TI 2025-074/III
Tinbergen Institute Discussion Paper

Exponentially weighted estimands and the exponential family: filtering, prediction and smoothing

Simon Donker van Heel[1]
Neil Shephard[2]

1 Harvard University, Tinbergen Institute

2 Harvard University

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and Vrije Universiteit Amsterdam.

Contact: discussionpapers@tinbergen.nl

More TI discussion papers can be downloaded at https://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 598 4580

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900

Exponentially weighted estimands and the exponential family: filtering, prediction and smoothing

Simon Donker van Heel and Neil Shephard*
Department of Economics and Department of Statistics,
Harvard University, Cambridge, MA 02138, USA

December 17, 2025

Abstract

We propose using a discounted version of a convex combination of the log-likelihood with the corresponding expected log-likelihood such that when they are maximized they yield a filter, predictor and smoother for time series. This paper then focuses on working out the implications of this in the case of the canonical exponential family. The results are simple exact filters, predictors and smoothers with linear recursions. A theory for these models is developed and the models are illustrated on simulated and real data.

Keywords: Exponential family; EWMA; Filtering; Likelihood; Time Series.

1 Introduction

This paper provides a new, simple way to carry out filtering, predicting and smoothing. This is based on a discounted version of a convex combination of a log-likelihood and the expected log-likelihood function. The expected log-likelihood term has many familiar interpretations, e.g. shrinkage and mean reversion. The discounting is carried out using an exponential function.

We focus on log-likelihoods based around the canonical exponential family. It has a similar structure to generalized linear regression models, but now for time series. Various properties of the resulting procedures are established and illustrated by simulation, estimation and empirical work.

Section 2 develops the new filter, predictor and smoother based on what we call the weighted estimands. Section 3 looks at the case of the canonical exponential family of models. The resulting filters, predictors and smoothers are derived from the weighting principle. These have to be found by a convex numerical optimization.

Section 4 looks at a flexible special case of the canonical exponential family which results in analytic expressions for the filters, predictors and smoothers. They do not need any form of iteration or

*We are grateful for the comments and encouragement from Dick van Dijk, Siem Jan Koopman, Rutger-Jan Lange and Bram van Os. Donker van Heel thanks the Fulbright Program for financial support under their scholarship program. We used `Refine.ink` to review an earlier draft.

approximation. This special case covers a vast number of familiar models, e.g. based on distributions such as, for example, the Gaussian, multinomial, Poisson, Pareto, exponential, von Mises and Dirichlet.

Section 5 develops a quasi-likelihood based estimation strategy for the hyperparameters of this time series model structure. This method is extended to allow for a two-step estimator, where an easy to use moment estimator reduces the dimension of the required numerical maximization.

Section 6 provides an empirical illustration based on a seven category Dirichlet model. The section uses the empirical results to provide simulation based evidence for the performance of the hyperparameter estimators.

Section 7 details some conclusions. Section 8 contains a proof.

2 Principle of weighting

Think of time series data $y_{1:T} = \{y_1, ..., y_T\}$ and a time series model for the corresponding random variables $Y_{1:T} = \{Y_1, ..., Y_T\}$. Here T is the length of the time series. The properties of the time series model $Y_{1:T}$ are determined by the abstract probability triple $(\Omega, \mathcal{F}, P)$.

Definition 1 *For the random variable Y_t write a frame*

$$\log L_t(\theta; y_t), \quad \theta \in \Theta, y_t \in \mathcal{Y}_t, \quad t = 1, ..., T,$$

for the specification of the filtering, smoothing and prediction estimands. The frame is called stable if it is of the form $\log L(\theta; y_t)$, i.e. not varying over time.

At no point do we regard this frame as true. It will index a class of filters, smoothers and predictors.

The form of the frame can vary over time, e.g. it could be based around the random variable $Y_t \sim \texttt{Multinomial}(n_t, p)$, where p is determined by θ and n_t is regarded as non-stochastic (e.g. we have formed the probability mass function for Y_t by conditioning on n_t) or $Y_t \sim \texttt{Pois}(e^{\theta^\intercal x_t})$ where x_t is a non-stochastic vector (or conditioned on).

At various points in the paper some assumptions are made about the form of the frame.[1] Here we list some of them.

Assumption 1 *The $\mathbb{E}[\log L_t(\theta; Y_t)]$ exist for each t and $\theta \in \Theta$.*

[1]We follow the statistical convention of taking the log-likelihood as ignoring terms which only involve data not also hyperparameters. One way of implementing this, for a model with a probability density function, for example, is to take $\log L_t(\theta; y_t) := \log f_t(y_t; \theta) - \log f_t(y_t; \theta_0)$ where $\theta_0 \in \Theta$ is some arbitrary fixed value as $\theta \in \Theta$ varies. As we need the expected log-likelihood to exist, removing these data terms lessens the requirements we need to prove the expected value exists.

The expectation is under the probability measure for Y_t, not the probability model that would rationalize the frame.

Assumption 2 *The parameter space is $\Theta = \mathbb{R}^d$. For each t and all $y \in \mathcal{Y}_t$, the $\log L_t(\theta; y)$ is concave and upper semicontinuous in θ, taking values in $\mathbb{R} \cup \{-\infty\}$. The $\mathbb{E}[\log L_t(\theta; Y_t)]$ is finite for all $\theta \in \mathbb{R}^d$, strictly concave in θ, and satisfies the coercivity condition that $\|\theta\| \to \infty$ implies that $\mathbb{E}[\log L_t(\theta; Y_t)] \to -\infty$.*

The core of this paper will be based around an exponential version of a weighting principle.[2]

The principle of weighting delivers three sequences of estimands through time: a filter, a predictor and a smoother. Notice that the estimands are random variables.

Definition 2 (Exponentially weighted estimands) *Let Assumptions 1-2 hold, $\lambda \in [0, 1]$ be a discount hyperparameter and $\alpha \in [0, 1]$ an anchoring hyperparameter. The exponentially weighted filter at time t, using data up to time t and the sequence of frames $\{\log L_t(\theta; y_t)\}_{t=1}^{T}$, is*

$$\widetilde{\theta}_t = \underset{\theta \in \Theta}{\arg\max} \; Q_{\lambda,t}(\theta),$$

where

$$Q_{\lambda,t}(\theta) = \sum_{j=1}^{t} \lambda^{t-j} \left\{(1-\alpha)\mathbb{E}[\log L_j(\theta; Y_j)] + \alpha \log L_j(\theta; Y_j)\right\}.$$

The time t exponentially weighted predictor based on data up to time $(t-s)$, for $s \geq 1$, is

$$\widetilde{\theta}_{t|t-s} = \underset{\theta \in \Theta}{\arg\max} \; Q_{\lambda,t|t-s}(\theta),$$

where

$$Q_{\lambda,t|t-s}(\theta) = (1-\alpha) \sum_{j=1}^{t} \lambda^{t-j} \mathrm{E}\left[\log L_j\left(\theta; Y_j\right)\right] + \alpha \sum_{j=1}^{t-s} \lambda^{t-j} \log L_j\left(\theta; Y_j\right).$$

The exponentially weighted smoother at time t, using data up to time $T \geq t$, is

$$\widetilde{\theta}_{t|T} = \underset{\theta \in \Theta}{\arg\max} \; Q_{\lambda,t|T}(\theta),$$

where

$$Q_{\lambda,t|T}(\theta) = \sum_{j=1}^{T} \lambda^{|t-j|} \left\{(1-\alpha)\mathbb{E}[\log L_j(\theta; Y_j)] + \alpha \log L_j(\theta; Y_j)\right\}.$$

Note that Definition 2, Assumptions 1-2 and $\alpha \in [0, 1)$ implies that $Q_{\lambda,t}(\theta)$, $Q_{\lambda,t|t-s}(\theta)$ and $Q_{\lambda,t|T}$ are globally strictly concave and so $\widetilde{\theta}_t$, $\widetilde{\theta}_{t|t-s}$ and $\widetilde{\theta}_{t|T}$ are guaranteed to exist and be unique.

[2]The extension to allow a prior density $\pi(\theta)$ for $\theta \in \Theta$, yielding a exponentially weighted MAP (i.e. mode of a type of posterior), means just adding $\log \pi(\theta)$ to our criteria function and then maximizing that with respect to θ for each t. This is particularly important if θ is high dimensional.

As the likelihood function is invariant to reparameterization, like the maximum likelihood estimator (MLE), the exponentially weighted filter, predictor and smoother are invariant under one-to-one reparameterization, i.e. if $\gamma = \gamma(\theta)$, then $\widetilde{\gamma}_t = \gamma(\widetilde{\theta}_t)$. This will turn out to be a very powerful property in practice.

Intuitively, $\lambda \in [0,1]$ controls the rate at which older information is downweighted, while $\alpha \in [0,1]$ controls the weight placed on the anchor $\mathbb{E}[\log L_j(\theta; Y_j)]$ terms.

Remark 1 *(a) When $\alpha = 1$ the resulting exponentially weighted maximum likelihood estimator (EWMLE) appears often in various literatures often under the label of local modeling or local regression, e.g. Chapter 8 of Fan and Yao (2005) focus on non-linear time series and Fan et al. (1995) who look at non-linear regression. There the localization is typically carried out using a kernel weighting function — our use of the exponential function is a special case of this. The Gaussian likelihood for an unknown mean yields the celebrated* `EWMA` *of Brown (1956).*

(b) Weighting likelihood contributions has appeared in some applied contexts, e.g. Dixon and Coles (1997) use an early version of exponentially weighted likelihoods over time to model football match outcomes, while Hu and Zidek (2002) and Blasques et al. (2016) weight over the cross-section.

(c) Luxenberg and Boyd (2024) look at exponentially weighted moving losses, replacing the likelihood with a loss function. Their work has some overlap with our approach, but focuses on penalization (which we do not), and does not consider the $\mathbb{E}[\log L_j(\theta, Y_j)]$ terms.

(d) When $\alpha < 1$ we do not know of a literature which uses this approach, including the works cited in parts (a)–(c). Of course, broadly, the expected log-likelihood can be thought as a form of shrinkage. Hence it relates to the vast literatures on empirical Bayes, ridge and Lasso regression, etc (e.g. Efron and Morris (1977), Efron (2012), Hoerl and Kennard (1970b,a), Tibshirani (1996) and Hansen (2016)).

(e) There is a large general Bayesian literature which raises the likelihood to a power less than one to obtain some "robustness" properties for the posterior. Some of the corresponding literature is discussed in Holmes and Walker (2017).

(f) When α is close to one, the estimands are mostly determined by past observations; when α is close to zero, the estimands are predominantly determined by the expected log-likelihoods $\mathbb{E}[\log L_j(\theta, Y_j)]$. The discount parameter $\lambda \in [0,1]$ determines how quickly past observations are downweighted, with larger values giving more weight to distant observations.

(g) In applications with naturally constrained parameters (e.g. $\sigma > 0$), we work with a smooth one-to-one reparametrization such as a log or logit transform so that the resulting parameter vector θ lives in $\mathbb{R}^d$. Since the exponentially weighted estimands are reparametrization invariant, this involves no loss

of generality and allows us to take $\Theta = \mathbb{R}^d$. Together with the upper semicontinuity of the realized log-frame and the coercivity and strict concavity of the expected log-frame in Assumption 2, this guarantees that all $Q_{\lambda,t}$, $Q_{\lambda,t|t-s}$ and $Q_{\lambda,t|T}$ are coercive, concave and upper semicontinuous, and hence that their maximizers exist and are unique.

3 Canonical exponential family case

The rest of this paper will focus entirely on the case where the frame is based on a member of the canonical exponential family. This family includes most of the important probability models used in applied statistics. In our context, crucially, it allows either an analytic solution or an easy to compute solution for each t.

3.1 Canonical exponential family

Start by recalling the definition of the canonical exponential family and three of its properties.

Definition 3 (Canonical exponential family) *Assume a random variable $Y \sim \texttt{CEF}(\theta, h, \psi)$ — that is its probability density function (or probability mass function for a discrete random variable) can be written as a member of the minimal canonical exponential family:*

$$f(y;\theta) = \exp\{\theta^{\intercal} h(y) - \psi(\theta)\}/b(y), \quad \theta \in \Theta = \{\theta : \psi(\theta) < \infty\}, y \in \mathcal{Y},$$

where $\psi(\theta)$ is infinitely continuously differentiable and strictly convex with respect to θ while $b(y) > 0$ for all $y \in \mathcal{Y}$, (e.g. Barndorff-Nielsen (1978), McCullagh and Nelder (1989) and Efron (2022)).

In the canonical exponential family the cumulant function $\log \mathbb{E}[e^{a^{\intercal} h(Y)}; \theta] = \psi(\theta + a) - \psi(\theta)$, so

$$\begin{aligned} \psi'(\theta) &= \frac{\partial \psi(\theta)}{\partial \theta} \\ &= \mathbb{E}[h(Y); \theta] \\ &= \mu, \end{aligned}$$

the finite expected value of $h(Y)$ under the $\texttt{CEF}(\theta, h, \psi)$ model and

$$\begin{aligned} \psi''(\theta) &= \frac{\partial^2 \psi(\theta)}{\partial \theta \partial \theta^{\intercal}} \\ &= \mathrm{V}(h(Y); \theta), \end{aligned}$$

the corresponding finite variance. The minimal canonical exponential family requires $\psi(\theta)$ to be strictly convex so $\mathrm{V}(h(Y); \theta)$ is positive definite under the model. It is helpful to go from $\mu = \mathbb{E}[h(Y)]$ to θ through the inverse function

$$\psi'^{-1}(\mu) = \theta.$$

This inverse has a unique solution (as ψ is strictly convex), but in some cases it has to be solved numerically — a classic task in statistics as this is isomorphic to computing the MLE of θ in the canonical exponential family. Famously, it can be carried out reliably and rapidly using a Newton-Raphson algorithm, which converges quadratically close to the solution. We state this formally as Algorithm 1 for reference later.

Algorithm 1 (Solving $\theta = \psi'^{-1}(\mu)$) *Starting from some point $\theta^{(0)}$, then iterate*

$$\theta^{(i)} = \theta^{(i-1)} - \left\{\psi''(\theta^{(i-1)})\right\}^{-1}\{\psi'(\theta^{(i-1)}) - \mu\}, \quad i = 1, 2, ...$$

until convergence, which is the solution θ. Recall here $\mu = \mathbb{E}[h(Y)]$.

There is a substantial literature on time series models based around the exponential family. Examples include Zeger and Qaqish (1988), Li (1994), Benjamin et al. (2003) and the review by Davis et al. (2021). The dynamic conditional score (DCS) filter and the generalized autoregressive score (GAS) filter of Harvey (2013) and Creal et al. (2013) can be applied in this context to deliver recursive filters. Our predictive version is closest to the generalized `ARMA` work of Benjamin et al. (2003), which is built around the exponential family and generalized linear models.

3.2 Filtering, prediction and smoothing

In what follows throughout we take the frame $\log L_t(\theta; y) = \log f_t(y; \theta)$ as coming from a member of the canonical exponential family $\texttt{CEF}(\theta, h_t, \psi_t)$. Then, for each t, (ignoring the implied $b_t(y)$ term as it has no impact on θ and dropping it means we do not need to make an assumption that $\mathbb{E}[\log b_t(Y_t)]$ exists)

$$(1-\alpha)\mathbb{E}[\log L_t(\theta; Y_t)] + \alpha \log L_t(\theta; Y_t) = \theta^{\mathtt{T}}\{(1-\alpha)\mathbb{E}[h_t(Y_t)] + \alpha h_t(Y_t)\} - \psi_t(\theta).$$

In the case where the data is assumed to be strictly stationary, the frame is stable and $\mathbb{E}[h(Y_t)]$ exists, then the right hand side of this expression simplifies to

$$\theta^{\mathtt{T}}\{(1-\alpha)\mathbb{E}[h(Y_1)] + \alpha h(Y_t)\} - \psi(\theta).$$

To compactly write the exponentially weighted filter it is helpful to denote the exponentially weighted moving sums

$$x_{\lambda,t} = \sum_{j=1}^{t} \lambda^{t-j}\mathbb{E}[h_j(Y_j)], \quad h_{\lambda,t} = \sum_{j=1}^{t} \lambda^{t-j} h_j(Y_j),$$

as well as the double sided exponentially weighted moving sums

$$x_{\lambda,t|T} = \sum_{j=1}^{T} \lambda^{|t-j|}\mathbb{E}[h_j(Y_j)], \quad h_{\lambda,t|T} = \sum_{j=1}^{T} \lambda^{|t-j|} h_j(Y_j).$$

Each of these terms can be computed recursively, implying each entire series, e.g. $h_{\lambda,1|T}, ..., h_{\lambda,T|T}$, can be computed in $O(\sum_{t=1}^{T} \dim(h_t(y_t)))$ computations.

Theorem 1 *Assume a minimal canonical exponential family frame* $\texttt{CEF}(\theta, h_t, \psi_t)$ *for each* t*. Then the exponentially weighted filter* ($\widetilde{\theta}_t$)*, predictor* ($\widetilde{\theta}_{t|t-s}$ *with* $s \geq 1$) *and smoother* ($\widetilde{\theta}_{t|T}$) *solve:*

$$x_{\lambda,t} + \alpha(h_{\lambda,t} - x_{\lambda,t}) = \sum_{j=1}^{t} \lambda^{t-j} \psi_j'(\widetilde{\theta}_t),$$

$$(1-\alpha)x_{\lambda,t} + \alpha\lambda^s h_{\lambda,t-s} = (1-\alpha)\sum_{j=1}^{t} \lambda^{t-j} \psi_j'\left(\widetilde{\theta}_{t|t-s}\right) + \alpha \sum_{j=1}^{t-s} \lambda^{t-j} \psi_j'\left(\widetilde{\theta}_{t|t-s}\right),$$

$$x_{\lambda,t|T} + \alpha(h_{\lambda,t|T} - x_{\lambda,t|T}) = \sum_{j=1}^{T} \lambda^{|t-j|} \psi_j'(\widetilde{\theta}_{t|T}),$$

respectively. Then write the filtered, predicted and smoothed mean and variance as

$$\widetilde{\mu}_t := \psi_t'(\widetilde{\theta}_t), \quad \widetilde{\mu}_{t|t-s} := \psi_t'(\widetilde{\theta}_{t|t-s}), \quad \widetilde{\mu}_{t|T} := \psi_t'(\widetilde{\theta}_{t|T}),$$

and

$$\widetilde{\Sigma}_t := \psi_t''(\widetilde{\theta}_t), \quad \widetilde{\Sigma}_{t|t-s} := \psi_t''(\widetilde{\theta}_{t|t-s}), \quad \widetilde{\Sigma}_{t|T} := \psi_t''(\widetilde{\theta}_{t|T}).$$

Proof. Given in the Appendix. ■

All the left hand side terms (e.g. $x_{\lambda,t}$ and h_{λ_t}) can be computed recursively. The right hand side, not so much, as they depend upon the value of θ that the ψ_j function is evaluated at. Typically $\widetilde{\theta}_t$, $\widetilde{\theta}_{t|t-s}$ and $\widetilde{\theta}_{t|T}$ have to be found by numerical root solving. Due to the strict convexity of ψ this is numerically straightforward, for each individual value of t, using Newton-Raphson, e.g. Algorithm 2 computes $\widetilde{\theta}_t$, the exponentially weighted filter.

Algorithm 2 (Computing $\widetilde{\theta}_t$) *Starting from some point* $\widetilde{\theta}_t^{(0)}$*, then iterate*

$$\widetilde{\theta}_t^{(i)} = \widetilde{\theta}_t^{(i-1)} - \left\{\sum_{j=1}^{t} \lambda^{t-j} \psi_j''(\widetilde{\theta}_t^{(i-1)})\right\}^{-1} \left\{\sum_{j=1}^{t} \lambda^{t-j} \psi_j'(\widetilde{\theta}_t^{(i-1)}) - \{x_{\lambda,t} + \alpha(h_{\lambda,t} - x_{\lambda,t})\}\right\}, \quad i = 1, 2, ...$$

until convergence, which is the solution $\widetilde{\theta}_t$*.*

A downside is that the numerical procedure has to be run separately for each t, so computing, for example, the time series $\widetilde{\theta}_1, ..., \widetilde{\theta}_T$ costs $O(\{\sum_{t=1}^{T} \dim(h_t(y_t))^3\}^2)$ calculations.

3.3 A $\texttt{CEF}(\theta, h_t, \psi_t)$ data generating process

We will use the following structure in our simulations later.

Assumption 3 (Model based data generating process) *(a) Use a frame based on* $\texttt{CEF}(\theta, h_t, \psi_t)$*, computing the sequence* $\widetilde{\theta}_{t|t-1}$ *for* $t = 1, 2, ..., T$*, recursively through Definition 2. (b) Generate the data as*

$$Y_t | Y_{1:t-1} \sim \texttt{CEF}(\widetilde{\theta}_{t|t-1}, h_t, \psi_t), \quad t = 1, 2, ..., T.$$

The stable frame version of this is where we set the frame to have $h_t(y) = h(y)$ *and* $\psi_t = \psi$ *for* $t = 1, ..., T$*.*

The use of the frame to derive a filter, which is then used as an input into a data generating process echos the DCS/GAS models of Harvey (2013) and Creal et al. (2013).

Some examples of these types of simulations will be given shortly.

4 Special case with analytic solution: $\psi_t = n_t \psi$

The following special case has an analytic solution: we will focus on it in the rest of this paper.

Example 1 *Use a minimal canonical exponential family frame* $\texttt{CEF}(\theta, h_t, n_t\psi)$ *for each* t *where* n_t *is a non-stochastic scalar. Then the exponentially weighted filter* ($\widetilde{\theta}_t$)*, predictor* ($\widetilde{\theta}_{t|t-s}$ *with* $s \geq 1$) *and smoother* ($\widetilde{\theta}_{t|T}$) *are:*

$$\widetilde{\theta}_t = \psi'^{-1}\left(\bar{m}_t\right), \quad \widetilde{\theta}_{t|t-s} = \psi'^{-1}\left(\bar{m}_{t|t-s}\right), \quad \widetilde{\theta}_{t|T} = \psi'^{-1}\left(\bar{m}_{t|T}\right),$$

respectively, where

$$\bar{m}_t = \frac{m_{\lambda,t}}{n_{\lambda,t}}, \quad \bar{m}_{t|t-s} = \frac{m_{\lambda,t|t-s}}{n_{\lambda,t|t-s}}, \quad \bar{m}_{t|T} = \frac{m_{\lambda,t|T}}{n_{\lambda,t|T}},$$

with

$$\begin{aligned} m_{\lambda,t} &= (1-\alpha)x_{\lambda,t} + \alpha h_{\lambda,t}, & n_{\lambda,t} &= \sum_{j=1}^{t} \lambda^{t-j} n_j, \\ m_{\lambda,t|t-s} &= (1-\alpha)x_{\lambda,t} + \alpha\lambda^s h_{\lambda,t-s}, & n_{\lambda,t|t-s} &= (1-\alpha)n_{\lambda,t} + \alpha\lambda^s n_{\lambda,t-s}, \\ m_{\lambda,t|T} &= (1-\alpha)x_{\lambda,t|T} + \alpha h_{\lambda,t|T}, & n_{\lambda,t|T} &= \sum_{j=1}^{T} \lambda^{|t-j|} n_j. \end{aligned}$$

Then

$$\widetilde{\mu}_t = n_t \bar{m}_t, \quad \widetilde{\mu}_{t|t-s} = n_t \bar{m}_{t|t-s}, \quad \widetilde{\mu}_{t|T} = n_t \bar{m}_{t|T},$$

and

$$\widetilde{\Sigma}_t = n_t \psi''(\widetilde{\theta}_t), \quad \widetilde{\Sigma}_{t|t-s} = n_t \psi''(\widetilde{\theta}_{t|t-s}), \quad \widetilde{\Sigma}_{t|T} = n_t \psi''(\widetilde{\theta}_{t|T}).$$

This Example covers many interesting models.

Remark 2 *(a) In the special case of a stable frame canonical exponential family and strictly stationary process then*

$$x_{\lambda,t} = n_{\lambda,t}\mathbb{E}[h(Y_1)], \quad x_{\lambda,t|T} = n_{\lambda,t|T}\mathbb{E}[h(Y_1)].$$

(b) For filtering, the $\widetilde{m}_{\lambda,t}$ is a convex combination of an `EWMA` *of the $\mathbb{E}[h(Y_1)], ..., \mathbb{E}[h(Y_t)]$ and the* `EWMA` *of $h(Y_1), ..., h(Y_t)$. For smoothing, the $\widetilde{m}_{\lambda,t|T}$ is a convex combination of the double sided* `EWMA` *of the $\mathbb{E}[h(Y_1)], ..., \mathbb{E}[h(Y_T)]$ and the double sided* `EWMA` *of $h(Y_1), ..., h(Y_T)$ for time t.*

(c) If $\lambda \in (0, 1)$, and $n_t = 1$ for all t, then

$$n_{\lambda,t} \to 1/(1-\lambda)$$

as $t \to \infty$. If $\lambda = 1$, then $n_{\lambda,t} = t$, while if $\lambda = 0$, then $n_{\lambda,t} = 1$. Likewise $n_{\lambda,t|T} \to (1+\lambda)/(1-\lambda)$ as $t \to \infty$. If $\lambda = 1$, then $n_{\lambda,t|T} = T$, while if $\lambda = 0$, then $n_{\lambda,t|T} = 1$.

(d) Famously the `EWMA` *can be computed recursively. In our case it follows*

$$\begin{aligned} n_{\lambda,t} &= n_t + \lambda n_{\lambda,t-1}, & n_{\lambda,0} &:= 0, \\ x_{\lambda,t} &= \mathbb{E}[h(Y_t)] + \lambda x_{\lambda,t-1}, & x_{\lambda,0} &:= 0, \\ h_{\lambda,t} &= h(Y_t) + \lambda h_{\lambda,t-1}, & h_{\lambda,0} &:= 0. \end{aligned}$$

Likewise, the double sided `EWMA` *can be computed recursively, going backwards*

$$\begin{aligned} h_{\lambda,t|T} &= h_{\lambda,t} + \lambda(h_{\lambda,t+1|T} - \lambda h_{\lambda,t}), \quad h_{\lambda,T|T} = h_{\lambda,T} \\ n_{\lambda,t|T} &= n_{\lambda,t} + \lambda(n_{\lambda,t+1|T} - \lambda n_{\lambda,t}), \quad n_{\lambda,T|T} = n_{\lambda,T}, \end{aligned}$$

using the output from the forward pass of the `EWMA`*.*

(e) The $m_{\lambda,t}$ and $m_{\lambda,t|t-1}$ can also be written recursively:

$$\begin{aligned} m_{\lambda,t} &= \mathbb{E}[h_t(Y_t)] + \alpha\{h_t(Y_t) - \mathbb{E}[h_t(Y_t)]\} + \lambda m_{\lambda,t-1}, \\ m_{\lambda,t|t-1} &= (1-\alpha)\mathbb{E}[h_t(Y_t)] + \alpha\lambda h_{t-1}(Y_{t-1}) + \lambda m_{\lambda,t-1|t-2}. \end{aligned}$$

In the steady state stable frame case with $n_t = 1$, the $\bar{m}_t = \widetilde{\mu}_t$ and $\bar{m}_{t|t-1} = \widetilde{\mu}_{t|t-1}$ become, collecting terms

$$\begin{aligned} \widetilde{\mu}_t &= (1-\alpha)(1-\lambda)\mathbb{E}[h(Y_1)] + \alpha(1-\lambda)h(Y_t) + \lambda\widetilde{\mu}_{t-1}, \\ \widetilde{\mu}_{t|t-1} &= \frac{(1-\alpha)(1-\lambda)}{1-\alpha(1-\lambda)}\mathbb{E}[h(Y_1)] + \frac{\alpha\lambda(1-\lambda)}{1-\alpha(1-\lambda)}h(Y_{t-1}) + \lambda\widetilde{\mu}_{t-1|t-2}. \end{aligned}$$

The $\widetilde{\mu}_{t|t-1}$ then relates to the generalized `ARMA` *model of Benjamin et al. (2003) who model a link function of the data as being linear in past conditional means and past link functions of the data.*

Define $U_t := h(Y_t) - \widetilde{\mu}_{t|t-1}$, then

$$\begin{aligned}
h(Y_t) &= \widetilde{\mu}_{t|t-1} + U_t \\
&= \frac{(1-\alpha)(1-\lambda)}{1-\alpha(1-\lambda)}\mathbb{E}[h(Y_1)] + \frac{\alpha\lambda(1-\lambda)}{1-\alpha(1-\lambda)}h(Y_{t-1}) + \lambda\widetilde{\mu}_{t-1|t-2} + U_t \\
&= \frac{(1-\alpha)(1-\lambda)}{1-\alpha(1-\lambda)}\mathbb{E}[h(Y_1)] + \frac{\alpha\lambda(1-\lambda)}{1-\alpha(1-\lambda)}h(Y_{t-1}) + \lambda\{h(Y_{t-1}) - U_{t-1}\} + U_t \\
&= \frac{(1-\alpha)(1-\lambda)}{1-\alpha(1-\lambda)}\mathbb{E}[h(Y_1)] + \frac{\lambda}{1-\alpha(1-\lambda)}h(Y_{t-1}) + U_t - \lambda U_{t-1}.
\end{aligned}$$

If the data has the property that $\mathbb{E}[h(Y_t)|Y_{1:t-1}] = \widetilde{\mu}_{t|t-1}$ and $\mathbb{E}[|Y_t|] < \infty$, then $\{U_t\}_{t=1}^T$ is a martingale difference (MD) sequence with respect to the filtration generated by the data. The steady state process is a vector `ARMA`(1,1)-MD process, with the autoregressive root being

$$\frac{\lambda}{1-\alpha(1-\lambda)} \in [\lambda, 1), \quad \textit{assuming} \quad \lambda, \alpha \in [0,1)^2,$$

with the $-\lambda$ moving average root. For example, if $\lambda = 0.93$ then the `AR`(1) root is roughly 0.978 and 0.996 when $\alpha = 0.7$ and $\alpha = 0.95$, respectively. Hence the process can have a great deal of memory, although sometimes the individual autocorrelations can be modest due to near root cancellation. If the data has the additional property that $\mathrm{V}(h(Y_t))$ exists and is time invariant, then in steady state the $\{U_t\}$ sequence is weak white noise (WN) and the $h(Y_t)$ sequence is a vector `ARMA`(1,1)-WN covariance stationary process.

(f) The $n_{\lambda,t}$ and $x_{\lambda,t}$ are free of α and data, while $h_{\lambda,t}$ is free of α (likewise for the smoothed versions).

(g) For prediction, $m_{\lambda,t|t-s}$ places non-negative weight $(1-\alpha)\lambda^{t-j}$ on each $\mathbb{E}[h_j(Y_j)]$ for $j = 1, \ldots, t$, and non-negative weight $\alpha\lambda^{t-j}$ on each $h_j(Y_j)$ for $j = 1, \ldots, t-s$. These weights sum to $n_{\lambda,t|t-s}$, so that $\bar{m}_{\lambda,t|t-s} = m_{\lambda,t|t-s}/n_{\lambda,t|t-s}$ has weights on all inputs which are non-negative and sum to one. This latter property also holds for $\bar{m}_{\lambda,t|t}$ and $\bar{m}_{\lambda,t|T}$. Hence Jensen's inequality can be applied to the filter, predictor and smoother, e.g. writing the predictor as $\bar{m}_{\lambda,t|t-s}(h_{1:t})$ to note its dependence on $h_1(Y_1), \ldots, h_t(Y_t)$, then

$$\varphi(\bar{m}_{\lambda,t|t-s}(h_{1:t})) \leq \bar{m}_{\lambda,t|t-s}(\varphi(h_1), \ldots, \varphi(h_t)),$$

where the right-hand side applies the linear operator $\bar{m}_{\lambda,t|t-s}$ to the component-wise transformed vector $(\varphi(h_1), \ldots, \varphi(h_t))$ with φ being any convex function. A simple version of this is where $h(y) = y$, and $\varphi(h(y)) = y^2$.

4.1 Simulating ten CEF examples: design

We illustrate the exponentially weighted predictor on ten examples from the canonical exponential family. Given $\widetilde{\mu}_{t|t-1}$, each yields a unique solution for $\widetilde{\theta}_{t|t-1} = (\psi')^{-1}(\widetilde{\mu}_{t|t-1})$. In cases 1-6 and 9, this can be solved analytically; in cases 7, 8, and 10, numerically via Algorithm 1. In cases 1-6, the canonical parameter θ is a scalar.

1. Bernoulli. For $Y_t \in \{0, 1\}$, the canonical parameterization is

$$h(y) = y, \quad \psi(\theta) = \log(1 + e^\theta), \quad \psi'(\theta) = \frac{e^\theta}{1 + e^\theta}, \quad (\psi')^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

2. Gaussian (known standard deviation). For $Y_t \in \mathbb{R}$ with known standard deviation σ, the canonical parameterization is

$$h(y) = y, \quad \psi(\theta) = \frac{\theta^2}{2\sigma^2}, \quad \psi'(\theta) = \frac{\theta}{\sigma^2}, \quad (\psi')^{-1}(\mu) = \sigma^2\mu.$$

The resulting predictor $\widetilde{\mu}_{t|t-1}$ will be the stationary version of the ARMA(1,1) process. We will compare the filter version with the result from the Kalman filter in Section 4.3.

3. Poisson. For $Y_t \in \{0, 1, 2, \ldots\}$, the canonical parameterization is

$$h(y) = y, \quad \psi(\theta) = e^\theta, \quad \psi'(\theta) = e^\theta, \quad (\psi')^{-1}(\mu) = \log(\mu).$$

4. Exponential. For $Y_t \in \mathbb{R}_{>0}$, the canonical parameterization is

$$h(y) = y, \quad \psi(\theta) = -\log(-\theta), \quad \psi'(\theta) = -\frac{1}{\theta}, \quad (\psi')^{-1}(\mu) = -\frac{1}{\mu}.$$

5. Gaussian (zero mean). For $Y_t \in \mathbb{R}$ with known zero mean, the canonical parameterization is

$$h(y) = y^2, \quad \psi(\theta) = -\frac{1}{2}\log(-2\theta), \quad \psi'(\theta) = -\frac{1}{2\theta}, \quad (\psi')^{-1}(\mu) = -\frac{1}{2\mu}.$$

In this notation $\mu = \mathbb{E}[Y^2] = \mathrm{V}(Y)$. The resulting predictor $\widetilde{\mu}_{t|t-1}$ will be the stationary version of the GARCH(1,1) model of Engle (1982) and Bollerslev (1986).

6. Pareto. For $Y_t \geq m > 0$, the canonical parameterization is

$$h(y) = \log(y), \quad \psi(\theta) = -\log(-\theta) + \theta\log(m), \quad \psi'(\theta) = -\frac{1}{\theta} + \log(m), \quad (\psi')^{-1}(\mu) = -\frac{1}{\mu - \log m}.$$

7. Beta. For $Y_t \in [0, 1]$, the canonical parameterization is

$$h(y) = \{\log(y), \log(1-y)\}^\intercal, \quad \psi(\theta) = \log B(\theta_1, \theta_2), \quad \psi'(\theta) = \begin{bmatrix} Dig(\theta_1) - Dig(\theta_1 + \theta_2) \\ Dig(\theta_2) - Dig(\theta_1 + \theta_2) \end{bmatrix},$$

where $B(\theta_1, \theta_2)$ is the beta function and Dig denotes the digamma function (the derivative of the log-gamma function). Given μ, the corresponding $\theta = (\psi')^{-1}(\mu)$ is unique and can found numerically via Algorithm 1.

8. Dirichlet. For the $Y_t \in \Delta_{d-1}$, a $d-1$ dimensional standard simplex, the canonical parameterization is

$$h(y) = \{\log(y_1), ..., \log(y_d)\}^{\mathtt{T}}, \quad \psi(\theta) = \log B(\theta_1, ..., \theta_d), \quad \psi'(\theta) = \begin{bmatrix} Dig(\theta_1) - Dig(\theta_1 + ... + \theta_d) \\ \vdots \\ Dig(\theta_d) - Dig(\theta_1 + \theta_d) \end{bmatrix},$$

Given μ, again the corresponding $\theta = (\psi')^{-1}(\mu)$ is unique and can found numerically via Algorithm 1. When $d = 2$ this reproduces the beta distribution case.

9. Gaussian. For $Y_t \in \mathbb{R}$, the canonical parameterization is

$$h(y) = \{y, y^2\}^{\mathtt{T}}, \quad \psi(\theta) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2), \quad \psi'(\theta) = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} \end{bmatrix}, \quad (\psi')^{-1}(\mu) = \begin{bmatrix} \frac{\mu_1}{\mu_2 - \mu_1^2} \\ -\frac{1}{2(\mu_2 - \mu_1^2)} \end{bmatrix}.$$

The resulting predictor $\widetilde{\sigma}^2_{t|t-1} = \widetilde{\mu}_{2,t|t-1} - \widetilde{\mu}^2_{1,t|t-1} \geq 0$ by invariance and then Jensen's inequality.

10. von Mises. For $Y_t \in [0, 2\pi]$, the canonical parameterization is

$$h(y) = \{\sin(y), \cos(y)\}^{\mathtt{T}}, \quad \psi(\theta) = \log I_0\left(\sqrt{\theta_1^2 + \theta_2^2}\right), \quad \psi'(\theta) = \frac{I_1\left(\sqrt{\theta_1^2 + \theta_2^2}\right)}{I_0\left(\sqrt{\theta_1^2 + \theta_2^2}\right)} \begin{bmatrix} \frac{\theta_1}{\sqrt{\theta_1^2 + \theta_2^2}} \\ \frac{\theta_2}{\sqrt{\theta_1^2 + \theta_2^2}} \end{bmatrix},$$

where I_0 is the modified Bessel function of the 1st kind and I_1 is the modified Bessel function of the 1st kind of order 1 (e.g. Abramowitz and Stegun (1970)). Given μ, again the corresponding $\theta = (\psi')^{-1}(\mu)$ is unique and can found numerically via Algorithm 1.

Following Assumption 3, we simulate each process using $Y_t | Y_{1:t-1} \sim \mathtt{CEF}(\widetilde{\theta}_{t|t-1}, h, \psi)$ with

$$\widetilde{\theta}_{t|t-1} = (\psi')^{-1}(\bar{m}_{t|t-1}) = (\psi')^{-1}\left(\frac{m_{\lambda,t|t-1}}{n_{\lambda,t}}\right),$$

where $m_{\lambda,t|t-1} = \mathbb{E}[h(Y_1)] + \lambda m_{\lambda,t-1}$, recalling Example 1. We impose stability by centering at $\mathbb{E}[h(Y_1)] = \mathbb{E}[Y_1] = 0.5$ (Bernoulli), $\mathbb{E}[h(Y_1)] = \mathbb{E}[Y_1] = 0$ (Gaussian with known variance $\sigma^2 = 1$), $\mathbb{E}[h(Y_1)] = \mathbb{E}[Y_1] = 1$ (Poisson and Exponential), and $\mathbb{E}[h(Y_1)] = \mathbb{E}[Y_1^2] = 1$ (Gaussian with zero mean), and $\mathbb{E}[h(Y_1)] = \mathbb{E}[\log(Y_1)] = 1/3$ (Pareto with scale $m = 1$, and shape $\theta_0 = -3$). For the bivariate cases, we take $\mathbb{E}[h(Y_1)] = \psi'(\theta_0)$ with initial values $\theta_0 = (2, 5)'$ for Beta (corresponding to $\mathtt{Beta}(2, 5)$), $\theta_0 = (0, -1/2)'$ for Gaussian (corresponding to mean 0, variance 1), and $\theta_0 = (0, -2)'$ for von Mises (corresponding to mean direction π, concentration 2).

4.2 Simulating ten CEF examples: results

Figures 1 and 2 show simulated time series Y_t and conditional predictors for time $t = 5, ..., T = 2000$ with $\lambda = 0.93$, $\alpha \in \{0.70, 0.95\}$, and $\widetilde{\theta}_{1|0} = (\psi')^{-1}(\mathbb{E}[h(Y_1)])$. Figure 1 displays the conditional mean $\mathbb{E}[\widetilde{Y_t | Y_{1:t-1}}] = \psi'(\widetilde{\theta}_{t|t-1})$ for Bernoulli, Gaussian (known variance $\sigma^2 = 1$), and Poisson distributions. Integer-valued observations (Bernoulli, Poisson) are jittered with $\mathrm{Unif}(-0.1, 0.1)$ noise for visualization.

Figure 2 shows the exponential distribution with conditional mean $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}] = \psi'(\widetilde{\theta}_{t|t-1})$, the Gaussian (zero mean) distribution with conditional standard deviation $\widetilde{\sigma}_{t|t-1} = \sqrt{\psi'(\widetilde{\theta}_{t|t-1})}$, and the Pareto distribution with conditional mean $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}] = \frac{\widetilde{\theta}_{t|t-1}}{\widetilde{\theta}_{t|t-1}+1}$ when $\widetilde{\theta}_{t|t-1} < -1$ and ∞ otherwise.

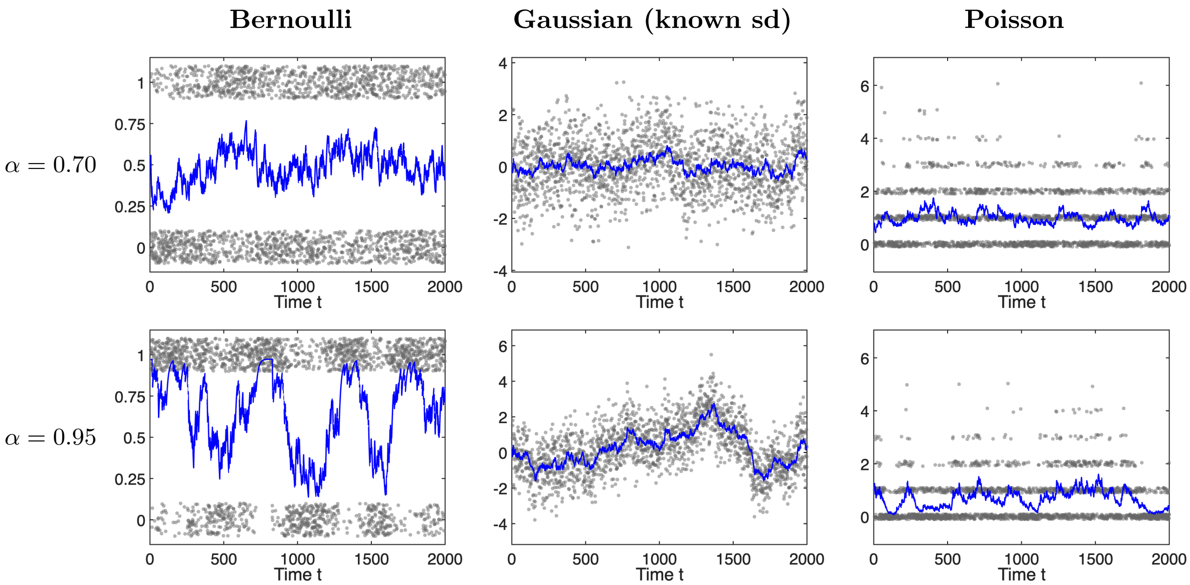

Figure 1: Simulation from $Y_t|Y_{1:t-1} \sim \texttt{CEF}(\widetilde{\theta}_{t|t-1}, h, \psi)$, where $\widetilde{\theta}_{t|t-1}$ follows the exponentially weighted predictor for time $t = 5, ..., T = 2000$ with discount parameter $\lambda = 0.93$. Top row has the anchoring parameter $\alpha = 0.7$, bottom row has $\alpha = 0.95$. Columns show the evolution of $\mathrm{E}[\widetilde{Y_t|Y_{1:t-1}}] = \psi'(\widetilde{\theta}_{t|t-1})$ in blue with observations Y_t as gray circles.

Figure 3 shows simulation results for the three bivariate cases: Beta, Gaussian with time-varying mean and variance, and von Mises. Each column displays the evolution of the conditional expectation $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}]$ with simulated observations. For the Beta distribution, the conditional mean is $\widetilde{\theta}_{1,t|t-1}/(\widetilde{\theta}_{1,t|t-1} + \widetilde{\theta}_{2,t|t-1})$. For Gaussian, it is $-\widetilde{\theta}_{1,t|t-1}/(2\widetilde{\theta}_{2,t|t-1})$ and the conditional standard deviation $\widetilde{\sigma}_{t|t-1} = \sqrt{-1/(2\widetilde{\theta}_{2,t|t-1})}$ is shown as a red line. For von Mises, we present the mean direction $\widetilde{\mu}_{t|t-1} = \text{atan2}(\widetilde{\theta}_{1,t|t-1}, \widetilde{\theta}_{2,t|t-1})$, modulo 2π to ensure $\widetilde{\mu}_{t|t-1} \in [0, 2\pi]$, where $\text{atan2}(y, x)$ is the four-quadrant arctangent of (x, y). For the anchoring parameter $\alpha = 0.70$, the predictors oscillate more around the centering values than for $\alpha = 0.95$, due to stronger mean reversion.

4.3 Gaussian case: comparison to Kalman filter

The Gaussian example with known standard deviation, case 2, links strongly to the Kalman filter for the univariate Gaussian local level model

$$Y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_\varepsilon^2),$$

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \overset{iid}{\sim} N(0, q\sigma_\varepsilon^2), \quad (\varepsilon_{1:T} \perp\!\!\!\perp \eta_{2:T})|\mu_1, \quad \mu_1 \sim N(a_{1|0}, \sigma_\varepsilon^2 P_1).$$

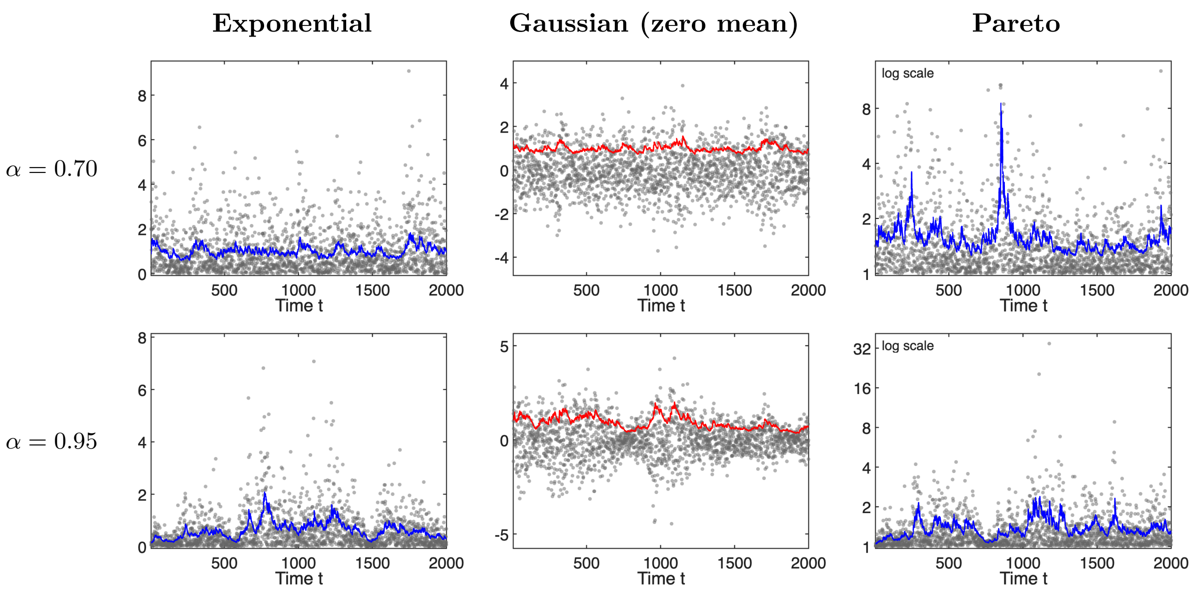

Figure 2: Simulation from $Y_t|Y_{1:t-1} \sim \texttt{CEF}(\widetilde{\theta}_{t|t-1}, h, \psi)$, where $\widetilde{\theta}_{t|t-1}$ follows the exponentially weighted predictor for time $t = 5, ..., T = 2000$ with discount parameter $\lambda = 0.93$. Top row has persistence parameter $\alpha = 0.7$, bottom row has $\alpha = 0.95$. Gray circles show observations Y_t. Blue lines show the conditional mean: $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}] = \psi'(\widetilde{\theta}_{t|t-1})$ for Exponential, and $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}] = \frac{\widetilde{\theta}_{t|t-1}}{\widetilde{\theta}_{t|t-1}+1}$ for Pareto (y-axis on log scale) when $\widetilde{\theta}_{t|t-1} < -1$ and ∞ otherwise. The red line for Gaussian (zero mean) shows the conditional standard deviation $\widetilde{\sigma}_{t|t-1} = \sqrt{\psi'(\widetilde{\theta}_{t|t-1})}$.

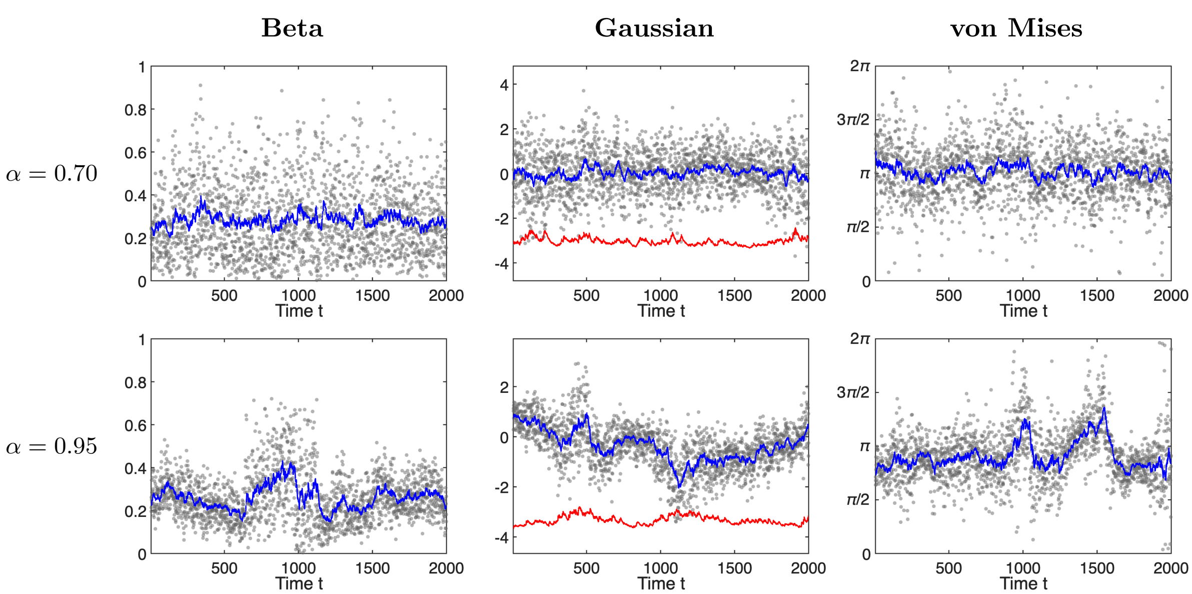

Figure 3: Simulation from $Y_t|Y_{1:t-1} \sim \texttt{CEF}(\widetilde{\theta}_{t|t-1}, h, \psi)$ following the exponentially weighted predictor for time $t = 5, ..., T = 2000$ with discount parameter $\lambda = 0.93$. Top row has anchoring parameter $\alpha = 0.7$, bottom row has $\alpha = 0.95$. Columns show the evolution of the conditional expectation $\mathbb{E}[\widetilde{Y_t|Y_{1:t-1}}]$ (blue line) with observations Y_t (gray circles) for Beta, Gaussian (time-varying mean and variance), and von Mises distributions. For the Gaussian distribution, the red line shows the conditional standard deviation $\widetilde{\sigma}_{t|t-1}$, shifted down by 4 units for visualization.

The Kalman filter implies $a_{t|t-1} = \mathrm{E}[\mu_t|Y_{1:t-1}] = \mathrm{E}[\mu_{t-1}|Y_{1:t-1}]$ is given by (e.g. Harvey (1989) and Durbin and Koopman (2012))

$$a_{t+1|t} = a_{t|t-1} + K_t(Y_t - a_{t|t-1}), \quad K_t = \frac{P_t}{P_t + 1}, \quad P_{t+1} = K_t + q,$$

where K_t is called the Kalman gain and $\sigma_\varepsilon^2 P_t = \mathrm{V}(\mu_t|Y_{1:t-1})$. If $q > 0$, in steady state, K_t and P_t converge to

$$K = \frac{P}{P+1}, \quad \text{and} \quad P = K + q.$$

Then $K^2 + qK - q = 0$, implying $K = \frac{-q+\sqrt{q^2+4q}}{2}$. In steady state, the one-step predictor is

$$a_{t+1|t} = (1-K)a_{t|t-1} + KY_t = K\sum_{j=0}^{\infty}(1-K)^j Y_{t-j},$$

and `EWMA` with discount hyperparameter

$$\lambda = 1 - K = \frac{2 + q - \sqrt{(2+q)^2 - 4}}{2}.$$

The discount rate monotonically declines as q increases, going from 1 to 0.

Our predictor in steady state, setting $\mathbb{E}[Y_1] = 0$ and $\alpha = 1$, implies

$$\widetilde{\mu}_{t+1|t} = \lambda\widetilde{\mu}_{t|t-1} + (1-\lambda)Y_t.$$

A more subtle comparison is out of steady state. For the Kalman filter, take $P_{1|0} = \infty$, then $K_1 = 1$. In our predictor,

$$\widetilde{\mu}_{t+1|t} = \left(1 - \frac{1}{n_{\lambda,t}}\right)\widetilde{\mu}_{t|t-1} + \frac{1}{n_{\lambda,t}}Y_t.$$

We note that if $\lambda > 0$, the $n_{\lambda,t} = \frac{1-\lambda^t}{1-\lambda}$. Hence it is interesting to plot

$$K_t \times n_{\lambda,t}$$

against t, when $P_{1|0} = \infty$, the so called diffuse initial conditions for a Kalman filter (e.g. Chapter 1 of Durbin and Koopman (2012)). If this product is less than one, then our recursion gives more weight on Y_t than the Kalman filter, and subsequently, less on the past.

Then $K_t n_{\lambda,t} = 1$ if $t = 1$. It is also 1 for every value of t when $q = 0$ for then $n_{\lambda,t} = t$. 1 is also the limit as $t \to \infty$ for $q > 0$ (the steady state case). However the product is not 1 for every t and $q > 0$.

The easiest way to see differences is to think of q as small, then, for example,

$$\lambda \approx 1 - q^{1/2}, \quad n_{\lambda,2} \approx 2 - q^{1/2}, \quad K_2 \approx \frac{1}{2} - q, \quad K_2 n_{\lambda,2} \approx 1 - q^{1/2}/2.$$

Figure 4 plots the product $K_t \times n_{\lambda,t}$ for $q \in \{0.001, 0.1, 0.3, 1, 2, 10\}$ against t. The worst case is when q is tiny, as it takes quite a large t for the effect to disappear (as we move to the steady state).

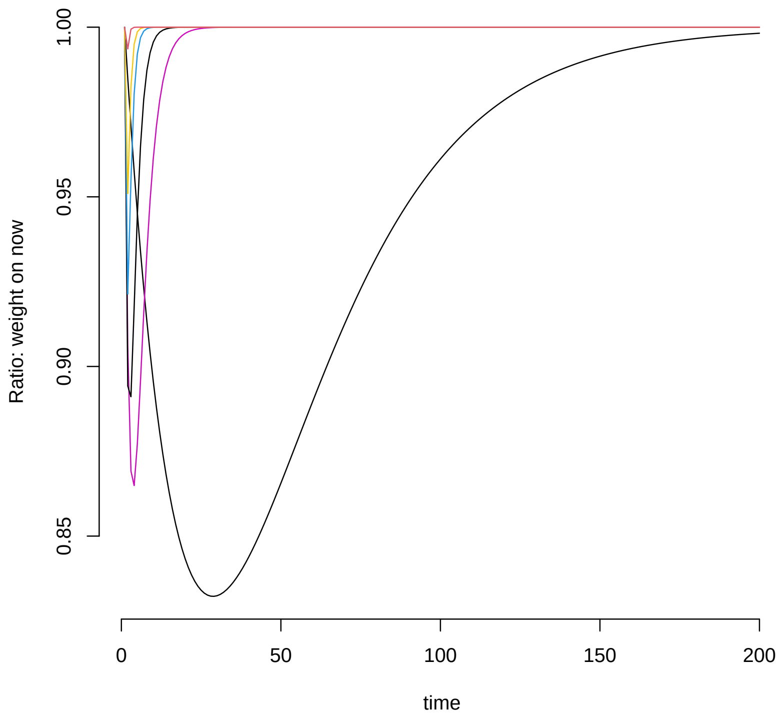

Figure 4: Product $K_t n_{\lambda,t}$, the ratio of the Kalman filter's weight (using diffuse initial conditions) on now, that is Y_t, to that of the EWMA scheme. All converge to 1, in steady state. Ratio is plotted against time, showing the impact of the initial conditions. Here the results for the cases $q \in \{0.001, 0.1, 0.3, 1, 2, 10\}$ are plotted against t. Biggest moves away from 1 are when q is small.

5 A quasi-likelihood approach to hyperparameter estimation

The analysis in this Section will be based on a working model

$$Y_t | Y_{1:t-1} \sim \texttt{CEF}(\widetilde{\theta}_{t|t-1}; h_t, \psi_t), \quad t = 1, ..., T,$$

where $\widetilde{\theta}_{t|t-1}$ comes from the exponentially weighted estimands. Later we will explicitly note the dependence of $\widetilde{\theta}_{t|t-1}$ on some hyperparameters, denoted ω.

Under the working model, write

$$\widetilde{\mu}_{t|t-1} := \psi_t'(\widetilde{\theta}_{t|t-1}), \quad \widetilde{\Sigma}_{t|t-1} := \psi_t''(\widetilde{\theta}_{t|t-1}),$$

the model's conditional mean and variance, respectively.

The working model yields a working log-likelihood via the prediction decomposition, ignoring constants,

$$l_t = \sum_{j=1}^{t} \Delta l_j, \quad \Delta l_t = h_t(y_t)^{\mathsf{T}} \widetilde{\theta}_{t|t-1} - \psi_t(\widetilde{\theta}_{t|t-1}).$$

To make sense of this working log-likelihood function we need some regularity assumptions.

Assumption 4 *Assume for a sequence $\{\widetilde{\theta}_{t|t-1}\}_{t=1}^{T}$ there exists constants $c_{1:2}, d_{1:2}$ such that for all t:*

(a) the $\mathbb{E}[|h(Y_t)^{\mathsf{T}}\widetilde{\theta}_{t|t-1}|] < c_1$ and $\mathbb{E}[|\psi_t(\widetilde{\theta}_{t|t-1})|] < d_1$.

(b) the $\mathrm{V}[h(Y_t)^{\mathsf{T}}\widetilde{\theta}_{t|t-1}] < c_2$ and $\mathrm{V}[\psi_t(\widetilde{\theta}_{t|t-1})] < d_2$.

5.1 A quasi-likelihood interpretation

As the working log-likelihood

$$l_t, \quad t = 1, ..., T,$$

comes from the $\mathtt{CEF}(\widetilde{\theta}_{t|t-1}, h_t, \psi_t)$, under weak regularity conditions, it will also be a quasi-likelihood — so long as the predictive mean is modeled correctly. This property does not depend upon the particular features of $\widetilde{\theta}_{t|t-1}$. We demonstrate this for canonical exponential family time series case. It then applies as a special case to when $\widetilde{\theta}_{t|t-1}$ is formed from the weighted predictor.

Now make some assumptions about the data.

Assumption 5 *Under the data, assume that:*

(a) $\mathbb{E}[|h(Y_t)|] < \infty$ and denote

$$\mathbb{E}[h_t(Y_t)|Y_{1:t-1}] := \mu_{t|t-1}, \quad t = 1, ..., T.$$

(b) $\mathrm{V}[h_t(Y_t)]$ exists, the

$$\Sigma_{t|t-1} := \mathrm{V}[h_t(Y_t)|Y_{1:t-1}], \quad t = 1, ..., T$$

and there exists a positive definite matrix C such that $\Sigma_{t|t-1} - C$ is positive semidefinite for every t.

Using strict convexity of ψ_t and Assumption 5(a) there exists a unique $\theta_{t|t-1}$ such that $\mu_{t|t-1} = \psi_t'(\theta_{t|t-1})$.

Definition 4 *Assume the sequence $\{\theta_{t|t-1}\}_{t=1}^{T}$ obeys Assumption 4(a) and $\theta_{t|t-1} = (\psi')^{-1}(\mu_{t|t-1})$ from Assumption 5(a). Then define the oracle log-likelihood:*

$$l_t^* = \sum_{j=1}^{t} \Delta l_j^*, \quad \Delta l_t^* = h_t(y_t)^{\mathsf{T}} \theta_{t|t-1} - \psi_t(\theta_{t|t-1}), \quad t = 1, ..., T.$$

Under Assumptions 4(a) and 5(a), define the two sequences

$$M_t = \sum_{j=1}^{t} \{l_j - \mathbb{E}[l_j^* | Y_{1:j-1}]\}, \quad t = 1, 2..., T,$$
$$C_t = \sum_{j=1}^{t} \{\mathbb{E}[l_j | Y_{1:j-1}] - \mathbb{E}[l_j^* | Y_{1:j-1}]\}.$$

then $\{M_t\}_{t=1}^T$ is a supermartingale with respect to the data's natural filtration. The $\{C_t\}_{t=1}^T$ is a previsible drift with increments

$$c_t = \psi'(\theta_{t|t-1})^{\intercal}\{\widetilde{\theta}_{t|t-1} - \theta_{t|t-1}\} - [\psi\{\widetilde{\theta}_{t|t-1}\} - \psi\{\theta_{t|t-1}\}] \leq 0,$$

by convexity of ψ from Definition 3 and then using p.69 of Boyd and Vandenberghe (2004). The equality is only obtained iff

$$\widetilde{\theta}_{t|t-1} = \theta_{t|t-1},$$

that is only if the working model has the correct predictive mean. Thus under Assumptions 4(a) and 5(a) the log-likelihood process $\{l_t\}_{t=1}^T$ can be view as a quasi-likelihood process.

If the magnitude of $\widetilde{\theta}_{t|t-1} - \theta_{t|t-1}$ is small, then

$$c_t \approx -\frac{1}{2}(\widetilde{\theta}_{t|t-1} - \theta_{t|t-1})^{\intercal}\psi_t''(\theta_{t|t-1})(\widetilde{\theta}_{t|t-1} - \theta_{t|t-1}).$$

More broadly, under Assumptions 4(a) and 5(a), the $\{M_t - C_t\}_{t=1}^T$ is a martingale. Further, under Assumptions 4(b) and Assumption 5(b) the angle bracket of the log-likelihood process

$$\begin{aligned}\langle l \rangle_t &:= \sum_{j=1}^{t} \mathrm{V}(l_j | Y_{1:j-1}), \quad t = 1, ..., T \\ &= \sum_{j=1}^{t} \widetilde{\theta}_{j|j-1}^{\intercal} \Sigma_{j|j-1} \widetilde{\theta}_{j|j-1}, \\ &= \langle M - C \rangle_t\end{aligned}$$

the angle-bracket process of the $\{M_t - C_t\}_{t=1}^T$ process. For all cases where $\langle M - C \rangle_T \to \infty$ then $(M_T - C_T)/\langle M - C \rangle_T \to 0$ almost surely, by the martingale strong law of large numbers.

Remark 3 *(a) This quasi-likelihood interpretation is not surprising, it falls in line with the storied history of, for example, generalized linear models, e.g. Wedderburn (1974) and McCullagh and Nelder (1989). More broadly quasi-likelihoods have played a vast role in modern statistics with the estimation theory going back at least to Cox (1961), Huber (1967), Gallant (1987) and White (1982).*

(b) The scaled $\{C_t(\omega)\}_{t=1}^T$ process measures how the scaled working log-likelihood l_T minus the scaled oracle l_T^* drifts downwards as T increases.

5.2 Maximum likelihood estimation

Now turn to the maximum likelihood estimator of the hyperparameters based on the above exponential family quasi-likelihood, where the predictor comes from exponential weighting and uses a canonical exponential family frame.

Assumption 6 *Based on a* $\texttt{CEF}(\theta, h_t, \psi_t)$ *frame, write* $\widetilde{\mu}_{t|t-1}(\omega)$ *as the predictor using some finite dimensional hyperparameter vector* $\omega \in \Omega$*. The hyperparameter vector may include any static parameters* $\phi \in \Phi$ *from the model density. Assume there exists a* ω^* *such that*

$$\widehat{\mu}_{t|t-1}(\omega^*) = \widetilde{\mu}_{t|t-1}, \quad \textit{for all} \quad t = 1, ..., T,$$

in the data and that $\widetilde{\mu}_{t|t-1}(\omega^*) \neq \widetilde{\mu}_{t|t-1}$ *for all* $\omega \in (\Omega \setminus \omega^*)$*. Then refer to* ω^* *as the pseudo-true or oracle value.*

The leading version of this is given in Example 2.

Example 2 *For a strictly stationary stochastic process* $\{Y_t\}_{t \geq 1}$*, assume that* $\mathbb{E}[h(Y_1)]$ *exists and lies in* $\mathcal{H} := h(\mathcal{Y}) \subseteq \mathbb{R}^k$*. For the exponentially weighted estimands under the stable frame minimal canonical exponential family* $\texttt{CEF}(\theta, h, \psi)$*, in steady state,*

$$\widetilde{\mu}_{t|t-1}(\omega) = \frac{(1-\alpha)(1-\lambda)}{1-\alpha(1-\lambda)}\mathbb{E}[h(Y_1)] + \frac{\alpha\lambda(1-\lambda)}{1-\alpha(1-\lambda)}h(Y_{t-1}) + \lambda\widetilde{\mu}_{t-1|t-2},$$

following Remark 2(e). Here the hyperparameters are

$$\omega := (\mathbb{E}[h(Y_1])^{\mathtt{T}}, \alpha, \lambda)^{\mathtt{T}} \in \Omega^* = \mathcal{H} \times (0,1) \times (0,1).$$

The $\widetilde{\mu}_{t|t-1}(\omega)$ *is infinitely differentiable with respect to* ω*, linear in transforms of the past data* $h(y_1), ..., h(y_{t-1})$*,* λ *and* $\mathbb{E}[h(Y_1)]$*, but nonlinear in* λ*. More generally, if the model density contains additional static parameters* ϕ *(e.g., a static variance parameter* σ^2 *in the Gaussian case that is not modeled by the exponentially weighted estimand), these can be included in the hyperparameter vector*

$$\omega := (\mathbb{E}[h(Y_1)]^{\mathtt{T}}, \alpha, \lambda, \phi^{\mathtt{T}})^{\mathtt{T}} \in \Omega^* = \mathcal{H} \times (0,1) \times (0,1) \times \Phi.$$

Such static parameters enter the $l_t(\omega)$ *but do not affect the structure of the predictor* $\widetilde{\mu}_{t|t-1}$*.*

The resulting maximum likelihood estimate formed from the quasi-likelihood is:

$$\hat{\omega}_{\texttt{MLE}} = \underset{\omega \in \Omega}{\arg\max} \; l_T(\omega),$$

noting explicitly how the quasi-likelihood is impacted by the choice of ω through $\{\widetilde{\theta}_{t|t-1}(\omega)\}_{t=1}^{T}$. Under Assumption 6, the corresponding ω^* is the pseudo-true value of ω for this quasi-likelihood function.

Throughout we will posit that $\widetilde{\theta}_{t|t-1}(\omega)$ is infinitely differentiable for all $\omega \in \Omega^*$ and we write

$$\widetilde{\mu}'_{t|t-1}(\omega) = \frac{\partial \widetilde{\mu}_{t|t-1}(\omega)^{\mathtt{T}}}{\partial \omega}.$$

The corresponding score up to time $1 \leq t \leq T$ is

$$\begin{aligned} S_t(\omega; y_{1:t}) &:= \frac{\partial l_t(\omega)}{\partial \omega} \\ &= \sum_{j=1}^{t} s_j(\omega), \quad s_t(\omega) := \frac{\partial l_t(\omega)}{\partial \omega} = \frac{\partial \widetilde{\theta}_{t|t-1}(\omega)^{\mathtt{T}}}{\partial \omega} \frac{\partial l_t(\omega)}{\partial \widetilde{\theta}_{t|t-1}(\omega)}. \end{aligned}$$

Assume that for all $\omega \in \Omega$ that $\widetilde{\Sigma}_{t|t-1}(\omega) = \psi''(\widehat{\theta}_{t|t-1}(\omega))$ is invertible. Then

$$s_t(\omega) = \widetilde{\theta}'_{t|t-1}(\omega)\{h(y_t) - \widetilde{\mu}_{t|t-1}(\omega)\}, \quad \text{where} \quad \widetilde{\theta}'_{t|t-1}(\omega) := \frac{\partial \widetilde{\theta}_{t|t-1}(\omega)^{\mathtt{T}}}{\partial \omega} = \widetilde{\mu}'_{t|t-1}(\omega)\widetilde{\Sigma}^{-1}_{t|t-1}(\omega),$$

recalling $\widetilde{\Sigma}_{t|t-1}(\omega) := \psi''(\widetilde{\theta}_{t|t-1}(\omega))$. The Hessian is

$$H_t(\omega) = -\sum_{j=1}^{t} l''_j(\omega), \quad \text{where} \quad l''_t(\omega) = \frac{\partial^2 l_t(\omega)}{\partial \omega \partial \omega^{\mathtt{T}}}.$$

Remark 4 *(a) Let* $s_t := s_t(\omega^*; Y_{1:t})$*, and* $S_t = \sum_{j=1}^{t} s_j$*. Then* $\{S_t\}_{t=1}^{T}$ *is a martingale sequence with respect to the data's natural filtration so long as* $\mathbb{E}[|s_t|] < \infty$*. Under the Brown (1971) martingale central limit theorem the*

$$\langle S, S\rangle_T^{-1/2} S_T \xrightarrow{D} N(0, I),$$

noting that $\langle S, S\rangle_T = \sum_{t=1}^{T} \widetilde{\theta}'_{t|t-1}\Sigma_{t|t-1}(\widetilde{\theta}'_{t|t-1})^{\mathtt{T}}$*, where* $\widetilde{\theta}'_{t|t-1} := \widetilde{\theta}'_{t|t-1}(\omega^*)$ *and* $\Sigma_{t|t-1} := \Sigma_{t|t-1}(\omega^*)$*.*

(b) By a multivariate mean value expansion $0 = S_T - \bar{H}_T(\hat{\omega}_{\mathtt{MLE}} - \omega^*)$*, where* $\bar{H}_T = \int_0^1 H_T(\omega^* + u(\hat{\omega}_{\mathtt{MLE}} - \omega^*))\, du$*. Hence if* $\langle S, S\rangle_T$ *is invertible then*

$$\begin{aligned} \{\langle S, S\rangle_T^{-1/2} \bar{H}_T\}(\hat{\omega}_{\mathtt{MLE}} - \omega^*) &= \langle S, S\rangle_T^{-1/2} S_T \\ &\xrightarrow{D} N(0, I). \end{aligned}$$

If $\bar{H}_T$ *is invertible, then*

$$W_T = \bar{H}_T^{-1}\langle S, S\rangle_T \bar{H}_T^{-1},$$

is an infeasible approximation to the variance-covariance matrix of the estimator.

(c) In practice we use the estimator

$$\hat{W}_T = H_T^{-1}(\hat{\omega}_{\mathtt{MLE}})[\hat{S}, \hat{S}]_T H_T^{-1}(\hat{\omega}_{\mathtt{MLE}}), \quad \textit{where} \quad [\hat{S}, \hat{S}]_T = \sum_{t=1}^{T} \hat{s}_t \hat{s}_t^{\mathtt{T}}, \quad \textit{with} \quad s_t = s_t(\hat{\omega}_{\mathtt{MLE}}; Y_{1:t}),$$

as a feasible approximation to the variance-covariance matrix of the estimator. Hence this is a sandwich matrix of the tradition we see in quasi-likelihood estimation, going back to at least Cox (1961) and Huber (1967).

5.3 Two step alternative to the MLE

An alternative to the MLE is a two step procedure (e.g. Newey and McFadden (1994), Engle and Mezrich (1996), Francq et al. (2013)), which has the following structure.

Definition 5 (2-step estimator) *For a strictly stationary stochastic process* $\{Y_t\}_{t\geq 1}$*, assume* $\mathbb{E}[h(Y_1)] \in \mathcal{H}$ *exists. For the exponentially weighted estimands under the stable frame minimal canonical exponential family* $\texttt{CEF}(\theta, h, \psi)$*, let* ϕ *denote any additional static parameters from the model density. Write*

$$\hat{\omega}_{\texttt{2Step}} = \begin{pmatrix} \widehat{\mathbb{E}[h(Y_1)]}_{\texttt{2Step}} \\ \widehat{\alpha}_{\texttt{2Step}} \\ \widehat{\lambda}_{\texttt{2Step}} \\ \widehat{\phi}_{\texttt{2Step}} \end{pmatrix}.$$

Then compute:

1. *The method of moments estimator:*

$$\widehat{\mathbb{E}[h(Y_1)]}_{\texttt{2Step}} = \frac{1}{T}\sum_{t=1}^{T} h(Y_t).$$

2. *The likelihood based estimator:*

$$\{\widehat{\alpha}_{\texttt{2Step}}, \widehat{\lambda}_{\texttt{2Step}}, \widehat{\phi}_{\texttt{2Step}}\} = \underset{\{\alpha,\lambda,\phi\}\in(0,1)^2\times\Phi}{\arg} \max \sum_{t=1}^{T} l_t(\widehat{\mathbb{E}[h(Y_1)]}_{\texttt{2Step}}, \alpha, \lambda, \phi),$$

Here the numerical optimization of the likelihood is only $2 + \dim(\phi)$ dimensional. This estimation strategy can be very attractive when $\mathbb{E}[h(Y_1)]$ is high dimensional.

The two step procedure can be viewed as a method of moments estimator (Pearson, 1894), based around a random function

$$S_t(\omega; y_{1:t}) := \sum_{j=1}^{t} s_j(\omega; y_{1:j}), \quad s_t(\omega; y_{1:t}) := \begin{pmatrix} h(Y_t) - \mathbb{E}[h(Y_1)] \\ \partial l_t(\omega)/\partial\alpha \\ \partial l_t(\omega)/\partial\lambda \\ \partial l_t(\omega)/\partial\phi \end{pmatrix}$$

for $t = 1, ..., T$, then $\hat{\omega}_{\texttt{2Step}}$ is a method of moments estimator which solves $S_T(\hat{\omega}_{\texttt{2Step}}; Y_{1:T}) = 0$, while $\mathbb{E}[s_t(\omega^*; Y_{1:t})] = 0$ for each $t = 1, ..., T$. The corresponding Hessian is

$$H_T(\omega) := -\sum_{t=1}^{T} \frac{\partial s_t(\omega)}{\partial \omega^{\mathsf{T}}} = \sum_{t=1}^{T} \begin{pmatrix} I \;\; 0 \;\; 0 \\ -\partial^2 l_t(\omega)/\partial\alpha\partial\omega^{\mathsf{T}} \\ -\partial^2 l_t(\omega)/\partial\lambda\partial\omega^{\mathsf{T}} \\ -\partial^2 l_t(\omega)/\partial\phi\partial\omega^{\mathsf{T}} \end{pmatrix}.$$

Remark 5 *(a) Write* $s_t = s_t(\omega^*; Y_{1:t})$ *and* $S_t = S_t(\omega^*; Y_{1:t})$*. The last three elements of* $\{s_t\}$ *are martingale differences. The remaining elements,* $h(Y_t) - \mathbb{E}[h(Y_1)]$*, have an unconditional zero mean but possibly substantial time series memory. Write out a corresponding CLT as*

$$V_T^{-1/2} S_T(\omega^*) \xrightarrow{D} N(0, I), \quad \textit{where} \quad V_T = \mathrm{V}(S_T(\omega^*)),$$

which needs the memory in $\{s_t\}$ to be controlled. The martingale difference elements have limited memory, the problem is the time series of $\{h(Y_t) - \mathbb{E}[h(Y_1)]\}$. A basic way of generating a CLT for these type of objects is to assume the $h(Y_t)$ series is strictly stationary and exhibits m-dependence (e.g. Janson (2021)).

(b) By a multivariate mean value expansion $0 = S_T - \bar{H}_T(\hat{\omega}_{\texttt{2Step}} - \omega^*)$, where $\bar{H}_T = \int_0^1 H_T(\omega^* + u(\hat{\omega}_{\texttt{2Step}} - \omega^*))\, du$. Thus

$$
\begin{aligned}
V_T^{-1/2}\bar{H}_T(\hat{\omega}_{\texttt{2Step}} - \omega^*) &= V_T^{-1/2} S_T \\
&\xrightarrow{D} N(0, I).
\end{aligned}
$$

So the infeasible variance matrix for the two step estimator is

$$
W_T = \bar{H}_T^{-1} V_T \bar{H}_T^{-1}.
$$

(c) In practice we use the estimator of the covariance matrix

$$
\hat{W}_T = H_T^{-1}(\hat{\omega}_{\texttt{2Step}})\hat{V}_T H_T^{-1}(\hat{\omega}_{\texttt{2Step}})^{\mathtt{T}},
$$

where $\hat{V}_T$ approximates $\mathrm{V}(\hat{S}_T)$. The latter can be estimated using T times a long-run variance of the time series $\hat{s}_1, ..., \hat{s}_T$, where $\hat{s}_t = s_t(\hat{\omega}_{\texttt{2Step}}; Y_{1:t})$.

6 Empirical example

6.1 Dirichlet based frame: household financial situation expectations

We apply the exponentially weighted predictor and smoother to monthly data on household financial situation expectations from the University of Michigan Survey of Consumers, covering January 1978 to September 2025 ($T = 573$ observations). The survey asks respondents two questions: first, whether they are better off or worse off financially than a year ago, and second, whether they expect to be better off or worse off a year from now. The combination of responses yields seven mutually exclusive categories:

(i) Continuous Increase (better off now and expect to be better off),

(ii) Intermittent Increase (one period better, one period same),

(iii) Remain Unchanged (same in both periods),

(iv) Intermittent Decline (one period worse, one period same),

(v) Continuous Decline (worse off now and expect to be worse off),

(vi) Mixed Change (improvement followed by decline or vice versa),

(vii) Don't Know/No answer.

For each month, only the sample proportions (percentages) falling into each category are reported, not the underlying individual responses or sample sizes.[3]

We model the observed proportions $y_t = (y_{1,t}, \ldots, y_{7,t})^\mathtt{T}$ as realizations from a time-varying seven category Dirichlet distribution

$$Y_t \mid Y_{1:t-1} \sim \text{Dirichlet}(\widetilde{\theta}_{t|t-1}), \quad \widetilde{\theta}_{t|t-1} = (\widetilde{\theta}_{1,t|t-1}, \ldots, \widetilde{\theta}_{7,t|t-1})^\mathtt{T},$$

with concentration parameters $\widetilde{\theta}_{j,t|t-1} > 0$ for all $j \in \{1, \ldots, 7\}$ and all t. Each observation is treated as a single draw of a proportion vector and the predictor is based on a Dirichlet frame.

The exponentially weighted predictor and smoother are computed via the recursions in Example 1 using the Dirichlet distribution's CEF parameterization given in Section 4.1.

Following the two-step quasi-likelihood procedure from Definition 5, we first estimate the centering parameter $\mathbb{E}[h(Y_1)]$ by the sample mean

$$\widehat{\mathbb{E}[h(Y_1)]} = T^{-1} \sum_{t=1}^{T} (\log y_{1,t}, \ldots, \log y_{7,t})^\mathtt{T}.$$

Next, we fix $\widehat{\mathbb{E}[h(Y_1)]}$ and estimate (α, λ) by maximizing the quasi-log-likelihood

$$\sum_{t=1}^{T} \log f(y_t; \widetilde{\theta}_{t|t-1}(\alpha, \lambda)).$$

For the household financial situation data with seven categories, we obtain the parameter estimates $(\widehat{\mathbb{E}[h(Y_1)]}^\mathtt{T}, \widehat{\alpha}, \widehat{\lambda}) \approx (-1.76, -1.41, -1.78, -1.77, -2.73, -2.23, -3.53, 0.95, 0.65)$. The high anchoring estimate $\widehat{\alpha} = 0.95$ indicates little long-run effect, suggesting that shifts in household financial expectations, once they occur, tend to persist a long time. The discount parameter $\widehat{\lambda} = 0.65$ implies a half-life of approximately 1.6 months for the exponential weights. Following Remark 2(e), the steady state process is a vector `ARMA`(1,1)-MD process, with autoregressive root $\widehat{\lambda}/\{1 - \widehat{\alpha}(1 - \widehat{\lambda})\} \approx 0.974$, and moving average root $-\widehat{\lambda} \approx -0.65$. Hence, the process exhibits substantial memory, with substantial root cancellation.

Figure 5 presents the estimation results. Each panel shows the evolution of one expectation category over the 47-year period. The gray circles represent observed monthly proportions $y_{k,t}$. The light colored lines show the exponentially weighted predictor, the dark colored lines the exponentially weighted smoother. Both successfully track major shifts in household financial expectations. Patterns that stand out are the sharp declines in both "Continuous Increase" and "Intermittent Increase" during the 2008 financial crisis, and, more strikingly, since around 2018. This deterioration in optimism about household

[3] Only for the period April 2024 to April 2025 the monthly sample sizes are given. We leave them out of our analysis.

finances is mirrored by corresponding increases in negative expectations: “Intermittent Decline” and “Continuous Decline” roughly track their inverse patterns. The categories “Remain Unchanged” and “Don’t Know/No Answer” stay relatively stable at 15-20% and 3%, respectively, throughout the sample.

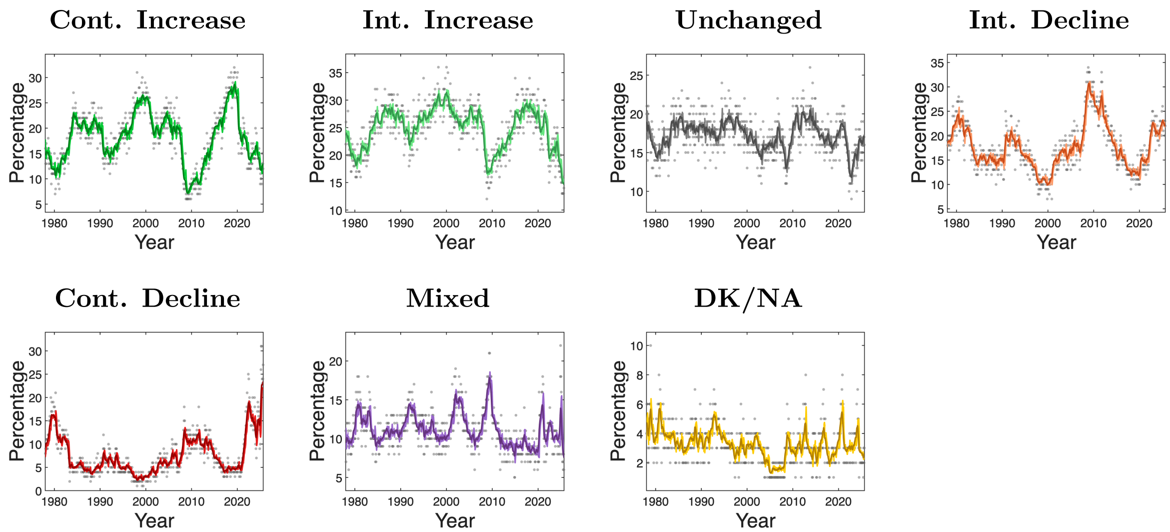

Figure 5: Household financial situation expectations from the University of Michigan Survey of Consumers (January 1978 to September 2025, $T = 573$ monthly observations). Categories combine responses about current versus past financial situation, and expectations for the year ahead (Cont. = Continuous, Int. = Intermittent, DK = Don’t Know, NA = No Answer): Continuous Increase means better off now and expecting better, Intermittent Increase indicates improvement in one period only, and analogously for declines. Each panel shows, for one response category, the observed monthly proportions $y_{k,t}$ (gray circles), the predictor $y_{k,t|t-1} = \widetilde{\theta}_{k,t|t-1} / \sum_{j=1}^{7} \widetilde{\theta}_{j,t|t-1}$ (light colored line), and the smoother $y_{k,t|T} = \widetilde{\theta}_{k,t|T} / \sum_{j=1}^{7} \widetilde{\theta}_{j,t|T}$ (dark colored line). Estimates: $(\widehat{\alpha}, \widehat{\lambda}) = (0.95, 0.65)$.

6.2 Monte Carlo assessment of estimator precision

To investigate the finite-sample precision of the quasi-likelihood estimator from Definition 5, we conduct a Monte Carlo study using the estimated model from Section 6.1 as the data-generating process (DGP). We generate time-varying proportion vectors following Assumption 3, that is,

$$Y_t \mid Y_{1:t-1} \sim \texttt{Dirichlet}(\theta_{t|t-1}), \quad \theta_{t|t-1} = (\theta_{1,t|t-1}, \ldots, \theta_{7,t|t-1})^\top,$$

where $\theta_{t|t-1}$ evolves according to the exponentially weighted predictor (Theorem 1), denoting $(\pi^*, \alpha^*, \lambda^*)$ for the hyperparameters in the DGP. We initialize each prediction path with $\widetilde{\theta}_{1|0} = (\psi')^{-1}(\pi^*)$ where $\pi^* = \widehat{\mathbb{E}[h(y_1)]}$ is the estimated centering parameter from Section 6.1.

To investigate how the anchoring α^* affects hyperparameter estimation, we consider four scenarios:

(i) $(\pi^*, \alpha^*, \lambda^*) = (\widehat{\pi}, 0.01, 0.65)$ (strong anchoring),

(ii) $(\pi^*, \alpha^*, \lambda^*) = (\widehat{\pi}, 0.30, 0.65)$ (moderate anchoring),

(iii) $(\pi^*, \alpha^*, \lambda^*) = (\widehat{\pi}, 0.60, 0.65)$ (weak anchoring),

(iv) $(\pi^*, \alpha^*, \lambda^*) = (\widehat{\pi}, 0.95, 0.65)$ (no anchoring, matching estimated values),

where $\pi^* = (-1.76, -1.41, -1.78, -1.77, -2.73, -2.23, -3.53)'$ and $\lambda^* = 0.65$ is fixed across all scenarios to match the estimated discount parameter from the household data.

Figure 6 displays the quasi log-likelihood surfaces and confidence regions for the estimated (α, λ). Confidence regions contract systematically as T increases, with maximum quasi-likelihood estimates (red stars) concentrating near the DGP values (blue circles). The pattern of contraction depends strongly on the anchoring hyperparameter α^*. When $\alpha^* = 0.01$ (first row), the confidence interval for λ remains wide even at $T = 10000$. This aligns with Remark 1(e): when α is close to zero, the frame is nearly stationary, and λ has minimal impact on the filter and predictor, making it difficult to estimate from the data. Contrastingly, when $\alpha^* \in \{0.60, 0.95\}$ (third and fourth rows), confidence intervals for both hyperparameters contract rapidly even at moderate sample sizes.

7 Conclusions

The exponentially weighted estimands combined with the canonical exponential family delivers simple exact filters, predictors and smoothers. They are in the same statistical spirit as generalized linear models (e.g. McCullagh and Nelder (1989)) and are relatively easy to fit to data using a quasi-likelihood approach either all at once or using a two step approach.

The exponentially weighted estimands framework can be used on non-exponentially family models and even for general loss functions. In such cases the optimization has to be carried out numerically or through a sequential approximation. This then relates to work on stochastic gradient descent (e.g. Robbins and Munro (1951), Toulis and Airoldi (2015, 2017), Toulis et al. (2016), Toulis et al. (2021)), Lange et al. (2024), Donker van Heel et al. (2024), as well as the DCS/GAS filter of Harvey (2013) and Creal et al. (2013).

References

Abramowitz, M. and I. A. Stegun (1970). *Handbook of Mathematical Functions*. New York: Dover Publications Inc.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Chichester: John Wiley and Sons.

Benjamin, M. A., R. A. Rigby, and D. M. Stasinopoulos (2003). Generalized autoregressive moving average model. *Journal of the American Statistical Association 98*, 214–223.

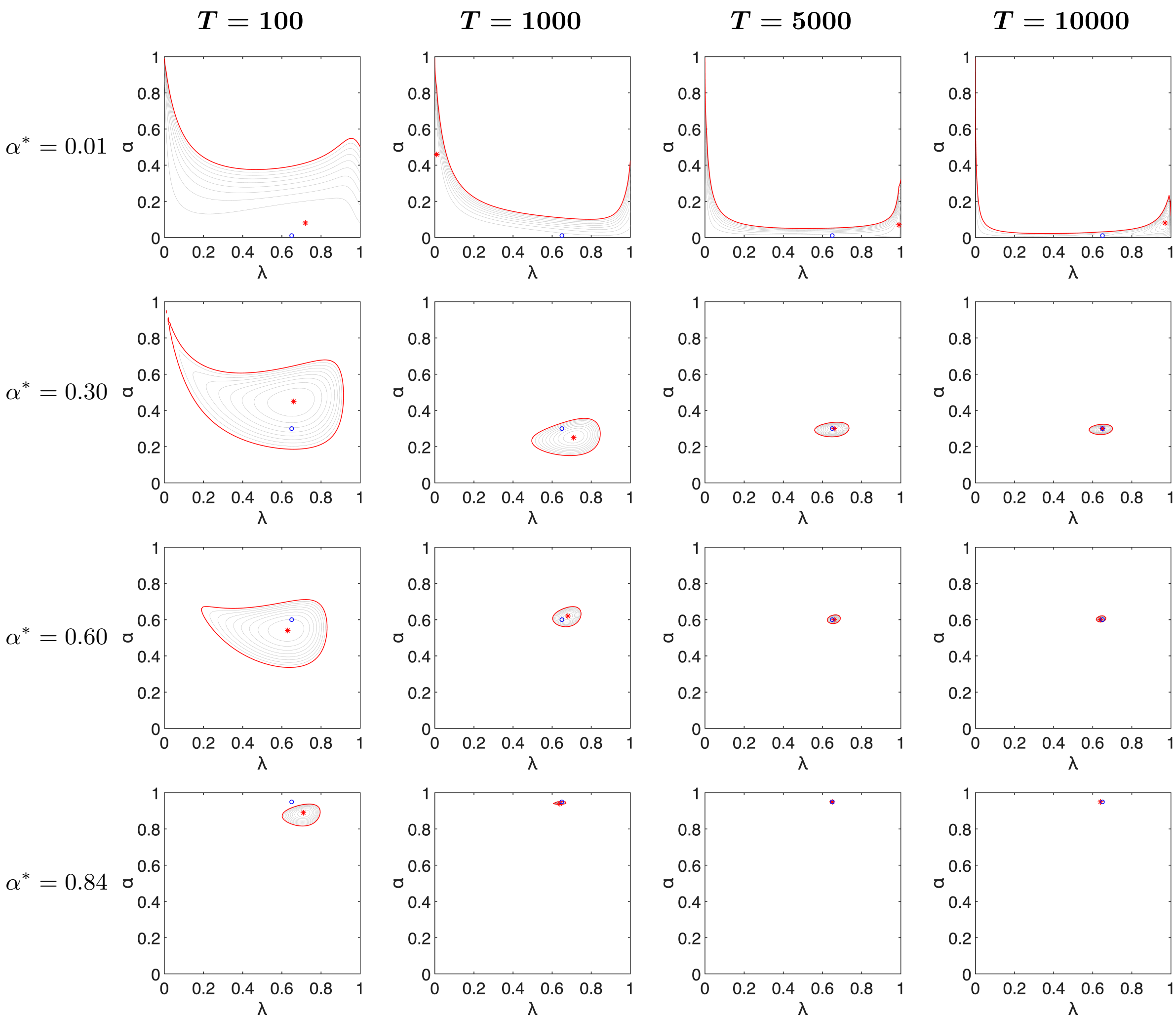

Figure 6: Monte Carlo assessment of quasi-likelihood estimator precision for household financial expectations across varying anchoring levels and sample sizes. Rows correspond to four DGP scenarios with anchoring hyperparameters: (i) $\alpha^* = 0.01$, (ii) $\alpha^* = 0.30$, (iii) $\alpha^* = 0.60$, and (iv) $\alpha^* = 0.95$ (matching estimated household data). All scenarios use the estimated centering parameter $\pi^* = \widehat{\mathbb{E}[h(y_1)]} = (-1.76, -1.41, -1.78, -1.77, -2.73, -2.23, -3.53)'$ and discount parameter $\lambda^* = 0.65$ from Section 6.1. Each panel displays the quasi log-likelihood surface $\ell(\alpha, \lambda)$ with the maximum quasi-likelihood estimate (red star) and DGP value (blue circle). The red contour traces the 99% confidence region. Columns show increasing sample sizes $T \in \{100, 1000, 5000, 10000\}$.

Blasques, F., S. J. Koopman, M. Mallee, and Z. Zhang (2016). Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data. *Journal of Econometrics 193*(2), 405–417.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics 51*, 307–327.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Brown, B. M. (1971). Martingale central limit theorems. *Annals of Mathematical Statistics 49*, 59–66.

Brown, R. G. (1956). *Exponential smoothing for predicting demand*. Cambridge, Massachusetts: Author D Little, Inc.

Cox, D. R. (1961). Tests of seperate families of hypotheses. *Proceedings of the Berkeley Symposium 4*, 105–123.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*, 777–795.

Davis, R. A., K. Fokianos, S. H. Holan, H. Joe, J. Livsey, R. Lund, V. Pipiras, and N. Ravishanker (2021). Count time series: A methodological review. *Journal of the American Statistical Association 116*(535), 1533–1547.

Dixon, M. J. and S. G. Coles (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 46*(2), 265–280.

Donker van Heel, S., R. J. Lange, B. van Os, and D. van Dijk (2024). Stability and performance guarantees for misspecified multivariate score-driven filters. Unpublished paper: Erasmus University.

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

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Cambridge University Press.

Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific American 236*, 119–127.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica 50*, 987–1007.

Engle, R. F. and J. Mezrich (1996). GARCH for groups. *Risk*, 36–40.

Fan, J., N. E. Heckman, and M. P. Wand (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association 90*, 141–150.

Fan, J. and Q. Yao (2005). *Nonlinear Time Series*. New York: Springer.

Francq, C., L. Horvath, and J.-M. Zakoïan (2013). Merits and drawbacks of variance targeting in garch models. *Journal of Financial Econometrics 9*, 619–656.

Gallant, A. R. (1987). *Nonlinear Statistical Models*. New York: John Wiley.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics 190*, 115–132.

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

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: With applications to financial and economic time series*. Cambridge University Press.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics 12*, 69–82.

Hoerl, A. E. and R. W. Kennard (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55—67.

Holmes, C. C. and S. G. Walker (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika 104*, 497–503.

Hu, F. and J. V. Zidek (2002). The weighted likelihood. *Canadian Journal of Statistics 30*(3), 347–371.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. University of California Press.

Janson, S. (2021). A central limit theorem for m-dependent variables. Unpublished paper: Department of Mathematics, Uppsala University.

Lange, R. J., B. van Os, and D. van Dijk (2024). Implicit score-driven filters for time-varying parameter models. Unpublished paper: Erasmus University.

Li, W. K. (1994). Time series models based on generalized linear models: Some further results. *Biometrics 50*, 506–511.

Luxenberg, E. and S. Boyd (2024). Exponentially weighted moving models. Unpublished paper: Stanford University.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2 ed.). London: Chapman & Hall.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *The Handbook of Econometrics, Volume 4*, pp. 2111–2245. North-Holland.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, Series A 185*, 71–110.

Robbins, H. and S. Munro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics 22*, 400–407.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B 58*, 267–288.

Toulis, P. and E. M. Airoldi (2015). Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Statistics and Computing 25*, 781–795.

Toulis, P. and E. M. Airoldi (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Annals of Statistics 45*(4), 1694–1727.

Toulis, P., T. Horel, and E. M. Airoldi (2021). The proximal Robbins–Monro method. *Journal of the Royal Statistical Society Series B: Statistical Methodology 83*(1), 188–212.

Toulis, P., D. Tran, and E. M. Airoldi (2016). Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pp. 1290–1298. PMLR.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton methods. *Biometrika 61*, 439–47.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Zeger, S. L. and B. Qaqish (1988). Markov regression models for time series, a quasi likelihood approach. *Biometrics 44*, 1019–1032.

8 Appendix: some proofs

8.1 Proof of Theorem 1

Focus on the filtering case. Now

$$Q_{\lambda,t}(\theta) = \theta^{\mathtt{T}}\{(1-\alpha)x_{\lambda,t} + \alpha h_{\lambda,t}\} - \sum_{j=1}^{t} \lambda^{t-j}\psi_j(\theta).$$

Differentiate $Q_{\lambda,t}(\theta)$ with respect to θ and solve. This yields the stated result.

The same argument applied to prediction and smoothing.

The remaining results follow by invariance.