $\mathbb{Z} \times \mathbb{R}^p \times \mathbb{R}$ to the space of positive semidefinite $s \times s$ matrices and $h(\cdot)$ is a function from the space of $s \times s$ matrices to the real numbers.

Ljung and Caines (1979) establishes, under minor regularity conditions, the asymptotic normality of the estimator obtained by optimizing (33). However the criterion (33) is too general to obtain any results regarding the efficiency of the estimators, and it is inherently difficult to obtain results for the optimal choices of l and h. In the PEM context, the following estimating equation seems an obvious choice

$$\partial_{\boldsymbol{\theta}^T} V_n(\boldsymbol{\theta}, \mathcal{F}_n) = \mathbf{0}, \tag{34}$$

where $\mathbf{0}$ is a p-dimensional vector of zeros. Thus it is difficult to relate the methods to one another in general. For simplicity assume that a one-dimensional parameter θ is to be estimated. If $\sigma_i^2(\theta) = \sigma_i^2$ then it is possible to compute a criterion function using EFL as follows

$$\begin{aligned} V_{1,n}(\theta, \mathcal{F}_n) &= \sum_{i=1}^{n} \int_{\Theta} \frac{\alpha_i'(\theta)}{\sigma_i^2}(X_i - \alpha_i(\theta))d\theta \\ &= \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \int_{\Theta} \alpha_i'(\theta)(X_i - \alpha_i(\theta))d\theta \end{aligned} \tag{35}$$

However, if σ_i^2 is allowed to depend on θ then the criterion function given by

$$V_{2,n}(\theta, \mathcal{F}_n) = \sum_{i=1}^{n} \int_{\Theta} \frac{\alpha_i'(\theta)}{\sigma_i^2(\theta)}(X_i - \alpha_i(\theta))d\theta \tag{36}$$

might not in general be integrable. If it is integrable the criterion function will depend on the specification of $\alpha_i(\theta)$ and $\sigma_i^2(\theta)$. This approach may be seen as a means of making the optimal choice of at least the function l in the PEM context.

In a number of special cases it is possible to make some comparisons. Assuming that the prediction errors $\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = (\varepsilon_1(\boldsymbol{\theta}), \ldots, \varepsilon_n(\boldsymbol{\theta}))^T$ are independent with covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ independent of $\boldsymbol{\theta}$, it is customary to consider the special case

$$l(i, \boldsymbol{\theta}, \boldsymbol{\varepsilon}(\boldsymbol{\theta})) = \boldsymbol{\varepsilon}^T(\boldsymbol{\theta})\Sigma^{-1}\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = \mathrm{tr}\, \boldsymbol{\varepsilon}(\boldsymbol{\theta})\boldsymbol{\varepsilon}^T(\boldsymbol{\theta})\Sigma^{-1}; \quad h(x) = x \tag{37}$$

This is the ordinary LS problem for which the LS and EFL methods provide the same optimal solution. The WLS problem is obtained by allowing Σ to depend on $\boldsymbol{\theta}$. Theorem 2.1 states that the WLS method may lead to biased and inconsistent estimates, whereas the EFL solution provided in Theorem 2.2 is optimal.

Assuming that $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$ is Gaussian with zero mean and covariance $\Sigma(\boldsymbol{\theta})$, it is convenient to consider the criterion

$$l(i, \boldsymbol{\theta}, \boldsymbol{\varepsilon}(\boldsymbol{\theta})) = \frac{1}{2}\boldsymbol{\varepsilon}^T(\boldsymbol{\theta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\varepsilon}(\boldsymbol{\theta}) + \frac{1}{2}\log\det \boldsymbol{\Sigma}(\boldsymbol{\theta}) \tag{38}$$

As the Gaussian distribution belongs to the exponential family, the ML, EFL and EFQ methods provide the same optimal solution, whereas the afore-mentioned comments regarding the LS and WLS methods still hold.

4 Examples

In this section the properties of the estimators provided by the PEM (LS and WLS), EFL and EFQ methods will be analyzed for three particular models using simulated data. In a fourth example only the formulae are given.

4.1 Example 1: The Poisson distribution

As a simple example, assume that $X_i \in \mathrm{Pois}(\theta i)$ for $i = 1, \ldots, n$, i.e. that X_i follows an inhomogeneous Poisson distribution with intensity θi. It follows that $\mathrm{E}[X_i] = \mathrm{V}[X_i] = \mathrm{E}[(X_i - \mathrm{E}[X_i])^3] = \theta i$ and $\mathrm{E}[(X_i - \mathrm{E}[X_i])^4] = (3\theta i + 1)\theta i$. The estimators and the variance of the estimators obtained by using PEM (LS and WLS), ML, EFL and EFQ are given in Table 1. It is noted that the estimator obtained using ML, EFL and EFQ are the same. This result holds for all distributions belonging to the exponential family of distributions (Barndorff-Nielsen, 1978), see e.g. (Nolsøe, 1999).

Method	PEM (LS)	PEM (WLS)	EFL/EFQ/ML
$\hat{\theta}$	$\frac{6}{n(n+1)(2n+1)}\sum_{i=1}^{n} iX_i$	$\pm\sqrt{\frac{2}{n(n+1)}\sum_{i=1}^{n}\frac{X_i^2}{i}}$	$\frac{2}{n(n+1)}\sum_{i=1}^{n} X_i$
$\mathrm{V}[\hat{\theta}]$	$\frac{9\theta}{(2n+2)^2}$	N/A	$\frac{2\theta}{n(n+1)}$

Table 1: The table contains estimators of θ using the PEM (LS and WLS) and EFL/EFQ/ML methods. The last three methods coincide for the Poisson distribution.

A simulation study has been repeated 500 times with $\theta = 0.1$ for $i = 1, \ldots, 100$, i.e. using 100 observations in each stochastically independent sample. The results reported in Table 2 have been obtained. It is seen that the WLS method provides a biased estimate, that the LS method provides a very inefficient estimate and that the EFL/EFQ/ML methods provide an unbiased estimate, which is also the most efficient. The computed values of the variance of the standardized estimating function $\mathrm{V}[G^{(s)}(\theta)]$ attains its maximum for the EFL/EFQ/ML method as expected. Based solely on the mean and variance of $\hat{\theta}$, it is not entirely clear whether one should accept the small bias of the WLS estimate given that it is more efficient than the PEM (LS) estimate or vice versa, but the values of $\mathrm{V}[G^{(s)}(\theta)]$ clearly illustrates the efficiency loss implied by the PEM (WLS) method. Thus in this case the LS method is preferable compared to the WLS method, albeit the observations exhibit heteroscedasticity. The asymptotic result $\mathrm{V}[\hat{\theta}] = \mathrm{V}[G^{(s)}(\hat{\theta})]^{-1}$ holds approximately except for the PEM (WLS) method. This result was to be expected according to Lemma 2.1. Histograms for the 500 estimates for each of the three methods are provided in Figure 2.

	PEM (LS)	PEM (WLS)	EFL/EFQ/ML
$\mathrm{E}[\hat{\theta}]$	0.1002	0.1094	0.1001
$\mathrm{V}[\hat{\theta}]$ $(\cdot 10^{-5})$	2.317	2.094	1.963
$\mathrm{V}[G^{(s)}(\hat{\theta})]$ $(\cdot 10^{4})$	4.336	0.232	5.193

Table 2: The mean and variance of the 500 estimates of θ are listed using the LS, WLS and EFL/EFQ/ML methods. The last three methods coincide for the Poisson distribution.

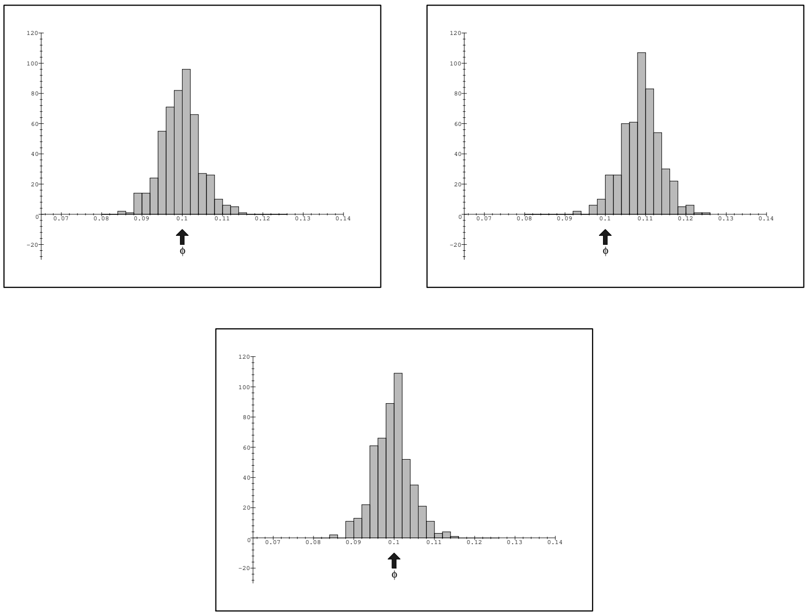

Figure 2: Histograms of the estimate obtained using the LS, WLS and EFL methods (left to right).

Figure 3 provides an illustration of the properties of the various estimators assuming that the distribution of X_i belongs to the exponential family. It is seen that the EFL, EFQ and ML provide the same estimator.

4.2 Example 2: The log-normal distribution

As an example of a distribution that does not belong to the exponential family consider the log-normal distribution. This example also illustrates that an unbiased EF need not yield an unbiased estimator (see the discussion on the bias-variance problem in Section 2.1). Assume that $X_i \in \mathrm{LN}(\theta, \beta^2)$ for $i = 1, \ldots, n$, where β is assumed known. It follows readily that $\mathrm{E}[X_i] = e^{\theta + \frac{1}{2}\beta^2}$, $\mathrm{V}[X_i] = e^{2\theta+\beta^2}(e^{\beta^2} - 1)$, $\mathrm{E}[(X_i - \mathrm{E}[X_i])^3] = e^{3\theta + \frac{3}{2}\beta^2}(e^{\beta^2} - 1)^2(e^{\beta^2} + 2)$ and $\mathrm{E}[(X_i - \mathrm{E}[X_i])^4] = e^{4\theta + 2\beta^2}(e^{\beta^2} - 1)^2(e^{4\beta^2} + 2e^{3\beta^2} + 3e^{2\beta^2} - 3)$.

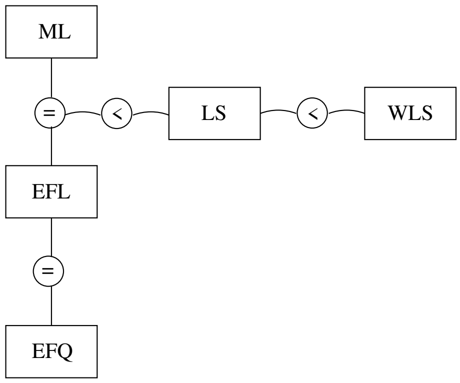

Figure 3: Hierarchy of estimation methods of parameters in the exponential family.

It is well-known that the unbiased ML estimator is given by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \ln(X_i). \tag{39}$$

It follows from (16) that the EFL estimator is given by

$$\hat{\theta} = -\frac{1}{2}\beta^2 + \ln\left(\sum_{i-1}^{n} X_i\right) - \ln(n). \tag{40}$$

This is also the PEM (LS) estimator. It may be shown that $\mathrm{E}[\hat{\theta}] = \theta - \frac{e^{\beta^2}-1}{2n}$, which implies that the EFL estimator is consistent, but biased for finite samples.

REMARK 4.1. Considering the transformed data $\ln(X_i)$ the EFL method will provide the ML estimator. ▼

The PEM (WLS)-estimator is found to be

$$\hat{\theta} = -\frac{1}{2}\beta^2 + \ln\left(\sum_{i=1}^{n} X_i^2\right) - \ln\left(\sum_{i=1}^{n} X_i\right), \tag{41}$$

which has the expected value $\mathrm{E}[\hat{\theta}] = \theta + \beta^2 + \frac{e^{\beta^2}-1}{2n} - \frac{e^{4\beta^2}-1}{2n^2}$, i.e. the PEM (WLS) estimator is inconsistent and biased.

The EFQ estimator should solve a second order equation in e^{θ}, which is left out for brevity, see (Nolsøe, 1999).

A simulation study has been repeated 500 times with $\theta = \beta = 1$ for $i = 1, \ldots, 100$, i.e. using 100 observations in each stochastically independent sample. The results reported in Table 3 have been obtained. It is seen that the ML method outperforms the other methods, and that the PEM method provides a grossly biased and inefficient estimate. These results give rise to Figure 4, which provides an illustration of the properties of the various estimators and their interrelations.

4.3 Example 3: AR(1)-ARCH(1) model

Consider a first-order AutoRegressive process

$$X_i = \phi X_{i-1} + \epsilon_i, \tag{42}$$

	PEM	EFL	EFQ	EFQ′	ML
$\mathrm{E}[\hat{\theta}]$	1.9009	0.9884	0.9906	0.9902	0.9920
$\mathrm{V}[\hat{\theta}]$	0.1210	0.0160	0.0125	0.0130	0.0094
$\mathrm{V}[G^{(s)}(\hat{\theta})]$	1.6260	60.0510	79.0818	76.9148	94.5909

Table 3: The mean and variance of the 500 estimates of θ are listed using the PEM (LS and WLS), EFL, EFQ and ML methods.

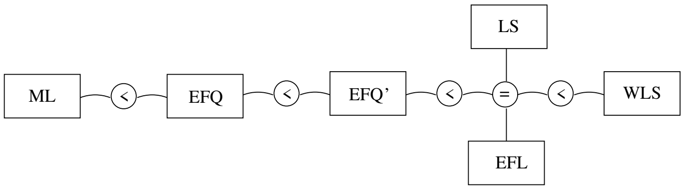

Figure 4: Hierarchy of estimators assuming that $X_i \in \mathrm{LN}(\theta, \beta^2)$.

where ϵ_i is a process exhibiting first order AutoRegressive Conditional Heteroscedasticity (ARCH), i.e. $\epsilon_i = \sigma_i \eta_i$ with $\eta_i \sim \mathrm{N}(0,1)$, and

$$\sigma_i^2 = \alpha_0 + \alpha_1 \epsilon_{i-1}^2, \tag{43}$$

where $\alpha_0 > 0$, $\alpha_0 \geq 0$. The unconditional variance of ϵ_i is $\sigma^2 = \alpha_0/(1-\alpha_1)$ provided that $\alpha_1 < 1$. In other words

$$\mathrm{E}[X_i|\mathcal{F}_{i-1}] = \phi X_{i-1} \tag{44a}$$
$$\mathrm{V}[X_i|\mathcal{F}_{i-1}] = \alpha_0 + \alpha_1 \epsilon_{i-1}^2 \tag{44b}$$

Note that the conditioning has been extended from X_{i-1} to the entire information set available at time $i-1$ denoted $\mathcal{F}_{i-1}$.

For simplicity, consider the case $\alpha_0 = 1$ and $\alpha_1 = 1 + \phi$. To obtain an estimate of ϕ using EFL (10), the following estimating equation should be solved

$$G_n(\phi) = \sum_{i=1}^{n} \frac{X_{i-1}}{1 + (1+\phi)\epsilon_{i-1}^2}[X_i - \phi X_{i-1}] \tag{45}$$

Using (9) the PEM (LS) estimator is given by

$$\phi = \frac{\sum_{i=1}^{n} X_{i-1} X_i}{\sum_{i=1}^{n} X_{i-1}^2} \tag{46}$$

Finally, using (3) the PEM (WLS) estimator is the root of the equation

$$\sum_{i=1}^{n} \left(\frac{2X_{i-1}(X_i - \phi X_{i-1})}{1 + (1+\phi)\epsilon_{i-1}^2} + \frac{\epsilon_{i-1}^2 (X_i - \phi X_{i-1})^2}{(1 + (1+\phi)\epsilon_{i-1}^2)^2} \right) = 0 \tag{47}$$

The results reported in Table 4 are obtained by simulating 500 stochastic independent samples each consisting of 500 observations for $\phi = -0.5$ and $\phi = 0.5$, respectively. Note that (43) is not stationary for $\phi = 0.5$. The 500 parameter estimates for each value of ϕ are illustrated in Figure 5. The LS

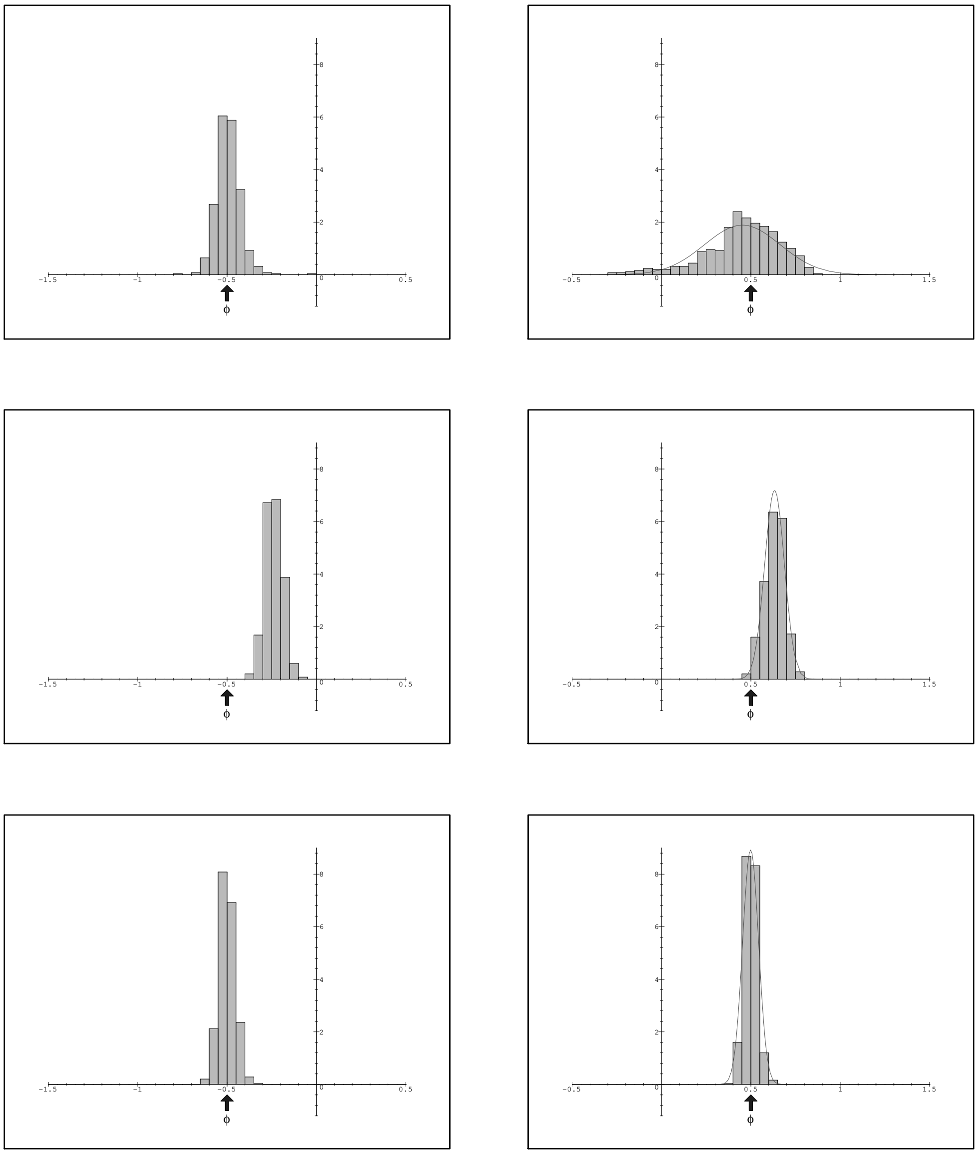

Figure 5: Estimation results for the LS, WLS and EFL methods (top to bottom). The left (right) panel shows the histogram and the empirical density (i.e. a normal density with the empirical mean and variance) of the same estimates for $\phi = -0.5$ ($\phi = 0.5$).

method provides an unbiased estimate that is mostly inefficient in the nonstationary case ($\phi = 0.5$). The WLS method provides the second most efficient, but grossly biased estimate. Indeed the biasedness is most pronounced in the stationary case ($\phi = -0.5$). It is seen that the EFL method provides the most efficient and unbiased estimate. Figure 5 shows the histograms of the parameter estimates along with normal distributions with the mean and variance listed in Table 4 in order to investigate the asymptotic distribution of the estimates (right panel only). Significant deviations from normality are not seen for any of the methods.

	$\phi = -0.5$			$\phi = 0.5$		
	LS	WLS	EFL	LS	WLS	EFL
$\mathrm{E}[\hat{\theta}]$	-0.4930	-0.2385	-0.4992	0.4330	0.6274	0.4989
$\mathrm{V}[\hat{\theta}]$	0.004632	0.002437	0.001916	0.0446	0.0032	0.0014
$\mathrm{V}[G^{(s)}(\hat{\theta})]$	267.79	0.7080	501.25	14.65	28.87	998.16

Table 4: The mean and variance of the 500 estimates of $\phi = -0.5$ and $\phi = 0.5$ are listed using the PEM (LS and WLS) and EFL methods.

4.4 Example 4: Stochastic variance models

Taylor (1986) proposes *Stochastic Variance (SV) models* (or *Stochastic Volatility models*) as an alternative to GARCH models, i.e.

$$Y_i = e^{\frac{1}{2}X_i}\varepsilon_i; \quad \varepsilon_i \sim \mathrm{N}(0,1), \tag{48}$$

where X_i is an AR(1)-process

$$X_i = \phi_0 + \phi_1 X_{i-1} + e_i; \quad e_i \sim \mathrm{N}(0, \sigma_e^2). \tag{49}$$

It is assumed that ε_i and e_i are mutually independent. Harvey, Ruiz and Shephard (1994) considers the special case, where (49) is a pure random walk ($\phi_0 = 0$ and $\phi_1 = 1$).
Eqs. (48)–(49) may be restated in stochastic state space form with $\ln \chi^2(1)$ measurement noise

$$\ln Y_i^2 = X_i + \xi_i - 1.27 \tag{50a}$$
$$X_i = \phi_0 + \phi_1 X_{i-1} + e_i \tag{50b}$$

with $\xi_i = \ln e_i^2 + 1.27$ such that $\mathrm{E}[\xi_i] = 0$ and $\mathrm{Var}[\xi_i] = \pi^2/2$, see e.g. (Abramowitz and Stegun, 1970, p. 943), where (50a) may be interpreted as a measurement equation and (50b) as a state space equation[1]. Harvey et al. (1994) proposes that the ξ_i be treated as $\mathrm{N}(0, \pi^2/2)$ and applies the Kalman filter in combination with a Quasi-Maximum Likelihood (QML) method. However, Sandmann and Koopman (1998) shows that the distribution $\mathrm{N}(-1.27, \pi^2/2)$ provides a poor approximation to the exact $\ln \chi^2(1)$ distribution.

REMARK 4.2. As in the last section it is assumed that the realization of the latent process is known, because it is outside the scope of the present paper to discuss the filtering problem. The recently proposed *Prediction-based Estimating Functions* (PEF) makes it possible to estimate the parameters in the latent process (49), but not the states X_i, see (Sørensen, 1999; Nolsøe, Nielsen and Madsen, 2000). ▼

[1]The exact mean of $\ln e_i^2$ is $-(\gamma + \ln 2)$, where $\gamma \approx 0.5772$ is Euler's constant.

It follows immediately that $\mathrm{E}[X_i|X_{i-1}] = \phi_0 + \phi_1 X_{i-1}$ and that $\mathrm{V}[X_i|X_{i-1}] = \sigma_e^2$. Due to the normality of η_i and X_i, and hence log-normality of $e^{\frac{1}{2}X_i}$, it holds that

$$\begin{aligned}
\mathrm{E}[Y_i^{2j+1}|\mathcal{F}_{i-1}] &= 0 \quad j = 0, 1, 2, \ldots & \text{(51a)}\\
\mathrm{E}[Y_i^{2j}|\mathcal{F}_{i-1}] &= \prod_{k=1}^{j}(2k-1)e^{j(\phi_0+\phi_1 X_{i-1})+\frac{1}{2}j^2\sigma_e^2} \quad j = 1, 2, 3, \ldots. & \text{(51b)}
\end{aligned}$$

Stochastic variance models provide a description of the conditional variance as do GARCH models, which implies that EFL cannot be used to estimate all the model parameters. Instead EFQ must be used, which requires explicit expressions for the following martingale differences with $\boldsymbol{\theta} = (\phi_0, \phi_1, \sigma_e^2)^T$

$$\begin{aligned}
\alpha_i(\boldsymbol{\theta}) &= \mathrm{E}[Y_i|\mathcal{F}_{i-1}; \boldsymbol{\theta}] = 0 & \text{(52a)}\\
\sigma_i^2(\boldsymbol{\theta}) &= \mathrm{E}[(Y_i - \alpha_i(\boldsymbol{\theta}))^2|\mathcal{F}_{i-1}; \boldsymbol{\theta}] = \mathrm{E}[Y_i^2|\mathcal{F}_{i-1}; \boldsymbol{\theta}] = e^{\phi_0+\phi_1 X_{i-1}+\frac{1}{2}\sigma_e^2} & \text{(52b)}\\
\eta_i(\boldsymbol{\theta}) &= \mathrm{E}[(Y_i - \alpha_i(\boldsymbol{\theta}))\left\{(Y_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta})\right\}|\mathcal{F}_{i-1}; \boldsymbol{\theta}] = 0 & \text{(52c)}\\
\psi_i(\boldsymbol{\theta}) &= \mathrm{E}[\left\{(Y_i - \alpha_i(\boldsymbol{\theta}))^2 - \sigma_i^2(\boldsymbol{\theta})\right\}^2|\mathcal{F}_{i-1}; \boldsymbol{\theta}] \\
&= \left(3e^{2\sigma_e^2} + e^{\sigma_e^2} + 2e^{\frac{3}{2}\sigma_e^2}\right) e^{2(\phi_0+\phi_1 X_{i-1})} & \text{(52d)}
\end{aligned}$$

The optimal EFQ is readily obtained by solving

$$\sum_{i=1}^{n}(1 \ X_{i-1} \ \frac{1}{2})^T \frac{\sigma_i^2(\boldsymbol{\theta})}{\psi_i(\boldsymbol{\theta})}(Y_i^2 - \sigma_i^2(\boldsymbol{\theta})) = \mathbf{0}. \tag{53}$$

It is also readily seen that the PEM (WLS) estimator is obtained by solving

$$\sum_{i=1}^{n}(1 \ X_{i-1} \ \frac{1}{2})^T \frac{Y_i^2}{\sigma_i^2(\boldsymbol{\theta})} = \mathbf{0}, \tag{54}$$

which can, however, only be used to estimate one of the parameters, i.e. only the elements of the sum containing X_{i-1} can take on all real values and, hence, sum to zero. Note that this does not imply that the other parameters cannot be estimated if a transformation of Y_i is applied.

REMARK 4.3. Assuming that σ_e^2 is known, the optimal EFL (31) may be determined from (50) by defining $Z_i = \ln Y_i^2 = X_i + \xi_i - 1.27$. It follows readily that $\mathrm{E}[Z_i|\mathcal{F}_{i-1}] = \phi_0 + \phi_1 X_{i-1} - 1.27$ and $\mathrm{V}[Z_i|\mathcal{F}_{i-1}] = \sigma_e^2 + \frac{\pi^2}{2}$, which yields the EF

$$\sum_{i=1}^{n}(1 \ X_{i-1})^T \frac{(Z_i - \phi_0 - \phi_1 X_{i-1} + 1.27)}{\sigma_e^2 + \frac{\pi^2}{2}} = \mathbf{0}, \tag{55}$$

which may be solved in closed form. ▼

5 Discussion and conclusion

The main result of this paper is that the prediction error method, in particular the special case of a (weighted) least squares method where the inverse of the (conditional) variance is applied as weights, leads to biased estimates provided that the (conditional) variance depends on the parameter vector. However, using the theory of estimating functions, it is possible to derive optimal estimators in the intuitive appealing sense of optimality considered in (Heyde, 1997). For estimating functions from the linear family only explicit expressions for the (conditional) mean and variance are needed, whereas for estimating functions from the quadratic family explicit martingale differences in terms of the third and fourth order moments are also needed. In both cases no distributional assumptions need be imposed.

Some comparisons between prediction error methods and estimating functions have been made. It is shown that the latter only fits in the former framework in a few special cases. However in one special case estimating functions from the linear family may be used to derive an optimal criterion function. Recent developments of socalled *prediction-based estimating functions* bridges the gap between these two methodologies and provides a means of choosing optimal weights in a reasonably general framework, see (Sørensen, 1999) for the general theory, and Nolsøe et al. (2000) for a generalization that allows for measurement noise.

References

Abramowitz, M. and Stegun, N. C. (1970), *Handbook of Mathematical Functions*, Dover Publications, Inc., New York.

Barndorff-Nielsen, O. (1978), *Information and Exponential Families in Statistical Theory*, Wiley, New York.

Bollerslev, T., Chou, R. Y. and Kroner, K. F. (1992), 'ARCH models in finance: A review of the theory and evidence', *Journal of Econometrics* **52**, 5–59.

Engle, R. F. (1982), 'Autoregressive conditional heteroscedasticity with estimates of the U.K. inflation', *Econometrica* **50**(4), 987–1008.

Godambe, V. P. (1960), 'An optimum property of regular maximum-likelihood estimation', *Annals of Mathematical Statistics* **31**, 1208–1211.

Godambe, V. P. E. (1991), *Estimating Functions*, Oxford Science Publications, Oxford.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994), 'Multivariate stochastic variance models', *Review of Economic Studies* **61**, 247–264.

Heyde, C. C. (1997), *Quasi-Likelihood And Its Applications*, Springer Series in Statistics, Springer, New York.

Ljung, L. and Caines, P. E. (1979), 'Asymptotic normality of predictor error estimators for approximate system models', *Stochastics* **3**, 29–46.

McLeish, D. L. and Small, C. G. (1988), *The Theory and Applications of Statistical Inference Functions*, Vol. 44 of *Lecture Notes in Statistics*, Springer, New York.

Nolsøe, K. (1999), Estimating functions and applications, Master's thesis, Institute of Mathematical Modelling, Technical University of Denmark.

Nolsøe, K., Nielsen, J. N. and Madsen, H. (2000), Optimal Weights in Prediction Error and Weighted Least Squares Methods. Submitted.

Sandmann, G. and Koopman, S. J. (1998), 'Estimation of stochastic volatility models via Monte Carlo maximum likelihood', *Journal of Econometrics* **87**, 271–301.

Sørensen, M. (1999), Prediction-Based Estimating Functions, Preprint 1999-5, Department of Theoretical Statistics, University of Copenhagen.

Taylor, S. J. (1986), *Modelling Financial Time Series*, Chicester.