Robust estimation of time-varying processes

A.V. den Boer

University of Twente, P.O. Box 217, 7500 AE Enschede

July 25, 2013

Abstract

We consider the question of optimally estimating a time-varying multivariate stochastic process, in order to minimize the expected squared estimation error. This is motivated by adaptive control problems under uncertainty in a changing environment. A distinguishing feature of our approach is that we do not need a completely specified model for the stochastic process under consideration. Instead, we merely assume that the process obeys certain predetermined assumptions, and subsequently derive the optimal min-max estimator w.r.t. these assumptions. This can be seen as a stochastically robust method to estimate a time-varying process. We provide tight upper bounds on the expected squared estimation error, and explicitly derive the optimal weighted least-squares estimator in several relevant examples.

Keywords: robust estimation, time-varying parameters, weighted least squares, exponential smoothing.

1 Introduction

In many optimization problems of practical interest, the optimal decision depends on certain real-world parameters with values that are unknown to the decision maker. If the optimization problem is static, in the sense that there is only a single moment at which a decision can be made, the field of robust optimization provides a framework for handling such parameter uncertainty. A typical approach is to assume that the unknown parameters are contained in a known uncertainty set, and to optimize with respect to the worst-case values in this uncertainty set. A comprehensive overview of robust optimization can be found in Ben-Tal et al. (2009).

Many decision problems are not static but sequential: the decision maker can revise her decision, periodically or continuously, according to changing circumstances or insights. In addition, in many of such problems, information on the unknown parameters becomes available

as time progresses, in the form of measurements or realizations of a random variable. This allows the decision maker to form estimates on the unknown parameter values, and adjust the optimal decision each time the parameter estimates are updated. Ideally, in such a setting, the parameter estimates converge to their true value as time progresses, and the taken decision to the optimal decision.

A complicating issue is that in practice, parameter values are usually not constant, but change over time. These changes may occur gradually or in "shocks", periodically or at random time periods, and may have a large impact on the optimal decision. Clearly, the decision maker should take these fluctuations into account while forming estimates. How this is best done, is the subject of this paper.

A common approach is to model the unknown parameter as a time series, and subsequently determine an estimator that is optimally tailored to this model. A disadvantage of such an approach is proneness to misspecification of the parameter process. Even before any data is available, the decision maker needs to make strong assumptions on the nature of the process (like "the parameter behaves as an ARMA(2,2) time series"). This is undesirable - if not impossible - in practice, and leaves the question open what happens if the time-series model is misspecified, or if the variables that characterize the time-series themselves change over time.

An alternative approach is to use exponential smoothing, or one of its many variants and extensions (Gardner, 1985, 2006). This method is reported to be quite robust to misspecification, but explicit performance bounds or statistical rationales to use this method are scarce. Exceptions are Muth (1960), Satchell and Timmermann (1995), Chatfield et al. (2001), who show that exponential smoothing is optimal for some specific models.

In this paper we propose a new framework for estimating a time-varying process, which overcomes the two disadvantages mentioned above: it does not need a detailed model of the parameter process as input, and it comes with explicit, tight upper bounds on the expected squared estimation error. The key idea is to optimize the weights of a weighted least-squares estimator, not with respect to a certain specific time-series model, but with respect to a whole class of processes. In this way we obtain an estimator that is robust to misspecification of the unknown process. The decision maker specifies an "uncertainty set" of processes, and subsequently determines the optimal weighted least-squares estimator in a min-max sense, in order to minimize the worst-case expected squared estimation error. This uncertainty set may be defined by very general conditions; for example, processes that are a.s. bounded in norm by some constant, or processes with one-step differences are a.s. bounded in norm by some constant. If desirable, the uncertainty set may also be chosen more specific, for example all stochastic processes that can be described by an ARMA$(2, 2)$ time-series.

For three types of uncertainty sets, we explicitly derive the optimal weighted least-squares estimator. If we assume that the process is a.s. bounded, it is optimal to give all available data equal weight; in other words, under these conditions, the ordinary least squares estimator is optimal. If we assume a bound on the maximum difference between two consecutive values of the process, the optimal weight function is the positive part of a linearly decaying function. We also show that when an exponential decaying weight function is used instead of the optimal linear one, the estimation error may increase up to seven percent. This shows that it may be rewarding to use and study the optimal estimator. Finally, if we assume that the process is a simple random walk, we show that the optimal weight function is contained between two exponentially decaying functions.

Our robust approach to estimating a time-varying process has several advantageous features. First, it removes the need to formulate an explicit, detailed (time-series) model for the unknown process, which in practice is difficult and prone to misspecification errors. Neither are any assumptions on the exact form of the distribution of measurement errors required. The approach can be applied to multivariate processes without any additional difficulties, and it not difficult to compute the optimal weight function for many other uncertainty sets than the ones that are explicitly considered in this paper. We derive explicit bounds on the expected squared estimation error, which is a useful result in sequential decision or control problems under uncertainty. In addition, our approach is applicable to several kinds of changes in the parameter process (e.g. both abrupt and gradual changes).

The rest of this paper is organized as follows. In Section 2 we introduce the weighted least-squares estimator, and mention some common choices. Section 3 discusses the robust min-max optimization of the weight function, with respect to assumptions on the process made by the decision maker. We derive a general upper bound on the expected squared estimation error, that depends on these assumptions. In Section 3.1 we obtain the optimal weight function assuming the process is bounded, in Section 3.2 under the assumption that the one-step differences are bounded, and in Section 3.3 assuming that the process is a simple random walk. A discussion of the methodology is provided in Section 4, together with several directions for future research. All proofs are contained in the Appendix.

Notation. For $x \in \mathbb{R}$ we write $(x)^+ = \max\{x, 0\}$. We write $\mathbb{R}_+ = [0, \infty)$. The function $\mathbf{1}(A)$ equals one if A is true, and zero otherwise.

2 Estimation of a time-varying stochastic process

Let $(Y(t))_{t\in\mathbb{N}}$ be a stochastic process taking values in $\mathbb{R}^m$, with $\sup_{t\in\mathbb{N}} E\left[||Y(t)||\right] < \infty$. At each time period $t \in \mathbb{N}$ a measurement $z(t) \in \mathbb{R}^m$ is observed of the form

$$z(t) = Y(t) + \epsilon(t). \tag{1}$$

Here $(\epsilon(t))_{t\in\mathbb{N}}$ is an m-dimensional stochastic process with $E\left[\epsilon(t)\right] = 0$ and $E\left[\epsilon_j(t)\epsilon_j(t')\right] = 0$, for all $t, t' \in \mathbb{N}$, $t \neq t'$, and $j = 1, \ldots, m$. In addition we assume $\sigma_j^2 = \sup_{t\in\mathbb{N}} E\left[\epsilon_j(t)^2\right] < \infty$, for all $j = 1, \ldots, m$. Note that we do not require $\epsilon(t)$ and $Y(t)$ to be independent.

Based on observations $z_1, \ldots, z_t$, we estimate the value of $Y(t)$ by minimizing the following weighted sum-of-squares criterion:

$$\hat{Y}[\varphi](t) = \min_{y\in\mathbb{R}^m} \sum_{i=1}^{t} ||z(i) - y||^2 \varphi(t-i). \tag{2}$$

Here $\varphi : \{0, 1, 2, \ldots\} \to [0, \infty)$, $\varphi \neq 0$ is a weight or kernel function, and $||\cdot||$ denotes the Euclidean norm. By taking the derivative w.r.t. y, it follows that $\hat{Y}[\varphi](t)$ is given by

$$\hat{Y}[\varphi](t) = \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-1} \sum_{i=1}^{t} z(i)\varphi(t-i). \tag{3}$$

provided $\sum_{i=1}^{t} \varphi(t-i) > 0$.

Common examples of weight functions are (i) $\varphi(i) = 1$ for all $i \in \mathbb{N}$, which gives all available data equal weight; (ii) an exponentially decaying weight function $\varphi(i) = \lambda^i$, for some $\lambda \in (0, 1)$, and (iii) a moving average weight function $\varphi(i) = \mathbf{1}(i < N)$ for some $N \in \mathbb{N}$, which means that only the N most recent observations are taken into account in the estimation, and all receive equal weight.

3 Optimal weight selection

Our goal is to determine a weight function φ that optimizes the quality of the estimate (3). Naturally, the quality of this estimate depends on the characteristics of the process $(Y(t))_{t\in\mathbb{N}}$, and thus the optimal choice of φ depends on properties of $(Y(t))_{t\in\mathbb{N}}$. A weight function φ may for example have a very good performance for rather volatile processes $(Y(t))_{t\in\mathbb{N}}$ but a mediocre performance if $(Y(t))_{t\in\mathbb{N}}$ is constant, or the other way around.

Optimization of φ is thus always done with respect to certain a priori made assumptions on

$(Y(t))_{t\in\mathbb{N}}$. We represent this in a formal manner by assuming that $(Y(t))_{t\in\mathbb{N}}$ is an element of a certain known subset $\mathcal{Y}$ of all stochastic processes on $\mathbb{R}^m$, and optimize φ w.r.t. $\mathcal{Y}$. In particular, we search for a weight function $\varphi^*_{\mathcal{Y}}$, dependent on $\mathcal{Y}$, that minimizes the following min-max criterion:

$$\min_{\varphi} \sup_{Y\in\mathcal{Y}} E\left[\left|\left|\hat{Y}[\varphi](t) - Y(t+1)\right|\right|^2\right], \tag{4}$$

for some $t \in \mathbb{N}$.

Remark 1 (Comparing with $Y(t+1)$ instead of $Y(t)$)**.** The reason that we compare $\hat{Y}[\varphi](t)$ with $Y(t+1)$, and not with $Y(t)$, is motivated by applications in adaptive control or sequential decision problems: suppose that for each $t \in \mathbb{N}$, a decision maker has to determine a decision x_t that maximizes some function $f(x, Y(t))$. If $x^*(Y(t+1))$ denotes the optimal decision at stage $t+1$, and $x^*(\hat{Y}[\varphi](t))$ the optimal decision based on the estimate $\hat{Y}[\varphi](t)$, then in many cases the expected loss caused by deviating from the optimal decision is related to the quantity $E[\| \hat{Y}[\varphi](t) - Y(t+1) \|^2$. Our analysis and outcomes would not differ much if we would consider the objective of minimizing $\sup_{Y\in\mathcal{Y}} E[\| \hat{Y}[\varphi](t) - Y(t) \|^2$.

Remark 2 (Cumulative prediction error)**.** We focus in (4) on the expected squared norm of the prediction error for a particular value of t. Instead, one could also consider a complete time horizon $t = 1, \ldots, T$ for some $T \in \mathbb{N}$, and study min-max optimal estimators for $\min_\varphi \sup_{Y\in\mathcal{Y}} \sum_{t=1}^{T} E\left[\left|\left|\hat{Y}[\varphi](t) - Y(t+1)\right|\right|^2\right]$, the cumulative expected squared norm of the prediction error. This is an important direction for future research.

As already mentioned, fluctuations in $(Y(t))_{t\in\mathbb{N}}$ impact the quality of the estimate $\hat{Y}[\varphi](t)$. This impact is measured by the function $\mathcal{I}_{t+1,Y}(\varphi)$, which is defined by

$$\mathcal{I}_{t+1,Y}(\varphi) = \sum_{j=1}^{m} \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-2} E\left[\left(\sum_{i=1}^{t} (Y_j(i) - Y_j(t+1))\varphi(t-i)\right)^2\right]. \tag{5}$$

Furthermore, for fixed $t \in \mathbb{N}$ and $\mathcal{Y}$, define the function

$$G_{t+1,\mathcal{Y}}(\varphi) := \sup_{Y\in\mathcal{Y}} \mathcal{I}_{t+1,Y}(\varphi) + \sigma^2 \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-2} \sum_{i=1}^{t} \varphi(t-i)^2, \tag{6}$$

where we write $\sigma^2 = \sum_{j=1}^{m} \sigma_j^2$.

The following proposition gives an upper bound on (4), in terms of the function $G_{t+1,\mathcal{Y}}$:

Proposition 1. *Let φ be a weight function and $t \in \mathbb{N}$. Then*

$$\sup_{Y\in\mathcal{Y}} E\left[\left|\left|\hat{Y}[\varphi](t) - Y(t+1)\right|\right|^2\right] \leq 2G_{t+1,\mathcal{Y}}(\varphi). \tag{7}$$

If the processes Y and ϵ are uncorrelated in the sense that $E[Y_j(i)\epsilon_j(i')] = 0$ for all $i, i' \in \mathbb{N}$, $j = 1, \dots, m$, then

$$\sup_{Y \in \mathcal{Y}} E\left[\left|\left|\hat{Y}[\varphi](t) - Y(t+1)\right|\right|^2\right] \leq G_{t+1,\mathcal{Y}}(\varphi), \tag{8}$$

and if in addition the noise terms are homoscedastic, i.e. $E[\epsilon_j(t)^2] = \sigma_j^2$ for all $t \in \mathbb{N}$, $j = 1, \dots, m$, then (8) is an equality.

Proposition 1 provides a recipe for optimally selecting φ: given $\mathcal{Y}$, characterize the maximum impact $\sup_{Y \in \mathcal{Y}} \mathcal{I}_{t+1,Y}(\varphi)$, and subsequently minimize $G_{t+1,\mathcal{Y}}(\varphi)$ w.r.t. φ. Note that $G_{t+1,\mathcal{Y}}(\alpha\varphi) = G_{t+1,\mathcal{Y}}(\varphi)$ for all $\alpha \neq 0$; in the minimization we may thus w.l.o.g. assume $\sum_{i=1}^{t} \varphi(t-i) = 1$. Then we are minimizing a continuous function over the t-dimensional simplex, which is compact, and thus a minimum of $G_{t+1,\mathcal{Y}}(\varphi)$ always exists. The following proposition shows that $G_{t+1,\mathcal{Y}}(\varphi)$ in fact is strictly convex:

Proposition 2. *For any fixed $t \in \mathbb{N}$ and non-empty $\mathcal{Y}$, $G_{t+1,\mathcal{Y}}(\varphi)$ is strictly convex in φ and has a unique minimizer φ^* on the simplex $\Delta_t = \{x \in \mathbb{R}_+^t \mid \sum_{i=1}^{t} x_i = 1\}$, provided $\sigma^2 > 0$.*

In the following subsections we explicitly calculate the optimal weight function $\varphi^*_{\mathcal{Y}}$ for the following three choices of $\mathcal{Y}$:

(3.1) $(Y(t))_{t \in \mathbb{N}}$ is bounded, in the sense that $\sup_{t \in \mathbb{N}} |Y_j(t)| \leq d_j$ a.s. for some $d_1, \dots, d_m > 0$ and all $j = 1, \dots, m$.

(3.2) One-step changes in $(Y(t))_{t \in \mathbb{N}}$ are bounded, in the sense that $\sup_{t \in \mathbb{N}} |Y_j(t) - Y_j(t+1)| \leq d_j$ a.s. for some $d_1, \dots, d_m > 0$ and all $j = 1, \dots, m$.

(3.3) $(Y(t))_{t \in \mathbb{N}}$ is a simple one-dimensional random walk.

3.1 Bounded processes

Let $d_1, \dots, d_m \in (0, \infty)$, and let $\mathcal{Y}$ be the set of all stochastic processes $(Y(t))_{t \in \mathbb{N}}$ such that $|Y_j(t)| \leq d_j$ a.s., for all $j = 1, \dots, m$.

The following theorem shows that, under these assumptions, it is optimal to give all available data equal weight.

Theorem 1. *$G_{t+1,\mathcal{Y}}(\varphi)$ is minimized if $\varphi(i) = 1/t$, for all $i = 0, \dots, t-1$.*

The proof makes use of the following auxiliary lemma.

Lemma 1.

$$\min_{x \in \mathbb{R}_+^t \setminus \{0\}} \Big(\sum_{i=1}^{t} x_i\Big)^{-2} \sum_{i=1}^{t} x_i^2 = \frac{1}{t},$$

and the minimum is attained at $x = \left(\frac{1}{t}, \ldots, \frac{1}{t}\right)$

At first sight it may seem somewhat surprising that in this setting, it is optimal to give all available data equal weight: $\varphi^*_{\mathcal{Y}}(0) = \ldots = \varphi^*_{\mathcal{Y}}(t-1) = 1/t$. The intuition is that the assumptions $|Y_j| \leq d_j$ are quite weak, and allow very volatile and frequently changing processes $(Y(t))_{t\in\mathbb{N}}$. "Tracking" the value of $Y(t)$ by $\hat{Y}[\varphi](t)$ may not be possible, and therefore the optimal weight function focuses on minimizing the effect of the disturbance terms $\epsilon(1), \ldots, \epsilon(t)$ on the estimation error. This is best done by giving all available equal weight.

3.2 Bounds on one-step changes

Let $d_1, \ldots d_m \in (0, \infty)$, write $d = (d_1, \ldots, d_m)$, and let $\mathcal{Y}$ be the set of all stochastic processes $(Y(t))_{t\in\mathbb{N}}$ such that $\sup_{t\in\mathbb{N}} |Y_j(t) - Y_j(t+1)| \leq d_j$ a.s., for all $j = 1, \ldots, m$.

The following theorem shows that the optimal weight function is the positive part of a linear decreasing function.

Theorem 2. *Up to multiplication by a strictly positive constant, there is a unique* $\varphi^*_{\mathcal{Y}}$ *that minimizes* $G_{t+1,\mathcal{Y}}(\varphi)$*, which is of the form* $\varphi^*_{\mathcal{Y}}(i) = (\alpha(t) - \beta(t)i)^+$*, for all* $i = 0, \ldots, t-1$*, and some* $\alpha(t) > 0$*,* $\beta(t) > 0$*. In addition, there is a* $t_d \in \mathbb{N}$*,* $t_d \leq 1 + (\sigma/\,||d||)^2$*, such that* $\alpha(t) = \alpha(t_d)$ *and* $\beta(t) = \beta(t_d)$ *for all* $t \geq t_d$*.*

The proof is based on the following auxiliary lemma.

Lemma 2. *Let* $a > 0$*,* $b > 0$*,* $t \in \mathbb{N}$*, and for all* $x \in \mathbb{R}^t_+ \backslash \{0\}$*, define*

$$f_t(x) = \Big(\sum_{i=1}^t x_i\Big)^{-2} \Big(a\Big(\sum_{i=1}^t i x_i\Big)^2 + b\sum_{i=1}^t x_i^2\Big).$$

There is a minimizer $x^*(t)$ *of* $f_t(x)$ *on* $\mathbb{R}^t_+ \backslash \{0\}$*, which is unique up to multiplication by a strictly positive constant, and* $x^*(t)$ *is of the form*

$$x^*_k(t) = (\alpha(t) - \beta(t)k)^+, \quad (k = 1, \ldots, t), \tag{9}$$

for some $\alpha(t) > 0$*,* $\beta(t) > 0$*. In addition, there is a* $t_d \in \mathbb{N}$*,* $t_d \leq 1 + b/a$*, such that for all* $t \geq t_d$*,* $(\alpha(t), \beta(t)) = (\alpha(t_d), \beta(t_d))$*.*

Remark 3 (Computation of $\varphi^*_{\mathcal{Y}}$)**.** In Lemma 2, the parametric form (9) of $x^*(t)$ allows efficient calculation of $\alpha(t_d)$ and $\beta(t_d)$. The value of $\varphi^*_{\mathcal{Y}}$ for $t \geq t_d$ then follows by taking $a = \sum_{j=1}^m d_j^2$, $b = \sigma^2$, and $\varphi^*(i) = x^*_{i+1}(t) = (\alpha(t_d) - \beta(t_d)(i+1))^+$ for all $i = 0, \ldots, t-1$.

In particular, write $\alpha = \alpha(t_d)$, $\beta = \beta(t_d)$, and let $N^* = \lfloor \alpha/\beta \rfloor$. Using $1 = \sum_{i=1}^N (\alpha - \beta i) =$

$\alpha N - \frac{1}{2}\beta N(N+1)$ and $\sum_{i=1}^{N} i^2 = N(N+1)(2N+1)/6$, we have

$$\begin{aligned} f_{t_d}(\alpha) &= a\left(\sum_{i=1}^{N} i(\alpha - \beta_i)\right)^2 + b\sum_{i=1}^{N}(\alpha - \beta i)^2 \\ &= a\left(\frac{1}{2}N(N+1)\alpha - 2(\alpha N - 1)(2N+1)/6\right)^2 \\ &+ b\left(\alpha^2 N - 2\alpha(\alpha N - 1) + 4(\alpha N - 1)^2 N^{-1}(N+1)^{-1}(2N+1)/6\right). \end{aligned}$$

This is a quadratic function in α, and for fixed N, the minimizer $\alpha_N := \min_{\alpha \geq 0} f_{t_d}(\alpha)$ can easily be computed. The corresponding β_N follows from $1 = \alpha_N N - \frac{1}{2}\beta_N N(N+1)$. Now, α_N and β_N should satisfy $\alpha_N - \beta_N N > 0$ and $\alpha_N - \beta_N(N+1) \leq 0$, i.e. $\alpha_N/\beta_N > N$ and $\alpha_N/\beta_N \leq (N+1)$. Using $\beta_N = 2(\alpha_N N - 1)N^{-1}(N+1)^{-1}$, this is equivalent to $\frac{\alpha_N}{2(\alpha_N N - 1)}(N+1) > 1$ and $\frac{\alpha_N}{2(\alpha_N N - 1)}N \leq 1$, i.e. $2/N \leq \alpha_N < 2/(N-1)$. Since $\lim_{N\to\infty} \alpha_N = 4$, there are only finitely many N s.t. $2/N \leq \alpha_N < 2/(N-1)$. Simply checking them all and evaluating f at the corresponding α_N and β_N yields the optimal values.

Remark 4 (Probabilistic interpretation)**.** Lemma 2 has a nice probabilistic interpretation: if X_f is a random variable taking value in $\mathbb{N}$, and f is its probability mass function, then Lemma 2 shows that the random variable X_f on $\mathbb{N}$ that minimizes

$$a(E[X_f])^2 + bE[f(X_f)], \quad (a > 0,\, b > 0),$$

has probability mass function $f(k) = P(X_f = k) = (\alpha - \beta k)^+$, for some strictly positive α and β.

Remark 5 (Comparison with exponential decaying weight function)**.** The optimal weight function is the positive part of a linear decreasing function. This has a significantly different shape than the exponential-decaying weight function $\varphi(i) = \lambda^i$, or the moving-average-type weight function $\varphi(i) = \mathbf{1}(i < N)$, that are both often used in practice. This raises the question what the effect is of using one of these weight functions instead of the optimal one. To gain some insight on this issue, we consider the exponentially decaying weight function $\varphi_{\text{exp}}(i) = \lambda^i$. We numerically calculate the optimal λ that minimizes $G_{t+1,\mathcal{Y}}(\varphi)$; let λ^* denote the solution, let $\varphi^*_{\text{exp}}(i) = (\lambda^*)^i$, and note that both φ^*_{exp} and the optimal weight function $\varphi^*_{\mathcal{Y}}$ only depend on the ratio $(\sigma/||d||)^2$. The relative loss of using φ^*_{exp} instead of $\varphi^*_{\mathcal{Y}}$ is measured by the ratio

$$\frac{G_{t+1,\mathcal{Y}}(\varphi^*_{\text{exp}})}{G_{t+1,\mathcal{Y}}(\varphi^*_{\mathcal{Y}})}. \tag{10}$$

Figure 1 shows a plot of (10), for different values of $(\sigma/||d||)^2$. The plot is calculated for $t = 100$, but larger values of t lead to a similar picture. The picture suggest that the loss

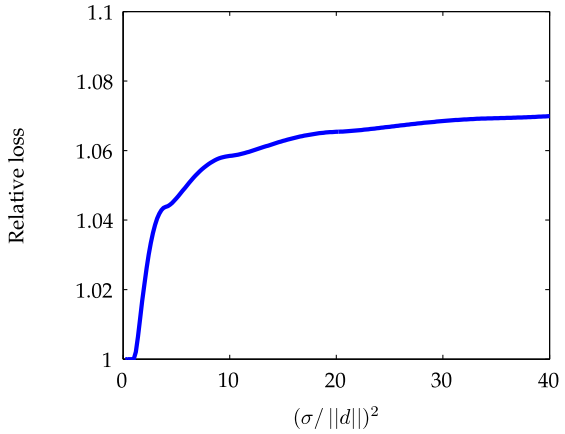

Figure 1: Relative loss of using $\varphi^*_{\exp}$ instead of $\varphi^*_{\mathcal{Y}}$

caused by using an exponentially decaying weight function instead of the optimal, linearly decaying weight function, may worsen the expected squared estimation error by as much as seven percent.

Remark 6 (Application to sequential decision problems)**.** The fact that the optimal weight function is independent of t, for all $t \geq t_d$, has interesting consequences for performance bounds in sequential optimization problems. Suppose that for each $t \in \mathbb{N}$, a decision maker has to make a decision x_t that minimizes some function $f(x, Y(t))$. For fixed values of y, let $x^*(y)$ be a minimizer of $f(x, y)$, and suppose $f(x^*(y'), y) - f(x^*(y), y) \leq C\,||y - y'||^2$ for some $C > 0$ and all y, y'. A myopic or passive learning policy is to always use the decision that is optimal with respect to the current estimate of $Y(t)$, i.e. $x_{t+1} = x^*(\hat{Y}[\varphi^*_{\mathcal{Y}}](t))$ for all t. The expected average loss caused by decisions $(x_t)_{t\in\mathbb{N}}$ over an infinite time horizon equals

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E\left[f(x_t, Y(t)) - f(x^*(Y(t)), Y(t))\right].$$

Now, for all parameter processes $(Y(t))_{t\in\mathbb{N}} \in \mathcal{Y}$, Proposition 1 and Theorem 2 provide a bound on the loss of the myopic policy, which offers the decision maker an explicit insight in the costs needed to hedge against changes in the parameter process:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} CE\left[\left|\left|\hat{Y}[\varphi^*_{\mathcal{Y}}](t) - Y(t+1)\right|\right|^2\right] \leq 2CG_{t_d,\mathcal{Y}}(\varphi^*_{\mathcal{Y}}).$$

3.3 Simple random walk

Let $\mathcal{Y}$ be the set of all one-dimensional stochastic processes $(Y(t))_{t\in\mathbb{N}}$ such that

$$Y(t) = \sum_{k=-\infty}^{t} e(k), \quad (t \in \mathbb{N}),$$

for some white-noise process $(e(t))_{t\in\mathbb{Z}}$ that satisfies $E[e(t)^2] \leq \varsigma^2$ for all $t \in \mathbb{Z}$ and some $\varsigma > 0$.

The following theorem characterizes the optimal weight function.

Theorem 3. *Up to multiplication by a strictly positive constant, there is a unique $\varphi^*_{\mathcal{Y}}$ that minimizes $G_{t+1,\mathcal{Y}}(\varphi)$, which satisfies*

$$\varphi^*_{\mathcal{Y}}(k-1) = \varphi^*_{\mathcal{Y}}(k) + \frac{\varsigma^2}{\sigma^2} \sum_{i=k}^{t-1} \varphi^*_{\mathcal{Y}}(i) \text{ for all } k = 1, \ldots, t-1. \tag{11}$$

The proof is based on the following auxiliary lemma.

Lemma 3. *Let $a > 0$, $b > 0$, $t \in \mathbb{N}$, and for all $x \in \mathbb{R}^t_+ \backslash \{0\}$, define*

$$f_t(x) = \Big(\sum_{i=1}^{t} x_i \Big)^{-2} \Big(a \sum_{i=1}^{t} \sum_{j=1}^{t} x_i x_j \min\{i,j\} + b \sum_{i=1}^{t} x_i^2 \Big).$$

There is a minimizer $x^*(t)$ of $f_t(x)$ on $\mathbb{R}^t_+ \backslash \{0\}$, which is unique up to multiplication by a strictly positive constant, and $x^*(t)$ satisfies

$$x^*_{k-1}(t) = x^*_k(t) + \sum_{i=k}^{t} x^*_i(t) a b^{-1} \text{ for all } k = 2, \ldots, t. \tag{12}$$

Remark 7. From equation (11) it can be shown that the optimal solution $\varphi^*_{\mathcal{Y}}$ with $\varphi^*_{\mathcal{Y}}(0) = 1$ satisfies $(2+(\varsigma/\sigma)^2)^{-k} \leq \varphi^*_{\mathcal{Y}}(k) \leq (1+(\varsigma/\sigma)^2)^{-k}$, for all $k = 0, \ldots, t-1$. Thus, $\varphi^*_{\mathcal{Y}}$ is contained between two exponentially decaying weight functions.

4 Discussion

In this paper we present a framework to estimate a multivariate time-varying process. It is based on the commonly used criterion of minimizing a weighted sum of the squared errors. The weight function is chosen to minimize the expected squared estimation error, in a robust sense: given a set of assumptions on the time-varying process, the weight function is selected that minimizes the worst-case expected squared estimation error. For three specific sets of

assumptions we calculate the optimal weight function explicitly: for bounded processes the optimal weight function is constant, for processes with bounds on the one-step changes the optimal weight function is linearly decaying, and for a simple random walk the optimal weight function is bounded from above and from below by an exponentially decaying function.

Our approach has three important advantages over existing methods. First, it does not require an explicit, detailed model (which is often prone to misspecification), but only some (possibly very general) assumptions on the parameter process. Second, it offers a mathematical foundation why a certain weight function should be selected; namely, because it minimizes the worst-case expected squared estimation error. Third, it comes with explicit bounds on the estimation error, which in several instances are tight. We show that for some sequential decision problems under uncertainty with time-varying parameters, these bounds directly translate into bounds on the performance of taken decisions (see Remark 6).

Our results point to several directions for future research. First, as already alluded to in Remark 2, a study of min-max optimal estimation functions w.r.t. the *cumulative* mean squared error is an important direction for future research, and would be useful in various time-series estimation problems.

Second, we note that although we optimize the weight function φ, the criterion of minimizing the weighted sum of squared errors in (2) is fixed in this paper. By differentiating (2) w.r.t. y, it follows that $\hat{Y}[\varphi](t)$ is the solution to the linear equation $G_t(y) = \sum_{i=1}^{t}(z(i)-y)\varphi(t-i) = 0$. One could study more general types of estimating functions, in the same spirit as the work of Godambe and Heyde (2010).

Third, in this paper we have elaborated the optimal weight function for three sets of assumptions. It would be interesting to elaborate the optimal $\varphi^*_{\mathcal{Y}}$ for other assumptions, for example by considering Markov chains with finite state-space, or by considering a class of time-series models.

Finally, from a practical perspective it is worthwhile to investigate if quantities like $\sigma/\left|\left|d\right|\right|$ and ς/σ, that determine the optimal weight functions in Section 3.2 and 3.3, can be estimated from data instead of imposing a certain value a priori.

Appendix: proofs

Proof of Proposition 1. By plugging (1) into (3) we obtain

$$\hat{Y}[\varphi](t) - Y(t+1) = \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-1} \sum_{i=1}^{t} \varphi(t-i)(Y(i) - Y(t+1) + \epsilon(i)).$$

Using $||a+b||^2 \leq 2(||a||^2 + ||b||^2)$,

$$E\left[\left|\left|\hat{Y}[\varphi](t) - Y(t+1)\right|\right|^2\right] \leq 2\mathcal{I}_{t+1,\mathcal{Y}}(\varphi) + 2E\left[\left|\left|\Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-1} \sum_{i=1}^{t} \varphi(t-i)\epsilon(i)\right|\right|^2\right].$$

Taking the supremum over $Y \in \mathcal{Y}$ and using $E[\epsilon_j(t)\epsilon_j(t') = 0]$ for all $t \neq t'$ yields (7). Equation (8) follows if $E[(Y_j(i) - Y_j(t+1))\epsilon_j(i)] = 0$, with equality if $E\left[\epsilon_j(i)^2\right] = \sigma_j^2$ for all $j = 1, \ldots, m$, $i \in \mathbb{N}$.

Proof of Proposition 2. For $\varphi \in \Delta_t$ we can write

$$\begin{aligned}\mathcal{I}_{t+1,Y}(\varphi) &= \sum_{j=1}^{m} \sum_{i=0}^{t-1} \sum_{k=0}^{t-1} E\left[(Y_j(t-i) - Y_j(t+1))(Y_j(t-k) - Y_j(t+1))\right] \varphi(i)\varphi(k) \\ &= \sum_{j=1}^{m} \varphi^T W(j) \varphi,\end{aligned}$$

where $W(j)$ is the $t \times t$ matrix $\left(E\left[(Y_j(t-i) - Y_j(t+1))(Y_j(t-k) - Y_j(t+1))\right]\right)_{1 \leq i,k \leq t}$. Writing $V(j) = (Y_j(t-0) - Y_j(t+1), Y_j(t-1) - Y_j(t+1), \ldots, Y_j(1) - Y_j(t+1))^T \in \mathbb{R}^t$, we have $W(j) = E\left[V(j)V(j)^T\right]$, which implies that $W(j)$ is positive definite for all $j = 1, \ldots, m$, and thus $\mathcal{I}_{t+1,Y}(\varphi)$ is convex on Δ_t. Then also $\sup_{Y \in \mathcal{Y}} \mathcal{I}_{t+1,Y}(\varphi)$ is convex on Δ_t, and since $\sigma^2\Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-2} \sum_{i=1}^{t} \varphi(t-i)^2$ is strictly convex in φ, it follows that $G_{t+1,\mathcal{Y}}(\varphi)$ is strictly convex on Δ_t and has a unique minimizer φ^*.

Proof of Theorem 1. Write $d = (d_1, \ldots, d_m)$. For all $Y \in \mathcal{Y}$,

$$\begin{aligned}\mathcal{I}_{t+1,Y}(\varphi) &= \sum_{j=1}^{m} \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-2} E\left[\left(\sum_{i=1}^{t} (Y_j(i) - Y_j(t+1))\varphi(t-i)\right)^2\right] \\ &\leq \sum_{j=1}^{m} \Big(\sum_{i=1}^{t} \varphi(t-i)\Big)^{-2} (2d_j)^2 \left(\sum_{i=1}^{t} \varphi(t-i)\right)^2 \\ &= 4\,||d||^2,\end{aligned}$$

with equality if $Y(1) = \ldots = Y(t) = d$, $Y_{t+1} = -d$, and thus

$$G_{t+1,\mathcal{Y}_d}(\varphi) = 4\,||d||^2 + \sigma^2\Big(\sum_{i=1}^{t}\varphi(t-i)\Big)^{-2}\sum_{i=1}^{t}\varphi(t-i)^2. \tag{13}$$

Application of Lemma 1 with $x_i = \varphi(i-1)$ then implies that for all fixed t, $\varphi(i) = 1/t$ $(i = 0, \ldots, t-1)$ minimizes $G_{t+1,\mathcal{Y}}(\varphi)$.

Proof of Lemma 1. Since $f_t(x) := \Big(\sum_{i=1}^{t} x_i\Big)^{-2}\sum_{i=1}^{t} x_i^2$ satisfies $f_t(cx) = f_t(x)$ for all $c \neq 0$, we have $\min_{x\in\mathbb{R}_+^t\setminus\{0\}} f_t(x) = \min_{x\in\mathbb{R}_+^t,\sum_{i=1}^t x_i=1}\sum_{i=1}^{t} x_i^2$. Simple algebra shows that the latter minimum is attained at $x = (1/t, \ldots, 1/t)$.

Proof of Theorem 2. For all $Y \in \mathcal{Y}$,

$$\begin{aligned}\mathcal{I}_{t+1,Y}(\varphi) &= \sum_{j=1}^{m}\Big(\sum_{i=1}^{t}\varphi(t-i)\Big)^{-2} E\left[\left(\sum_{i=1}^{t}(Y_j(i) - Y_j(t+1))\varphi(t-i)\right)^2\right] \\ &\leq \sum_{j=1}^{m}\Big(\sum_{i=1}^{t}\varphi(t-i)\Big)^{-2} d_j^2\left(\sum_{i=1}^{t}(t+1-i)\varphi(t-i)\right)^2,\end{aligned}$$

with equality if $Y(i) = d\cdot i$ for all $i = 1, \ldots, t+1$, and thus

$$G_{t+1,\mathcal{Y}}(\varphi) = \Big(\sum_{i=1}^{t}\varphi(t-i)\Big)^{-2}\left(||d||^2\left(\sum_{i=1}^{t}(t+1-i)\varphi(t-i)\right)^2 + \sigma^2\sum_{i=1}^{t}\varphi(t-i)^2\right). \tag{14}$$

The assertions of the theorem follow from Lemma 2, with $a = ||d||^2$, $b = \sigma^2$, and $x_i = \varphi(i-1)$ for $i = 1, \ldots, t$.

Proof of Lemma 2. Since $f_t(cx) = f_t(x)$ for all $c \neq 0$, we have

$$\min_{x\in\mathbb{R}_+^t\setminus\{0\}} f_t(x) = \min_{x\in\mathbb{R}_+^t\setminus\{0\},\sum_{i=1}^t x_i=1} g_t(x),$$

where $g_t(x) = a\Big(\sum_{i=1}^{t} i x_i\Big)^2 + b\sum_{i=1}^{t} x_i^2$. This is the minimum of a continuous, strict convex function on a compact convex set. The minimum is thus attained, and there is a unique minimizer $x^*(t) = (x_1^*(t), \ldots, x_t^*(t))$ that satisfies $\sum_{i=1}^{t} x_i^*(t) = 1$.

Observe that $x_j^*(t) \geq x_{j+1}^*(t)$ for all $j = 1, \ldots, t-1$, since if not then interchanging the j-th and $(j+1)$-th component strictly decreases f_t.

Suppose $x_k^*(t) > 0$ for some k. Then $\frac{\partial f_t}{\partial x_k}(x^*(t)) = 0$. Since

$$\frac{\partial f_t}{\partial x_k}(x) = 2\frac{a(\sum_{i=1}^t ix_i)k + bx_k - f_t(x)(\sum_{i=1}^t x_i)}{(\sum_{i=1}^t x_i)^2},$$

this implies $x_k^*(t) = b^{-1}f_t(x^*(t))(\sum_{i=1}^t x_i^*(t)) - b^{-1}a(\sum_{i=1}^t ix_i^*(t))k$, and $x_k^*(t) > 0$ implies $f_t(x^*(t)) > a(\sum_{i=1}^t ix_i^*(t))(\sum_{i=1}^t x_i^*(t))^{-1}k \geq ak$. For $k \geq 1 + ba^{-1}$ this contradicts the minimality of $f_t(x^*(t))$, since $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}_+^t$ satisfies $f_t(e_1) = a + b \leq ak < f_t(x^*(t))$. Thus, for all $t \geq 1 + ba^{-1}$ and $k \geq 1 + ba^{-1}$, $x_k^*(t) = 0$.

In fact, the following holds: if $x_t^*(t) = 0$, then the minimizer of f_{t-1} in $\{x \in \mathbb{R}_+^{t-1} \mid \sum_{i=1}^{t-1} x_i = 1\}$ is $(x_1^*(t), \ldots, x_{t-1}^*(t))$. This can be seen as follows: let $x_t^*(t) = 0$, and suppose $f_{t-1}(x^*(t-1)) < f_{t-1}(x_1^*(t), \ldots, x_{t-1}^*(t))$. Then

$$\begin{aligned} f_t(x_1^*(t-1), \ldots, x_{t-1}^*(t-1), 0) &= f_{t-1}(x_1^*(t-1), \ldots, x_{t-1}^*(t-1)) \\ &< f_{t-1}(x_1^*(t), \ldots, x_{t-1}^*(t)) = f_t(x_1^*(t), \ldots, x_{t-1}^*(t), 0) = f_t(x^*(t)), \end{aligned}$$

contradicting the minimality of $f_t(x^*(t))$ on $\{x \in \mathbb{R}_+^t \mid \sum_{i=1}^t x_i = 1\}$.

This implies that there is a t_d such that $x_{t_d}^*(t_d) > 0$, and

$$x^*(t) = (x_1^*(t_d), x_2^*(t_d), \ldots, x_{t_d}^*(t_d), 0, \ldots, 0) \in \mathbb{R}^t, \quad \text{for all } t \geq t_d.$$

Since $\frac{\partial f_t}{\partial x_k}(x^*(t)) = 0$ for all $k = 1, \ldots, t$, $t \leq t_d$, we have

$$x_k^*(t) = b^{-1}f_t(x^*(t)) - b^{-1}a(\sum_{i=1}^t ix_i^*(t))k,$$

for all $k = 1, \ldots, t$, $t \leq t_d$, and thus for all $t \in \mathbb{N}$, $x^*(t)$ is of the form

$$x_k^*(t) = (\alpha(t) - \beta(t)k)^+,$$

with $\alpha(t) = b^{-1}f_t(x^*(t))$ and $\beta(t) = b^{-1}a(\sum_{i=1}^t ix_i^*(t))$.

Proof of Theorem 3.

We have $Y(i) - Y(t+1) = -\sum_{k=i+1}^{t+1} e(k)$ for $i < t$, and thus

$$\begin{aligned}
&E[(Y(i) - Y(t+1))(Y(j) - Y(t+1))] \\
&= \sum_{k=i+1}^{t+1} \sum_{l=j+1}^{t+1} E[e(k)e(l)] \leq \sum_{k=i+1}^{t+1} \sum_{l=j+1}^{t+1} \varsigma^2 \mathbf{1}_{k=l} \\
&= \sum_{k=\max\{i,j\}+1}^{t+1} \varsigma^2 = (t+1-\max\{i,j\})\varsigma^2
\end{aligned}$$

for all $1 \leq i, j \leq t$. Equality holds in the equations above if $E[e(k)^2] = \varsigma^2$ for all k.

Then

$$\begin{aligned}
\mathcal{I}_{t+1}(\varphi) &= (\sum_{i=1}^{t} \varphi(t-i))^{-2} \sum_{i=1}^{t} \sum_{j=1}^{t} \varphi(t-i)\varphi(t-j) E[(Y(i) - Y(t+1))(Y(j) - Y(t+1))] \\
&\leq \varsigma^2 (\sum_{i=1}^{t} \varphi(t-i))^{-2} \sum_{i=1}^{t} \sum_{j=1}^{t} \varphi(t-i)\varphi(t-j)(t+1-\max\{i,j\}),
\end{aligned}$$

with equality if $(e(k))_{k\in\mathbb{Z}}$ is homoscedastic, and thus

$$\begin{aligned}
G_{t+1,\mathcal{Y}}(\varphi) = \left(\sum_{i=1}^{t} \varphi(t-i)\right)^{-2} \Bigg(&\varsigma^2 \sum_{i=1}^{t} \sum_{j=1}^{t} \varphi(t-i)\varphi(t-j)(t+1-\max\{i,j\}) \\
&+ \sigma^2 \sum_{i=1}^{t} \varphi(t-i)^2 \Bigg).
\end{aligned}$$

The assertions of the theorem follow from Lemma 3, with $a = \varsigma^2$, $b = \sigma^2$, and $x_i = \varphi(i-1)$ for $i = 1, \ldots, t$.

Proof of Lemma 3. Since $f_t(cx) = f_t(x)$ for all $c \neq 0$, we have

$$\min_{x \in \mathbb{R}_+^t \setminus \{0\}} f_t(x) = \min_{x \in \mathbb{R}_+^t \setminus \{0\}, \sum_{i=1}^t x_i = 1} g_t(x),$$

where $g_t(x) = a \sum_{i=1}^t \sum_{j=1}^t x_i x_j \min\{i,j\} + b \sum_{i=1}^t x_i^2$. This is the minimum of a continuous, strict convex function on a compact convex set. The minimum is thus attained, and there is a unique minimizer $x^*(t) = (x_1^*(t), \ldots, x_t^*(t))$ that satisfies $\sum_{i=1}^t x_i^*(t) = 1$.

Claim. $x_j^*(t) > 0$ for all $j = 1, \ldots, t$.

Proof of claim. Let k be the largest integer in $\{1, \ldots, t\}$ such that $x^*(t) > 0$, and suppose $k < t$. Observe that $x_j^*(t) \geq x_{j+1}^*(t)$ for all $j = 1, \ldots, t-1$, since if not then interchanging the j-th and $(j+1)$-th component strictly decreases f_t. This implies $x_j^*(t) > 0$ for all $j = 1, \ldots, k$. Moreover, $(x_1^*(t), \ldots, x_k^*(t)) = (x_1^*(k), \ldots, x_k^*(k))$, since if not then

$g_t(x_1^*(k),\dots,x_k^*(k),0,\dots,0) = g_k(x^*(k)) < g_k(x_1^*(t),\dots,x_k^*(t)) = g_t(x^*(t))$, contradicting the minimality of $g_t(x^*(t))$.

Because $x_k^*(k) > 0$ we have

$$\frac{\partial f_k}{\partial x_k}(x) = 2\frac{a\sum_{i=1}^k x_i \min\{i,k\} + bx_k - f_k(x)(\sum_{i=1}^k x_i)}{(\sum_{i=1}^k x_i)^2} = 0.$$

This implies $x_k^*(k) = b^{-1}f_k(x^*(k))(\sum_{i=1}^k x_i^*(k)) - b^{-1}a(\sum_{i=1}^k x_i^*(k)\min\{i,k\}) > 0$, and thus $f_k(x^*(k)) > a(\sum_{i=1}^k x_i^*(k))^{-1}(\sum_{i=1}^k x_i^*(k)\min\{i,k\})$; in particular, $g_k(x^*(k)) > a\sum_{i=1}^k x_i^*(k)i$.

For $\gamma \in [0,1]$ let

$$x(\gamma) = (\gamma x_1^*(k),\dots,\gamma x_k^*(k),0,\dots,0,1-\gamma) \in \mathbb{R}^t.$$

Note that $\sum_{i=1}^t x_i(\gamma) = 1$. We have

$$g_t(x(\gamma)) = \gamma^2 g_k(x^*(k)) + 2a\gamma(1-\gamma)\sum_{i=1}^k i x_i^*(k) + (1-\gamma)^2(at+b).$$

The derivative $\frac{\partial}{\partial\gamma}g_t(x(\gamma))$ of $g_t(x(\gamma))$ with respect to τ, evaluated at $\tau = 1$, is equal to $2g_k(x^*(k)) - 2a\sum_{i=1}^k i x_i^*(k)$, which is strictly larger than zero. This implies that there is a $\gamma \in [0,1)$ such that $g_t(x(\gamma)) < g_t(x(1)) = g_t(x^*(t))$, contradicting the minimality of $g_t(x^*(t))$. **End of proof claim.**

Having proved $x_k^*(t) > 0$ for all $k = 1,\dots,t$, it follows from

$$\frac{\partial f_t}{\partial x_k}(x) = 2\frac{a\sum_{i=1}^t x_i \min\{i,k\} + bx_k - f_t(x)(\sum_{i=1}^t x_i)}{(\sum_{i=1}^t x_i)^2}$$

that $a\sum_{i=1}^t x_i^*(t)\min\{i,k\} + bx_k^*(t) - f_t(x^*(t))(\sum_{i=1}^t x_i^*(t)) = 0$ for all $k = 1,\dots,t$. In particular, this implies

$$\begin{aligned} x_{k-1}^*(t) &= b^{-1}f_t(x^*(t))(\sum_{i=1}^t x_i^*(t)) - ab^{-1}\sum_{i=1}^t x_i^*(t)\min\{i,k-1\} \\ &= x_k^*(t) + ab^{-1}\sum_{i=k}^t x_i^*(t), \end{aligned}$$

for all $2 \le k \le t$.

Acknowledgements

I thank Neil Walton for providing constructive comments and suggestions. Part of this research was done while the author was affiliated with Centrum Wiskunde en Informatica (CWI), Amsterdam, Eindhoven University of Technology, and University of Amsterdam.

References

A. Ben-Tal, L. El Ghaoui, and A. S. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton and Oxford, 2009.

C. Chatfield, A. B. Koehler, J. K. Ord, and R. D. Snyder. A new look at models for exponential smoothing. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 50(2):147–159, 2001.

E. S. Gardner. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4(1):1–28, 1985.

E. S. Gardner. Exponential smoothing: The state of the art - part II. *International Journal of Forecasting*, 22 (4):637–666, 2006.

V. P. Godambe and C. C. Heyde. Quasi-likelihood and optimal estimation. In R. Maller, I. Basawa, P. Hall, and E. Seneta, editors, *Selected Works of C.C. Heyde*, chapter 49, pages 386–399. Springer, New York, 2010.

J. F. Muth. Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290):299–306, 1960.

S. Satchell and A. Timmermann. On the optimality of adaptive expectations: Muth revisited. *International Journal of Forecasting*, 11(3):407–416, 1995.