# Estimating Functions with Prior Knowledge, (EFPK) for diffusions

Kim Nolsøe Mathieu Kessler Henrik Madsen

December 16 2003

**Abstract**

In this paper a method is formulated in an estimating function setting for parameter estimation, which allows the use of prior information. The main idea is to use prior knowledge of the parameters, either specified as moments restrictions or as a distribution, and use it in the construction of an estimating function. It may be useful when the full Bayesian analysis is difficult to carry out for computational reasons. This is almost always the case for diffusions, which is the focus of this paper, though the method applies in other settings.

**Keywords:** Small sample size, Estimating Functions, Diffusion Process, Cox Ingersoll & Ross (CIR) Process, Ornstein-Uhlenbeck Process.

## 1 Introduction

Diffusion processes are widely used within engineering, physics, biology and finance since they in many cases are able to give a good description of data using a limited number of parameters. Most importantly the continuous time formulation enables a direct use of any prior physical knowledge in the model formulation, and a direct interpretation of the estimated parameters. Data is in general observed discretely, hence classical time series analysis might be the initial idea for modelling, however irregular sampled data is difficult or impossible to handle using classical time series analysis.

Working with diffusions one is confronted with this problem of determining parameters contained in the coefficients of the stochastic differential equation. Applying maximum likelihood estimation (MLE) the transition density is required. However the transition density is only tractable in a very few special cases in practice. Different numerical approximations have been suggested to deal with the problem of the untractable transition density when doing inference. In [Pedersen, 1995] by applying the approximating Euler scheme, in [Jensen and Poulsen, 2002] by approximating the solution to the Fokker-Planck equation, or by a truncated Hermite type expansion, see [Aït-Sahalia, 2002]. Yet another idea avoiding the often untractable transition density is suggested by estimating functions. Estimating functions turn out to provide an alternative to (MLE) which yields a consistent and asymptotically normal estimator. General theory for estimating functions can be found in [Heyde, 1997]. Introduction to parameter estimation for discretely observed diffusions applying estimating functions is available in by [Bibby and Sørensen, 1995], and [Kessler and Sørensen, 1999].

In the case of few measurements any prior knowledge about the parameters is useful in order to obtain an acceptable precision of the estimates. This calls for Bayesian methods for solving the parameter estimation problem. However in the Bayesian framework we are still left with the problem of determining the often untractable likelihood. In [Johannes and Polson, 2004] and [Cano et al., 2003] approximations to posterior density for diffusions are investigated applying Bayesian inference.

To suggest a method not applying the likelihood and still using a prior, Zellner investigated the Bayesian Method Of Moments (BMOM) [Zellner, 1996]. The main idea is to specify some moment restrictions when the likelihood can not be determined properly, and thereby determining a posterior distribution applying maximum entropy as the optimal information processing rule.

Here we will introduce yet an other alternative, where the aim is not to determine a posterior distribution as suggested by Zellner. Instead the idea is inspired by estimating function theory, we will compensate the possibly poor estimators given few observations by getting as close as possible to the posterior score. This idea yields estimators not being unbiased in the classical sense but having a more reasonable precision.

## 1.1 Framework and notation

In this paper we consider one-dimensional diffusions characterized by

$$dX_t = a(X_t;\theta)dW_t + b(X_t;\theta)dt, X_0 = x_0$$

where $W_t$ is the one-dimensional standard Brownian motion. The state space of $X$ is denoted $\Omega$, $\Omega \subseteq \mathbb{R}$, $\theta$ is a $p$ dimensional vector from the parameter space $\Theta \subseteq \mathbb{R}^p$, the true value of $\theta$ is denoted $\theta_0$, the functions, $a : \Omega \times \Theta \mapsto \mathbb{R}$ and $b : \Omega \times \Theta \mapsto \mathbb{R}$ are known apart from the parameter $\theta$.
$X_0 = x_0$ indicates that the process is known at $t_0$.
We consider data of the form $(X_1, \ldots, X_n) = X_{1:n}$, the density of $X_{1:n}$ is $f(x_{1:n};\theta)$, $f : \Omega^n \times \Theta \mapsto \mathbb{R}$, the prior density of $\theta$ is denoted $\pi(\theta)$, $\pi : \Theta \mapsto \mathbb{R}^p$, the distribution of $X_t$, given $X_s = x$, $t > s$ is denoted $p(t-s, x, y;\theta)$, $p : \mathbb{R} \times \Omega^2 \times \Theta \mapsto \mathbb{R}$.
The diffusion process is a Markov process hence

$$f(x_{1:n};\theta) = \prod_{i=1}^{n} p(\Delta_i, x_i, x_{i-1};\theta)$$

where $\Delta_i = t_i - t_{i-1}$.

Differentiating a function we implicitly assume that the function is differentiable, and when looking for a minimum we assume it exists uniquely. It is also assumed whenever integration and differentiation are interchanged that it is allowed to do so.
The following notation is used for the mean operator

$$\begin{aligned} E_{[\theta,.]}[.] &= \int (.) f(x_{1:n};\theta) dx_{1:n} \\ E_{[.,x]}[.] &= \int (.) \pi(\theta) d\theta \\ E_{[.,.]}[.] &= \int (.) d(f(x_{1:n};\theta)\pi(\theta)), \end{aligned}$$

or equivalently the notation for $\|.\|^2_{L^2}$

$$\begin{aligned} \|.\|^2_{L^2(f(x_{1:n};\theta)dx_{1:n})} &= < .,. >_{L^2(f(x_{1:n};\theta)dx_{1:n})} = \int |.|^2 f(x_{1:n};\theta) dx_{1:n} \\ \|.\|^2_{L^2(\pi(\theta)d\theta)} &= \int |.|^2 \pi(\theta) d\theta \\ \|.\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} &= \int |.|^2 d(f(x_{1:n};\theta)\pi(\theta)), \end{aligned}$$

The F-Optimal estimating function from a fixed sample size we denote, $G^*(X_{1:n};\theta) \in \mathcal{G}$, same notation as in [Heyde, 1997], also the standardized estimating function notation from the same book is used $G^{(s)}(.) = -E_{[\theta,.]}[\partial_\theta G(.)]^T E_{[\theta,.]}[G(.)G(.)^T]G$, where $T$ indicate the transposed and $\partial_\theta G(.)$ indicate that $G(.)$ is differentiated with respect to $\theta$

$$\partial_\theta f = (\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_p})^T.$$

We define three classes of estimating functions, that are assumed to be closed under addition

$$\begin{aligned}
\text{Class}\, \mathcal{H} &= \{H : (\theta; X_{1:n}) \mapsto H(\theta; X_{1:n}) \\
\text{Class}\, \mathcal{F} &= \{F : \theta \mapsto F(\theta) \quad \text{s.t. F is of a form}\, F(\theta) = k(\theta) \\
\text{Class}\, \mathcal{G} &= \{G : (\theta; X_{1:n}) \mapsto G(\theta; X_{1:n}) \quad \text{s.t.}\, \exists\, H \in \mathcal{H},\, \exists\, F \in \mathcal{F} \\
&\quad\ \text{with}\, G(\theta; X_{1:n}) = H(\theta; X_{1:n}) + F(\theta).
\end{aligned}$$

### 1.2 Overview of the paper

The basic idea behind Estimating Function with Prior Knowledge (EFPK) is introduced in section 2. In section 3 the optimality criterion is described. In section 4 the parameter estimation for the Ornstein-Uhlenbeck process using (EFPK) from the linear family is dealt with and results are compared to those obtained using classical linear estimating functions and (MAP) estimators. Similar comparisons are found in section 5 for the well known (CIR) process from finance. Finally section 6 contains concluding remarks.

## 2 Introducing Estimating Function with Prior Knowledge

This paper describes a method for parameter estimation, refereed to as Estimating Functions with Prior Knowledge (EFPK). The method yields estimates where prior knowledge is incorporated. The idea behind (EFPK) is inspired by theory from estimation functions and theory from Bayesian analysis. Estimates are determined without having to fully specify the often untractable transition density while the estimating function is as close as possible to the posterior score (MAP).

Using (EFPK) perhaps little is gained having many observations and limited prior information, which for example often will be the case in finance, however, when only a few samples are available, which likely is the case when observations are difficult or costly to get, the estimates applying ordinary frequentistic analysis, can be very unreliable compared to methods which takes some prior knowledge into account, and in this case (EFPK) could be a method worth considering.

Another area where (EFPK) for parameter estimation might be worth considering is in a population setup. Such as a situation where little data is available about an individual but data from somehow similar individuals are available. In this case (EFPK's) yields estimates which combines knowledge from the population with knowledge from the individual's. We will shortly demonstrate how to apply idea behind (EFPK) for a setup where observations are normal distributed and the prior knowledge of the parameter being estimated is normal distributed.

In the classical setting estimating functions are created such that $H(\theta; X_{1:n}) \in \mathcal{H}$ and

$$E_{[\theta,.]}[H(\theta; X_{1:n})] = 0, \quad \theta = \theta_0$$

and the estimator is found by solving the estimating equation

$$H(\hat{\theta}; X_{1:n}) = 0.$$

Basically what we want to do is to extend the estimating equation

$$H(\hat{\theta}; X_{1:n}) = 0$$

to

$$G(\hat{\theta}; X_{1:n}) = 0, G(\theta; X_{1:n}) \in \mathcal{G}$$

such that prior knowledge is incorporated into the estimating equation. For this equation it is no longer true that $E_{[\theta,.]}[G(\theta; X_{1:n})] = 0$ is fulfilled for $\theta = \theta_0$, instead by construction $E_{[.,.]}[G(\theta; X_{1:n})] = 0$ is correct.

To motivate the idea behind (EFPK's) consider the following mixed effect model where both the individual parameters and the population parameter are normal distributed, i.e.

$$\begin{aligned} X_i|\mu &\sim N(\mu, \sigma^2) \\ \mu &\sim N(\mu_1, \sigma_1^2). \end{aligned}$$

It is well known that estimates of the individual parameters can benefit from using prior knowledge about the population. Given the prior density for $\mu$ the Best Linear Unbiased Prediction estimator (BLUP) can be shown to be

$$\hat{\mu} = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{\sum_{i=1}^n X_i}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{n}{\sigma^2}} \tag{1}$$

simply by maximizing the posterior score, see [Karunamuni, 2002] for a details.

Using (EFPK) the aim is to construct an estimating function where the weights $A(\mu)$ and $B$ somehow are chosen in an optimal way

$$G^*_{EFPKL}(\mu) = \sum_{i=1}^n A^*(\mu)(X_i - \mu) + B^*(\mu - \mu_1).$$

Indeed choosing $A^*(\mu) = \frac{1}{\sigma^2}$ and $B^* = -\frac{1}{\sigma_1^2}$ and solving the estimating equation $G^*_{EFPKL}(\hat{\mu}) = 0$, the same estimator for $\hat{\mu}$ is obtained as in (1).

In a similar way estimators for diffusions will be investigated by expanding the estimating functions such that prior knowledge is taken into account, optimal estimating equations will be derived both for the The Ornstein-Uhlenbeck process $G^*_{EFPKL}(\theta)$ and the Cox Ingersoll & Ross (CIR) process $G^{\dagger^*}_{EFPKL}(\theta)$ by determining the optimal weights $A^*(\Delta_i, X_{i-1}, \theta), B^*$ and $A^{\dagger^*}(\Delta_i, X_{i-1}, \theta), B^{\dagger^*}$ in respectively

$$\begin{aligned} G^*_{EFPKL}(\theta) &= B^*(\theta - \alpha) + \sum_{i=1}^n A^*(\Delta_i, X_{i-1}, \theta)(X_i - X_{i-1}e^{-\theta\Delta_i}) \\ G^{\dagger^*}_{EFPKL}(\theta) &= B^{\dagger^*}(\theta - \alpha) + \sum_{i=1}^n A^{\dagger^*}(\Delta_i, X_{i-1}, \theta)(X_i - X_{i-1}e^{-\theta\tau} - \alpha(1 - e^{-\theta\tau})). \end{aligned}$$

A simulation study of the Ornstein-Uhlenbeck process and the Cox Ingersoll & Ross (CIR) process is carried out to justify the claim that (EFPK) outperforms ordinary frequentistic inference in certain cases as described above.

# 3 Optimality Criterion

This section will describe the optimality criterion used for (EFPK) and derive an expression for the optimal (EFPK). The main idea is to minimize the $L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)$ distance to the posterior score instead of as traditionally to minimize the $L^2(f(x_{1:n};\theta)dx_{1:n})$ distance to the score function. In subsection 3.1 the basic idea behind optimal estimating functions in the classical sense is briefly sketched.

## 3.1 Optimality Criterion, in the classical setting, a brief review

In the classical setting, estimating functions are constructed such that

$$E_{[\theta,.]}[G(\theta;X_{1:n})] = 0,$$

for $\theta = \theta_0$, estimating functions with this property are called unbiased.

Definition 3.1. *$G^*(\theta;X_{1:n}), \in \mathcal{H}$ is F-Optimal in $\mathcal{H}$ if*

$$\frac{E_{[\theta,.]}[G(\theta;X_{1:n})^2]}{(E_{[\theta,.]}[\partial_\theta G(\theta;X_{1:n})])^2} \geq \frac{E_{[\theta,.]}[G^*(\theta;X_{1:n})^2]}{(E_{[\theta,.]}[\partial_\theta G^*(\theta;X_{1:n})])^2}$$

*for all $\theta \in \Theta$ and for all $G(\theta;X_{1:n}) \in \mathcal{H}$.*

It can be shown that the $G^{**}(\theta;X_{1:n}), \in \mathcal{H}$ with the shortest $L^2(f(x_{1:n};\theta)dx_{1:n})$ distance to the score function, i.e.

$$\|G^{**}(\theta;X_{1:n}) - U(\theta;X_{1:n})\|^2_{L^2(f(x_{1:n};\theta)dx_{1:n})} \leq \|G(\theta;X_{1:n}) - U(\theta;X_{1:n})\|^2_{L^2(f(x_{1:n};\theta)dx_{1:n})}, \quad (2)$$

$\forall G(\theta;X_{1:n}) \in \mathcal{H}, \forall\theta \in \Theta$ is F-Optimal hence $G^{**}(\theta;X_{1:n}) = G^*(\theta;X_{1:n})$ see [Godambe and Heyde, 1987].

For Markov processes we will chose $\mathcal{H} = \sum_{j=1}^{N} \beta_j h_j(\Delta, x, y; \theta)$ with the normal conditions fulfilled s.t.

$$H^*(\theta;X_{1:n}) = \sum_{i=1}^{n} g^*(\Delta_i, X_i, X_{i-1};\theta),$$

where $g^* = (g_1^*, \ldots, g_p^*)$ and $g_i^*$ is the orthogonal projection w.r.t. $< .,. >$ of $y \mapsto \partial_\theta logp(\Delta, x, y; \theta)$ onto $\mathcal{H}$. $H^*(\theta;X_{1:n})$ is F-Optimal see [Kessler, 1995]. With the simplest possible choice of $\sum_{j=1}^{N} \beta_j h_j(\Delta_i, x, y; \theta) = \alpha(\Delta_i, X_{i-1}; \theta) h(\Delta_i, X_i, X_{i-1}; \theta)$

$$g^*(\Delta_i, X_i, X_{i-1}; \theta) = g_1^*(\Delta_i, X_i, X_{i-1}; \theta) = \alpha^*(\Delta_i, X_{i-1}; \theta) h(\Delta_i, X_i, X_{i-1}; \theta),$$

with

$$\alpha^*(\Delta_i, X_{i-1}; \theta) = -\frac{E_{[.,\theta]}[\partial_\theta h(\Delta_i, X_i, X_{i-1}; \theta)^T]}{E_{[.,\theta]}[h(\Delta_i, X_i, X_{i-1}; \theta) h(\Delta_i, X_i, X_{i-1}; \theta)]}. \tag{3}$$

## 3.2 Optimal Estimating Functions with Prior Knowledge

For (EFPK) the optimal one is found by minimizing the $L^2(f(x_{1:n}; \theta)\pi(\theta) dx_{1:n} d\theta)$ distance to the posterior score, i.e.

Definition 3.2. *The optimal (EFPK) $G^*(\theta; X_{1:n}) \in \mathcal{G}$ is the one which satisfies*

$$\|G^*(\theta; X_{1:n}) - U(\theta; X_{1:n})\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \leq \|G(\theta; X_{1:n}) - U(\theta; X_{1:n})\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \tag{4}$$

$\forall G(\theta; X_{1:n}) \in \mathcal{G}, \forall \theta \in \Theta.$

Proposition 3.1. *The optimal Estimating Function with Prior Knowledge is*

$$G^*(\theta; X_{1:n}) = H^*(\theta; X_{1:n}) + F^*(\theta)$$

*where $H^*(\theta; X_{1:n})$ minimizes (5)*

$$\|H(\theta; X_{1:n}) - \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}\|^2_{L^2(f(x_{1:n};\theta)dx_{1:n})} \tag{5}$$

$\forall \theta \in \Theta$, $\forall H(\theta; X_{1:n}) \in \mathcal{H}$, *and $F^*(\theta)$ minimizes (6)*

$$\|F(\theta) - \frac{\partial_\theta \pi(\theta)}{\pi(\theta)}\|^2_{L^2(\pi(\theta)d\theta)} \tag{6}$$

$\forall \theta \in \Theta$, $\forall F(\theta) \in \mathcal{F}$.

**Note** Explicit expression of $H^*(\theta; X_{1:n})$ and $F^*(\theta)$ are well known to be $H^*(\theta; X_{1:n}) = \sum_{i=1}^{n} -\frac{E_{[\theta,.]}[\partial_\theta h(\Delta_i, X_i, X_{i-1}; \theta)^T]}{E_{[\theta,.]}[h(\Delta_i, X_i, X_{i-1}; \theta) h(\Delta_i, X_i, X_{i-1}; \theta)]} h(\Delta_i, X_i, X_{i-1}; \theta)$ and $F^*(\theta) = -E_{[.,x]}[\partial_\theta k(\theta)^T](E_{[.,x]}[k(\theta)k(\theta)^T])^{-1} k(\theta)$, for Markov processes with a simple choice $g^*$, see above.

*Proof.* Let $G(\theta; X_{1:n})$ be an estimating function from $\mathcal{G}$

$$G(\theta; X_{1:n}) = H(\theta; X_{1:n}) + F(\theta)$$

where $F(\theta)$ does not depend on $X_{1:n}$.

First we will rewrite the expression of the posterior score

$$U(X_{1:n}; \theta) = \partial_\theta ln(f(x_{1:n}; \theta)\pi(\theta)) = \partial_\theta ln(l(\theta; X_{1:n})\pi(\theta)) = \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})} + \frac{\partial_\theta \pi(\theta)}{\pi(\theta)}, \quad (7)$$

inserting the expression from (7) in

$$\|G(\theta; X_{1:n}) - U(X_{1:n}; \theta)\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)}$$

yields

$$\begin{aligned}
& \|H(\theta; X_{1:n}) - \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \\
& +\|F(\theta) - \frac{\partial_\theta \pi(\theta)}{\pi(\theta)}\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \\
& +2 < H(\theta; X_{1:n}) - \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}, F(\theta) - \frac{\partial_\theta \pi(\theta)}{\pi(\theta)} >_{L^2(f(x_{1:n};\theta)dx_{1:n})\pi(\theta)d\theta)} \Rightarrow \\
\|G(\theta; X_{1:n}) \quad - \quad & U(X_{1:n}; \theta)\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \\
= \quad & \|H(\theta; X_{1:n}) - \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \\
& +\|F(\theta) - \frac{\partial_\theta \pi(\theta)}{\pi(\theta)}\|^2_{L^2(\pi(\theta)d\theta)} \qquad (8)
\end{aligned}$$

since

$$E_{[.,x]}[H(\theta; X_{1:n})] = E_{[.,x]}[\frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}] = 0.$$

From (8) it is concluded that

$$G^*(\theta; X_{1:n}) = H^{**}(\theta; X_{1:n}) + F^*(\theta),$$

where $H^{**}(\theta; X_{1:n})$ minimizes (9) for all $\theta \in \Theta$ and all $H(\theta; X_{1:n}) \in \mathcal{G}$

$$\|H(\theta; X_{1:n}) - \frac{\partial_\theta l(\theta; X_{1:n})}{l(\theta; X_{1:n})}\|^2_{L^2(f(x_{1:n};\theta)\pi(\theta)dx_{1:n}d\theta)} \quad (9)$$

and $F^*(\theta)$ minimizes (6) for all $\theta \in \Theta$ and all $F(\theta) \in \mathcal{F}$. Next we need to prove that

$$H^{**}(\theta; X_{1:n}) = H^*(\theta; X_{1:n})$$

This is however straightforward since if $H^*(\theta; X_{1:n})$ solves (2) for all $\theta \in \Theta$ and all $H(\theta; X_{1:n}) \in \mathcal{G}$, the inequality is still fulfilled integrating both sides w.r.t. $\pi(\theta)d\theta$ □

# 4 The Ornstein-Uhlenbeck Process

Consider the Ornstein-Uhlenbeck process given by the stochastic differential equation

$$dX_t = -\theta(X_t - \alpha)dt + \sigma dW_t, \tag{10}$$

which often is used in the literature to model an exponentially decaying function with process noise. An explicit solution to (10) is readily obtained by using standard Itô integration

$$\begin{aligned} X_{t+\tau} &= (X_t - \alpha)e^{-\theta\tau} + \alpha + \sigma \int_t^{t+\tau} e^{-\theta(s-\tau-t)} dW_s \\ &\sim N((X_t - \alpha)e^{-\theta\tau} + \alpha, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta\tau})). \end{aligned}$$

Having the solution, standard maximum likelihood theory can be applied to estimate the parameters in the diffusion. For the (O-U) process estimates applying maximum likelihood theory and (EFL) turn out to result in the same estimator. In the following subsections a description of the simulation study is presented. Next equations to determine estimates using (EFL), (EFPKL) and (MAP) are shown. Finally results from the simulation study is listed in Table 2.

Obviously when the (MAP) is easily derived which is the case for the Ornstein-Uhlenbeck process the (MAP) will be preferred to (EFPKL), however from the simulation study we note that estimators using (EFPKL) approximates (MAP) rather well, see Table 2.

## 4.1 The simulation study

Consider now a population experiment where the Ornstein-Uhlenbeck process is used to describe the evolution in time of the concentration of some injected chemical drug. In each experiment the population size of 10 individuals is chosen, the process describing each individual is created using the same parameters except for the parameter $\theta_j$ which differs for each process. In each experiment data is created by first drawing a $\theta_j$ from the prior and then simulating from (10) using $\theta_j$, $\alpha$, and $\sigma$. In Figure 1 (top, left) and (top, right) two different experiments are presented. For the experiment shown in Figure (top, left), $\theta_j$ is chosen from a prior with smaller variance compared to the prior in the experiment in Figure (top, right). As a consequence the processes in (top, left) are more alike.

Finally $\theta_j$ is estimated for each process $j$ and each experiment applying the methods mentioned above. To compare the methods we have chosen to calculate the sum of the squared differences between the estimates and the true value of $\theta_j$. For the data presented in Figure 1 (top, left) and (top, right) the estimates applying the different methods described previously is presented respectively (bottom, left) and (bottom, right). Table 2 contains results for a more thorough investigation, here the different estimators are compared for 24 different experiments.

## 4.2 Estimating Functions from the Linear family (EFL)

In order to estimate $\theta_j$ using estimating function from the linear family we determine $E[X_{t+\tau}|X_t]$ and $Var[X_{t+\tau}|X_t]$ which is straightforward having the solution i.e.

$$E[X_{t+\tau}|X_t] = (X_t - \alpha)e^{-\theta\tau} + \alpha \tag{11}$$

$$Var[X_{t+\tau}|X_t] = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta\tau}). \tag{12}$$

The F-Optimal linear estimating function from the process defined by (10), is

$$G^*_{EFL}(\theta_j) = \sum_{i=1}^{n} \frac{-\Delta_i(X_{i-1} - \alpha)e^{-\theta\Delta_i}}{\frac{\sigma^2}{2\theta_j}(1 - e^{-2\theta_j\Delta_i})}(X_i - (X_{i-1} - \alpha)e^{-\theta_j\Delta_i} - \alpha),$$

hence the $\hat{\theta}_{j\,EFL}$ is

$$\hat{\theta}_{j\,EFL} = \frac{1}{\Delta} ln(\frac{\sum_{i=1}^{n}(X_{i-1} - \alpha)(X_{i-1} - \alpha)}{\sum_{i=1}^{n-1}(X_i - \alpha)(X_{i-1} - \alpha)})$$

when $\Delta_i = \Delta$, otherwise an explicit expression is not possible to derive.

## 4.3 Estimating Functions with Prior Knowledge from the Linear family (EFPKL)

Next (EFPKL) is applied on the data sets from Figure 1 (top, left) and (top, right). We will assume that the prior knowledge of the parameter $\theta_j$ is $\theta_j \sim N(\theta_0, \sigma_1^2)$. Inserting the expressions of the conditional moments from (11) and (12) in (5) and (6) the following expression is obtained

$$G^*_{EFPKL}(\theta_j) = \frac{(\theta_j - \theta_0)}{\sigma_1^2} + \sum_{i=1}^{n} \frac{-\Delta_i X_{i-1} e^{-\theta_j \Delta_i}}{\frac{\sigma^2}{2\theta_j}(1 - e^{-2\theta_j \Delta_i})}(X_i - X_{i-1}e^{-\theta_j \Delta_i}).$$

As it appears is not possible to find an explicit expression of $\hat{\theta}_{j_{EFPKL}}$.

## 4.4 Maximum Posterior estimates (MAP)

Since an explicit expression of the solution to (10) is available an equation maximizing the posterior score can be found, after some trivial calculations the following equation is obtained

$$\begin{aligned}
\frac{\partial lnp}{\partial \theta_j} &= \frac{(\theta_j - \theta_0)}{\sigma_1^2} + \frac{(n-1)}{2\theta_j} + \\
&\sum_{i=1}^{n} \frac{\Delta_i \, e^{-2\,\theta_j\,\Delta_i}}{(e^{-2\,\theta_j\,\Delta_i} - 1)} + 2\,\frac{\left(X_i - (X_{i-1} - \alpha)\, e^{-\theta_j\,\Delta_i} - \alpha\right)\theta_j \; (X_{i-1} - \alpha)\, \Delta_i \, e^{-\theta_j\,\Delta_i}}{\sigma^2\,(e^{-2\,\theta_j\,\Delta_i} - 1)} \\
&\frac{\left(X_i - (X_{i-1} - \alpha)\, e^{-\theta_j\,\Delta_i} - \alpha\right)^2 (2\theta_j\,\Delta_i\, e^{-2\,\theta_j\,\Delta_i} + \left(e^{-2\,\theta_j\,\Delta_i} - 1\right))}{\sigma^2\,(e^{-2\,\theta_j\,\Delta_i} - 1)^2} = 0.
\end{aligned}$$

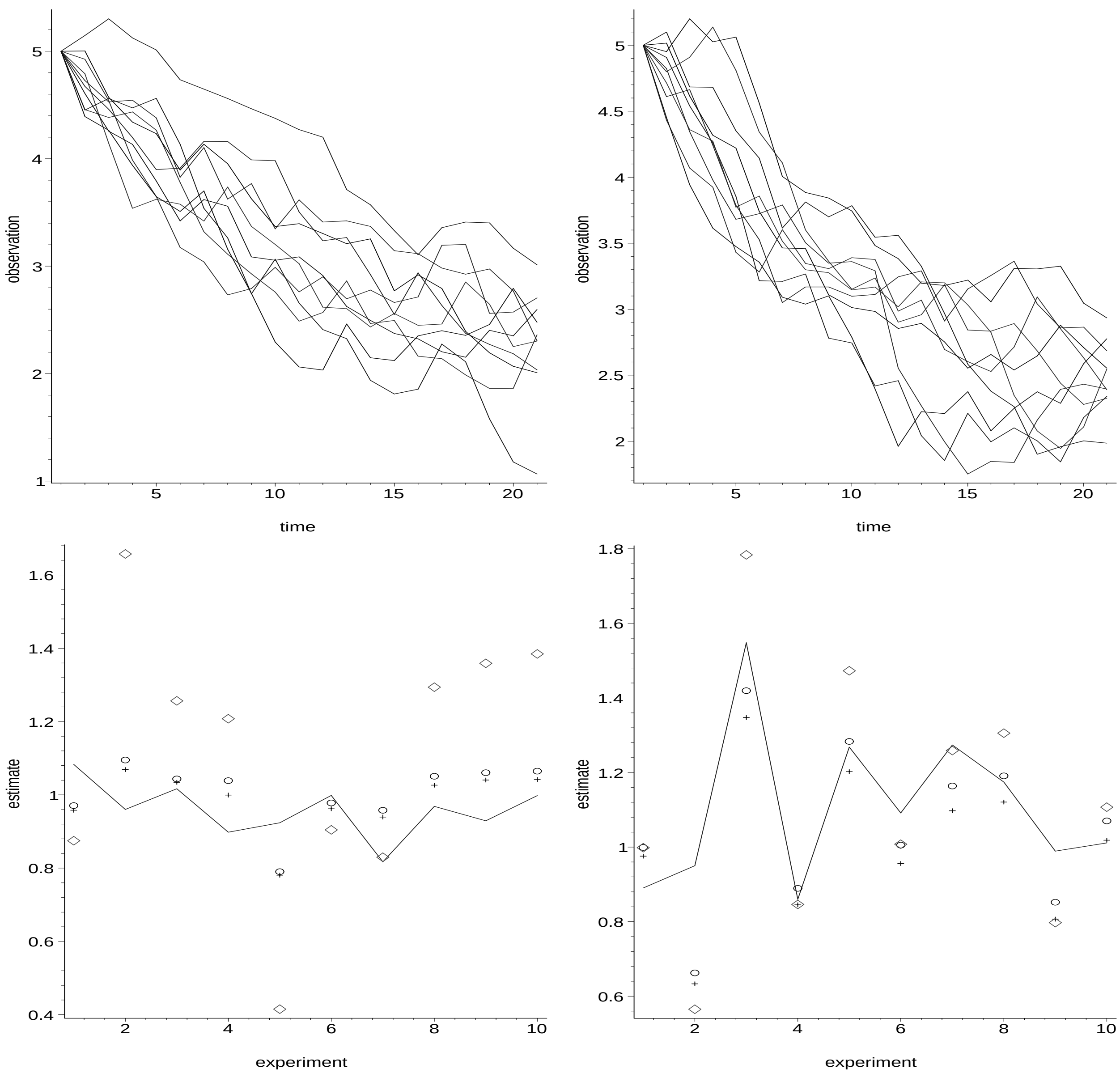


Table 1: (top, left) represents 10 trajectories generated from the (O-U) process described by (10) with the parameters: number of observations=20, $\Delta_i = 0.1$, $\alpha = 2$, $\sigma^2 = 0.09$, $\sigma_1^2 = 0.005$ and $\theta_i \sim N(1, \sigma_1^2)$. (top, right) represents 10 trajectories generated from the (O-U) process described by (10) with the parameters: number of observations=20, $\Delta_i = 0.1$, $\alpha = 2$, $\sigma^2 = 0.09$, $\sigma_1^2 = 0.04$ and $\theta_i \sim N(1, \sigma_1^2)$. (bottom, left) and (bottom, right) represents estimates obtained using the data from the (top) Figures. The following symbols has been used (EFL) ($\diamond$), (EFPKL) ($\circ$) and maximization of the posterior score (+). Solid lines ending points represent the true value of $\theta_j$.

| n | $\Delta$ | $\sigma^2$ | $\sigma_1^2$ | $\sum_{j=1}^{m}(\hat{\theta}_{j\,EFL} - \theta_j)^2$ | $\sum_{j=1}^{m}(\hat{\theta}_{j\,EFPKL} - \theta_j)^2$ | $\sum_{j=1}^{m}(\hat{\theta}_{j\,MAP} - \theta_j)^2$ |
|---|---|---|---|---|---|---|
| 5 | 0.1 | 0.01 | 0.005 | .13134 | .06668 | .06836 |
| | | | 0.04 | .31546 | .27171 | .27231 |
| | | 0.09 | 0.005 | 1.35790 | .10775 | .10618 |
| | | | 0.04 | 1.25224 | .40624 | .38911 |
| | 0.5 | 0.01 | 0.005 | .05606 | .03579 | .03563 |
| | | | 0.04 | .02425 | .02015 | .02000 |
| | | 0.09 | 0.005 | .39042 | .04523 | .04262 |
| | | | 0.04 | .13845 | .09696 | .09712 |
| 20 | 0.1 | 0.01 | 0.005 | .14033 | .06218 | .06181 |
| | | | 0.04 | .12693 | .11796 | .12286 |
| | | 0.09 | 0.005 | .55852 | .03351 | .03883 |
| | | | 0.04 | .32418 | .23436 | .19109 |
| | 0.5 | 0.01 | 0.005 | .05372 | .01512 | .01379 |
| | | | 0.04 | .04980 | .04708 | .04719 |
| | | 0.09 | 0.005 | .34615 | .04525 | .04411 |
| | | | 0.04 | .43681 | .37527 | .36348 |
| 50 | 0.1 | 0.01 | 0.005 | .06107 | .04617 | .04520 |
| | | | 0.04 | .31646 | .29114 | .27900 |
| | | 0.09 | 0.005 | .56771 | .04689 | .04947 |
| | | | 0.04 | .49285 | .23729 | .47473 |
| | 0.5 | 0.01 | 0.005 | .03777 | .02780 | .02569 |
| | | | 0.04 | .08272 | .05237 | .04882 |
| | | 0.09 | 0.005 | .25874 | .02176 | .02791 |
| | | | 0.04 | .93075 | .37642 | .24393 |

Table 2: Comparison between the different estimators for the Ornstein-Uhlenbeck process for 24 different experiments. The parameters used to create data for each experiment is presented in the first 4 columns, the next 3 columns contains the sum of squared difference between the estimate $\hat{\theta}_j$ and the true value $\theta_j$ for each estimator in each experiment.

# 5 The Cox, Ingersoll and Ross (CIR) process

Consider the Cox, Ingersoll and Ross (CIR) diffusion process given by the stochastic differential equation

$$dX_t = -\theta(X_t - \alpha)dt + \sigma\sqrt{X_t}dW_t, \tag{13}$$

which often in the literature is used to model short term interest rates, see [Cox et al., 1985]. It can be shown that the transition density is a non central chi-square distribution with non-integer parameters. However, applying (EFL) is straightforward to obtain estimates since analytical expression for conditional moments in the (CIR) process is straightforward to find. After some calculations we obtain

$$E[X_{t+\tau}|X_t] = X_t e^{-\theta\tau} + \alpha(1 - e^{-\theta\tau}) \tag{14}$$

$$Var[X_{t+\tau}|X_t] = \frac{\sigma^2}{2\theta}(1 - e^{-\theta\tau})(\alpha(1 - e^{-\theta\tau}) + 2X_t e^{-\theta\tau}). \tag{15}$$

## 5.1 Estimating functions from the Linear family (EFL)

The F-Optimal linear estimating function for sampled realizations from (13), is

$$\begin{aligned} G^*_{EFL}(\theta_j) &= \sum_{i=1}^{n} \frac{-\Delta_i(X_{i-1} - \alpha)e^{-\theta_j\Delta_i}(X_i - (X_{i-1} - \alpha)e^{-\theta_j\Delta_i} - \alpha)}{\frac{\sigma^2}{2\theta_j}(1 - e^{-\theta_j\Delta_i})(\alpha(1 - e^{-\theta_j\Delta_i}) + 2X_{i-1}e^{-\theta_j\Delta_i})} \\ &= \sum_{i=1}^{n} \frac{(X_{i-1} - \alpha)(X_i - (X_{i-1} - \alpha)e^{-\theta_j\Delta_i} - \alpha)}{\alpha(1 - e^{-\theta_j\Delta_i}) + 2X_{i-1}e^{-\theta_j\Delta_i}}. \end{aligned} \tag{16}$$

Given this expression it is not possible to derive an explicit expression of the estimator of $\theta_j$. Estimating the parameter $\alpha$ is also straightforward applying the same procedure we obtain the estimating equation

$$\begin{aligned} G^*_{EFL}(\alpha_j) &= \sum_{i=1}^{n} \frac{(1 - e^{-\theta\Delta_i})(X_i - (X_{i-1} - \alpha_j)e^{-\theta\Delta_i} - \alpha_j)}{\frac{\sigma^2}{2\theta}(1 - e^{-\theta\Delta_i})(\alpha_j(1 - e^{-\theta\Delta_i}) + 2X_{i-1}e^{-\theta\Delta_i})} \\ &= \sum_{i=1}^{n} \frac{(X_{i-1} - \alpha_j)(X_i - (X_{i-1} - \alpha_j)e^{-\theta\Delta_i} - \alpha_j)}{\alpha_j(1 - e^{-\theta\Delta_i}) + 2X_{i-1}e^{-\theta\Delta_i}}. \end{aligned} \tag{17}$$

Estimating $\sigma^2$ is not possible applying (EFL) since the first conditional moment does not depend on $\sigma^2$. However applying estimating functions from the Quadratic Family (EFQ) it is possible to create an estimating function to estimate $\sigma^2$ see [Bibby and Sørensen, 1995]. Creating the estimating equation from the EFQ it is possible to estimate $\theta_j$, $\alpha_j$ and $\sigma^2$ simultaniously see [Bibby and Sørensen, 1995] however these estimating equations do not yield equations from where explicit expressions for the estimators can be found. Keeping the Martingale property but using weights not being F-Optimal, as investigated in [Bibby and Sørensen, 1995], equations are created from where explicit expressions can be found. In [Pedersen, 2000] explicit expressions is derived for parameters in a (CIR) process using Martingale estimating functions not being F-Optimal, we will not investigate these estimating functions further. Also note that it is straightforward to do the calculations in [Bibby and Sørensen, 1995] in a (EFPK) framework estimating all parameters simultaneously incorporating prior knowledge.

## 5.2 Estimating Functions with Prior Knowledge from the Linear family (EFPKL)

Assume that the prior knowledge of is $\theta_j \sim N(\theta_0, \sigma_1^2)$ and is $\alpha_j \sim N(\alpha_0, \sigma_2^2)$. Inserting the expression for the conditional moments from (14) and (15) in (5) and (6) yields

$$\begin{aligned}
G^*_{EFPKL}(\theta_j) &= \frac{(\theta_j - \theta_0)}{\sigma_1^2} + \frac{-\Delta_i e^{-\theta_j \Delta_i}}{\frac{\sigma^2}{2\theta_j}(1 - e^{-\theta_j \Delta_i})} \sum_{i=1}^{n} \frac{(X_{i-1} - \alpha)(X_i - (X_{i-1} - \alpha)e^{-\theta_j \Delta_i} - \alpha)}{\alpha(1 - e^{-\theta_j \Delta_i}) + 2X_{i-1}e^{-\theta_j \Delta_i}} \\
G^*_{EFPKL}(\alpha_j) &= \frac{(\alpha_j - \alpha_0)}{\sigma_2^2} + \sum_{i=1}^{n} \frac{(1 - e^{-\theta \Delta_i})(X_i - (X_{i-1} - \alpha_j)e^{-\theta \Delta_i} - \alpha_j)}{\frac{\sigma^2}{2\theta}(1 - e^{-\theta \Delta_i})(\alpha_j(1 - e^{-\theta \Delta_i}) + 2X_{i-1}e^{-\theta \Delta_i})} \\
&= \frac{(\alpha_j - \alpha_0)}{\sigma_2^2} + \frac{(1 - e^{-\theta \Delta_i})}{\frac{\sigma^2}{2\theta}(1 - e^{-\theta \Delta_i})} \sum_{i=1}^{n} \frac{(X_i - (X_{i-1} - \alpha_j)e^{-\theta \Delta_i} - \alpha_j)}{(\alpha_j(1 - e^{-\theta \Delta_i}) + 2X_{i-1}e^{-\theta \Delta_i})}. \qquad (18)
\end{aligned}$$

(18) is not possible to solve with respect to $\alpha_j$ but a comparison between the estimators obtained from (17) and (18) has been carried out by simulation in the same manner as for the Ornstein-Uhlenbeck process, see Figure 3

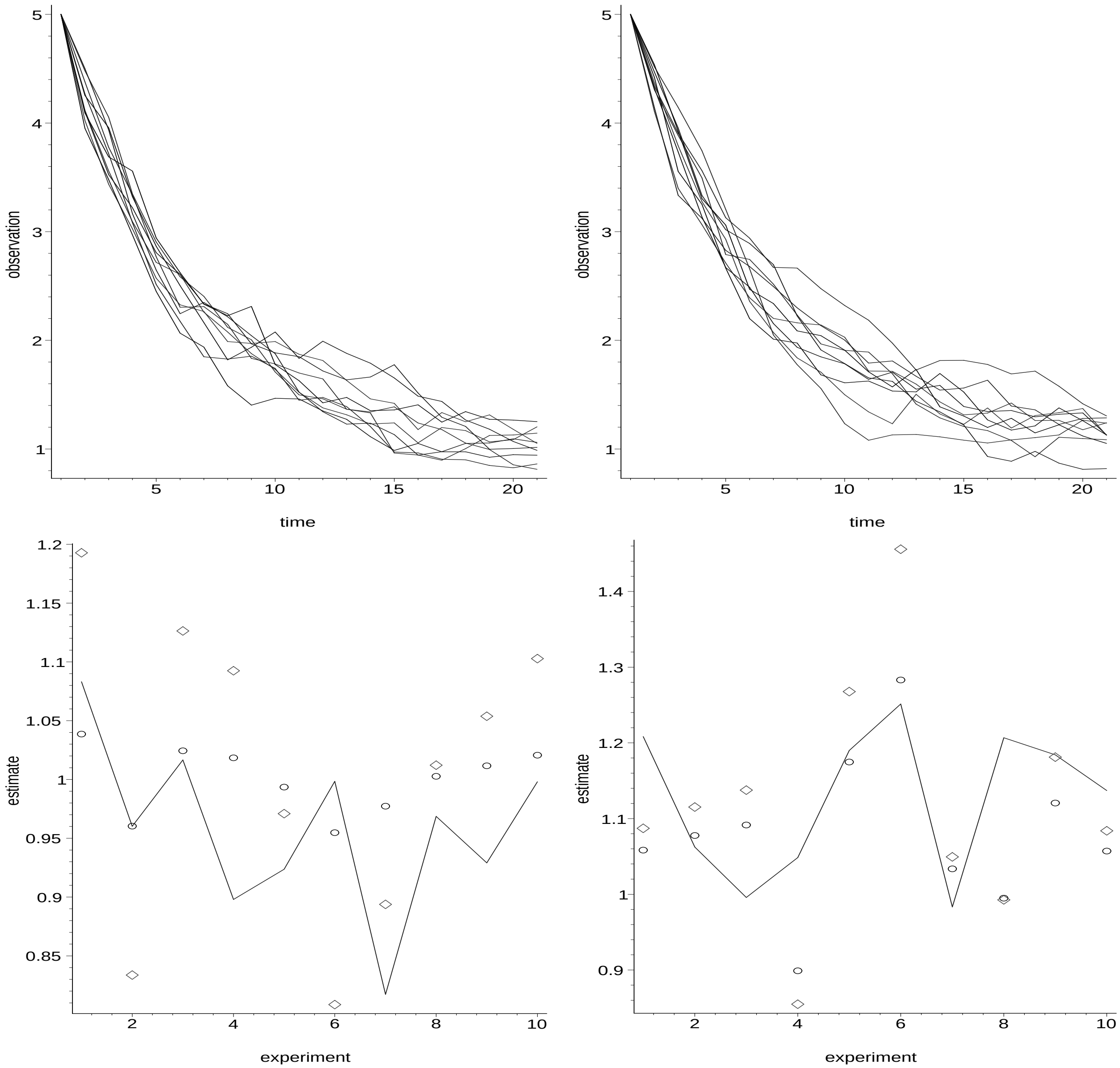


Table 3: (top, left) represents 10 trajectories generated from the (CIR) process described by (13) with the parameters: number of observations=20, $\Delta_i = 0.1$, $\theta = 2$, $\sigma^2 = 0.09$, $\sigma_2^2 = 0.005$ and $\alpha_i \sim N(1, \sigma_2^2)$. (top, right) represents 10 trajectories generated from the (CIR) process described by (13) with the parameters: number of observations=20, $\Delta_i = 0.1$, $\theta = 2$, $\sigma^2 = 0.09$, $\sigma_2^2 = 0.04$ and $\alpha_i \sim N(1, \sigma_2^2)$. (bottom, left) and (bottom, right) represents estimates obtained using the data from the (top) Figures. The following symbols has been used (EFL) ($\diamond$), (EFPKL) ($\circ$). Solid lines ending points represent the true value of $\theta_j$.

| n | $\Delta$ | $\sigma^2$ | $\sigma_1^2$ | $\sum_{j=1}^{m}(\hat{\theta}_{j\,EFL} - \theta_j)^2$ | $\sum_{j=1}^{m}(\hat{\theta}_{j\,EFPKL} - \theta_j)^2$ |
|---|---|---|---|---|---|
| 5 | 0.1 | 0.01 | 0.005 | .06926 | .02631 |
| | | | 0.04 | .17051 | .11770 |
| | | 0.09 | 0.005 | 1.51666 | .06233 |
| | | | 0.04 | 1.97042 | .34325 |
| | 0.5 | 0.01 | 0.005 | .03002 | .02536 |
| | | | 0.04 | .01596 | .01335 |
| | | 0.09 | 0.005 | .13827 | .03564 |
| | | | 0.04 | .09817 | .10801 |
| 20 | 0.1 | 0.01 | 0.005 | .01907 | .00747 |
| | | | 0.04 | .02619 | .02625 |
| | | 0.09 | 0.005 | .14172 | .02714 |
| | | | 0.04 | .18824 | .07697 |
| | 0.5 | 0.01 | 0.005 | .00298 | .00258 |
| | | | 0.04 | .00726 | .00725 |
| | | 0.09 | 0.005 | .04661 | .01357 |
| | | | 0.04 | .07527 | .06393 |
| 100 | 0.1 | 0.01 | 0.005 | .00584 | .00450 |
| | | | 0.04 | .00251 | .00255 |
| | | 0.09 | 0.005 | .02229 | .01532 |
| | | | 0.04 | .06257 | .03467 |
| | 0.5 | 0.01 | 0.005 | .00168 | .00153 |
| | | | 0.04 | .00072 | .00073 |
| | | 0.09 | 0.005 | .00991 | .00797 |
| | | | 0.04 | .00776 | .00740 |

Table 4: Comparison between the different estimators for the (CIR) process for 24 different experiments. The parameters used to create data for each experiment is presented in the first 4 columns, the next 2 columns contains the sum of squared difference between the estimate $\hat{\theta}_j$ and the true value $\theta_j$ for each estimator in each experiment.

# 6 Conclusion

The proposed method Estimating Functions with Prior Knowledge (EFPK) constitutes a method for parameter estimation which incorporates prior knowledge in the estimates.

This is done by adding an additional term to the ordinary estimating equation. Adding this term in the estimating function results in equations from where explicit expressions of estimators in general are more difficult to derive. Also as a consequence of the structure of the estimating equations derived from (EFPK) the method can be applied whenever ordinary (EF) is applicable.

The Estimating Functions with Prior Knowledge approach is in particular useful for small sample sizes since the classical estimates in this situation might be very unreliable, the incorporated prior pools estimates towards the prior and thereby ”remove” extreme estimates, and reduce the variation of the estimates. The basic idea behind (EFPK) is to create estimating function which are maximal correlated with the posterior score, contrary to the classical setup where we try to imitate the score function. The idea is formalized by minimizing the $L^2$ distance to the posterior score.

We have demonstrated how to implement (EFPK) for parameter estimation for discretely observed diffusions. As case studies the Ornstein-Uhlenbeck Process and the Cox, Ingersoll and Ross (CIR) process were chosen. For Ornstein-Uhlenbeck Process an expression for the posterior score is readily found and we saw from the simulations that (EFPK) clearly out preforms (EFL) and get reasonably close to the (MAP) estimators. For the Cox, Ingersoll and Ross (CIR) process the method illustrates is worth, again it clearly in many cases out performs (EFL) and on the other hand the (MAP) estimator is not easy to derive.

# 7 Acknowledgement

Most of this research was done while Kim Nolsøe stayed at the Department of statistics, University of Cartagena. This stay was supported by the European Commission through the Research Training Network DYNSTOCH under the Human Potential Programme.

# References

[Aït-Sahalia, 2002] Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1):223–262.

[Bibby and Sørensen, 1995] Bibby, B. M. and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1(1-2):17–39.

[Cano et al., 2003] Cano, J., Kessler, M., and Salmeron, D. (2003). Approximation of the posterior density for diffusions processes. *Technical report*, (57):1–9.

[Cox et al., 1985] Cox, J. C., Ingersoll, Jr., J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407.

[Godambe and Heyde, 1987] Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.*, 55(3):231–244.

[Heyde, 1997] Heyde, C. C. (1997). *Quasi-likelihood and its application.* Springer Series in Statistics. Springer-Verlag, New York. A general approach to optimal parameter estimation.

[Jensen and Poulsen, 2002] Jensen, B. and Poulsen, P. (2002). Transition densities of diffusion processes. *Numerical Comparison of Approximation Techniques, Journal of Derivatives*, (9):18–32.

[Johannes and Polson, 2004] Johannes, M. and Polson, N. (2004). Mcmc methods for financial econometrics. *in Ait-Sahalia Y and Hansen L. P. eds., Handbook of financial econometrics. Forthcoming.*

[Karunamuni, 2002] Karunamuni, R. J. (2002). An empirical Bayes derivation of best linear unbiased predictors. *Int. J. Math. Math. Sci.*, 31(12):703–714.

[Kessler, 1995] Kessler, M. (1995). Martingale estimating functions for a markov chain. preprint. *Laboratoire de Probabilitès, Universitè Paris VI.*

[Kessler and Sørensen, 1999] Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, 5(2):299–314.

[Pedersen, 2000] Pedersen, A. (2000). Estimating the nitrous oxide emission rate from the soil surface by means of a diffusion model. *Scandinavian Journal of Statistics*, 27:385–403.

[Pedersen, 1995] Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, 22(1):55–71.

[Zellner, 1996] Zellner, A. (1996). Bayesian method of moments (BMOM) analysis of mean and regression models. In *Modelling and prediction (Hsinchu, 1994)*, pages 61–74. Springer, New York.