Forecasting The Boat Race

G. Mesters[(a,b,c)] *and S.J. Koopman*[(b,c,d)]

[(a)] Netherlands Institute for the Study of Crime and Law Enforcement,
[(b)] Department of Econometrics, VU University Amsterdam,
[(c)] Tinbergen Institute,
[(d)] CREATES, Aarhus University

May 15, 2014

Abstract

We study the forecasting of the yearly outcome of the Boat Race between Cambridge and Oxford. We compare the relative performance of different dynamic models for forty years of forecasting. Each model is defined by a binary density conditional on a latent signal that is specified as a dynamic stochastic process with fixed predictors. The out-of-sample predictive ability of the models is compared between each other by using a variety of loss functions and predictive ability tests. We find that the model with its latent signal specified as an autoregressive process cannot be outperformed by the other specifications. This model is able to correctly forecast 31 out of 40 outcomes of the Boat Race.

Keywords: Binary time series, Predictive ability, Non-Gaussian state space model

1 Introduction

The Boat Race between the universities of Cambridge and Oxford is an annual rowing event that started in 1829 and is by now the most popular one day sports event in the world. Each year, a crew from the University of Cambridge rows against a crew from the University of Oxford. The rowing crews consist of eight rowers and one cox. The race takes place on the river Thames between Putney and Mortlake. We consider a class of models for the Boat Race outcomes and investigate the ability of the models to forecast the next outcome of the Boat Race. More specifically, for a variety of dynamic binary model specifications, we assess their ability to forecast the outcome of the race between 1971 and 2010 using information that is available to us just prior to the race. The relative out-of-sample forecasting performance of the models is compared for different loss functions and equal predictive ability tests; see Diebold & Lopez (1996), Diebold & Mariano (1995) and West (1996).

Forecasting the Boat Race is interesting from a number of viewpoints. First, bookmakers and gamblers may increase their expected profits by using the forecasts. Especially, if the forecasts from the dynamic models are significantly better than the forecasts resulting from simple procedures, such as flipping a coin or setting the forecast equal to the winner of the previous race. Second, from an econometric forecasting perspective, the historical outcomes of the Boat Race form an interesting binary time series. While the number of observations can be limited, the observations themselves can be spread over a long period of time and with many missing entries. Binary time series occur in many fields of research such as finance, criminology and computer science. The generation of accurate forecasts for events (such as winning a race, intervening in a market or closing a computer network) can be of vital importance for many instances.

The presentation of our empirical results for the Boat Race forecasting study in this volume provides us with an opportunity to praise the innovative and empirically motivated time series studies of Andrew Harvey. Although most of his influential work is motivated by statistical and econometric challenges in economics and finance, he also has shown interest in tackling problems in the analysis of time series that have an even more direct impact on society. A good illustration is his extended study with James Durbin concerning the effect of the seatbelt legislation on the monthly numbers of road passengers killed or seriously injured in Great Britain, see Harvey & Durbin (1986). His enthusiasm for the structural time series approach becomes even more apparent when he illustrates the methodology in fields as diverse as crime ("Purses snatched in Hyde Park area of Chicago"), environment ("Rainfall in Fortaleza, north-east Brazil" and "Ice volumes measured at intervals of 2,000 years") and many others ("Australian telephone calls to Britain" and "Consumption of spirits in UK"). His work with Cristiano Fernandes on the exponential family of structural time series models for count data or qualitative observations has also inspired Andrew to work on some more colourful illustrations. For example, in Harvey & Fernandes (1989) a convincing analysis is presented of the results of international football matches played between England and Scotland. The binary time series of the yearly Boat Race has also not escaped their attention. In the years when the second author was just starting to work on his PhD under the superb guidance of Andrew, Cristiano and Andrew already completed a full time series analysis of the Boat Race data using their methodology in Fernandes & Harvey (1990).

In our current study we first investigate the in-sample performance of different binary models. The models are summarized by a binary density for the outcome of the race that is defined conditional on a signal that depends on an observable deterministic component and on a latent, stochastic and dynamic component. The deterministic component includes a constant and a vector of predictors, such as the average difference in weight between the boats, the winner of the toss and the previous winning distance. The latent stochastic component is modeled by a Gaussian dynamic processes such as random walk, autoregressive, autoregressive fractionally integrated and stochastic cycle processes. The parameters of the models are estimated by the method of Monte Carlo maximum likelihood. In particular, for the stochastic processes that imply a short-memory process we implement the simulation-based methods of Shephard & Pitt (1997), Durbin & Koopman (1997) and Jungbacker &

Koopman (2007). These methods are based on the importance sampling technique. When the stochastic component is a long-memory process we alter the estimation method based on the procedures discussed in Mesters, Koopman & Ooms (2011).

We further perform an out-of-sample forecasting study. Each model is used to compute a probability forecast for the event that Cambridge wins the next Boat Race. This is done repeatedly for a period of forty years (1971-2010), where we only use information that would have been available just before the race starts. Forecasting in this manner is referred to as "pseudo-out-of-sample" forecasting by Stock & Watson (2003). The relative out-of-sample predictive performance of the different models is compared by different loss functions, or score statistics. We verify whether the models possess significantly different predictive abilities as implied by different loss functions, see Diebold & Mariano (1995) and West (1996).

The remainder of our paper is organized as follows. We continue the introduction with a short outline of the history of the Boat Race and describe the data that is obtained from the history of the race. Section 2 discusses the binary models that we consider for forecasting the race. Section 3 discusses the forecasting methodology and the predictive ability tests. It further presents and discusses the results of forty years of out-of-sample forecasting. Section 4 provides a brief summary of our study.

1.1 History

The first race between the universities of Oxford and Cambridge was organized in 1829. The idea came from two friends; Charles Merivale, a student at Cambridge, and Charles Wordsworth, a student at Oxford. On 12 March 1829 Cambridge sent a challenge to Oxford and the history of the race started. The first race was at Henley-on-Thames and was won by Oxford. In 1839 the race was relocated to London, by now the race had become an annual event taking place between Westminster and Putney. However, the increased crowds interested in the race made it necessary to move yet again. In 1845 the race was first held on the course between Putney and Mortlake, which is also on the river Thames. This course of the Boat Race is displayed in Figure 1 which is reprinted from Drinkwater (1939). In 1836 the Oxford crew selected the color dark blue to race in and the Cambridge crew selected the "duck egg blue" of Eton to race in. In 1849, after a gap of two years and for the first and only time, there were two races. The first in March from Putney to Mortlake was won by Cambridge. Oxford challenged Cambridge to a second race in December. This second race of 1849 is the only time in the series to date that the Boat Race was decided by a disqualification, following a foul by Cambridge at the finish.

The race outcome were about even until 1861, when Oxford started the first winning streak in the history lasting nine years. In 1877, the race was declared a dead heat for the first and only time in its history - although legend has it that the judge, "Honest John" Phelps, was asleep under a bush when the crews reached the finishing line. The 1912 event witnessed another Boat Race first when both boats sank and the race had to be re-run the next day, with Oxford claiming the honors at the second attempt. The race was not held between 1915 and 1919 due to the First World War. When it resumed in 1920, Cambridge

embarked on a lengthy period of domination. They would win the race 13 years running between 1924 and 1936, the longest winning streak in the race's history.

There was another break between 1940 and 1945 because of the Second World War, although four unofficial races were held during this time, all outside London. The 1952 contest witnessed perhaps the most extreme weather in Boat Race history, with Oxford prevailing in the midst of a blizzard. The dark blues also won the 100th Boat Race in 1954. In 1981, Sue Brown became the first female to enter the Boat Race, acting as cox for Oxford. The following year, Hugh and Rob Clay of Oxford became the first twins to win the race. The dark blues dominated throughout the Eighties, as Cambridge suffered a series of misfortunes. The biggest of these came in 1984, when they managed to write off their boat before the start of the race. Controversy engulfed Oxford at the 1987 race when a section of the crew rose up in mutiny against the president over team selection policy. However, the dispute, which was chronicled in the book and film 'True Blue', did not prevent them from winning the race again.

Cambridge regained their pride in 1993 by ending Oxford's domination. They subsequently won the race seven years running, the highlight coming in 1998 when they broke the course record by a massive 26 seconds. The last race that we consider was won by Cambridge in 2010.

1.2 Data

The dependent variables in our study are denoted by the elements of the $n \times 1$ vector $\mathbf{y} = (y_1, \ldots, y_n)'$, which indicate the outcome path of the Boat Race. All data is freely obtained from `http://www.theboatrace.org`. For each year t, where $t = 1$ refers to the year 1829 and $t = n$ to the year 2010, we let $y_t = 1$ indicate that Cambridge has won the race in year t, whereas $y_t = 0$ indicates that Oxford has won the race in year t. For some years the race was not held due to a variety of circumstances, for these years we consider the corresponding outcomes as missing at random. In particular, the years 1830-1835, 1837, 1838, 1843, 1844, 1847, 1848, 1850, 1851, 1853, 1855, 1877 (dead heat), 1915-1919 (WWI) and 1940-1945 (WWII) are considered missing at random. Only the first race of the year 1849 (the March race) is included as the second race ended by disqualification. In total we have $n = 182$ observations, of which 28 are missing. By 2010, Cambridge lead the series, with 80 wins compared to 74 wins for Oxford.

During the history of the race the average weight of the rowers has increased substantially. This has led to the suspicion that increasing the weight of the rowers leads to faster boats, due to the presence of more muscle power. The contrary could however also be possible, by reasoning that a lighter boat has less water resistance and would therefore be faster. We include the average log difference in weight between the rowers in the Cambridge and Oxford boats as the first predictor.

The second time-varying predictor that we include is the outcome of the coin toss. The club presidents toss a coin before the race for the right to choose which side of the river (station) they will row on. Their decision is based on the weather conditions and how the

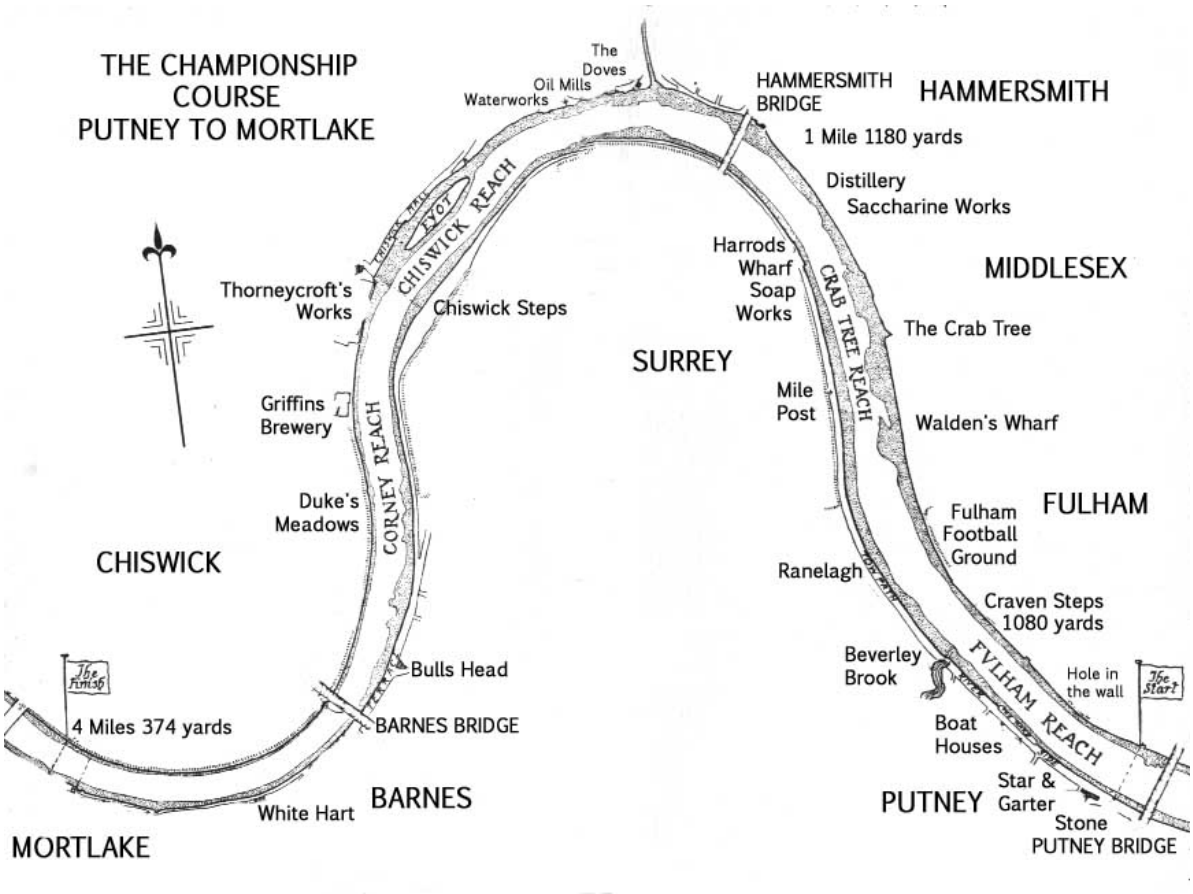

Figure 1: The course of The Boat Race (reprinted from George Carr Drinkwater's book "The Boat Race").

various bends in the course might favor the pace of their crew. The north station (Middlesex) has the advantage of the first and last bends, whereas the south (Surrey) station has the longer middle bend. It is generally believed that the winner of the coin toss has an improved chance of winning. In most years, the betting odds change severely after the toss outcome has been made public.

The third and final covariate that we include is the distance by which the previous race was won. This winning distance may be viewed as a proxy for the technological and physical gap in ability between the two rowing teams. As the crew members, coaches and technological advantages may overlap subsequent years, larger winning distances might measure structural advantages. The variable is constructed from the perspective of Cambridge. It is positive when Cambridge has won the previous race and negative when Oxford has won the previous race. The magnitude is measured in terms of boat lengths.

The $n \times 3$ matrix $\mathbf{X}$ includes the time-varying predictors, where element $x_{t,1}$ is the average log difference in weight between the rowers of Cambridge and Oxford in year t, $x_{t,2} = 1$ if Cambridge wins the toss in year t and zero otherwise and $x_{t,3}$ is the winning distance in boat lengths from the previous race.

2 Models and parameter estimation

The models that we use as our basis to forecast the outcome path of the race can be split into two parts; a conditional binary density for the observations and a signal that includes all dynamics and predictors. The conditional binary density is given by

$$p(y_t|\pi_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t}, \qquad t = 1, \ldots, n, \tag{1}$$

where π_t denotes the winning probability for Cambridge in year t. Time-varying probability π_t is unknown to us, but is clearly restricted between zero and one. Therefore we specify the transformed probability as $\theta_t = \log(\pi_t/(1-\pi_t))$ where the transformation is performed by the canonical link function for binary models; see Cox & Snell (1989). We refer to θ_t to as the signal for year t. The conditional density for the observations, given in (1) can be rewritten in terms of the latent signal as

$$p(y_t|\theta_t) = \exp[y_t\theta_t - \log(1+\exp\theta_t)], \qquad t = 1, \ldots, n, \tag{2}$$

which is assumed independent over $t = 1, \ldots, n$. It holds that $p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{t=1}^{n} p(y_t|\theta_t)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$. We decompose the signal in an observed deterministic part and a latent stochastic part. The deterministic part includes the intercept and three predictors (difference in weight of boat, winner of toss and the previous winning distance). The latent stochastic part captures the possible dynamic effects which reflect that the events of previous years also affect outcome of the race this year. The signal decomposition is given by the regression equation

$$\theta_t = \mu + \mathbf{x}_t\boldsymbol{\beta} + u_t, \qquad t = 1, \ldots, n, \tag{3}$$

where μ is the intercept, the 3×1 parameter vector $\boldsymbol{\beta}$ measures the effect of the observable predictors $\mathbf{x}_t$ and u_t is the latent stochastic part. The 1×3 vector $\mathbf{x}_t$ is the t^{th} row of matrix $\mathbf{X}$. The models that we consider differ in the way that the latent process u_t is specified. Possible specifications that we include range from setting it to zero, which leads to a logistic regression model, to a complete dynamic specification. In the next subsection we discuss all specifications in some detail.

2.1 Signal specifications

Many dynamic specifications for u_t are possible. In general, we can expect to have a serial correlation structure between outcomes in successive years due to overlapping crews and coaches. For example, during the first winning streak of Oxford, between 1862 and 1869, the crew had the same coach George Morrison for six years who was the former oarsman of the crew. Furthermore, technological developments resulting from training methods or boat construction may provide comparative advantages to a team that can last for a number years. We focus on models for u_t that are able to pick up this kind of dynamic effects.

(i) The simplest possible specification is the "constant" specification given by

$$u_t = 0 \qquad t = 1, \ldots, n, \tag{4}$$

which results in a signal that is completely determined by the intercept and the three observed predictors. This is the baseline specification, which assumes that the entire signal is observable. In effect the model for y_t is reduced to a logistic regression model.

(ii) The first specification that we consider to capture serial correlation for u_t is the random walk as given by

$$u_{t+1} = u_t + \eta_{t+1}, \qquad \eta_{t+1} \sim N(0, \sigma_\eta^2), \qquad t = 2, \ldots, n, \tag{5}$$

where the next value of u_t is equal to the old value plus an error term. The initial state, u_1, has an unknown distribution and is therefore fixed at zero. This corresponds to a fifty percent winning probability for Cambridge without covariates.

(iii) The dependence of the outcome on previous realisations of u_t can be tampered by using an autoregressive (AR) specification. The coefficients of the AR(p) polynomial $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ enable the adjustment of observations from the past, where B denotes the back shift operator as defined by $B^s y_t = y_{t-s}$ for any integer $s > 0$. The latent process u_t is given by

$$\phi(B) u_{t+1} = \eta_{t+1}, \qquad \eta_{t+1} \sim N(0, \sigma_\eta^2), \qquad t = p + 1, \ldots, n. \tag{6}$$

When we require to impose a stationary process for u_t, as opposed to the non-stationary random walk process (5), we need to assume that the roots of the autoregressive

polynomial $\phi(B)$ lie outside the unit circle. The process is initialized for $u_1, \ldots, u_p$ by the unconditional distribution. For example, when $p = 1$ we have

$$u_1 \sim N\left\{0, \sigma_\eta^2/(1-\phi_1^2)\right\}.$$

The resulting model with the AR signal can be alternatively be viewed as a dynamic logit model. That is the model where the signal is given by $u_t = 0$ and $\mathbf{x}_t$ includes $y_{t-1}, \ldots, y_{t-p}$.

(iv) Since some of the winning streaks are relatively long, we also consider a long memory specification for the latent process u_t. In particular, we investigate the autoregressive fractionally integrated (ARFI) model; see Granger & Joyeux (1980) and Palma (2007). The model for u_t can be expressed by

$$\phi(B)(1-B)^d u_{t+1} = \eta_{t+1}, \qquad \eta_{t+1} \sim N(0, \sigma_\eta^2), \qquad t = p+1, \ldots, n, \tag{7}$$

where the fractional integration part can be expressed by the binomial expansion

$$(1-B)^d = \sum_{k=0}^{\infty} \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}(-1)^k B^k.$$

The fractional parameter d is a real valued constant in the range of $-1 < d < 0.5$. This assumption together with the previous assumption on the autoregressive polynomial ensures that the process is stationary, invertible and causal; see Palma (2007) for a proof.

(v) Another possible component that may be present in some time series is the stochastic cycle process of Harvey (1989). To have a stochastically time-varying cycle component for u_t, we consider the following specification

$$\begin{bmatrix} u_{t+1} \\ u_{t+1}^* \end{bmatrix} = \phi \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} u_t \\ u_t^* \end{bmatrix} + \begin{bmatrix} \eta_{t+1} \\ \eta_{t+1}^* \end{bmatrix}, \qquad t = 2, \ldots, n, \tag{8}$$

where the errors term η_t and η_t^* are serially and mutually uncorrelated with mean zero and variance σ_η^2. The parameter λ gives the frequency of the cycle and is estimated together with the other parameters. The period of the cycle is $2\pi/\lambda$. Restrictions on the damping term, $0 \leq \phi \leq 1$, ensure that the process u_t is stationary. The cycle disturbance variance for both η_t and η_t^* is specified as $\sigma_\eta^2 = (1-\phi^2)\sigma_u^2$ where the cycle variance σ_u^2 is for u_t and is treated as the parameter to estimate. The initial distribution of the cycle is then given by

$$u_1 \sim N\left(0, \sigma_u^2\right), \qquad u_1^* \sim N\left(0, \sigma_u^2\right), \qquad E(u_1 u_1^*) = 0.$$

It follows that a constant but stochastic cycle is obtained with $\phi = 1$ which remains properly defined; see Harvey & Koopman (1997).

The models that we consider can be summarized by the observational density (2), the signal (3) and one of the specification for u_t in (4), (5), (6), (7) and (8). For each model specification, the unknown parameters are collected in the vector $\boldsymbol{\psi}$.

2.2 Parameter estimation

Different estimation procedures may be considered for different model specifications. For exposition and comparison purposes we use the same Monte Carlo maximum likelihood estimation procedure for the estimation of the parameter vector $\boldsymbol{\psi}$ in all cases. In particular, we adopt an estimation procedure that is based on the simulation methods developed in Shephard & Pitt (1997) and Durbin & Koopman (1997). Alternatively, other simulation based estimation methods may be modified to facilitate the estimation of the different models specifications, see for example West (1985), Chib & Jeliazkov (2006) Czado & Song (2008) and Richard & Zhang (2007).

The loglikelihood, for all models, is defined as $\ell(\boldsymbol{\psi}) = \log p(\mathbf{y};\boldsymbol{\psi})$ where $p(\mathbf{y};\boldsymbol{\psi})$ is the joint density of all observations. The terms after the semi-colon in $p(\cdot;\cdot)$ are the fixed, possibly unknown, arguments of the density function $p(\cdot;\cdot)$. The joint density for models defined by density (2) and signal (3) is given by

$$p(\mathbf{y};\boldsymbol{\psi}) = \int_{\boldsymbol{\theta}} p(\mathbf{y},\boldsymbol{\theta};\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta} = \int_{\mathbf{u}} p(\mathbf{y},\mathbf{u};\mathbf{X},\boldsymbol{\psi})\mathrm{d}\mathbf{u} = \int_{\mathbf{u}} p(\mathbf{y}|\mathbf{u};\mathbf{X},\boldsymbol{\psi})p(\mathbf{u};\boldsymbol{\psi})\mathrm{d}\mathbf{u}, \tag{9}$$

where $\mathbf{u} = (u_1,\ldots,u_n)'$ and $p(\cdot)$, $p(\cdot,\cdot)$ and $p(\cdot|\cdot)$ denote marginal, joint and conditional density functions, respectively. Since $\mathbf{X}$ and $\boldsymbol{\psi}$ are fixed we only need to integrate the stochastic process u_t from the joint density $p(\mathbf{y},\mathbf{u};\mathbf{X},\boldsymbol{\psi})$. We further have

$$p(\mathbf{y}|\mathbf{u};\mathbf{X},\boldsymbol{\psi}) = p(\mathbf{y}|\boldsymbol{\theta};\boldsymbol{\psi}) = \prod_{t=1}^{n} p(y_t|\theta_t;\boldsymbol{\psi}),$$

and

$$p(\mathbf{u};\boldsymbol{\psi}) = p(u_1,\ldots,u_p;\boldsymbol{\psi}) \prod_{t=p+1}^{n} p(u_t|u_1,\ldots,u_{t-1};\boldsymbol{\psi}).$$

A closed form solution for integral (9) is not available if u_t is stochastic and if $p(y_t|\theta_t;\boldsymbol{\psi})$ is binary. Instead we rely on numerical methods to solve the integral. In particular, we adopt the importance sampling method as it is discussed in, for example, Ripley (1987). It holds that

$$p(\mathbf{y};\boldsymbol{\psi}) = \int_{\mathbf{u}} \frac{p(\mathbf{y}|\mathbf{u};\mathbf{X},\boldsymbol{\psi})p(\mathbf{u};\boldsymbol{\psi})}{g(\mathbf{u}|\mathbf{y})} g(\mathbf{u}|\mathbf{y})\mathrm{d}\mathbf{u} = g(\mathbf{y}) \int_{\mathbf{u}} \frac{p(\mathbf{y}|\mathbf{u};\mathbf{X},\boldsymbol{\psi})p(\mathbf{u};\boldsymbol{\psi})}{g(\mathbf{y}|\mathbf{u})g(\mathbf{u})} g(\mathbf{u}|\mathbf{y})\mathrm{d}\mathbf{u}, \tag{10}$$

where $g(\mathbf{u}|\mathbf{y}) = g(\mathbf{y}|\mathbf{u})g(\mathbf{u})/g(\mathbf{y})$ is referred to as the importance density, which must be proportional to $p(\mathbf{y},\mathbf{u};\mathbf{X},\boldsymbol{\psi})$, easy to sample from and easy to compute. In our approach, we take the importance density as Gaussian with a linear relation between y_t and $\theta_t = \mu + \mathbf{x}_t\boldsymbol{\beta} + u_t$ for $t = 1,\ldots,n$ and for which it holds that $g(\mathbf{y}|\mathbf{u}) = \prod_{t=1}^{n} g(y_t|u_t)$. The mean and variance of

the importance density is chosen such that the corresponding first two moments of $p(y_t|\theta_t;\boldsymbol{\psi})$ and $g(y_t|u_t)$ with respect to u_t are set equal. The resulting set of nonlinear equations can be solved iteratively using Kalman filter and smoothing methods; see Durbin & Koopman (2012, Section 10.6). A Monte Carlo estimator for (10) is given by

$$\hat{p}(\mathbf{y};\boldsymbol{\psi}) = g(\mathbf{y})M^{-1}\sum_{i=1}^{M}\frac{p(\mathbf{y}|\mathbf{u}^{(i)};\mathbf{X},\boldsymbol{\psi})p(\mathbf{u}^{(i)};\boldsymbol{\psi}))}{g(\mathbf{y}|\mathbf{u}^{(i)})g(\mathbf{u}^{(i)})}, \tag{11}$$

where samples $\mathbf{u}^{(i)}$ are drawn from $g(\mathbf{u}|\mathbf{y})$ for $i = 1,\ldots,M$ using simulation smoothing methods.

For any importance density, it holds that $\hat{p}(\mathbf{y};\boldsymbol{\psi}) \to p(\mathbf{y};\boldsymbol{\psi})$ as $M \to \infty$, which is implied by the law of large numbers. The Lindeberg-Levy central limit theorem implies a $\sqrt{M}$ convergence rate if draws from the importance sampler are independent and if importance weights

$$w(\mathbf{u}^{(i)},\mathbf{y};\mathbf{X},\boldsymbol{\psi}) = \frac{p(\mathbf{y}|\mathbf{u}^{(i)};\mathbf{X},\boldsymbol{\psi})p(\mathbf{u}^{(i)};\boldsymbol{\psi}))}{g(\mathbf{y}|\mathbf{u}^{(i)})g(\mathbf{u}^{(i)})} \qquad i = 1,\ldots,M. \tag{12}$$

have finite mean and variance; see the discussions in Geweke (1989). Failure of this condition to hold can lead to slow and unstable convergence of the estimator.

Further technical details of importance sampling methods for this class of time series models are discussed in Jungbacker & Koopman (2007). All our estimation results are based on this implementation. Mesters et al. (2011) have extended their procedures to allow for the long-memory (ARFI) specification of the signal in (7).

Missing Values

In the years that the Boat Race was not held the outcomes are considered missing at random. This seems reasonable as there exists a variety of reasons for which the race was not held, see the discussion in Section 1. The treatment of missing values within our estimation framework is relatively straightforward. The likelihood estimator (11) consists of two components: the likelihood of the approximating model $g(\mathbf{y})$ is rescaled by the mean of the importance sampling weights $w(\mathbf{u}^{(i)},\mathbf{y};\mathbf{X},\boldsymbol{\psi})$. When missing values are present in the time series, the likelihood $g(\mathbf{y})$ can still be computed by the Kalman filter as argued by Durbin & Koopman (2012, §4.8). The importance weights are only evaluated for observed elements in $\mathbf{y}$. The realised samples from the importance density $u^{(i)}$ are computed by simulation smoothing methods that also account for missing values, see Durbin & Koopman (2002).

2.3 In-sample estimation results

Our sample of observations $\mathbf{y}$, the Boat Race outcome from 1829 until 2010, is used to optimize the simulated loglikelihood function based on (11) but subject to a bias-correction for the log transformation. For each model, we have taken $M = 1000$ draws from the constructed importance density and make use of three antithetic variables to relocate and

rescale the weights. The simulated loglikelihood function is evaluated for different values of $\boldsymbol{\psi}$ during the numerical optimisation by using the same set of random input values to obtain a smooth loglikelihood function in $\boldsymbol{\psi}$. Most of such standard implementation details are discussed in Durbin & Koopman (1997).

In our discussion we distinguish models that include the two predictors in $\mathbf{x}_t$ and models that depend only on the dynamics in u_t. It allows us to assess whether including dynamic structures, such as the autoregressive structure, improves the model fit when compared to only having the intercept in the model. We include the following model specifications from Section 2.1 for the latent stochastic component u_t: constant, random walk, AR(1), ARFI(0,d), ARFI(1,d) and the stochastic cycle.

The parameter estimates of $\boldsymbol{\psi}$ for each specification are presented in Table 1. The top panel shows the results for models that include the predictors in $\mathbf{x}_t$. For all models, the estimated long-run mean μ is not significantly different from zero. Hence we can conclude that the models do not significantly predict Cambridge or Oxford as the winner for the Boat Race in the long run. The coefficient estimates for μ are all positive. This finding corresponds to the lead up to 2010 in the series for Cambridge (80 wins against 74 for Oxford).

The difference in weight between the two boats is a strong predictor. The estimated corresponding coefficient is positive and significant for all model specifications. Hence we may conclude that having more muscle power in the boat is more important than having less water resistance. Ideally, we would like to investigate this weight difference in more detail to verify whether the muscle power causes this effect or whether the weight difference presents is a proxy for a different set of variables. This requires additional information that is yet unavailable to us.

The outcome of the toss does not come forward as an important predictor. It is insignificant in all models with small coefficient values. This is rather surprising, given that the gambling odds usually change after the toss is made. It may have some impact when we account for certain weather conditions under which it may signify some advantage when a team starts at a specific side of the river. The winning distance by which the previous years race was won does not provide further predictability. It is insignificant in all models, also with small coefficient values.

The parameters that account for the dynamic effects in the model, including autoregressive coefficients, improve the model fit. For example, the parameter ϕ_1 (or ϕ) is estimated between 0.75 and 0.85 for the AR(1), ARFI(1,d) and stochastic cycle models. The long-memory parameter d is only positive and significant for the ARFI(0,d) model but it cancels out when the ϕ_1 parameter is also included as it is the case for the ARFI(1,d) model.

The stochastic cycle model shows overall the best in-sample fit according to the estimated loglikelihood value and the AIC criteria. The cycle frequency is estimated at 0.316, which indicates a cycle period of nearly 18 years. This is a long cycle period and is mainly caused by the long winning streaks in the sample.

The parameter estimates for the model that do not include predictors are shown in the bottom panel of Table 1. We have also removed the insignificant mean μ for most of the models. The results are very similar when we compare these for models including predictors.

Overall they perform less well in terms of likelihood and AIC values. The estimated dynamic parameters are however similar.

Figure 2 presents the smoothed estimate of the probabilty π_t which is the logit transformation of the signal θ_t. The estimate is conditional on all data (the smoothed estimate) and is computed using the importance sampling technique as discussed in Durbin & Koopman (2012, Section 11.4). The top panel shows them for models that include covariates. Here the estimated signals are mainly determined by the log difference in weight covariate. We have experimented with different transformations for this variable, but the jumps in the weight differences have remained large. The bottom panel for models without predictors presents much smoother estimates of the signal θ_t.

Model diagnostics

To guarantee a $\sqrt{M}$ convergence rate for the Monte Carlo likelihood estimator (11), we must ensure that the weight function $w(\mathbf{u}^{(i)}, \mathbf{y}; \mathbf{X}, \boldsymbol{\psi})$ in (12) has a finite variance as argued by Geweke (1989). To verify whether this is the case for our model specifications, we have implemented the Wald test as discussed by Koopman, Shephard & Creal (2009) and based on extreme value theory. The test statistic is evaluated at the estimated parameters in Table 1 and is constructed by the following step-wise procedure:

(i) Given a model and its corresponding estimated parameters $\hat{\boldsymbol{\psi}}$, we generate 100,000 weight values $w(\mathbf{u}^{(i)}, \mathbf{y}; \mathbf{X}, \hat{\boldsymbol{\psi}})$;

(ii) Given a threshold value $w^{\min}$, we only consider weights that surpass the threshold;

(iii) The resulting s exceedences $z_1, \ldots, z_s$ are modelled by the generalized Pareto density $f(w_i; a, b) = -\log b - (1 + a^{-1}) \log(1 + ab^{-1} z_i)$, for $i = 1, \ldots, s$; the coefficients a and b are estimated by maximum likelihood straightforwardly; the estimates are denoted by $\hat{a}$ and $\hat{b}$, respectively. The null hypothesis of variance existence is given by $H_0 : a \leq 0.5$.

(iv) The test statistic $t_w = \hat{b}^{-1}\sqrt{s/3}(\hat{a} - 0.5)$ is under the null hypothesis asymptotically standard normal. We reject the null for a sufficiently positive value of t_w.

In Figure 3 we present the test statistics for the models in the top panel of Table 1. The test statistics are computed for a multiple of thresholds that result in sets of exceedence weights (with numbers of 1% to 50% of all weights). The test statistics are never rejected which indicate that the variance exists for many different treshold values. A similar panel of graphics can be produced for models without predictors.

Forecasting outlook

The likelihood values and the AIC criteria in Table 1 imply that the in-sample performance of our range of models is quite similar. It does not imply that the out-of-sample forecasting properties are also similar. We can illustrate this by presenting in Figure 4 the theoretical autocorrelation functions (ACF) of the estimated latent processes u_t. The ACFs are implied

	μ	ϕ	σ_η	λ	d	β_1	β_2	β_3	$\log \hat{L}(\boldsymbol{\psi}, \mathbf{y})$	AIC
Including Predictors										
Constant	0.148 $_{0.242}$					0.225 $_{0.063}$	0.099 $_{0.356}$	0.030 $_{0.022}$	-238.967	485.935
RW	0.366 $_{0.876}$		0.161 $_{0.162}$			0.221 $_{0.066}$	0.103 $_{0.387}$	0.019 $_{0.030}$	-238.737	487.474
AR(1)	0.377 $_{0.524}$	0.754 $_{0.096}$	1.264 $_{0.430}$			0.260 $_{0.094}$	-0.115 $_{0.520}$	-0.047 $_{0.045}$	-234.44	480.881
ARFI(0,d,0)	0.424 $_{1.177}$		2.120 $_{0.662}$		0.331 $_{0.127}$	0.335 $_{0.124}$	-0.037 $_{0.618}$	-0.018 $_{0.044}$	-236.837	485.674
ARFI(1,d,0)	0.489 $_{0.619}$	0.778 $_{0.145}$	1.943 $_{0.683}$		-0.079 $_{0.253}$	0.311 $_{0.125}$	-0.173 $_{0.624}$	-0.080 $_{0.059}$	-233.068	480.135
Cycle	0.390 $_{0.573}$	0.858 $_{0.053}$	1.655 $_{0.467}$	0.316 $_{0.082}$		0.304 $_{0.114}$	-0.109 $_{0.637}$	-0.098 $_{0.090}$	-231.273	476.546
Only Dynamics										
Constant	0.078 $_{0.161}$								-248.144	498.289
RW			0.386 $_{0.295}$						-246.894	495.787
AR(1)		0.753 $_{0.101}$	1.075 $_{0.368}$						-240.626	485.253
ARFI(0,d,0)			1.413 $_{0.454}$		0.391 $_{0.104}$				-243.803	491.605
ARFI(1,d,0)		0.771 $_{0.159}$	1.100 $_{0.407}$		-0.038 $_{0.266}$				-240.616	487.232
Cycle		0.872 $_{0.066}$	0.764 $_{0.288}$	0.294 $_{0.086}$					-239.391	484.781

Table 1: Parameter estimation results for the Boat Race from 1829 until 2010 ($n = 182$). The standard errors of the estimates are given in small print. The method of Monte Carlo maximum likelihood is based on importance sampling. The number of importance simulations for likelihood evaluation is $M = 1000$. The regression parameters β_1, β_2 and β_3 correspond to the effect of the log average difference in weight between the rowers of the boats of Cambridge and Oxford, the effect of Cambridge winning the toss and the log winning distance from the previous years race, respectively. The Akaike information criteria is computed as $2\mathsf{k} - 2\log \hat{p}(\mathbf{y}; \hat{\boldsymbol{\psi}})$, where k denotes the number of parameters present in model k.

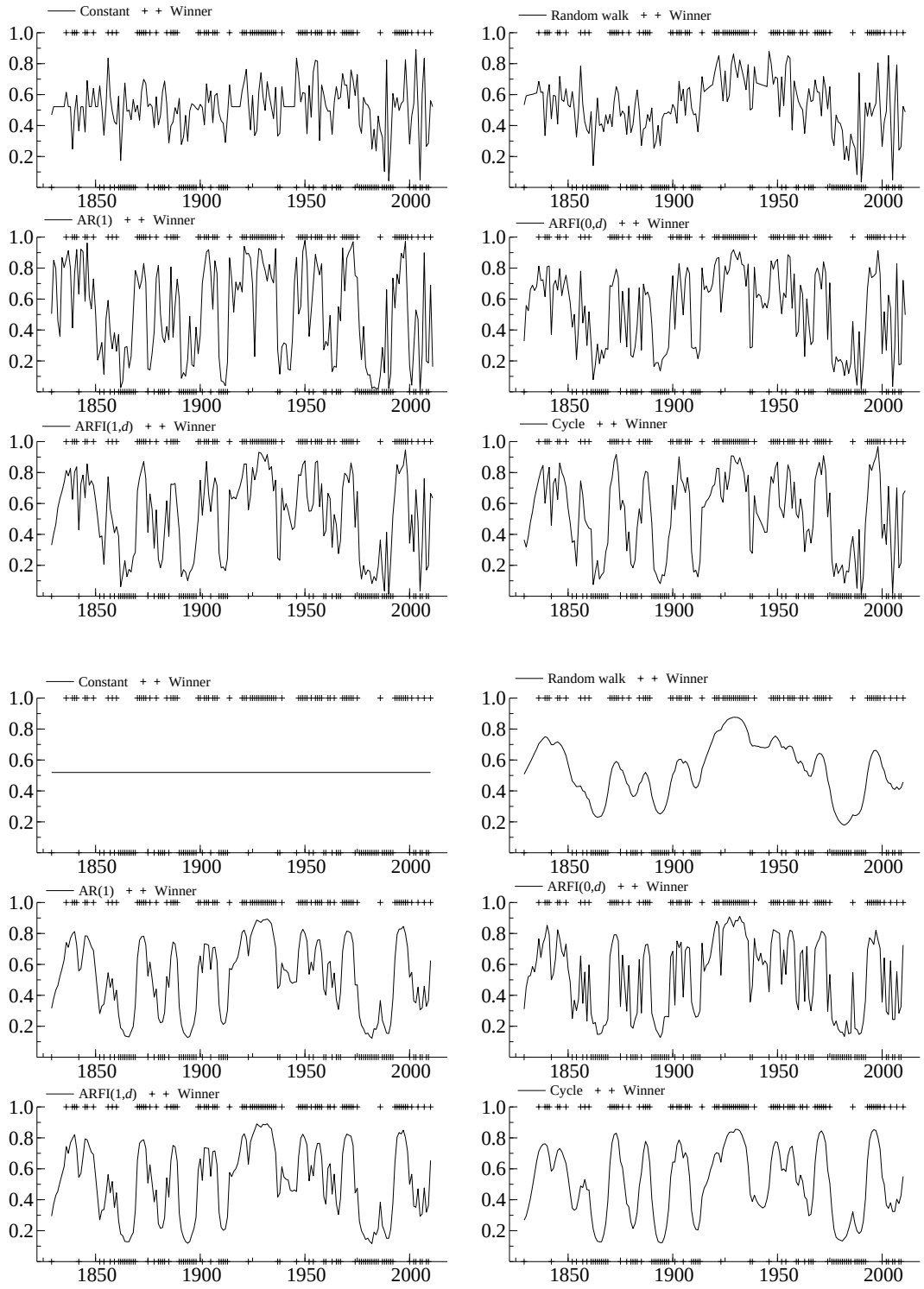

Figure 2: Estimated probabilities, $\mathrm{E}_p(\pi_t|\mathbf{y};\mathbf{X},\hat{\boldsymbol{\psi}})$ where $\pi_t = \exp(\theta_t) \,/\, (1+\exp(\theta_t))$, for all models from Table 1. The top panel presents the probabilities for models including predictors $\mathbf{X}$ while the bottom panel shows the probabilities for models that only include dynamics.

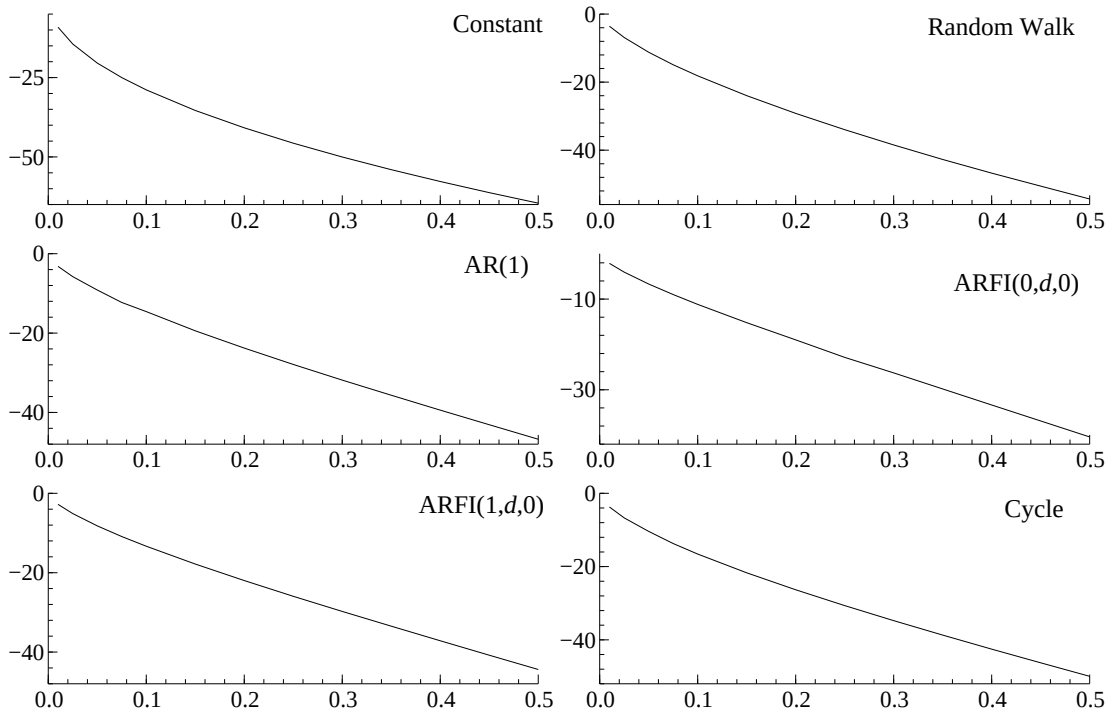

Figure 3: Importance sampling test statistics t_w for all models in the top panel of Table 1. The test statistics are based on $100,000$ simulated weights $w(\mathbf{u}^{(i)}, \mathbf{y}; \mathbf{X}, \boldsymbol{\psi})$. For each model we compute the test statistics for different thresholds $w^{\min}$. The threshold value is implied by the number of exceedence values s. The x-axis refers to the proportion of exceedences $s/100,000$ and the y-axis displays t_w. All test statistics are sufficiently negative for variance existence.

by the model specification with the parameters replaced by their estimates as given in the top panel of Table 1.

The constant model takes no serial correlation into account; its ACF is zero for all lags and is not included in Figure 4. All past outcomes of the Boat Race have an equally weighted impact on the forecast of the outcome this year. The random walk model is a non-stationary dynamic process and its ACF is not properly defined. But for the random walk process, all outcomes have no impact on the current forecast except the last, most recent outcome. The AR(1) and ARFI(1,d) models have nearly identical autocorrelation functions. This is not surprising since the d and ϕ parameters in the ARFI(1,d) cancel out against each other to yield the short-memory AR(1) model. In both cases, only the more recent outcomes have an impact on the current forecast. The ACF of the ARFI(0,d) model shows hyperbolic decay, which is the characteristic of the long-memory process. The autocorrelation function for the stochastic cycle model is first positive for almost 5 lags after which it becomes significantly negative for another 5 lags. This corresponds to the pattern of winning streaks that we observe in the data.

To summarize the in-sample findings. The overall in-sample performance of the models are similar: no large differences in the fit of the models are obtained. The observed average difference in the weights of the two boats is important as a covariate or predictor. At the time that the forecast is usually made, the weights of the boats are known. The models with an autoregressive structure perform relatively better. The out-of-sample forecasting performance is likely to give different results because the implied ACFs are different for different models as indicated in Figure 4.

3 A forty year forecasting assessment

In this section we discuss the out-of-sample forecasting results. All models are used to forecast the Cambridge winning probability repeatedly for a period of forty years. The first forecast is made for the race in 1971 and the final forecast is made for the race in 2010. The forecasts are compared across models based on various loss functions and predictive ability tests as discussed by Diebold & Mariano (1995). Further details of our design of the forecasting study are discussed below.

Empirical evidence from out-of-sample forecast performance is generally more relevant for bookmakers and gamblers than evidence based on in-sample performances such as the results presented in Section 2. For example, gamblers may wish to know how many times the Boat Race would have been correctly predicted by a particular model for a number of years. A comparison of the models based on these historical forecasting performance yields insight in there comparative advantages. We refer to further discussions by White (2000) and Diebold (2012) where the pros and cons of in-sample versus out-of-sample model comparison are given in more detail.

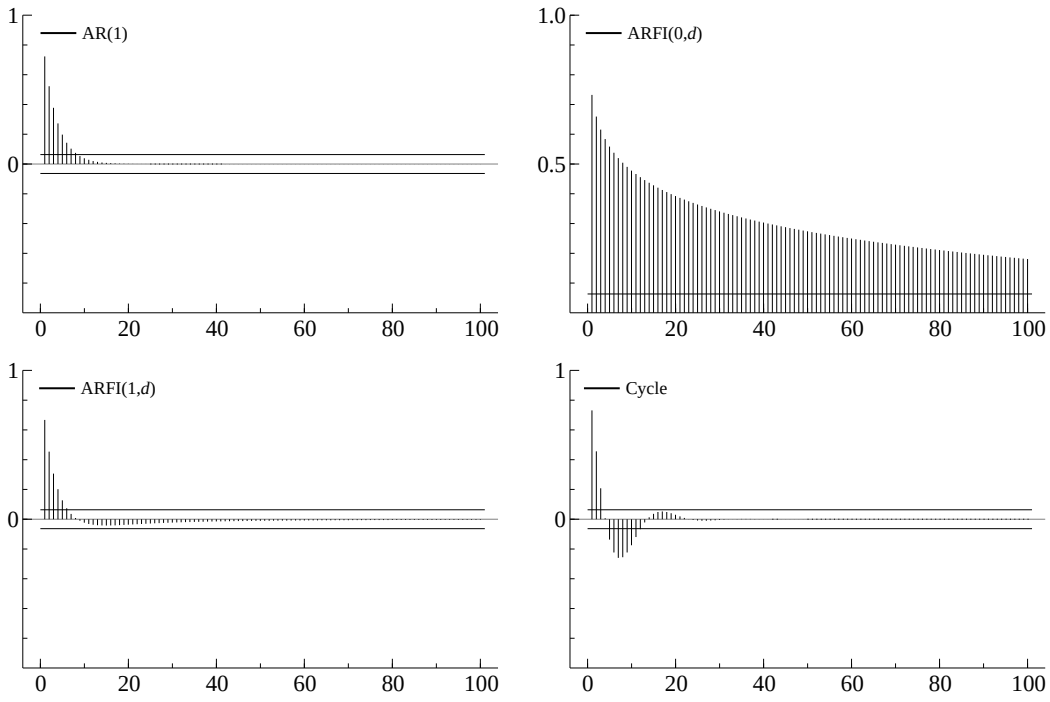

Figure 4: Theoretical autocorrelation functions based on the estimated parameters in the top panel of Table 1 for a selection of the stochastic processes u_t.

3.1 Forecasting procedure

We perform an out of sample forecast study where we repeatedly forecast the probability that Cambridge wins the Boat Race based on information that is available just before the race takes place. Forecasts for years 1971 until 2010 are computed using a rolling forecast window. The choice for the rolling window is motivated by assumptions underlying the test statistics employed.

To illustrate our procedure in detail, consider the first subsample. It consists of the outcomes corresponding to the years 1829 to 1970. Using this subsample we produce a forecast for 1971. The next subsample contains outcomes from years 1830 to 1971 and produces a forecast for 1972. The final forecast, for 2010, is based on outcomes from 1869 until 2009. In total, we construct $m = 40$ one-year-ahead real-time forecasts for each model.

Let the integer N denote the year 1970, which corresponds to the sample split point and the 114$^{\text{nd}}$ Boat Race (142 - 28 missing), respectively. The forecasts are computed for $t = N + 1, \ldots, N + m$ based on subsamples of the observations $y_j, \ldots, y_{N+j-1}$, for $j = 1, \ldots, m$. We estimate the parameter vector ψ for each subsample $y_j, \ldots, y_{N+j-1}$ and denote

the estimate by $\hat{\boldsymbol{\psi}}_j$. The predicted Cambridge winning probability is computed by

$$\hat{\pi}_{N+j|j,\ldots,N+j-1} = \frac{\sum_{i=1}^{M} \exp(\theta^{(i)}_{N+j|j,\ldots,N+j-1}) \;/\; (1+\exp(\theta^{(i)}_{N+j|j,\ldots,N+j-1}))w(\mathbf{u}^{(i)},\mathbf{y};\mathbf{X},\hat{\boldsymbol{\psi}}_j)}{\sum_{i=1}^{M} w(\mathbf{u}^{(i)},\mathbf{y};\mathbf{X},\hat{\boldsymbol{\psi}}_j)}, \tag{13}$$

where

$$\theta^{(i)}_{N+j|j,\ldots,N+j-1} = \hat{\mu} + \mathbf{x}_{N+j}\hat{\beta} + u^{(i)}_{N+j|j,\ldots,N+j-1},$$

for $j = 1,\ldots,m$. The one-year-ahead prediction for the stochastic component $u^{(i)}_{N+j|j,\ldots,N+j-1}$ is based on the linear Gaussian importance model $g(\mathbf{y}|\mathbf{u})g(\mathbf{u})$ and is computed by the Kalman filter; see Durbin & Koopman (2012, Chapter 11) for further details.

Let the set of models $\mathcal{M}$ include all models. The models in $\mathcal{M}$ are indexed by k, for $k = 1,\ldots,K$, where K is the number of models in set $\mathcal{M}$. We adjust the notation for the one-year-ahead forecast to reflect that $\hat{\pi}^k_{N+j|j,\ldots,N+j-1}$ is the forecast for sample $j = 1,\ldots,m$ and model $k = 1,\ldots,K$.

3.2 Loss functions

To evaluate the predictive accuracy of the different models, we rely on loss functions. We can refer to these as scores when evaluating probability forecasts. In the literature a large variety of loss functions exist and, as discussed by Diebold (1993), appropriate loss functions depend on the situation at hand. From a gamblers perspective, the loss function should have the betting odds in each year as an argument. This would allow us to measure the loss in terms of lost investments in a similar way as when trading rules are evaluated; see Hsu & Kuan (2005). However, this information is not available to us.

In our study we rely on more general loss functions which we have computed as follows. For $j = 1,\ldots,m$ and $k = 1,\ldots,K$, we compute

(i) Brier loss:

$$L^{1,k}_{N+j} = 2(y_{N+j} - \hat{\pi}^k_{N+j|j,\ldots,N+j-1})^2.$$

(ii) Predictive log likelihood loss:

$$L^{2,k}_{N+j} = -y_{N+j}\log(\hat{\pi}^k_{N+j|j,\ldots,N+j-1}) - (1-y_{N+j})\log(1-\hat{\pi}^k_{N+j|j,\ldots,N+j-1}).$$

(iii) Incorrect loss:

$$L^{3,k}_{N+j} = \begin{cases} 1 & \text{if} \quad \hat{\pi}^k_{N+j|j,\ldots,N+j-1} > 0.5 \quad \text{and} \quad y_{N+j} = 0 \\ 1 & \text{if} \quad \hat{\pi}^k_{N+j|j,\ldots,N+j-1} \leq 0.5 \quad \text{and} \quad y_{N+j} = 1 \\ 0 & \text{else} \end{cases}$$

The Brier loss function, $L^{1,k}_{N+j}$, is proposed by Brier (1950) and is well known for its use in evaluating weather forecasts. It may be viewed as the mean squared error loss function for

probabilistic forecasts. The difference is that $\hat{\pi}^{k}_{N+j|j,\ldots,N+j-1}$ is not a forecast for y_{N+j}, but rather a probability statement for the event $y_{N+j} = 1$. The predictive log likelihood loss function, $L^{2,k}_{N+j}$, uses the negative of the log likelihood as a measure of accuracy, which has optimal value zero. The incorrect loss function, $L^{3,k}_{N+j}$, measures the loss arising from event forecasting. It takes values of one or zero, indicating loss or success, respectively.

Using these loss functions we define the relative performance of a model l against another model k, with $k, l \in \mathcal{M}$, by

$$d^{s,lk}_{j} \equiv L^{s,l}_{N+j} - L^{s,k}_{N+j}, \qquad s = 1, 2, 3, \qquad l \neq k \qquad k, l = 1, \ldots, K, \qquad j = 1, \ldots, m. \quad (14)$$

All predictive ability tests that we consider use functions of the relative performance vectors, $\mathbf{d}^{s,lk} = (d^{s,lk}_{1}, \ldots, d^{s,lk}_{m})'$, as test statistics. It is easy to show that the vector series $\mathbf{d}^{s,lk}$ is stationary. Let the first moment of the the relative performance indicators be denoted by $\xi^{s,lk} = E[d^{s,lk}_{j}]$, which we assume independent from j, for all $l, k \in \mathcal{M}$ and $s = 1, 2, 3$. For notational convenience, we drop the dependence on the type of loss function s, when discussing the test statistics below.

3.3 Equal predictive ability

Equal predictive ability (EPA) tests are based upon the null hypothesis that their is no difference in accuracy between the competing models. For comparing models l and k this results in the null hypothesis $H_0 : \xi^{lk} = 0$; see Diebold & Mariano (1995) and West (1996).

The widely used EPA t-type test statistic, first proposed in Diebold & Mariano (1995), is given by

$$T^{EPA}_{lk} \equiv \frac{m^{1/2}\bar{d}^{lk}}{\sqrt{\hat{\mathrm{Var}}(d^{lk})}}, \quad (15)$$

where $\bar{d}^{lk} = m^{-1}\sum_{j=1}^{m} d^{lk}_{j}$ and $\hat{\mathrm{Var}}(d^{lk})$ is a consistent estimate for the long run variance matrix. Several approaches for estimating this variance matrix exist. A convenient choice is given by the Newey & West (1987) estimator. In our empirical work we follow the implementation suggested in Diebold & Mariano (1995).

When the models are nested under the null hypothesis of equal predictive ability, the asymptotic theory from West (1996) may not apply. For our models the nesting relationships are non-trivial. It is unclear given the nonlinearities in the observation density, as well as in some of the signal processes, whether the models are nested in the conventional sense. Clark & McCracken (2001) derive asymptotic properties and critical values for the EPA test statistics when models are nested. We view our EPA test statistic as indicative and we assume that it converges to a standard normal distribution as m goes to infinity as shown by Clark & McCracken (2001). We report the right-sided p-value for the probability that model l outperforms model k.

3.4 Results

We discuss the empirical results for our forty years of out-of-sample forecasting and testing. All models from Section 2 with the parameter estimates given in Table 1 are used to obtain forecasts for the Boat Race for the years from 1971 until 2010. We do not use the winner of the toss and the winning distance as possible predictors as their values were found insignificant for all models. Further, we do not show the results for the ARFI(1,d) model as its results were found indistinguishable from the AR(1) model results.

We do include a two state Markov Chain model as a simple competitor from a different model class. The Markov Chain model is described by the transition probabilities $p_{11} = p(y_t = 1|y_{t-1} = 1)$, $p_{01} = p(y_t = 0|y_{t-1} = 1) = 1 - p_{11}$, $p_{00} = p(y_t = 0|y_{t-1} = 0)$ and $p_{10} = p(y_t = 1|y_{t-1} = 0) = 1 - p_{00}$. These probabilities are estimated from the data. In particular, we estimate μ_i and β_i, where $p_{ii} = \exp(\mu_i + \mathbf{x}_t\beta_i)/[1 + \exp(\mu_i + \mathbf{x}_t\beta_i)]$, for $i = 1, 2$. The likelihood for the Markov Chain model is known in closed form.

Also, we include a number of non-parametric forecasting rules for comparison purposes. In particular, we include forecasts that select: last years race winner (the team is in a flow), last years race loser (the team is highly motivated), always Cambridge (duck egg blue supporter) and always Oxford (dark blue supporter). The goal is to show that, at least a selection, of the parametric models of Section 2 are able to outperform these simple forecasting procedures. The forecasts for the parametric models of Section 2 are computed using a rolling forecasting window. For each model we compute forty forecasts using equation (13) with $M = 1000$ number of draws from the importance density.

Forecasting results

The average loss functions $m^{-1}\sum_{j=1}^{m} L_{N+j}^{s,k}$ for models $k = 1, \ldots, K$ are shown in Table 2. The non-parametric forecasting rules are not evaluated for the log loss function, $s = 2$, as they do not provide probability forecasts.

First we discuss the models that include the average log difference in weight between the boats as a predictor. For all loss functions, the binary model with the stochastic cycle component of Harvey (1989) as the signal u_t provides the lowest loss scores. Also the model that includes the AR(1) component in the signal θ_t yields comparably low scores. The incorrect loss function reveals that our best model (Cycle) is able to predict 31 out of 40 races correctly. This is based on predicting Cambridge as the winner if the predicted probability is larger than one half and visa versa. We notice that the forecasting difference between the different parametric signal specification is small. For example, for the Brier loss function, an additional mean squared error of only 0.04 is found between the worst and the best model. The Markov Chain model performs comparable to the models with autoregressive dynamics in the state.

For the models that do not include the difference in the weight of the two boats as a predictor, the differences between the models are larger. This is illustrated in Figure 5, where we present the predicted probability paths for all parametric models. The difference in weight predictor clearly dominates the forecasts when it is included. When it is not included,

Model	Brier	LOG	Incorrect
	Including predictors		
Constant	0.417	0.610	0.375
Random walk	0.446	0.642	0.375
AR(1)	0.383	0.564	0.275
ARFI(0,d,0)	0.396	0.581	0.325
CYCLE	0.371	0.555	0.225
MCHAIN	0.388	0.563	0.350
	Only dynamics		
Constant	0.522	0.715	0.625
Random walk	0.489	0.706	0.350
AR(1)	0.431	0.623	0.350
ARFI(0,d,0)	0.455	0.648	0.350
CYCLE	0.430	0.620	0.350
MCHAIN	0.479	0.674	0.350
	Non-parametric rules		
Winner	0.700	-	0.350
Loser	1.300	-	0.650
Cambridge	1.200	-	0.600
Oxford	0.800	-	0.400

Table 2: Average loss functions for 40 year forecasting for the Boat Race from 1971 until 2010 ($m = 40$ forecasts). Each forecast is based on 142 observations. Brier corresponds to loss function $m^{-1}\sum_{j=1}^{m} L_j^{1,k}$, LOG to $m^{-1}\sum_{j=1}^{m} L_j^{2,k}$ and Incorrect to $m^{-1}\sum_{j=1}^{m} L_j^{3,k}$. The highlighted numbers indicate the lowest loss per category per loss function.

the constant and random walk models perform much worse, whereas the other parametric models are less affected. Overall we may conclude that the inclusion of the weight predictor is important. The incorrect loss function is not able to distinguish between the models when the log weight predictor is not included.

The average loss functions for the non-parametric forecasting rules are presented in the lower panel of Table 2. The Brier loss function attributes very high losses to these models. It is shown that the "always Cambridge" and "loser" rules predict less than 50 % of the races correctly. It seems that always forecasting last year's race winner is a reasonable strategy.

Predictive ability results

In Table 3 we present the p-values for the EPA test statistics for the models that include the log difference in weight as a predictor. Each model is compared against all other models for all loss functions. When the p-values are smaller than α, it indicates that the model listed in the row is outperformed by the model listed in the column, with significance level α which we take as $\alpha = 0.1$ in our results. For the Brier loss function, the EPA tests give two main results. First, all non-parametric forecasting rules are significantly outperformed by all parametric models. Especially the rules that predict the "loser" and "always Cambridge"

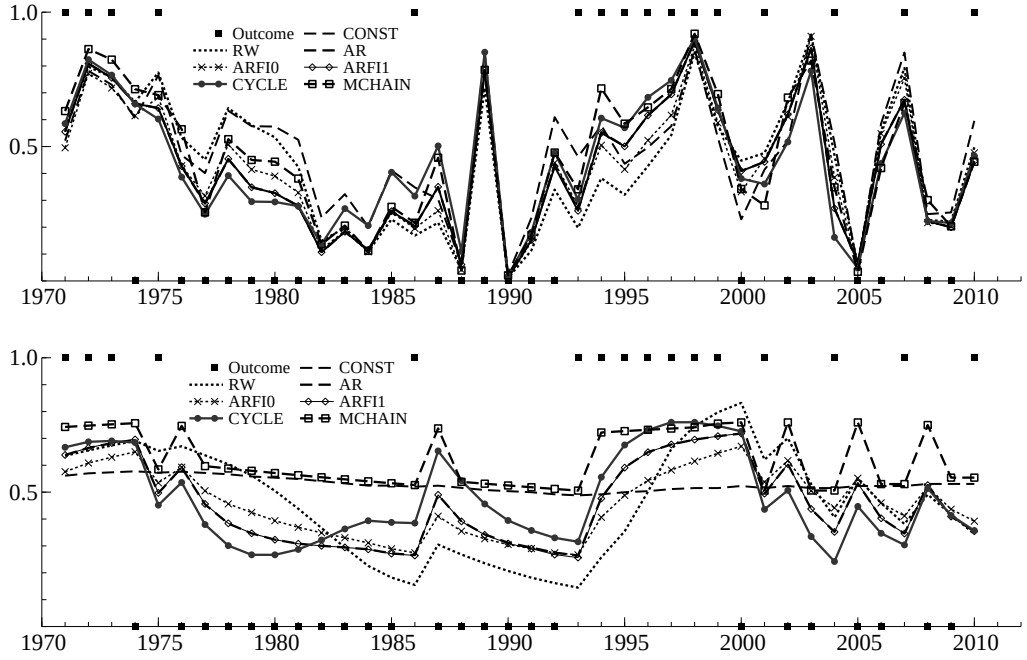

Figure 5: 40 year forecasts for the Boat Race from 1971 until 2010 ($m = 40$ forecasts). Each forecast is based on 142 observations. The top panel shows the forecasts for the parametric models of Section 2 including the weight difference of the two boats as a predictor. The bottom panel shows the forecast for models including only dynamics

appear to be unsatisfactory when aiming to produce accurate forecasts. Second, the model that includes a random walk process in the signal is outperformed by models that include a more advanced stochastic process in the signal, such as the AR(1), ARFI(0,d) and stochastic cycle processes. Also the random walk model is outperformed by the Markov Chain model. Perhaps surprisingly, the logistic regression model (constant) is not outperformed by the more advanced dynamic models that take the serial correlation into account. However, the AR(1) and stochastic cycle based models are most unlikely to be outperformed.

The EPA tests based on the log loss function confirm most of the results that were found for the Brier loss function. Again the random walk model is statistically outperformed by the models that include more advanced stochastic processes. The EPA test based on the log loss function is also not able to distinguish between the forecasting performance of the other parametric models. In other words, no evidence in found in favor of either the logistic regression (constant), AR(1), ARFI(0,d),stochastic cycle, or Markov Chain models.

The EPA tests based on the incorrect loss function show somewhat different results. The logistic regression (constant) model is now significantly outperformed by the AR(1), stochastic cycle and Markov Chain models. The random walk model is also outperformed by the AR(1), stochastic cycle and Markov Chain models. Interestingly, the forecasting rule that predicts the last years race winner seems to perform much better based on the incorrect loss function. Only the stochastic cycle model is able to reject this model. It also rejects the Markov Chain model.

In Table 4 we show the results for the models that only depend on dynamics and do not include predictors. Here the differences between the models are more clearly visible for the Brier and log loss functions. The constant, random walk and all non-parametric forecasting rules are significantly outperformed by the other models for the Brier and log loss functions. For the incorrect loss function the opposite is found. Here only the constant model and the "always Cambridge" and "loser" prediction rules are outperformed.

When summarizing the testing results for EPA tests, we may conclude that the parametric models predict the Boat Race significantly better when compared to the non-parametric methods based on ad-hoc rules. Furthermore we have provided substantial evidence in favor of models that include an autoregressive component in the signal. These are the models with AR(1) or stochastic cycle process in the signal. There is little evidence against the ARFI(0,d) and Markov Chain models but they do not significantly outperform the other models.

4 Conclusion

The forecasting of the outcomes of the yearly Boat Race between Cambridge and Oxford over the forty year period 1971-2010 is extensively evaluated in a real-time study. The accuracy of the forecasts is measured by different loss functions and by equal predictive ability tests. The overall finding is that parametric models predict the outcome of the Boat Race significantly better than ad-hoc methods. Furthermore, models with a latent autoregressive component

Equal predictive ability; Including Regressors

Bench. / Alt.	Constant	Random walk	AR(1)	ARFI(0,d)	CYCLE	MCHAIN	Winner	Loser	Cambridge	Oxford
					Brier loss					
Constant	-	0.723	0.174	0.226	0.132	0.212	0.954	1.000	1.000	0.988
Random walk	0.277	-	0.015	0.042	0.050	0.080	0.956	1.000	1.000	0.986
AR(1)	0.826	0.985	-	0.818	0.279	0.596	0.986	1.000	1.000	0.996
ARFI(0,d)	0.774	0.958	0.182	-	0.207	0.393	0.976	1.000	1.000	0.995
CYCLE	0.868	0.950	0.721	0.793	-	0.749	0.989	1.000	1.000	0.996
MCHAIN	0.788	0.920	0.404	0.607	0.251	-	0.988	1.000	1.000	0.994
Winner	0.046	0.044	0.014	0.024	0.011	0.012	-	0.977	0.984	0.692
Loser	0.000	0.000	0.000	0.000	0.000	0.000	0.023	0.000	0.308	0.016
Cambridge	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.692	-	0.098
Oxford	0.012	0.014	0.004	0.005	0.004	0.006	0.308	0.984	0.902	-
					Log loss					
Constant	-	0.697	0.138	0.185	0.142	0.147				
Random walk	0.303	-	0.017	0.045	0.079	0.063	-	-	-	-
AR(1)	0.862	0.983	-	0.824	0.381	0.484	-	-	-	-
ARFI(0,d)	0.815	0.955	0.176	-	0.272	0.304	-	-	-	-
CYCLE	0.858	0.921	0.619	0.728	-	0.588	-	-	-	-
MCHAIN	0.853	0.937	0.516	0.696	0.412	-	-	-	-	-
Winner	-	-	-	-	-	-	-	-	-	-
Loser	-	-	-	-	-	-	-	-	-	-
Cambridge	-	-	-	-	-	-	-	-	-	-
Oxford	-	-	-	-	-	-	-	-	-	-
					Incorrect loss					
Constant	-	0.500	0.073	0.238	0.023	0.098	0.409	0.995	0.986	0.586
Random walk	0.500	-	0.045	0.205	0.012	0.045	0.391	0.988	0.975	0.596
AR(1)	0.927	0.955	-	0.927	0.073	0.500	0.820	0.999	0.998	0.906
ARFI(0,d)	0.762	0.795	0.073	-	0.018	0.156	0.609	0.997	0.991	0.782
CYCLE	0.977	0.988	0.927	0.982	-	0.927	0.958	1.000	1.000	0.970
MCHAIN	0.902	0.955	0.500	0.844	0.073	-	0.844	0.999	0.999	0.892
Winner	0.591	0.609	0.180	0.391	0.042	0.156	-	0.977	0.984	0.692
Loser	0.005	0.012	0.001	0.003	0.000	0.001	0.023	-	0.308	0.016
Cambridge	0.014	0.025	0.002	0.009	0.000	0.001	0.016	0.692	-	0.098
Oxford	0.414	0.404	0.094	0.218	0.030	0.108	0.308	0.984	0.902	-

Table 3: Equal predictive ability tests for comparing 40 year forecasting results for the Boat Race from 1971 until 2010 ($m = 40$ forecasts) for models that include predictors. The t^{EPA} is constructed for different loss functions as discussed in Section 3.3. The highlighted numbers are significant for $\alpha = 0.1$.

Equal predictive ability; Only dynamics

Bench. / Alt.	Constant	Random walk	AR(1)	ARFI(0,d)	CYCLE	MCHAIN	Winner	Loser	Cambridge	Oxford
					Brier loss					
Constant	-	0.328	0.025	0.044	0.027	0.189	0.880	1.000	1.000	0.955
Random walk	0.672	-	0.065	0.205	0.151	0.435	0.965	1.000	1.000	0.976
AR(1)	0.975	0.935	-	0.941	0.495	0.965	0.991	1.000	1.000	0.993
ARFI(0,d)	0.956	0.795	0.059	-	0.241	0.757	0.974	1.000	1.000	0.990
CYCLE	0.973	0.849	0.505	0.759	-	0.962	0.991	1.000	1.000	0.990
MCHAIN	0.811	0.565	0.035	0.243	0.038	-	0.984	1.000	1.000	0.975
Winner	0.120	0.035	0.009	0.026	0.009	0.016	-	0.977	0.984	0.692
Loser	0.000	0.000	0.000	0.000	0.000	0.000	0.023	-	0.308	0.016
Cambridge	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.692	-	0.098
Oxford	0.045	0.024	0.007	0.010	0.010	0.025	0.308	0.984	0.902	-
					Log loss					
Constant	-	0.461	0.031	0.051	0.031	0.209				
Random walk	0.539	-	0.066	0.164	0.125	0.337				
AR(1)	0.969	0.934	-	0.936	0.470	0.963				
ARFI(0,d)	0.949	0.836	0.064	-	0.236	0.764				
CYCLE	0.969	0.875	0.530	0.764	-	0.965				
MCHAIN	0.791	0.663	0.037	0.236	0.035	-				
Winner							-			
Loser								-		
Cambridge									-	
Oxford										-
					Incorrect loss					
Constant	-	0.003	0.007	0.003	0.007	0.007	0.007	0.596	0.327	0.059
Random walk	0.997	-	0.500	0.500	0.500	0.500	0.500	0.988	0.984	0.692
AR(1)	0.993	0.500	-	0.500	0.500	0.500	0.500	0.980	0.980	0.704
ARFI(0,d)	0.997	0.500	0.500	-	0.500	0.500	0.500	0.985	0.984	0.692
CYCLE	0.993	0.500	0.500	0.500	-	0.500	0.500	0.980	0.984	0.692
MCHAIN	0.993	0.500	0.500	0.500	0.500	-	0.500	0.977	0.984	0.692
Winner	0.993	0.500	0.500	0.500	0.500	0.500	-	0.977	0.984	0.692
Loser	0.404	0.012	0.020	0.015	0.020	0.023	0.023	-	0.308	0.016
Cambridge	0.673	0.016	0.020	0.016	0.016	0.016	0.016	0.692	-	0.098
Oxford	0.941	0.308	0.296	0.308	0.308	0.308	0.308	0.984	0.902	-

Table 4: Equal predictive ability tests for comparing 40 year forecasting results for the Boat Race from 1971 until 2010 ($m = 40$ forecasts) for models that only include dynamics. The t^{EPA} is constructed for different loss functions as discussed in Section 3.3. The highlighted numbers are significant for $\alpha = 0.1$.

in the signal produce the most accurate forecasts. Although this study has been mostly fun for us, we do believe that statistical dynamic models have a serious role to play in event forecasting. While other events may have a more serious impact on us than outcomes of the Boat Race, the ability to forecasts binary time series accurately is important. The formulation of dynamic models, the development of estimation and forecasting procedures, and the assessment of significant outperformance in forecasting accuracy in the context of binary time series may provide many interesting research questions.

References

Brier, G. W. (1950), 'Verification of forecasts expressed in terms of probabilities', *Monthly Wheather Review* **78**, 1–3.

Chib, S. & Jeliazkov, I. (2006), 'Inference in semiparametric dynamic models for binary longitudinal data', *Journal of the American Statistical Association* **101**, 685–700.

Clark, T. E. & McCracken, M. W. (2001), 'Tests of equal forecast accuracy and encompassing for nested models', *Journal of Econometrics* **105**, 85–110.

Cox, D. R. & Snell, E. J. (1989), *Analysis of binary data*, Chapman and Hall, London.

Czado, C. & Song, P. X. K. (2008), 'State space mixed models for longitudinal observations with binary and binomial responses', *Statistical Papers* **49**, 691–714.

Diebold, F. X. (1993), 'On the limitations of comparing mean squared forecast errors: Comment', *Journal of Forecasting* **12**, 641–642.

Diebold, F. X. (2012), 'Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective one the Use and Abuse of Diebold-Mariano Tests', **13**, 253–265. Working Paper at http://www.ssc.upenn.edu/f̃diebold.

Diebold, F. X. & Lopez, J. A. (1996), Forecast Evaluation and Combination, *in* G. S. Maddala & C. R. Rao, eds, 'Handbook of Statistics', Elsevier, Amsterdam, North-Holland, pp. 241–268.

Diebold, F. X. & Mariano, R. (1995), 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics* **13**, 253–265.

Drinkwater, G. C. (1939), *The Boat Race*, Blackie and Son, London.

Durbin, J. & Koopman, S. J. (1997), 'Monte Carlo maximum likelihood estimation of non-Gaussian state space models', *Biometrika* **84**, 669–684.

Durbin, J. & Koopman, S. J. (2002), 'A simple and efficient simulation smoother for state space time series analysis', *Biometrika* **89**, 603–616.

Durbin, J. & Koopman, S. J. (2012), *Time Series Analysis by State Space Methods*, 2 edn, Oxford University Press, Oxford.

Fernandes, C. & Harvey, A. C. (1990), Modelling binary time series. Department of Statistics, London School of Economics.

Geweke, J. (1989), 'Bayesian inference in econometric models using Monte Carlo integration', *Econometrica* **57**, 1317–1339.

Granger, C. W. J. & Joyeux, R. (1980), 'An introduction to long-memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–29.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A. C. & Durbin, J. (1986), 'The effects of seat belt legislation on British road casualties: A case study in structural time series modelling, (with discussion)', *Journal of the Royal Statistical Society, Series A* **149**, 187–227.

Harvey, A. C. & Fernandes, C. (1989), 'Time series models for count data or qualitative observations', *Journal of Business and Economic Statistics* **7**, 407–17.

Harvey, A. C. & Koopman, S. J. (1997), Multivariate structural time series models, *in* C. Heij, H. Schumacher, B. Hanzon & C. Praagman, eds, 'Systematic dynamics in economic and financial models', John Wiley & Sons, Chichester, pp. 269–98.

Hsu, P. H. & Kuan, C. M. (2005), 'Reexamining the profitability of technical analysis with data snooping checks', *Journal of Financial Econometrics* **3**, 606–628.

Jungbacker, B. & Koopman, S. J. (2007), 'Monte Carlo estimation for nonlinear non-Gaussian state space models', *Biometrika* **94**, 827–839.

Koopman, S. J., Shephard, N. & Creal, D. D. (2009), 'Testing the assumptions behind importance sampling', *Journal of Econometrics* **149**, 2–11.

Mesters, G., Koopman, S. J. & Ooms, M. (2011), 'Monte Carlo Maximum Likelihood Estimation for Generalized Long-Memory Time Series Models', *Tinbergen Institute working paper: TI 090/4* .

Newey, W. & West, K. D. (1987), 'A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica* **55**, 703–708.

Palma, W. (2007), *Long Memory Time Series*, John Wiley & Sons, New York.

Richard, J. F. & Zhang, W. (2007), 'Efficient High-Dimensional Importance Sampling', *Journal of Econometrics* **141**, 1385–1411.

Ripley, B. D. (1987), *Stochastic Simulation*, John Wiley & Sons, New York.

Shephard, N. & Pitt, M. K. (1997), 'Likelihood analysis of non-Gaussian measurement time series', *Biometrika* **84**, 653–667.

Stock, J. H. & Watson, M. W. (2003), 'Forecasting output and inflation: The role of asset prices', *Journal of Economic Literature* **41**, 788–829.

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **65**, 1067–1084.

West, M. (1985), 'Dynamic generalized linear models and Bayesian forecasting', *Journal of the American Statistical Association* **80**, 73–83.

White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**, 1097–1126.