Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

SHIFTING PARADIGMS IN INFERENCE

C.C. Heyde
Australian National University and Columbia University

Abstract

Some personal perspectives on changing paradigms in inference are presented. The topics discussed include the changes from independence to dependence, estimators to estimating functions and from adhoc methods to Fisher information based methods. Recent trends in time series, general theory of inference, estimating functions and information based techniques are discussed.

1 Introduction

The advent of the new millenium gives us a particularly good excuse to take stock of changing paradigms in inference, or more particularly inference for stochastic processes. Our subject can reasonably be thought of as roughly a century old, and it was strongly practical and model based from the outset. Our distinguished ancestors such as Pyotr En'ko in 1889 with the chain binomial model for epidemics, Louis Bachelier in 1900 with Brownian motion and the modelling of the sharemarket and Filip Lundberg in 1903 with collective risk for insurance application, were very much motivated by the scientific needs of their times.

Any list of paradigms is, of course, rather subjective. The ones that I will treat in this paper are undoubtedly important, and are ones which have influenced me personally. But there are arguably others, and certainly one other that I would have liked to discuss. That is the advent of the computer as a tool for model exploration and simulation. I have learned a lot from the newly available technologies. A lot about bad models, poor behaviour of limit theorems, slow rates of convergence etc. But the constraints of the occasion have precluded discussion of these issues. So let me pass to the topics which I will discuss. These are the changes:

1. Independence ⇒ Dependence.
2. Estimators ⇒ Estimating Functions.
3. Ad hoc methods ⇒ Fisher Information based methods.

In connection with the first of these, I should remark that I am not seeking to minimize the role of independence in stochastic models. Regeneration in stochastic models is a key phenomenon and our capacity to simulate is strongly tied to independence. My focus is on what we can do in inference.

2 Independence to Dependence

The history of Inference for Stochastic Processes has two major strands of development:

(1) General theory of inference.

(2) Inference for time series.

We shall examine the impediments to the development of each of these.

2.1 Time Series

Time series as an autonomous area of study basically dates from the 1920s. There were many contributors but I will particularly mention the name of George Udny Yule who wrote papers in 1926, 1927 which laid the foundation of autoregressive process theory. The idea of a linear model in terms of a finite past of the process together with a stochastic disturbance was natural and immediately successful in a wide range of problems. The subject exolved into the autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) forms and entered the modern era in large part through the computational implementation of Box and Jenkins (1970).

From the outset there was a special focus on the case of Gaussian time series. Indeed, the theory was developed for second order stationary Gaussian processes, which are fully characterized by their means and covariances $\{\gamma(k) = cov(X_n, X_{n+k})\}$. Also, maximum likelihood (ML) and least squares (LS) estimation procedures were used from the outset, and continue to be used.

Amongst the most influential early results was the following:

Wold Decomposition (1938). If $\{X_j\}$ is a purely non-deterministic (physically realizable) stationary process with zero mean and finite variance it is representable in the form

$$X_n = \sum_{j=0}^{\infty} \alpha_j \epsilon_{n-j}$$

where $\{\epsilon_j\}$ is a stationary, uncorrelated, zero mean process and $\sum_{j=0}^{\infty} \alpha_j{}^2 < \infty$. If the $X's$ are normally distributed the $\epsilon's$ can be taken as independent and identically distributed (iid) and normally distributed.

This theorem implicitly suggested that all second order stationary processes could be reasonably approximated by ARMA processes of sufficiently high order, a suggestion which was further reinforced by results such as the following:

Theorem If $\{\gamma(.)\}$ is *any* covariance function such that $\gamma(k) \to 0$ as $k \to \infty$, then there is a causal AR(K) process whose autocovariance function at lags $0, 1, ..., K$ coincides with $\gamma(j), j = 0, 1, ...K$.

The results lulled users into a false sense of security concerning the breadth of applicability of ARIMA models, and it was not for some decades that the need to deal with long-range dependent processes, and various nonlinear phenomena, could no longer be denied. In the meantime, the subject proceeded via development of inference for the Gaussian case and then the Gaussian assumption was dropped and replaced by that of iid innovations ϵ. There was complete reliance on the Strong Law of Large Numbers (SLLN) and Lindeberg-Feller Central Limit Theorem (CLT) for sums of independent random variables to develop the consistency and asymptotic normality results which underpinned a useful inferential theory. This is where the subject stood at the end of the 1960s.

The 1970s saw the development of limit theory, in particular the SLLN and CLT, for martingales, subsuming the earlier results for independent random variables. With the martingale theory it became possible to treat issues such as when is a linear model appropriate.

There is, indeed, a simple answer to this question. Consider a stationary finite variance process $\{X_n\}$ and write

$$\epsilon_j = X_j - E(X_j | \mathcal{F}_{j-1}),$$

where the ϵ_j are the prediction errors, $\{\mathcal{F}_n\}$ are the past history σ-fields and $E(X_j|\mathcal{F}_{j-1})$ is the best one-step predictor of X_j. Then, it turns out that the *best linear predictor is the best predictor if and only if the ϵ's are martingale differences* (Hannan and Heyde (1972)).

In recent times there has been an increasing realization of the role of non-linear models, but much of the development has been coming from other professions, such as physicists (see for example Kantz and Schreiber (1997)). Dynamical systems, often with striking associated properties such as chaos, have attracted much attention and proponents of deterministic theory have thrown out a challenge to the stochastic community to which there has been all too little in the way of a reasoned response.

2.2 General Theory of Inference

This was developed in a setting of a random sample of iid rv (and is still typically taught in that setting!). Much of the theory rests on asymptotic normality (or mixed normality) of estimators, and when it was developed there were nice CLT results only for independent rv. The first attempts at a discussion of inference for stochastic processes in a general setting came only in the 1960s and 1970s. This can be seen in the books of Billingsley (1961) and Roussas (1972) in a setting of stationary ergodic Markov chains.

Soon thereafter, general central limit results for martingales became available subsuming independence results such as the Lindeberg-Feller CLT. Only then did it become possible to give a very general discussion of inference for stochastic processes in the traditional likelihood based setting.

The basic framework is as follows. We have a sample $\{X_1, X_2, ...X_n\}$ whose distribution depends on a parameter θ (which we take as scalar for convenience). The likelihood $L_n(\theta)$ is assumed to be differentiable with respect to θ. Then, ordinarily, the *score function*

$$U_n(\theta) = dlogL_n(\theta)/d\theta = \sum_{i=1}^{n} u_i(\theta)$$

is a martingale. The whole classical theory of the maximum likelihood estimator (MLE) carries over in its entirety to the general setting. One uses martingale limit theory on $U_n(\theta)$ and local linearity with a Taylor expansion in the neighbourhood of the MLE. Details are given in Hall and Heyde (1980, Chapter 6).

The essence of the results is as follows. If $I_n(\theta) = \sum_{i=1}^{n} E({u_i}^2 | \mathcal{F}_{i-1})$ is generalized Fisher information, and if $I_n(\theta) \xrightarrow{p} \infty$, and

$$I_n(\theta)/EI_n(\theta) \xrightarrow{up} \eta^2(\theta)$$

for some $\eta(\theta) > 0$ a.s., 'up' denoting uniform convergence in probability, then with little else one has optimality of the MLE in terms of producing minimum size asymptotic confidence intervals for θ and the classical theory is nicely subsumed.

Martingale theory provides the natural setting but - 20 years later - these things have regrettably not yet become part of the statistical consciousness. The consequence is that many contemporary developments in inference, for example on the "general" linear model, missing data and the EM algorithm, multiple roots of the score function,... - are carried out in an independence setting while a much more general treatment is possible. Martingales are not yet part of the statistical mainstream. They are still regarded as belonging to the domain of the probabilists. It is notable that, by contrast, the Econometricians have not been reluctant to embrace the theory. See, for example, Davidson (1994, p. xiii).

3 Estimators to Estimating Functions

An estimating function (EF) is a function of data and parameter, typically with mean zero, which when equated to zero gives a parameter estimator as its root.

The use of estimating functions is close to universal in statistical practice. It is just that there has been little focus to date on EF's themselves. Their usage dates back at least to Karl Pearson's *method of moments* (1894). For example, if X_i are iid with $EX_i = \mu$ and $varX_i = \sigma^2$, then

$$\sum_{i=1}^{n}(X_i - \mu), \sum_{i=1}^{n}((X_i - \mu)^2 - \sigma^2)$$

are estimating functions for $\theta = (\mu, \sigma^2)'$.

Maximum likelihood (ML) and least squares and its variants (LS, WLS) are basically EF methods, the parallel between them being shown below.

ML	**LS/WLS**
Likelihood $L(\theta)$	Sum(weighted sum) of squares $S(\theta)$
Form the score $dlogL(\theta)/d\theta$	Form $dS(\theta)/d\theta$
Equate to zero and solve	Equate to zero and solve

The score function is a *benchmark* (eg. Godambe (1960)). *It is the score function, rather than the MLE which comes from it, which is fundamental.* Indeed, the optimality properties which we ascribe to the MLE are really optimality properties of the score function. For example:

- Fisher information is an EF property (Fisher information is $varU$).
- The Cramér-Rao inequality is an EF property. It gives $varU$ as a bound on the variances of standardized estimating functions.

EF's have significant advantages over the estimators derived therefrom.

- EF's with information about an unknown parameter can be readily combined.
- EF's usually have straightforward asymptotics. That for the estimator is derived therefrom using local linearity plus regularity.

For a discussion of optimal inference it is best to choose an EF setting and to focus on optimality of the EF and *not* optimality of an estimator derived therefrom. The origins of this theory go back to the 1960s but it began a serious surge of development in the mid 1980s, much of the impetus being provided by Godambe's 1985 paper. A detailed treatment of the subject has been provided in book form in Heyde (1997). The theory, labelled as quasi-likelihood since it closely mimics the features of classical likelihood theory, is outlined below.

3.1 General QL Principles

The setting is of a sample $\{Z_t, t \in T\}$ from some stochastic system whose distribution involves θ. The θ to be efficiently estimated is a vector of dimension p.

The approach is via a chosen family of EFs

$$\mathcal{G} = \{G_T(\theta) = G_T(\{Z_t, t \in T\}, \theta)\},$$

the G_T being vectors of dimension p with $EG_T(\theta) = 0$ and the $p \times p$ matrices

$$E\dot{G}_T = (E\partial G_{T,i}/\partial\theta_j), EG_T G_T{}'$$

being assumed nonsingular.

Comparisons are made using an *information criterion* (generalized Fisher information)

$$\mathcal{E}(G_T) = (E\dot{G})'(EG_T G_T{}')^{-1}(E\dot{G}_T)$$

for $G_T \in \mathcal{G}$. We choose $G^*{}_T \in \mathcal{G}$ to maximize $\mathcal{E}(\mathcal{G}_T)$ in the partial order of non-negative definite matrices. (This amounts to a reformulation of the Gauss-Markov theorem.) Such a $G^*{}_T$ is called a *quasi-score* estimating function (QSEF) within $\mathcal{G}$.

It should be emphasized that the choice of the family $\mathcal{G}$ is open and should be tailored to the particular application.

The estimator $\theta^*{}_T$ obtained from $G^*{}_T(\theta^*{}_T) = 0$, termed a *quasi-likelihood estimator*, has under broad conditions, minimum size asymptotic confidence zone properties for θ, at least within $\mathcal{G}$. The basic properties are those of the MLE, but restricted to $\mathcal{G}$. The theory does *not* require a parametric setting, let alone the existence of a score function $U_T(\theta)$.

Important features of the theory include:

- It applies to general stochastic systems.
- It allows for the control of the problem of misspecification. This control is in the hands of the experimenter. No more than means and variances are required in many contexts.
- It carries with it all the classical theory of ML and LS.

There is no extra baggage required for the discussion of inference for stochastic processes. But for this setting the detailed asymptotics does require modern limit theory (especially that for martingales).

QSEFs can usually be found with the aid of the following result (Heyde (1997), Theorem 2.1, p.14):

Proposition $G^*{}_T \in \mathcal{G}$ is a QSEF within $\mathcal{G}$ if

$$(E\dot{G}_T)^{-1} EG_T G^*{}_T{}' = C_T \tag{3.1}$$

for all $G_T \in \mathcal{G}$, where C_T is a fixed matrix. Conversely, if $\mathcal{G}$ is convex and $G^*{}_T$ is a QSEF then (3.1) holds.

3.2 Finding Useful Families of EFs

Statistical models can generally, perhaps after suitable transformation, be described as

$$\text{data} = \text{signal} + \text{noise}$$

where the signal is a predictable trend term and the noise is a zero-mean stochastic disturbance. This can then be conveniently reformulated in terms of a special semimartingale representation as

$$X_t = X_0 + A_t(\theta) + M_t(\theta)$$

where A_t is a predictable finite variation process and M_t is a local martingale. This provides a natural route to estimating parameters in the signal. Discrete time processes and most continuous time processes with finite means have this kind of representation, via a suitable rewrite if necessary. Thus, for example, if $\{X_t = \sum_{i=1}^t x_i\}$ is a discrete time process, with past history σ-fields $\{\mathcal{F}_t\}$, it can be rewritten in special semimartingale form as

$$X_t = \sum_{i=1}^{t} E(x_i|\mathcal{F}_{i-1}) + \sum_{i=1}^{t}(x_i - E(x_i|\mathcal{F}_{i-1})).$$

A general strategy is to try the Hutton-Nelson family of EFs

$$\mathcal{G} = \{G_T : G_T = \int_0^T \alpha_s(\theta) dM_s(\theta) = \int_0^T \alpha_s(\theta) d(X_s - A_s(\theta))\}$$

for which the QSEF is

$$\sum_1^T (E(\dot{m}_s|\mathcal{F}_{s-1}))'(E m_s m_s'|\mathcal{F}_{s-1})^- m_s$$

where $M_t = \sum_{s=1}^t m_s$ in the discrete time case and

$$\int_0^T (E(d\dot{M}_s|\mathcal{F}_{s-}))'(d\langle M\rangle_s)^- dM_s$$

in the continuous time case. Here $\langle M\rangle_t$ is the quadratic characteristic and the minus superscript denotes the generalized inverse.

Example. The membrane potential V across a neuron is well described by a stochastic differential equation

$$dV(t) = (-\rho V(t) + \lambda)dt + dM(t)$$

(eg Kallianpur (1983)), where $M(t)$ is a martingale with a (centered) generalized Poisson distribution. Here $\langle M\rangle_t = \sigma^2 t, \sigma > 0$.

The QSEF for the Hutton-Nelson family on the basis of a single realization $\{V(t), 0 \leq t \leq T\}$ is

$$\int_0^T (-V(t)\ 1)'(dV(t) - (-\rho V(t) + \lambda)dt).$$

The estimators $\hat{\rho}$ and $\hat{\lambda}$ are then obtained from the estimating equations

$$\int_0^T V(t)dV(t) = \int_0^T (-\hat{\rho}V(t) + \hat{\lambda})V(t)dt$$

$$V(T) - V(0) = \int_0^T (-\hat{\rho}V(t) + \hat{\lambda})dt.$$

For more details about this subject, including such things as how to deal with parameters in the noise component of the semimartingale model, see Heyde (1997, Chapter 2).

An important role for semimartimgales in inference for stochastic processes has been evident since the papers of Hutton and Nelson (1986) and Thavenaswaran and Thompson (1986). There are few contexts where these methods cannot play a vital role.

Now there is a book on inference associated with semimartingale models (Prakasa Rao (1999)), although not with a focus of the kind that has been outlined above.

Amongst the (rare) processes which are not semimartingales is the fractional Brownian motion $B_H(t), 0 < H < 1, H \neq \frac{1}{2}$. The case $H = \frac{1}{2}$ corresponds to ordinary Brownian motion, which is a semimartingale. Fractional Brownian motion has a Gaussian distribution and the self-similarity property

$$B_H(ct) \stackrel{d}{=} c^H B_H(t), c > 0.$$

It has been widely used to model possible long-range dependence (see for example Beran (1994)).

An example where this process has been used is in modelling departures from the standard geometric Brownian motion model of Black and Scholes for the price of a risky asset which may now exhibit long-range dependence. Here the price P_t of the asset at time t is modelled by the stochastic differential equation (sde)

$$dP_t = P_t[\mu dt + \sigma dB_H(t)]$$

where $B_H(t), 0 < H < 1$, is a fractional Brownian motion process. The standard model corresponds to the case $H = \frac{1}{2}$ and the nonstandard model, $H \neq \frac{1}{2}$, is not amenable to the analysis described above. Substitute methods, however, have recently been developed. See, for example, Mikosch and

Norvaiša (2000) for a discussion of the above sde, and Norros, Valkeila and Virtamo (1999) for a discussion of the parameter estimation.

All the methods of inference, semimartingale based or not, make use, in some sense, of information or empirical information. Consistency results can generally be obtained via the martingale SLLN and (asymptotic) confidence intervals via the martingale CLT. For the latter, the most general results deal with the self-normalized case $[M]^{-\frac{1}{2}}M$, $[M]$ being the quadratic variation (Heyde (1997)). Ideas on information are the subject of the next, and last, section.

4 Fisher Information as a Statistical "Law of Nature"

The essential points that I wish to make are:

(1) The role of Fisher information in the general theory of inference has already been described (in the previous section).

(2) Fisher information has a key role as a scientific tool (for example in Physics).

(3) Many, perhaps most, statistical procedures rely on Fisher information in ways which have not hitherto been acknowledged.

Recently there has been some quite striking work in Physics based on the idea of Fisher information. The book Frieden (1998) caught the interest of the science journalist community. For example, it led to an article in *New Scientist* (Matthews (1999)). On the basis of this I bought the book and I found it both fascinating and frustrating. I wove consideration of it into a seminar course which I gave at Columbia University in the Fall of 1999. The ideas certainly warrant very serious consideration by the statistical community.

Frieden's thesis is that:

• All physical laws may be unified under the umbrella of measurement theory.

• With each phenomenon there is an associated Lagrangian, natural to the field. All Lagrangians consist entirely of two forms of Fisher information - data information and phenomenological information.

An informal explanation is that each context requires solution to some extremum problem. At the basis of this is a scalar function called the Lagrangian (like the likelihood). The solution of the problem can be phrased in terms of a pde involving the Lagrangian.

The parallels with statistics look good at face value. But the reality is much more complex. Much of the book treats physical systems where information decreases over time. Of course, statistical problems are typically ones where information increases over time - corresponding to the collection of more data. The sort of context where information decreases over time is, say, when the position of a particle is observed subject to noise. Over time the particle moves and its position is known with decreasing precision.

4.1 Procedures Involving Information

Most statistical procedures seem to be associated, directly or indirectly, with measures of information, and it is arguably of value to make the connection explicit as an aid to the development of useful methods. Also, it is important to note that there is a close connection between comparisons of information content and statistical distance. For example, the formulation of a quasi-score estimating function in terms of maximizing generalized Fisher information can be equivalently recast into a formulation in terms of minimizing dispersion distance (from the (generally unknown) score function) (Heyde (1997), p. 12). A new book focusing on the use of statistical distance is Lindsay and Markatou (2001).

We now proceed to examine two applications in which information based ideas are not immediately apparent, in order to see the role that they can play.

Choosing the order of an autoregression

The most widely used procedure, AIC, involves choosing the order k to minimize:

$$AIC(k) = -2logL(\hat{\theta}_k) + 2k \tag{4.1}$$

where $\hat{\theta}_k$ is the MLE of θ restricted to R^k and $2k$ is a penalty function. It is assumed that the order $k \leq K$ for some fixed K.

Akaike's original proof uses the Kullback-Liebler entropy KL given by

$$KL = -\int p(x) ln \frac{p(x)}{r(x)} dx$$

measuring the distance between two pdf's p, r and it should be noted that Fisher information can be thought of as a form of local entropy (Friedan (1998), pp. 31-32)).

The proof shows that, asymptotically, the order minimizing $EK(\theta, \hat{\theta}_k)$ is the same as the one minimizing (4.1). It proceeds via the likelihood ratio statistic for testing the null hypothesis $H_0 : \theta \in R^k$ versus the alternative $H_1 : \theta \in R^K - R^k$ and suggests a QL generalization of AIC based on generalized Fisher information.

Note the route to treating problems where one does not have estimating functions differentiable with respect to the parameter of interest. Here the variable in question is discrete.

Stochastic Resonance

The core idea here is of a weak signal operating in a noisy environment which is normally undetectable. However, by suitably increasing the noise a "resonance" can be set up making the signal apparent. Resonance may be a very important phenomenon scientifically. It has, for example, been proposed as a possible explanation for the ice ages. There is a burgeoning literature on the phenomenon which can be conveniently accessed via the the stochastic resonance web site http://www.umbrars.com/sr based in Perugia, Italy, which in turn has links to similar web sites in San Diego, USA and Saratov, Russia.

A very simple example concerns the tunable model

$$dX_t = (Asin\omega t)dt + \sigma dW(t) \tag{4.2}$$

where W is standard Brownian motion and the amplitude A is subthreshold (ie $A \leq A_0$). We want to estimate ω and the issue is the optimum choice of σ.

When estimating a frequency from discretely observed data $x_t = X_t - X_{t-1}$, the conventional wisdom is to work with the periodogram

$$I_p = \frac{2}{N} |\sum_1^N x_t e^{-i\omega_p t}|^2$$

where $\omega_p = 2\pi p/N, p = 0, 1, ...[N/2]$. The theory, which originated back with Fisher (1929), tells us to use the estimator ω corresponding to the ω_p for which $\max_p I_p$ obtains. Consistency and rate of convergence results are available.

An information formulation can proceed as follows. From the semimartingale representation

$$x_t e^{-i\omega_p t} = \mu_p e^{-i\omega_p t} + N_t e^{-i\omega_p t},$$

say, derived from (4.2), we obtain the estimating function

$$G_p = \sum_1^N (x_t e^{-i\omega_p t} - \mu_p e^{-i\omega_p t})$$

and the empirical information associated with this is $||G_p||^2$. Asymptotically, $||G_p||^2$ and I_p are maximized for the same p.

Of course in the stochastic resonance problem we do not observe the $\{x_t\}$ process, but rather the censored process $\{T_t = x_t I(|x_t| \geq A_0)\}$. But it seems that the periodogram based approach is still appropriate.

As a general conclusion, it seems profitable to think about statistical problems in a setting of maximizing an information. I see this as an important unifying principle.

References

Beran, J. (1994). *Statistics for Long-Memory Processes*, Chapman & Hall, New York.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*, Univ. Chicago Press, Chicago.

Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco.

Davidson, J. (1994). *Stochastic Limit Theory.* Oxford University Press Advanced Texts in Econometrics, Oxford.

Fisher, R.A. (1929). Tests of significance in harmonic analysis. *Proc. Roy. Soc. Ser. A* **125**, 54-59.

Frieden, B.R. (1998). *Physics from Fisher Information. A Unification*, Cambridge U. Press, Cambridge.

Godambe, V.P. (1960). An optimum property of regular maximum- likelihood estimation. *Ann. Math. Statist.* **31**, 1208-1211.

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

Hall, P.G. and Heyde, C.C. (1980). *Martingale Limit Theory and its Application*, Academic Press, New York.

Hannan, E.J. and Heyde, C.C. (1972). On limit theorems for quadratic functions of discrete time series. *Ann. Math. Statist.* **43**, 2058-2066.

Heyde, C.C. (1997). *Quasi-Likelihood and its Application. A General Approach to Optimal Parameter Estimation*, Springer, New York.

Hutton, J.E. and Nelson, P.I. (1986). Quasi-likelihood estimation for semimartingales. *Stochastic Process. Appl.* **22**, 245-257.

Kallianpur, G.P. (1983). On the diffusion approximation to a discontinuous model for a single neuron. In P.K. Sen, Ed., *Contributions to Statistics: Essays in Honor of Norman L. Johnson.* North-Holland, Amsterdam, 247-258.

Kantz, H. and Schreiber, T. (1997). *Nonlinear Time Series Analysis.* Cambridge University Press, Cambridge.

Lindsay, B.G. and Markatou, M. (2001). *Statistical Distances: A Global Framework for Inference.* Springer, New York, to appear.

Matthews, R. (1999). *I* is the law. *New Scientist*, 30 January 1999, 24-28.

Mikosch, T. and Norvaiša, R. (2000). Stochastic integral equations without probability. *Bernoulli* **6**, 401-434.

Norros, I., Valkeila, E. and Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motion. *Bernoulli* **5**, 571-587.

Prakasa Rao, B.L.S. (1999). *Semimartingales and their Statistical Inference.* Chapman & Hall/CRC, Boca Ratan.

Roussas, G.G. (1972). *Contiguity of Probability Measures*, Cambridge Univ. Press, London and New York.

Thavaneswaran, A. and Thompson, M.E. (1986). Optimal estimation for semimartingales. *J. Appl. Prob.* **23**, 409-417.