3-2016

Power Weighted Densities for Time Series Data

Daniel McCarthy
University of Pennsylvania

Shane T. Jensen
University of Pennsylvania

Recommended Citation
McCarthy, D., & Jensen, S. T. (2016). Power Weighted Densities for Time Series Data. *The Annals of Applied Statistics, 10* (1), 305-334.
http://dx.doi.org/10.1214/15-AOAS893

Power Weighted Densities for Time Series Data

Abstract
While time series prediction is an important, actively studied problem, the predictive accuracy of time series models is complicated by nonstationarity. We develop a fast and effective approach to allow for nonstationarity in the parameters of a chosen time series model. In our power-weighted density (PWD) approach, observations in the distant past are down-weighted in the likelihood function relative to more recent observations, while still giving the practitioner control over the choice of data model. One of the most popular nonstationary techniques in the academic finance community, rolling window estimation, is a special case of our PWD approach. Our PWD framework is a simpler alternative compared to popular state–space methods that explicitly model the evolution of an underlying state vector. We demonstrate the benefits of our PWD approach in terms of predictive performance compared to both stationary models and alternative nonstationary methods. In a financial application to thirty industry portfolios, our PWD method has a significantly favorable predictive performance and draws a number of substantive conclusions about the evolution of the coefficients and the importance of market factors over time.

Keywords
Time series analysis, power prior, forecasting, finance

Disciplines
Physical Sciences and Mathematics

Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

POWER WEIGHTED DENSITIES FOR TIME SERIES DATA

By Daniel McCarthy and Shane T. Jensen

University of Pennsylvania and University of Pennsylvania

While time series prediction is an important, actively studied problem, the predictive accuracy of time series models is complicated by non-stationarity. We develop a fast and effective approach to allow for non-stationarity in the parameters of a chosen time series model. In our power-weighted density (PWD) approach, observations in the distant past are down-weighted in the likelihood function relative to more recent observations, while still giving the practitioner control over the choice of data model. One of the most popular non-stationary techniques in the academic finance community, rolling window estimation, is a special case of our PWD approach. Our PWD framework is a simpler alternative compared to popular state-space methods that explicitly model the evolution of an underlying state vector. We demonstrate the benefits of our PWD approach in terms of predictive performance compared to both stationary models and alternative non-stationary methods. In a financial application to thirty industry portfolios, our PWD method has a significantly favorable predictive performance and draws a number of substantive conclusions about the evolution of the coefficients and the importance of market factors over time.

1. Introduction and Motivation. An increasingly prominent area of statistical application is the modeling of data that is ordered over time, either as a single time series or multiple time series, with the goal being the prediction of future time series data. It is often unrealistic to assume stationarity, whereby the underlying parameters of the chosen model are constant over time. Rather, it may be preferred to allow the parameters of the model to evolve over time, which complicates modeling efforts. We propose a general methodology which may be used to improve the predictive accuracy of time series models by addressing possible non-stationarity in model parameters.

In the time series application that we focus upon, the issue of non-stationarity is particularly acute – estimation of the sensitivity of stock returns to market factors. Fama and French (1993) introduced the popular three factor model in asset pricing, which relates the return on stock portfo-

MSC 2010 subject classifications: Primary 37M10, Secondary 62P05
Keywords and phrases: Time series analysis; power prior; forecasting; finance

lios to their valuation, size and sensitivity to the overall market. Specifically, the returns $y_{j,t}$ of a stock portfolio j at time t were modeled as a linear function of three factors,

$$y_{j,t} = \alpha_j + \beta_j^m \cdot m_t + \beta_j^s \cdot s_t + \beta_j^v \cdot v_t + \epsilon_{j,t} \tag{1.1}$$

where m_t represents excess return on the market portfolio ('MKT'), s_t represents excess return of small capitalization stocks over large capitalization stocks ('SMB'), v_t represents the excess return of value stocks over growth stocks ('HML') and $\epsilon_{j,t}$ is a noise term. Since then, hundreds of papers have been written trying to explain cross-sectional heterogeneity in asset price returns through the inclusion of additional factors. The overarching goal of this literature is to explain variation in returns across stocks through a relatively small number of market factors, which is equivalent to predicting stock returns using contemporaneous predictors in a time series regression.

Time series regression problems like this one are notoriously challenging because the parameters of the regression model are unlikely to be stationary over time. The sensitivity of parameters should be allowed to evolve over time (e.g. $\beta_{j,t}^m$ rather than β_j^m, $\beta_{j,t}^s$ rather than β_j^s, etc., in Equation 1.1). The question here, and in many other applied settings, is how to address potential non-stationarity in the parameters of a chosen model? Throughout the remainder of this paper, we will use the term 'non-stationarity' to mean that the parameters of the true underlying process generating the observed data are potentially varying over time.

Our methodological objective is to produce the best possible *predictions at the next time point*, conditional upon the model the practitioner has chosen. If we are unwilling to assume stationarity over time for the model parameters, the consequence is that not all historical data will be equally relevant to the prediction of future outcomes. With prediction as our ultimate goal, we will propose statistical methodology for a principled differential weighting of historical data that is simple and efficient relative to traditional methods that focus on estimation of the underlying parameter evolution. While this paper explores an application to market factor sensitivies, our *power-weighted densities* (PWD) approach can be applied to any time series setting where the underlying data generating process is believed to be non-stationary over time.

Financial data are an interesting case study for time series methods as many assets have been tracked for a relatively long time period. In this paper, we will model the monthly returns of 30 industry portfolios (Kenneth French). The time series begin in July 1926 and end in December 2014, which gives us 1062 time points for each of 30 stock portfolios.

However, the long length of these time series is deceptive due to non-stationarity in the underlying data generating process. Acknowledging this non-stationarity, practitioners usually employ some sort of data truncation, ignoring data which is 'old enough' under the assumption that market conditions make data prior to that point irrelevant or even harmful to the predictive accuracy of their model.

In the finance literature, non-stationary is usually addressed by estimating asset models using *rolling windows*, i.e. assuming a stationary model in a fixed window of data closest to the current time point. The key question is how long should one make the rolling window length? Petkova and Zhang (2005) chose a 5 year rolling window while Fama and French (1993) chose a 30 year rolling window. As part of their comparison of equity risk premium theories, Welch and Goyal (2008) use an expanding rolling window: at each time point t, they use all data up to and including time point t. While *explicit* data truncation via rolling window estimation is very frequently employed, *implicit* data truncation may be at least as prevalent, by pre-specifying the date range over which analysis will be performed. We seek a more principled approach to addressing non-stationarity in time series without relying on *ad hoc* decisions of how to truncate the data.

In the general approach to time-ordered data, a practitioner has chosen a model $p(y_t|\boldsymbol{\theta})$ that links the observed data $\mathbf{y}_{1:\mathrm{T}} = (y_1, \cdots, y_\mathrm{T})$ to underlying parameters $\boldsymbol{\theta}$. The practitioner may also have prior beliefs summarized in the prior distribution $p_0(\boldsymbol{\theta})$. The simplest Bayesian approach to modeling $\mathbf{y}_{1:\mathrm{T}}$ would be to assume that $\boldsymbol{\theta}$ is stationary over time and estimate the posterior distribution assuming the observed y_t's are exchangeable,

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}) \propto \prod_{t=1}^{\mathrm{T}} p(y_t|\boldsymbol{\theta})\, p_0(\boldsymbol{\theta}). \tag{1.2}$$

However, as we discussed above, stationarity is not always a reasonable assumption and so we need to allow for the underlying parameters of the model to evolve over time, i.e. $\boldsymbol{\theta}_{1:\mathrm{T}} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_{\mathrm{T}-1}, \boldsymbol{\theta}_\mathrm{T})$.

A standard Bayesian approach to non-stationarity specifies an additional level of the model for this parameter evolution (i.e. $\boldsymbol{\theta}_t$ given $\boldsymbol{\theta}_{1:t-1}$) such as the dynamic state-space model (West and Harrison, 1998). In addition to these extra modeling decisions, implementation is much more involved since the posterior distribution for an entire time-varying series of parameters,

$$p(\boldsymbol{\theta}_{1:\mathrm{T}}|\mathbf{y}_{1:\mathrm{T}}) \propto \prod_{t=1}^{\mathrm{T}} p(y_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{1:t-1})\, p_0(\boldsymbol{\theta}). \tag{1.3}$$

must be estimated. Under the simplifying assumption of normality, Carter and Kohn (1994) outline a Markov Chain Monte Carlo implementation for estimating the posterior distribution of a dynamic state-space model. More recent work has offered implementations for more complicated dynamic state-space models (Paez and Gamerman, 2013). However, all of these modeling approaches are inherently complicated (and usually computationally intensive) because the entire time-varying parameter vector $\boldsymbol{\theta}_{1:\mathrm{T}}$ must be estimated.

In contrast, we propose an alternative *power-weighted densities* (PWD) approach that avoids the direct specification of an evolution model for the parameter vector $\boldsymbol{\theta}_{1:\mathrm{T}}$. We leave the practitioners' chosen model as is, but differentially weight the contribution of individual observations to the likelihood function, so that more recent observations are more informative in the posterior distribution of the parameters at the current time point. Specifically, as we will see in Section 2,

$$p_\alpha(\boldsymbol{\theta}_\mathrm{T}|\mathbf{y}_{1:\mathrm{T}}) \propto p_0(\boldsymbol{\theta}_\mathrm{T}) \prod_{i=0}^{\mathrm{T}-1} p(y_{\mathrm{T}-i}|\boldsymbol{\theta}_\mathrm{T})^{\alpha_i}, \quad \alpha \in [0,1], \tag{1.4}$$

where α_i are weights placed on the lagged observations $y_{\mathrm{T}-i}$ away from the current time point T. These weights are estimated from the data in order to optimize the one step ahead predictive likelihood of the observed data.

Our PWD approach leaves intact the basic form of the model, $p(y_t|\boldsymbol{\theta})$ and $p_0(\boldsymbol{\theta})$, which makes our approach complementary to whatever data model is preferred by the practitioner. In contrast with the dynamic state-space model (1.3), our PWD approach does not require estimation of the entire parameter vector $\boldsymbol{\theta}_{1:\mathrm{T}}$ in order to infer the posterior distribution of the terminal time-point or to make predictions of future time points, which is the primary objective of our study.

As we will see in Section 2, rolling windows correspond to a specific set of lag-dependent PWD weights. While rolling windows also leave the choice of the data model up to the practitioner, we will see that the performance of rolling window approaches can be erratic in practice. In contrast, our PWD approach avoids the pre-specification of a fixed window length by differentially down-weighting *all* previous observations to optimize the predictive likelihood of the observed data. We present the details of our general power-weighted densities approach to time series data and compare our approach to state space models and other time series methods in in Section 2.

Our financial application consists of time series for 30 separate stock portfolios, which motivates extending our PWD approach to a hierarchical linear

regression model in Section 2.3. This extension permits sharing of information between the Fama and French (1993) three-factor models (1.1) for each stock portfolio while addressing non-stationarity within each stock portfolio time series. A hierarchical model is motivated by the central tendency of the market beta for a large number of stocks, often referred to as 'beta decay' by financial practitioners.

As much has been written about model uncertainty in stock return prediction, we will also incorporate uncertainty about our model choices by outlining a Bayesian Model Averaging ('BMA') extension of our PWD approach in Section 2.4. Our general PWD methodology for time series will be made available via a R package on CRAN.

In Section 3, we compare the operating characteristics of our PWD approach for hierarchical linear regression to alternative methods in synthetic data settings that mimic aspects of our financial data. In Section 4, we apply our PWD approach to hierarchical linear regression to the monthly returns of 30 industry portfolios (Kenneth French). In both real and synthetic data, our PWD approach performs significantly better in terms of predictive accuracy than models that assume stationarity in the underlying parameters, as well as competing non-stationary approaches such as dynamic state-space models and rolling windows. We will also demonstrate the computational convenience of our PWD approach.

There are a number of substantive implications of our results for financial practitioners. First, our results suggest a considerable amount of variation over time in the sensitivity of industry portfolio returns, particularly in the time periods around 1960 and 2000. Second, we observe a 'self-fulfilling prophecy' effect: the publication/acceptance of the importance of a market factor is followed by an increase in the importance of that market factor for prediction.

2. Power Weighted Densities for Time Series Data. The idea of differentially weighting historical data has been explored previously. Ibrahim and Chen (2000) introduced "power priors" as a way to integrate historical data with more recent data. Denoting the historical data by H, current data by $\mathbf{y}$, parameters of interest by $\boldsymbol{\theta}$ and a fixed power $\alpha \in [0, 1]$, the posterior distribution from their power prior model is

$$p(\boldsymbol{\theta}|\mathbf{y}, H, \alpha) \propto p(\mathbf{y}|\boldsymbol{\theta})p(H|\boldsymbol{\theta})^{\alpha}p(\boldsymbol{\theta}) \tag{2.1}$$

By setting $\alpha = 1$, the historical data is exchangeable with the current data, while $\alpha = 0$ implies the historical data is not used at all. Power priors have been applied in several clinical and epidemiological studies, including

Berry and Stangl (1996), Berry et al. (2010), Hobbs et al. (2011) and Tan et al. (2002). Brian (2010) applied power priors to pediatric quality of care evaluation.

In this paper, we are extending the power prior idea of Ibrahim and Chen (2000) to the modeling of time-ordered observations,

$$\mathbf{y} \stackrel{\Delta}{=} \mathbf{y}_{1:\mathrm{T}} = (y_1, y_2, \cdots, y_{\mathrm{T}-1}, y_{\mathrm{T}})$$

motivated by the assumption that older data may not be as relevant as more recent data when predicting future time series outcomes. We estimate the posterior distribution for $\boldsymbol{\theta}_{\mathrm{T}}$ at terminal time point T by raising the densities of each observation y_t to a different power,

$$p_\alpha(\boldsymbol{\theta}_{\mathrm{T}}|\mathbf{y}_{1:\mathrm{T}}) \propto p_0(\boldsymbol{\theta}_{\mathrm{T}}) \prod_{i=0}^{\mathrm{T}-1} p(y_{\mathrm{T}-i}|\boldsymbol{\theta}_{\mathrm{T}})^{\alpha_i}, \quad \alpha_i \in [0,1]. \tag{2.2}$$

which extends the power prior idea to place a lag-specific weight α_i on each i-th lagged historical data point away from the current time point. We still encode any prior beliefs we have regarding $\boldsymbol{\theta}_{\mathrm{T}}$ through the prior $p_0(\boldsymbol{\theta}_{\mathrm{T}})$.

The density (2.2) uniquely minimizes the convex sum of Kullback-Leibler divergences over a T-simplex representing all possible poolings of the historical data (further details in Supplement A). The popular *rolling window* strategy for model estimation in the financial literature corresponds to a special case of our PWD weights, where $\alpha_i = 1$ if $i < \tau$ and $\alpha_i = 0$ otherwise, with stopping time τ being pre-specified by the practitioner.

By avoiding the estimation of the entire time-series of underlying parameters $\boldsymbol{\theta}_{1:\mathrm{T}}$, our PWD approach should be less computationally intensive than the usual dynamic state-space model, but only if the extra weight parameters α_i can be estimated efficiently. We simplify this estimation task by imposing additional structure on the weight parameters.

Throughout this paper, we will restrict our weight parameters to an *exponentially-decreasing* function, $\alpha_i = \alpha^i$ of the lag i, parameterized by a single weight parameter $\alpha \in [0,1]$. Under this constraint, our *power-weighted densities* (PWD) posterior distribution for $\boldsymbol{\theta}_{\mathrm{T}}$ at current time point T becomes

$$p_\alpha(\boldsymbol{\theta}_{\mathrm{T}}|\mathbf{y}_{1:\mathrm{T}}) \propto p_0(\boldsymbol{\theta}_{\mathrm{T}}) \prod_{i=0}^{\mathrm{T}-1} p(y_{\mathrm{T}-i}|\boldsymbol{\theta}_{\mathrm{T}})^{\alpha^i}, \quad \alpha \in [0,1], \tag{2.3}$$

with a single weight parameter α that will be estimated from the data. This exponentially-decreasing regime of weights imposes a monotonicity constraint $\alpha_i \geq \alpha_{i+1}$ so that with $\alpha \in [0,1]$, more recent observations (those

with smaller lags i away from the current time point) have increased relevance relative to more distant observations.

There are many alternatives to our exponentially-decreasing weight regime, with the most obvious alternative being linearly-decreasing weights - we show in Supplement A that linearly decaying weights also perform well in practice. The exponentially-decreasing regime has the advantage of leading to simple posterior and posterior predictive distributions when used with exponential family likelihoods.

As an illustrative example, consider a single time series $\mathbf{y}_{1:\mathrm{T}}$ that is normally distributed, $y_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$, with unknown and possibly non-stationary mean μ_t and variance σ_t^2. We employ the prior $p(\mu_t, \sigma_t^2) \propto \sigma_t^{-2}$ suggested by Gelman et al. (2003) (p. 74). Combining this data and prior model with our exponentially-weighted PWD approach (2.3) gives the conditional posterior distribution for the terminal mean,

$$\mu_{\mathrm{T}} \,|\, \mathbf{y}, \alpha, \sigma_{\mathrm{T}}^2 \quad \sim \quad \mathcal{N}\Big(\hat{y}_{\alpha,\mathrm{T}}\,,\, \frac{\sigma_{\mathrm{T}}^2}{\mathrm{T}_\alpha}\Big), \tag{2.4}$$

and the marginal posterior distribution for the terminal variance,

$$\sigma_{\mathrm{T}}^2 \,|\, \mathbf{y}, \alpha \quad \sim \quad \mathrm{InvGamma}\Big(\frac{\mathrm{T}_\alpha - 1}{2}\,,\, \frac{\mathrm{T}_\alpha}{2}(\widehat{y_{\alpha,\mathrm{T}}^2} - \hat{y}_{\alpha,T}^2)\Big) \tag{2.5}$$

where

$$\mathrm{T}_\alpha = \sum_{i=0}^{\mathrm{T}-1} \alpha^i\,, \quad \hat{y}_{\alpha,\mathrm{T}} = \frac{\sum_{i=0}^{\mathrm{T}-1} \alpha^i y_{\mathrm{T}-i}}{\mathrm{T}_\alpha}\,, \quad \text{and} \quad \widehat{y_{\alpha,\mathrm{T}}^2} = \frac{\sum_{i=0}^{\mathrm{T}-1} \alpha^i y_{\mathrm{T}-i}^2}{\mathrm{T}_\alpha}.$$

The posterior distribution for μ_{T} is centered at $\hat{y}_{\alpha,\mathrm{T}}$, the exponentially weighted moving average (EWMA) of the observations $\mathbf{y}_{1:\mathrm{T}}$, which is a common estimator used by practitioners to accommodate non-stationary data. We interpret T_α as the "scaled count" of the number of observations in $\mathbf{y}_{1:\mathrm{T}}$, scaled by the weighting parameter α.

With prediction as our primary goal, the posterior predictive distribution of future observation $y^\star$ under our PWD approach is

$$y^\star \,|\, \mathbf{y}_{1:\mathrm{T}}, \alpha \quad \sim \quad t_{\mathrm{T}_\alpha - 1}\left(\hat{y}_{\alpha,\mathrm{T}}\,,\, \frac{\mathrm{T}_\alpha + 1}{\mathrm{T}_\alpha} S_{\alpha,\mathrm{T}}\right) \tag{2.6}$$

where

$$S_{\alpha,\mathrm{T}} = \frac{\mathrm{T}_\alpha}{\mathrm{T}_\alpha - 1}\left(\widehat{y_{\alpha,\mathrm{T}}^2} - \hat{y}_{\alpha,\mathrm{T}}^2\right). \tag{2.7}$$

The posterior distributions (2.4)-(2.6) reduce to the standard posterior distributions for a stationary model when $\alpha = 1$ whereas when $\alpha < 1$, data far in the past will be less relevant to the terminal time point and prediction of future observations.

The posterior predictive distribution (2.6) has a very simple form that can be used to make predictions of future data $y^\star$ while avoiding the need to estimate the non-stationarity in the underlying parameters μ_t and σ_t^2 directly. These results are conditioned on a known value of the weighting parameter α but in Section 2.2 we will discuss strategies for estimating α from the data.

2.1. *Related Time Series Approaches.* Our PWD approach for a normally distributed time series, $y_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$, closely mimics the first order state space model of West and Harrison (1998),

$$\begin{aligned} y_t &= \theta_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, V) \\ \theta_t &= \theta_{t-1} + \omega_t, \quad \omega_t \sim \mathcal{N}(0, W_t), \end{aligned}$$

with observation variance V constant over time but state variance W_t varying over time. This state space model is estimated through recursive equations culminating in a normal posterior distribution for the terminal mean, $\theta_\mathrm{T}|\mathbf{y}_{1:\mathrm{T}} \sim \mathcal{N}(m_\mathrm{T}, C_\mathrm{T})$ with

$$m_\mathrm{T} = m_{\mathrm{T}-1} + \frac{C_{\mathrm{T}-1} + W_\mathrm{T}}{C_{\mathrm{T}-1} + W_\mathrm{T} + V}(y_\mathrm{T} - m_{\mathrm{T}-1}) \quad \text{and} \quad C_\mathrm{T} = \frac{C_{\mathrm{T}-1} + W_\mathrm{T}}{C_{\mathrm{T}-1} + W_\mathrm{T} + V}$$

West and Harrison (1998) also provide a *discounted* alternative to their model with *discount factor* $\delta \in [0, 1]$ that downweights more distant observations in the time series by inflating the posterior variance of θ_t at each time step t,

$$m_\mathrm{T} = \frac{\sum_{i=0}^{\mathrm{T}-1} y_{\mathrm{T}-i}\, \delta^i}{\sum_{i=0}^{\mathrm{T}-1} \delta^i} \quad \text{and} \quad C_\mathrm{T} = \frac{V}{\sum_{i=0}^{\mathrm{T}-1} \delta^i}, \tag{2.8}$$

which are equivalent to our power-weighted densities (PWD) approach in (2.4)-(2.5). However, this equivalence is specific to normally distributed data and does not hold for the more general PWD approach in (2.2).

There is a similar connection between dynamic state space models and rolling windows approaches if the discount factor δ is allowed to vary over time in a lag-specific way with the following values,

$$\{\delta_1, \delta_2, \cdots, \delta_\mathrm{T}\} = (\underbrace{0, 0, \cdots, 0}_{\mathrm{T}-p}, \underbrace{1, 1, \cdots, 1}_{p}).$$

This representation highlights two issues with rolling windows. It is difficult to interpret rolling windows as a data generating process, since the normal model with a rolling window of length p implies a posterior distribution for θ_t with infinite variance at all time points $t \in \{1, 2, \ldots, \mathrm{T} - p\}$. It is also not clear how to estimate the optimal length p of the rolling window.

We will see superior predictive performance of our PWD approach over discounted state space models and rolling windows in our stock market analysis in Section 4. That said, we can still borrow insight from the discounted state space model of West and Harrison (1998) in terms of their estimation of the discount factor δ. In particular, they select δ which maximizes the one-step-ahead predictive likelihood of the data, and in Section 2.2, we will employ a similar strategy for the estimation of our weight parameter α.

Our PWD approach for a normally distributed time series also bears similarity to the exponentially weighted moving average model (EWMA), also known as an autoregressive integrated moving average process, ARIMA (0,1,1), in which the first differences of the data are modeled as

$$y_t - y_{t-1} = \epsilon_t + \rho\,\epsilon_{t-1} \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \quad \text{and} \quad \rho \in (-1, 1) \tag{2.9}$$

Recursively applying equation 2.9 and letting $\rho \equiv -\alpha$ gives

$$y_t = \epsilon_t + (1 - \alpha) \sum_{i=0}^{t-2} \alpha^i y_{t-i-1}, \tag{2.10}$$

which has a similar mean for y_t as the provided by our PWD approach in equation 2.4. The α parameter is estimated via (in-sample) maximum likelihood estimation in the usual EWMA procedure, whereas in Section 2.2, we propose estimating α by maximizing the *one-step-ahead predictive likelihood* of the data. We will show substantial gains in terms of accuracy and computational cost of our PWD approach compared to EWMA in synthetic settings in Section 3.1. In addition, our PWD approach generalizes more naturally to hierarchical linear regression (Section 2.3) which is needed for our financial application as well as allowing for other decay specifications (such as rolling windows and linearly decaying weights).

Smith (1979) and Smith (1981) introduce a Power Steady Model (PSM) which produces posterior distributions similar to our PWD approach for a general class of likelihoods with exponentially decaying weights. Grunwald, Raftery and Guttorp (1993) extend Smith (1981)'s framework to data which is conditionally Dirichlet-distributed. However, in this approach both the likelihood and the prior distribution are power-weighted, whereas our PWD approach only power-weights the likelihood term. It is also not clear how to extend this PSM model to non-exponential decays or hierarchical models.

Chen and Singpurwalla (1994) create a state-space model for data with a Gamma likelihood that includes a parameter for discounting older data in an exponential manner. Shephard (1994) derives state-space models with normal or exponential likelihoods where a scale parameter evolves over time. Both of these approaches are distribution-specific and the entire evolution of the state variable is estimated, whereas our PWD approach is intended as a fast and simple alternative to full state-space estimation when the goal is out-of-sample prediction.

2.2. *Estimation of Weight Parameter α.* Our estimation method for the weighting parameter α of our power-weighted densities approach mimics a method proposed by West and Harrison (1998) (p. 58) for their local level state-space model. We select the value $\alpha^\star$ that maximizes the *one-step-ahead predictive likelihood*,

$$\alpha^\star \;=\; \underset{\alpha}{\operatorname{argmax}}\; p^\star(\alpha \,|\, \mathbf{y}) \;\stackrel{\Delta}{=}\; \underset{\alpha}{\operatorname{argmax}}\; p_0(\alpha) \prod_{t=2}^{\mathrm{T}} p(y_t \,|\, \mathbf{y}_{1:t-1}, \alpha) \tag{2.11}$$

with $p(y_t \,|\, \mathbf{y}_{1:t-1}, \alpha)$ being the one-step-ahead predictive densities,

$$p(y_t \,|\, \mathbf{y}_{1:t-1}, \alpha) = \int p(y_t \,|\, \boldsymbol{\theta}_t)\, p_\alpha(\boldsymbol{\theta}_t \,|\, \mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_t)\, d\boldsymbol{\theta}_t \tag{2.12}$$

based on the power-weighted densities $p_\alpha(\boldsymbol{\theta}_t \,|\, \mathbf{y}_{1:t-1})$ from (2.3). This procedure is consistent with our primary goal: prediction of the next time point. The maximal value $\alpha^\star$ can be found with minor computational cost by a grid evaluation of $p^\star(\alpha \,|\, \mathbf{y}_{1:\mathrm{T}})$ over $\alpha \in [0, 1]$, though it is often easier to maximize the logarithm of (2.11) instead.

Note that the predictive likelihood (2.11) includes a prior distribution $p_0(\alpha)$ that can reflect any prior beliefs that a practitioner may have about the relative probability of particular values of α. In this paper, we will assume that all values of α are equally likely *a priori*.

Our predictive likelihood approach is related to the model selection procedure of Gelfand and Dey (1994). Assuming all models in a set of candidate models are equally likely *a priori*, they propose selecting the model with the best C-fold cross-validated out-of-sample *forecasting accuracy*. This strategy is also similar to the prequential approach of Dawid (1992) where preference is given to estimators with the smallest predictive loss.

For the illustrative normal model estimated by (2.4)-(2.6), we select the

$\alpha^\star$ that maximizes

$$\log p^\star(\alpha \,|\, \mathbf{y}) \;=\; \log p_0(\alpha) + \sum_{t=2}^{\mathrm{T}-1} \log \Gamma\left(\frac{t_\alpha + 1}{t_\alpha}\right) - \frac{1}{2}\Big(\log(t_\alpha + 1) + \log S_{\alpha,t}\Big)$$

$$-\Big(\frac{t_\alpha + 1}{2}\Big) \log\Big(1 + \frac{(y_{t+1} - \hat{y}_{\alpha,t+1})^2}{(t_\alpha + 1) S_{\alpha,t}}\Big), \tag{2.13}$$

where $t_\alpha = \sum_{i=0}^{t-1} \alpha^i$, $\hat{y}_{\alpha,t} = \sum_{i=0}^{t-1} y_{t-i}\, \alpha^i / t_\alpha$ and

$$S_{\alpha,t} = \frac{t_\alpha}{t_\alpha - 1}\left(\widehat{y^2_{\alpha,t}} - \hat{y}^2_{\alpha,t+1}\right) \quad \text{with} \quad \widehat{y^2_{\alpha,t}} = \frac{\sum_{i=0}^{t-1} \alpha^i y^2_{t-i}}{t_\alpha}$$

While the computation required for equation 2.13 may seem daunting, we show in Supplement A that evaluation of this expression scales linearly with the length of the time series. We will see in Section 3.1 that this linear time algorithm has computing times which are 5 to over 10 times faster than built-in `R` functions exponential weighted moving average and state space implementations. We will provide an `R` package for our PWD approach so that practitioners may benefit from our fast implementation.

One could also consider a fully Bayesian approach where we obtain posterior samples of α which would allow us to summarize the posterior variability in the weight parameter. However, the estimated posterior distribution of α tends to favor $\alpha \to 0$ since small values of α correspond to individual parameters $\boldsymbol{\theta}_t$ for each observation y_t, since there is no penalty for over-parameterization when fitting the entire time series *in sample* through the posterior distribution. For this reason, we prefer our one-step-ahead predictive likelihood approach (2.11), since its out-of-sample nature inherently protects against over-parameterization. If desired, we still can incorporate the variability in our weight parameter by instead sampling α from our one-step-ahead predictive likelihood $p^\star(\alpha \,|\, \mathbf{y})$. In Supplement A, we present a simulation study that suggests a sampling approach for α does not lead to better predictive performance than using the point estimate $\alpha^\star$ from (2.11).

2.3. *Power Weighted Densities for Hierarchical Linear Regression.* In this section, we extend our power-weighted densities approach for a hierarchical linear regression model, which is necessary for our application to monthly industry portfolio returns in Section 4. For that analysis, we need to model multiple time series each with potentially differing degrees of non-stationarity, while sharing information hierarchically across the multiple stock portfolios.

We consider the general setting of J different time series with outcome $y_{j,t}$ and p covariates $\mathbf{X}_{j,t}$ at each time point t in group j. We specify a separate regression model for each group j,

$$y_{j,t} = \mathbf{X}_{j,t}\,\boldsymbol{\beta}_{j,t} + \epsilon_{j,t}, \quad \epsilon_{j,t} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{j,t}) \tag{2.14}$$

with time varying coefficients $\boldsymbol{\beta}_{j,t}$ and residual variances $\sigma^2_{j,t}$. We share information across groups via a common prior distribution,

$$\boldsymbol{\beta}_{j,t} \sim \mathcal{N}_p(\boldsymbol{\beta}_0, \Sigma_0), \tag{2.15}$$

where Σ_0 is a diagonal matrix with diagonal entries $\boldsymbol{\tau}^2$. Note that by using a diagonal matrix Σ_0, we are assuming *a priori* independence of the components of $\boldsymbol{\beta}_{j,t}$, but this still allows for *a posteriori* dependence. We use non-informative prior distributions $p(\beta_{0,k}, \tau^2_k) \propto (\tau^2_k)^{-1/2}$ for our global parameters and $p(\sigma^2_{t,j}) \propto (\sigma^2_{t,j})^{-1}$ for the residual variances.

We can implement this hierarchical linear regression model using the Gibbs sampler (Geman and Geman (1984)). Denoting $\boldsymbol{\theta}_{-a}$ as all parameters excluding a, the conditional distributions of the global parameters for each covariate $k = 1, 2, \ldots, p$ are

$$\begin{aligned}
\beta_{0,k} | \,\boldsymbol{\theta}_{-\beta_{0,k}}, \mathbf{y} &\sim \mathcal{N}\left(\frac{\sum_{j=1}^J \beta_{j,k}}{J}\,,\, \frac{\tau^2_k}{J}\right), \\
\tau^2_k | \,\boldsymbol{\theta}_{-\tau^2_k}, \mathbf{y} &\sim \text{InvGamma}\left(\frac{J}{2}\,,\, \frac{1}{2}\sum_{j=1}^J (\beta_{j,k} - \beta_{0,k})^2\right)
\end{aligned} \tag{2.16}$$

If our hierarchical regression model was assumed to be *stationary* (i.e. $\boldsymbol{\beta}_{j,t} = \boldsymbol{\beta}_j$ and $\sigma^2_{j,t} = \sigma^2_j$), we would have the following conditional distributions for the group-specific parameters,

$$\begin{aligned}
\boldsymbol{\beta}_j | \,\boldsymbol{\theta}_{-\boldsymbol{\beta}_j}, \mathbf{y} &\sim \mathcal{N}_p(\hat{\boldsymbol{\beta}}_j\,,\, \hat{V}_j) \\
\sigma^2_j | \,\boldsymbol{\theta}_{-\sigma^2_j}, \mathbf{y} &\sim \text{InvGamma}\left(\frac{\mathrm{T}}{2}\,,\, \frac{1}{2}\sum_{i=1}^{\mathrm{T}} (y_{j,i} - \mathbf{X}_{j,i}\boldsymbol{\beta}_j)^2\right)
\end{aligned} \tag{2.17}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_j &= \hat{V}_j\left((\sigma^2_j)^{-1}\mathbf{X}'_{j,1:\mathrm{T}}\mathbf{y}_{j,1:\mathrm{T}} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right) \quad \text{and} \\
\hat{V}_j &= \left((\sigma^2_j)^{-1}\mathbf{X}'_{j,1:\mathrm{T}}\mathbf{X}_{j,1:\mathrm{T}})^{-1} + \Sigma_0^{-1}\right)^{-1}.
\end{aligned}$$

However, in our financial application (and for many other time series), the assumption of stationary in the group-specific parameters is not realistic.

Rather, we can use our exponentially-decreasing PWD approach (2.3) to address potential non-stationarity in our model parameters,

$$p_\alpha(\boldsymbol{\theta}_{j,\mathrm{T}} \,|\, \mathbf{y}_{j,1:\mathrm{T}}) \;\propto\; p_0(\boldsymbol{\theta}_{\mathrm{T}}) \prod_{i=0}^{\mathrm{T}-1} p(y_{j,\mathrm{T}-i}|\boldsymbol{\theta}_{j,\mathrm{T}})^{\alpha_j^i} \qquad \alpha_j \in [0,1] \tag{2.18}$$

where by using different weight parameters α_j we allow for differing degrees of non-stationarity in each time series j. Under this PWD approach, the conditional distributions of the *time-varying* group-specific parameters at terminal time point T are

$$\begin{aligned}
\boldsymbol{\beta}_{j,\mathrm{T}}|\, \boldsymbol{\theta}_{-\boldsymbol{\beta}_{j,\mathrm{T}}}, \mathbf{y} &\sim \mathcal{N}_p(\hat{\boldsymbol{\beta}}_{\alpha,j}\,,\, \hat{V}_{\alpha,j}) \\
\sigma^2_{j,\mathrm{T}}|\, \boldsymbol{\theta}_{-\sigma^2_{j,\mathrm{T}}}, \mathbf{y} &\sim \mathrm{InvGamma}\left(\frac{\mathrm{T}_{\alpha_j}}{2}\,,\, \frac{1}{2}\sum_{i=0}^{\mathrm{T}-1} \alpha_j^i (y_{j,\mathrm{T}-i} - \mathbf{X}_{j,T-i}\boldsymbol{\beta}_{j,\mathrm{T}})^2\right)
\end{aligned}$$

where

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\alpha,j} &= \hat{V}_{\alpha,j}\left((\sigma^2_{j,\mathrm{T}})^{-1}\mathbf{X}'_{j,1:\mathrm{T}}\,\mathbf{A}_{j,\mathrm{T}}\,\mathbf{y}_{j,1:\mathrm{T}} + \Sigma_0^{-1}\boldsymbol{\beta}_0\right) \quad \text{and} \\
\hat{V}_{\alpha,j} &= \left((\sigma^2_{j,T})^{-1}\mathbf{X}'_{j,1:\mathrm{T}}\,\mathbf{A}_{j,\mathrm{T}}\,\mathbf{X}_{j,1:\mathrm{T}} + \Sigma_0^{-1}\right)^{-1}
\end{aligned}$$

with weighting matrix $\mathbf{A}_{j,\mathrm{T}} \triangleq diag(1, \alpha_j, \alpha_j^2, \cdots, \alpha_j^{\mathrm{T}-1})$ and $\mathrm{T}_{\alpha_j} = \sum_{i=0}^{\mathrm{T}-1} \alpha_j^i$.

Comparing to the stationary model (2.17), our PWD approach acts through the weight matrix $\mathbf{A}_{j,\mathrm{T}}$ to downweight observations that are farther away from terminal time point T. The global parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\tau}^2$ can still be sampled using (2.16).

The model implementation above is conditional upon knowing the weight parameters α_j for each group. Our usual estimation procedure for the weight parameters (Section 2.2) would be to select the α_j which maximizes the one step ahead predictive likelihood for each group j:

$$\alpha_j^\star = \underset{\alpha_j}{\mathrm{argmax}}\; p_0(\alpha_j) \prod_{t=2}^{\mathrm{T}} p_{\alpha_j}(y_{j,t} \,|\, \mathbf{y}_{j,1:t-1})$$

This requires the evaluation of each one-step-ahead predictive density

$$p_{\alpha_j}(y_{j,t} \,|\, \mathbf{y}_{j,1:t-1}) = \int p(y_{j,t} \,|\, \boldsymbol{\theta}_{j,t})\, p_{\alpha_j}(\boldsymbol{\theta}_{j,t} \,|\, \mathbf{y}_{j,1:t-1})\, p_0(\boldsymbol{\theta}_{j,t})\, d\boldsymbol{\theta}_{j,t} \tag{2.19}$$

at each time point t by integrating over posterior samples of $\boldsymbol{\theta}_{j,t}$, which becomes computationally intensive if there are many groups J.

For that reason, we prefer the following approximate approach based on *plug-in estimators* of $\boldsymbol{\theta}$ which is very fast and performs well in practice. Specifically, we estimate each α_j as

$$\alpha_j^\star = \underset{\alpha_j}{\text{argmax}}\ p_0(\alpha_j) \prod_{t=2}^{\mathrm{T}} p_{\alpha_j, approx}(y_{j,t} \,|\, \mathbf{y}_{j,1:t-1}, \widehat{\boldsymbol{\theta}}) \tag{2.20}$$

where $p_{\alpha_j, approx}(y_{j,t} \,|\, \mathbf{y}_{j,1:t-1}, \widehat{\boldsymbol{\theta}})$ is the predictive likelihood of $y_{j,t}$ using plug-in estimators of the model parameters. For the hierarchical linear regression model, this predictive likelihood is

$$y_{j,t} \sim t_{t_\alpha - p - 1}\left(\mathbf{X}_{j,t}\tilde{\beta}_{j,t}\,,\ \tilde{\sigma}^2_{j,t} + \tilde{V}_{j,t}\right) \tag{2.21}$$

where

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_{j,t} &= \tilde{V}_{j,t}\left((\tilde{\sigma}^2_{j,t})^{-1}\mathbf{X}'_{j,1:(t-1)}\,\mathbf{A}_{j,t-1}\,\mathbf{y}_{j,1:(t-1)} + \tilde{\Sigma}_0^{-1}\tilde{\boldsymbol{\beta}}_0\right), \\
\tilde{V}_{j,t} &= \left((\tilde{\sigma}^2_{j,t})^{-1}\mathbf{X}'_{j,1:(t-1)}\,\mathbf{A}_{j,t-1}\,\mathbf{X}_{j,1:(t-1)} + \tilde{\Sigma}_0^{-1}\right)^{-1}, \\
\tilde{\sigma}^2_{j,t} &= \left(\sum_{i=0}^{t-1}\alpha_j^i (y_{j,t-i} - \mathbf{X}_{j,t-i}\tilde{\beta}_{j,t})^2\right)/(\mathrm{T}_{\alpha_j} - p), \\
\tilde{\boldsymbol{\beta}}_0 &= \sum_{j=1}^{J}\tilde{\boldsymbol{\beta}}_{j,t}/J, \quad \text{and} \quad \tilde{\Sigma}_0 = \sum_{j=1}^{J}(\tilde{\boldsymbol{\beta}}_{j,t} - \tilde{\boldsymbol{\beta}}_0)^2/(J-1).
\end{aligned}$$

with $t_{\alpha_j} = \sum_{i=0}^{t-1}\alpha_j^i$ and weighting matrix $\mathbf{A}_{j,t-1} = diag(1, \alpha_j, \alpha_j^2, \cdots, \alpha_j^{t-1})$. Since each of the above plug-in estimators is a function of α_j, we must iterate between:

1. Updating the plug-in estimators $\tilde{\beta}_{j,t}$, $\tilde{\sigma}^2_{j,t}$, $\tilde{V}_{j,t}$, $\tilde{\boldsymbol{\beta}}_0$ and $\tilde{\Sigma}_0^{-1}$ based on the current estimate of α_j.
2. Optimizing α_j in (2.20) using the predictive likelihood (2.21) based on the updated values of the plug-in estimators.

In Supplement A, we show that evaluation of these expressions scale linearly with the length of the time series. We assume convergence is achieved when the change in any α_j falls below a pre-specified threshold (in practice, we set this to be .005). This plug-in method performs quite well in practice and has a much lower computation cost than the evaluation of integral (2.19) when estimating many group-specific α_j's.

2.4. *Bayesian Model Averaging with Power Weighted Densities.* When modeling time series data, there is often uncertainty over the correct model to use in addition to the issue of non-stationarity within a particular model. For example, in our financial application we consider the Fama and French (1993) three factor model, but other popular alternatives are using no factors (Welch and Goyal, 2008), the CAPM model, and the four factor model of Carhart (1997) which adds an fourth momentum ('MOM') factor. More generally, we may want to allow for any of the $2^4 = 16$ combinations of these four factors in our model for industry portfolios in Section 4, and incorporate uncertainty about our model choices into our predictions.

Bayesian model averaging ('BMA') is a popular way of allowing for model uncertainty (Kass and Raftery, 1995), where the posterior densities of model parameters $\boldsymbol{\theta}$, are weighted by the probability of each model, M_k ($k = 1, \ldots, K$),

$$P(\boldsymbol{\theta}|D_{1:\mathrm{T}}) = \sum_{k=1}^{K} P(\boldsymbol{\theta}|D_{1:\mathrm{T}}, M_k)P(M_k|D_{1:\mathrm{T}}). \tag{2.22}$$

with the weights proportionate to by the marginal likelihood of the data under each alternative model,

$$P(M_k|D_{1:\mathrm{T}}) = \frac{P(D_{1:\mathrm{T}}|M_k)}{\sum_{l=1}^{K} P(D_{1:\mathrm{T}}|M_l)}. \tag{2.23}$$

with $D_{1:\mathrm{T}}$ denoting the data available up to and including time point T.

We adopt a predictive likelihood-based analog to BMA to allow for model uncertainty within our PWD approach. Similar to how our PWD approach selects the value of α which maximizes the marginal one-step-ahead predictive likelihood of the observed data, our PWD-BMA approach weighs the posterior density of parameters $\boldsymbol{\theta}$ under each alternative models by their respective marginal one-step-ahead predictive likelihoods. In other words, instead of Equation (2.23), we use

$$P_\alpha(M_k|D_{1:\mathrm{T}}) = \frac{\prod_{t=2}^{\mathrm{T}} P(D_t|D_{1:t-1}, \alpha_k^\star, M_k)}{\sum_{l=1}^{K} \prod_{t=2}^{\mathrm{T}} P(D_t|D_{1:t-1}, \alpha_l^\star, M_l)}, \tag{2.24}$$

where $\alpha_k^\star$ maximizes the one-step-ahead marginal predictive likelihood of the data under model M_k:

$$\alpha_k^\star = \mathrm{argmax}_\alpha \prod_{t=2}^{\mathrm{T}} P(D_t|D_{1:t-1}, M_k). \tag{2.25}$$

BMA-based estimators have many favorable qualities (Raftery and Zheng, 2003) and tend to perform well in terms of out-of-sample performance (Madigan and Raftery, 1994; Hoeting, Raftery and Madigan, 2002). In the finance literature, Avramov (2002) shows that BMA improves predictive regression forecast errors. Rapach, Strauss and Zhou (2009) accommodates model uncertainty in a financial setting based upon Stock and Watson (2004), by combining models using weights which are a function of their previous forecasting ability but with a discount factor which assigns greater weight to more recent forecasting accuracy. Aiolfi and Timmermann (2006) also address model uncertainty in a predictive regression setting, but Rapach, Strauss and Zhou (2009) showed that their performance may be uneven when used to predict monthly equity returns. Dangl and Halling (2012) applied BMA to a state-space linear regression model and outperformed alternatives which do not allow for time-variation in the regression coefficients in a financial prediction setting.

In summary, our PWD approach to model uncertainty is a variation of BMA where we, as in Avramov (2002), weight each model by its predictive fitness, emphasizing more recent predictions more than older predictions which performed well in Rapach, Strauss and Zhou (2009). We implement our approach in our financial application to industry stock portfolios in Section 4, which leads to both favorable performance and several implications for the importance of the model factors over time.

3. Simulation Evaluation of our PWD approach. We use several synthetic data settings to evaluate the predictive and computational performance of our PWD approach relative to other methods. We first consider a "null" setting where the data are normally distributed with an underlying scalar mean that is *stationary* over time. We then consider a *non-stationary* hierarchical regression setting that emulates the characteristics of our financial application in Section 4. We also compare several variants of our PWD approach in simple non-stationary data settings in Supplement A.

3.1. *Stationary Normal Mean Setting.* While methods which allow for parameter evolution are expected to perform better when there is actual non-stationarity in those parameters, it is also important to evaluate performance of those methods when the underlying parameters are, in fact, stationary. In this stationary case, non-stationary methods may lose predictive accuracy and have a higher computational cost.

We generate synthetic data for a univariate time series of length $T = 500$,

where the true underlying mean of the time series is constant over time:

$$y_t = \beta + \epsilon_t \quad \text{where} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2). \tag{3.1}$$

We set the true mean $\beta = 2$ and variance $\sigma^2 = 1$. We generate 4,000 datasets under this setting and use the first $\mathrm{T}-1$ time points of each dataset (holding out the terminal observation y_T) to train the following models:

1. `Stationary`: Assume mean is stationary and predict y_T with the simple average of the first $\mathrm{T}-1$ time points of each dataset.
2. `PWD`: Predict y_T with the mean of the posterior predictive distribution from Equation 2.6 using $\alpha^\star$ that maximizes Equation 2.13.
3. `EWMA`: Use `R`'s `ARIMA` function within the `stats` package to fit an ARIMA (0,1,1) model. The prediction of y_T is an *exponentially weighted moving average* of the first $\mathrm{T}-1$ time points.
4. `State-Space`: Use `R`'s `StructTS` function within the `stats` package to fit a local level state space model via maximum likelihood. The prediction of y_T is the mean of the one-step-ahead predictive distribution.

In Table 1, we compare these four methods in terms of root mean square prediction error (RMSE) for the held-out terminal observation y_T , the standard error (SE) over datasets of the RMSE, and the mean computing time in milliseconds (Time (ms)).

	`Stationary`	`EWMA`	`PWD`	`State-Space`
RMSE	.045	.064	.054	.064
SE	.000	.001	.001	.001
Time (ms)	.01	6.13	1.14	11.52

TABLE 1
Comparison of Methods in Stationary Setting

Since the underlying mean is stationary in this setting, `Stationary` should have an advantage and this is indeed the case, with the RMSE for `Stationary` approximately 20% lower than `PWD`. However, `PWD` has an RMSE approximately 20% smaller than both `State-Space` and `EWMA`, which suggests that our PWD approach is not as easily misled compared to these other methods when the underlying data generating process is truly stationary. The small SEs suggest that all of these RMSE differences are statistically significant.

Moreover, `PWD` was approximately 5 times faster than `EWMA` and approximately 10 times faster than `State-Space`. This dramatic speedup is impressive given that the `arima` and `structTS` functions, as part of the `stats` package within `R`, have been optimized for speed. In Figure 1, we further emphasize the reduced computational cost of our PWD approach by plotting

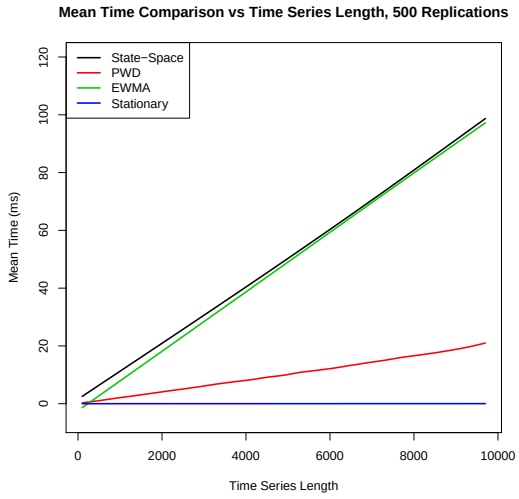

Fig 1: Mean Computing Time of 4 Models: `Stationary`, `PWD`, `EWMA` and `State-Space` as a function of time series length.

the mean computing time in milliseconds (averaged over 2000 replications) for the four methods as a function of time series length. All methods have computing times which scale linearly with time series length, but the slope associated with that linear scaling is much smaller for `PWD` compared to `EWMA` and `State-Space`.

3.2. *Non-Stationary Hierarchical Linear Regression Setting.* We generate synthetic data in a regression setting that represents a simplified version of our financial application in Section 4. Specifically, each synthetic dataset consists of a set of J portfolios where the return on each portfolio $y_{j,t}$ is a linear function of the return of the overall market m_t,

$$y_{j,t} = \beta_{j,t}\, m_t + \epsilon_{j,t} \qquad \epsilon_{j,t} \sim \mathcal{N}(0, \sigma^2) \tag{3.2}$$

with portfolio-specific sensitivities $\beta_{j,t}$ to the overall market that evolve over time t. This synthetic data model is analogous to the celebrated CAPM model (Fama and French, 1989). The market factor is generated as $m_{j,t} \sim \mathcal{N}(\mu_m, \sigma_m^2)$ where we set $\mu_m = .047$ and $\sigma_m^2 = .045^2$ based on historical monthly stock market data from Shiller (2014). The sensitivity of stocks to the market is *non-stationary* in that $\beta_{j,t}$ is centered upon its value from the prior period, $\beta_{j,t-1}$, plus a disturbance term,

$$\beta_{j,t} = \beta_{j,t-1} + \eta_{j,t}. \tag{3.3}$$

The evolution of $\beta_{j,t}$ is also *group mean reverting* in that the disturbance term $\eta_{j,t}$ pulls $\beta_{j,t}$ towards the group average of the prior period:

$$\eta_{j,t} = \phi_j(\bar{\beta}_{j,t-1} - \beta_{j,t-1}) + \zeta_{j,t}, \tag{3.4}$$

where ϕ_j represents the magnitude of stock j's mean reversion towards the overall group average, $\bar{\beta}_{t-1}$ is the group average β at time point $t-1$, and $\zeta_{j,t}$ is white noise:

$$\phi_j \sim \text{Beta}(a, b), \quad \bar{\beta}_{t-1} = \sum_{j=1}^{J} \beta_{j,t-1}/J, \quad \text{and} \quad \zeta_{j,t} \sim \mathcal{N}(0, \tau^2). \tag{3.5}$$

We set $\sigma^2 = .04^2$, $\tau^2 = .08^2$, $a = 3$ and $b = 97$ which leads to a strong correlation between portfolios and the overall market and meaningful evolution of $\beta_{j,t}$ over time, as well as mild mean reversion, shrinking the market sensitivity of portfolios towards the group average of the prior time point, consistent with the notion of "beta decay" amongst finance practitioners.

We examine two different data settings using this particular data generating process. `Setting 1` consists of a large number of groups ($J = 100$) that each contain a short time-series ($\text{T} = 10$). `Setting 2` consists of a small number of groups ($J = 10$) that each contain a long time-series ($\text{T} = 100$). `Setting 2` is more similar to our application to industry portfolios in Section 4, as that data contains a small number of relatively long time series.

We generated 500 synthetic datasets under both settings. For each approach that we consider, we train the model on the first $\text{T} - 1$ observations of each time series, $\mathbf{y}_{j,0:\text{T}-1}$, as well as the market return over that same time period, $\mathbf{m}_{0:\text{T}-1}$, and then predict the terminal observation, $y_{j,\text{T}}$ using the return on the market from the final time point, m_{T}. Performance of each method is judged based on the RMSE of that prediction.

In these evaluations, we consider two variants of our PWD approach that differ in the modeling of the weighting parameters and group-specific means: 1. `Hier-PWD` where we model all portfolios simultaneously using the hierarchical linear regression model outlined in Section 2.3, and `Sep-PWD` where we model each portfolio separately without any sharing between portfolios.

We compare these two PWD variants to three alternative approaches:

1. `Stationary:` estimate the parameters in (3.2) using standard OLS regression applied separately to each portfolio time series, assuming that the coefficients are stationary over time, i.e. $\beta_{j,t} = \beta_j$
2. `Stationary-Hier:` estimate the parameters in (3.2) simultaneously across portfolios using the hierarchical linear regression model (3.2)-(3.4), but still assume the coefficients are stationary over time, i.e. $\beta_{j,t} = \beta_j$

3. `State-space-LR:` estimate the coefficients $\beta_{j,t}$ in (3.2) using a local level dynamic linear regression model estimated via maximum likelihood (Petris, Petrone and Campagnoli, 2009).

Table 2 compares performance of our PWD variants to the three alternative approaches in `Setting 1` where we have a large number of groups ($J = 100$) that each contain a short time-series (T = 10). In this setting, there is limited data available within each portfolio time series to estimate non-stationary parameters, and so the hierarchical methods should benefit from borrowing strength between portfolios.

	Hier-PWD	Sep-PWD	State-Space-LR	Stationary	Stat-Hier
Mean(RMSE)	19.00	22.00	28.11	21.43	19.04
SE(RMSE)	0.18	0.32	0.82	0.30	0.18
t-test p-value		0.000	0.000	0.000	0.878

TABLE 2

Comparison of Methods in `Setting 1`*: Large Number of Short Time Series*

In Table 2, we evaluate each approach using the average RMSE of the terminal time point prediction across the 500 datasets, as well as the standard error of that average RMSE[1] . Observing that `Hier-PWD` had the best average RMSE, we also provide the p-value from a two-sided t-test (assuming unequal variances) of the difference between the RMSE of `Hier-PWD` and the RMSE of each method.

We see in Table 2 that `Hier-PWD` performed significantly better (at the 1% level) than all other methods except for `Stationary-Hier`. The fact that `Stationary-Hier` was the only method competitive with `Hier-PWD` suggests a benefit from sharing information across groups but perhaps not enough data within each group to benefit from allowing non-stationarity. We note the particularly poor performance of `State-Space-LR` in this data setting where we have a large number of short time series.

Table 3 compares performance of our PWD variants to the three alternatives in `Setting 2` where we have a small number of groups ($J = 10$) that each contain a long time-series (T = 100), which more closely emulates our financial application in Section 4 where we have long time series for a relatively small number of portfolios.

Comparing between our two PWD variants, we see that the pooling induced by `Hier-PWD` did not lead to as much of a gain in predictive performance as seen in `Setting 1`. The situation of few groups with substantial amounts of data within each group limits the benefit of hierarchically sharing information between groups. In this long time series setting where there

[1]RMSE(Mean) and RMSE(SE) are re-scaled by a factor of 10^4.

	Hier-PWD	Sep-PWD	State-Space-LR	Stationary	Stat-Hier
Mean(RMSE)	18.52	19.14	18.83	20.67	20.52
SE(RMSE)	0.27	0.29	0.27	0.37	0.37
t-test p-value		0.114	0.412	0.000	0.000

TABLE 3
RMSE: Small Group Count, Long Time Series; 500 Datasets

is ample data for estimating the non-stationary evolution of the underlying $\beta_{j,t}$'s, we see that the stationary models `Stationary` and `Stat-Hier` performed poorly relative to the non-stationarity methods. Among the non-stationary methods, `Hier-PWD`, `Sep-PWD` and `State-Space-LR` did not have significant differences in their predictive accuracy.

Our evaluation of both `Setting 1` and `Setting 2` suggests our power-weighted densities approach is robust to different data conditions, and is especially beneficial in situations where information sharing between groups is important, as in financial markets. The `State-Space-LR` approach was less robust: it performed competitively in `Setting 2` but performed significantly worse in `Setting 1` where less data was available in each time series.

We also observed dramatic benefits of our PWD approach in terms of computational cost in both `Setting 1` and `Setting 2`. Comparing the variant of our PWD approach most similar to the state-space model, `Sep-PWD`'s average computing time was 20-40 times faster than `State-Space-LR`. `Sep-PWD` had an average computing time for all groups of 50 milliseconds in `Setting 1` (compared with 2099 milliseconds for `State-Space-LR`) and 340 milliseconds in `Setting 2` (compared with 7660 milliseconds for `State-Space-LR`). Indeed, our `Sep-PWD` variant may strike the best balance between computing speed and predictive accuracy for practitioners.

4. Application to Prediction of Industry Portfolios. The ability to accurately estimate the sensitivity of portfolios to market factors is very important to financial practitioners since it enables firms to more accurately 'hedge' or decrease risk through offsetting financial positions. Dynamic hedging forms the basis for the pricing of financial derivatives, and the expected cost of the dynamic replication of a financial derivative (as well as the variability) drives the cost that a financial institution will charge to sell that derivative (Wilmott, 1995), directly tied to the notion of basis risk (Figlewski, 1984). In this section, we apply our power-weighted densities (PWD) approach for hierarchical linear regression (Section 3.2) to estimate the sensitivity of industry stock portfolios to market factors over time, and compare with several alternative methods.

Our data consists of 49 stock portfolios formed based upon industry, avail-

able on Kenneth French's website (Kenneth French). Of those 49 industries, we restricted our attention to the portfolios with the longest time series: there are 30 industry portfolios with monthly data starting December 1932 and running through December 2014. Using monthly data is the general convention in the CAPM and factor model literature (e.g. Fama and French (1989) and Lewellen and Nagel (2006)).

In total, we have $J = 30$ stock portfolios and $\mathrm{T} = 985$ monthly time points per stock portfolio with no missing data over that period. This data provides us with a representative cross section of market returns for many different asset classes and is a similar setting to the "few groups of long time series" synthetic Setting 2 of Section 3.2.

The celebrated work of Fama and French (1993) predicted the return $y_{j,t}$ on a stock portfolio j at a time t with a linear three factor model,

$$y_{j,t} = \alpha_{j,t} + \beta^m_{j,t} \cdot m_t + \beta^s_{j,t} \cdot s_t + \beta^v_{j,t} \cdot v_t + \epsilon_{j,t} \qquad \epsilon_{j,t} \sim \mathcal{N}(0, \sigma^2_{j,t}) \tag{4.1}$$

where m_t is the excess return on the market (MKT), s_t is the excess return of small capitalization stocks over large capitalization stocks (SMB), and v_t is the excess return of value stocks over growth stocks (HML). Compared to equation (1.1), we are now specifying normally-distributed errors and allowing for coefficients that are possibly time-varying (e.g. $\beta^m_{j,t}$ rather than β^m_j, etc). In the usual matrix notation, (4.1) is

$$y_{j,t} = \mathbf{X}_t \cdot \boldsymbol{\beta}_{j,t} + \epsilon_{j,t} \quad \text{where} \quad \epsilon_{j,t} \sim \mathcal{N}(0, \sigma^2_{j,t}) \tag{4.2}$$

with $\mathbf{X}_t = [1 \; m_t \; s_t \; v_t]$ and $\boldsymbol{\beta}_{j,t} = [\alpha_{j,t} \; \beta^m_{j,t} \; \beta^s_{j,t} \; \beta^v_{j,t}]$.

It is reasonable to believe that the $\boldsymbol{\beta}$'s for individual portfolios will have some central tendency, which suggests that sharing information across portfolios may be useful. We share information between our set of $J = 30$ portfolios through a global prior distribution at each time point,

$$\boldsymbol{\beta}_{j,t} \sim \mathcal{N}(\boldsymbol{\beta}_{0,t}, \Sigma_{0,t}) \qquad j = 1, \ldots, J \tag{4.3}$$

with $\boldsymbol{\beta}_{0,t} = [\alpha_{0,t} \; \beta^m_{0,t} \; \beta^s_{0,t} \; \beta^v_{0,t}]$ and $\Sigma_{0,t}$ being a diagonal matrix with diagonal elements $(\tau^2_{\alpha,t} \; \tau^2_{m,t} \; \tau^2_{s,t} \; \tau^2_{v,t})$. We use non-informative priors $p(\boldsymbol{\beta}_{0,t}, \Sigma_{0,t}) \propto (\tau^2_{\alpha,t} \, \tau^2_{m,t} \, \tau^2_{s,t} \, \tau^2_{v,t})^{-1/2}$ for the global parameters as well as $p(\sigma^2_{t,j}) \propto (\sigma^2_{t,j})^{-1}$ for the residual variances.

Even with a hierarchical structure on the parameters, this model is difficult to estimate unless we make a strong assumption of *stationarity* over time, i.e. $\boldsymbol{\beta}_{j,t} = \boldsymbol{\beta}_j$ for all $t = 1, \ldots, \mathrm{T}$. However, stationarity is not a reasonable assumption in most financial applications, and the standard approach

in the literature (e.g. Fama and French (1993)) is to estimate time-varying coefficients using a rolling window.

As an alternative to rolling windows, we will apply our power-weighted density (PWD) approach for hierarchical linear regression models (Section 2.3) to this set of 30 industry portfolio time series. Our PWD approach allows the regression coefficients $\boldsymbol{\beta}_{j,t}$ to evolve over time for each portfolio j but avoids estimating the entire parameter vector $\boldsymbol{\beta}_{j,1:\mathrm{T}}$ when constructing the posterior distribution for the terminal coefficients $\boldsymbol{\beta}_{j,\mathrm{T}}$ that are used to predict the future return $y_{j,\mathrm{T}+1}$. As outlined in Section 2.3, we estimate portfolio-specific weighting parameters α_j so that influence of past observations can vary between different portfolios.

We apply three variants of our PWD approach. The first two variants, `Hier-PWD` and `Sep-PWD`, were employed in our synthetic data evaluation in Section 3.2. We also consider a third variant, `Sep-PWD-BMA`, where we combine Bayesian model averaging with our PWD approach as described in Section 2.4. This BMA variant explores 16 different linear models that are the combinations of inclusion/exclusion of the three Fama-French factors and the extra momentum factor of Carhart (1997). We will evaluate the quality of these different models for each industry stock portfolio j at each time point t. The fast computational speed of our PWD approach greatly aids the practical implementation of `Sep-PWD-BMA`.

We will compare our three PWD variants to several alternative time series methods. Three of these alternatives were also evaluated in Section 3: `Stationary`, `Stationary-Hier` and `State-space-LR`. We will also evaluate a rolling window approach, `Window-5`, which estimates the coefficients $\boldsymbol{\beta}_{j,t}$ in (4.2) at each time point t with a standard OLS regression using only the 5 years prior to time point t, same as in Petkova and Zhang (2005). Rolling windows are the standard approach to non-stationarity in the financial literature (Welch and Goyal, 2008).

We will evaluate the predictive performance of each method by the rolling cumulative evaluation of their forecast errors, as done in Welch and Goyal (2008). Specifically, for a particular model M and a specific portfolio j up to time point t, we calculate the squared prediction error between the actual return at time $t+1$ and predicted return given all information up to time t,

$$\begin{aligned} SPE(M)_{j,t+1} &= (y_{j,t+1} - \hat{y}_{j,t+1})^2 \\ &= \left(y_{j,t+1} - \widehat{\alpha}_{j,t} - \widehat{\beta^m}_{j,t}\, m_{j,t+1} - \widehat{\beta^s}_{j,t}\, s_{j,t+1} - \widehat{\beta^v}_{j,t}\, v_{j,t+1}\right)^2 \end{aligned} \tag{4.4}$$

where $(\widehat{\alpha}_{j,t}, \widehat{\beta^m}_{j,t}, \widehat{\beta^s}_{j,t}, \widehat{\beta^v}_{j,t})$ are estimated by model M using all data up to time point t. We aggregate the squared prediction errors across all $J = 30$

stock portfolios to get the *cumulative sum of squared prediction errors* for a particular model M up to any time point t,

$$SSPE(M)_{1:t} = \sum_{i=1}^{t} \sum_{j=1}^{J} SPE(M)_{j,i}.$$

For each model, we evaluate this cumulative sum of squared prediction errors at each monthly time point, starting in November 1937 when all competing methods are able to provide predictions, and ending in December 2014. As in Welch and Goyal (2008), we select the `Stationary` model as a benchmark for our comparison since it represents the simplest approach to estimating the Fama and French (1993) three-factor model. Relative to the benchmark `Stationary` model, we can calculate the difference in our cumulative sum of squared prediction errors up to time point t,

$$\Delta SSPE(M, \texttt{Stationary})_{1:t} = SSPE(M)_{1:t} - SSPE(\texttt{Stationary})_{1:t}$$

In Figure 2, these differences in the cumulative sum of squared prediction errors (relative to `Stationary`) are plotted over time for our three PWD variants and our alternative models.

The most striking feature of Figure 2 is that the non-stationary methods (`Window-5`, `Sep-PWD`, `Hier-PWD` and `Sep-PWD-BMA`) show much better predictive performance than the baseline `Stationary` model, with rolling cumulative prediction errors $\Delta SSPE$ that grow increasingly negative over time. `State-space-LR` model shows the worst predictive performance among the non-stationary methods. `Stationary-Hier` offers even less improvement over the stationary model, although we do see some gains predictive performance from the hierarchical version of the stationary model.

Among the non-stationary methods, the three variants of our power-weighted densities approach, `Sep-PWD`, `Hier-PWD` and `Sep-PWD-BMA`, have the best predictive performance with increasingly lower cumulative prediction errors than the rolling window (`Window-5`) and dynamic linear model (`State-space-LR`) methods. The outperformance of our PWD approach is not isolated to any one period of time, though the time period around 2000-01 saw a sharp jump in the gains for all non-stationary methods.

In Table 4 we evaluate each model M by its squared prediction error, $SSPE(M)$. We provide the mean $SSPE(M)$ averaged over time points and portfolios as well as its standard error across portfolios. Observing that `PWD-BMA` had the smallest mean $SSPE(M)$, we also provide the p-value for a t-test on the difference between the `PWD-BMA` mean $SSPE$ and the mean $SSPE$ of each other method.

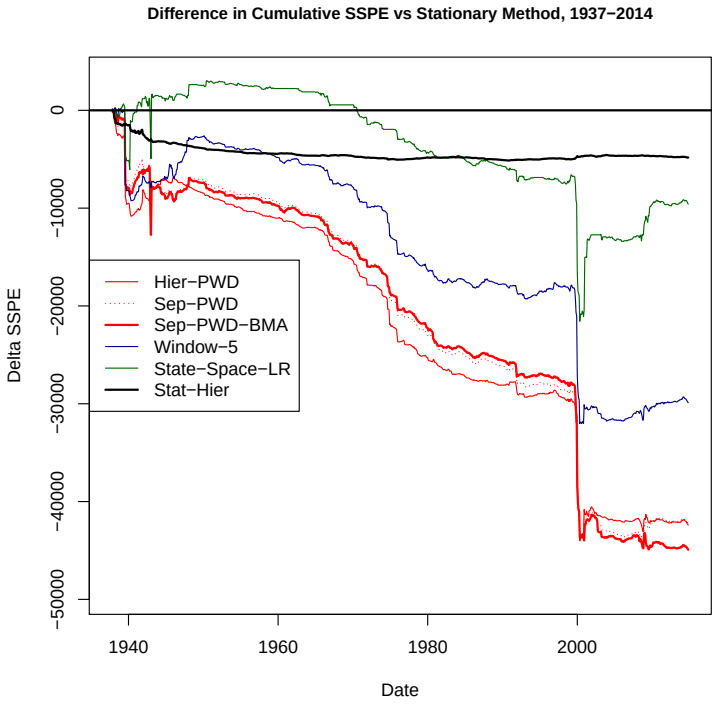

Fig 2: Rolling $\Delta SSPE$ relative to `Stationary` model of six models: `Hier-PWD`, `Sep-PWD`, `Sep-PWD-BMA`, `Window-30`, `State-space-LR`, and `Stat-Hier`.

Table 4 implies that `PWD-BMA`, `Hier-PWD` and `Sep-PWD` significantly improved upon the performance of `State-space-LR`, `Window-5`, `Stationary`, `Stat-Hier` and `Stat-BMA`. We see that `PWD-BMA` outperformed all other methods, achieving the smallest mean squared prediction error as well as one of the smaller standard errors. Financial practitioners value improvement in both the mean and the variance of squared prediction error because both reduce the amount of capital a financial practitioner would need to hold aside to maintain a hedge position over time. As we will see shortly, it appears `PWD-BMA` was able to adapt to secular cycles in the importance of the different market factors.

4.1. *Evolution of $\alpha_j^\star$ and β_j^m over time.* Our power-weighted densities approach provides some additional insight when we compare the estimated weight parameters $\alpha_j^\star$ for each of the 30 industry portfolios and their implication for the evolution of the sensitivities to changes in the overall stock

Statistic	PWD-BMA	H-PWD	Sep-PWD	SS-LR	W-5	Stat	H-Stat
Mean	13392	13476	13481	14570	13893	14889	14729
Std. Error	367	358	363	392	378	360	353
p-value		0.619	0.233	0.001	0.000	0.000	0.000

TABLE 4

Industry Portfolio Performance Comparison: Squared Prediction Error Mean, Standard Error, and p-value of Difference in Mean versus PWD-BMA. For compactness, we denote Hier-PWD by H-PWD, State-space-LR by SS-LR, Window-5 by W-5, Stationary by Stat and Hier-Stat by H-Stat

market ("market beta").

In Figure 3, we compare the estimated $\alpha_j^\star$ for the two industry portfolios with the lowest average $\alpha_j^\star$ to the two industry portfolios with the highest average $\alpha_j^\star$, as well as the average $\alpha_j^\star$ across all portfolios. Each $\alpha_j^\star$ is plotted as a smoothed trend over time, where the value of $\alpha_j^\star$ at time t is estimated using data for that portfolio up to time t.

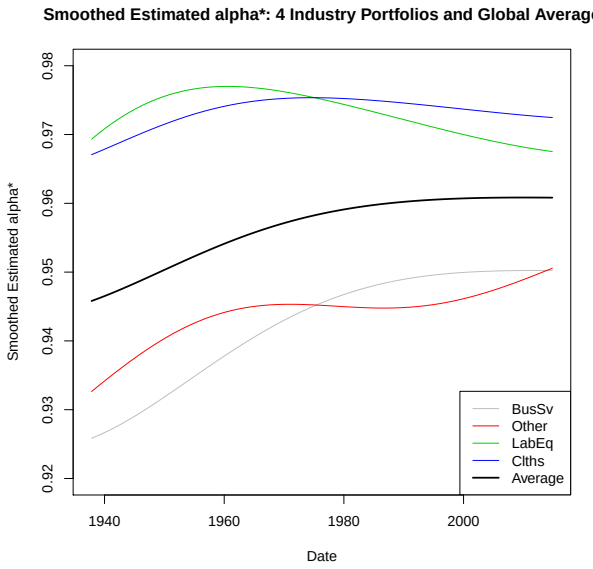

Fig 3: Smoothed estimated $\alpha_j^\star$ for highest and lowest empirical average estimated α_j^* over time. A local linear kernel bandwidth smoother over a dense grid of 600 grid points was used for the smoothing.

The average $\alpha^\star$ across industries has been trending slightly upwards over time. Business services and other industries ("BusSv" and "Other") have the highest amount of non-stationarity (lowest $\alpha_j^\star$ values) whereas lab equip-

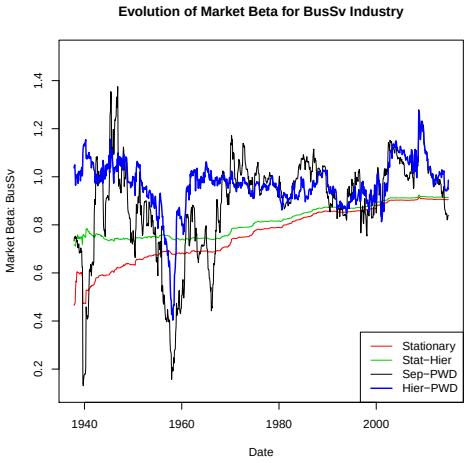

Fig 4: Estimated $\hat{\beta}_j^m$ for BusSv industry over time from 4 models

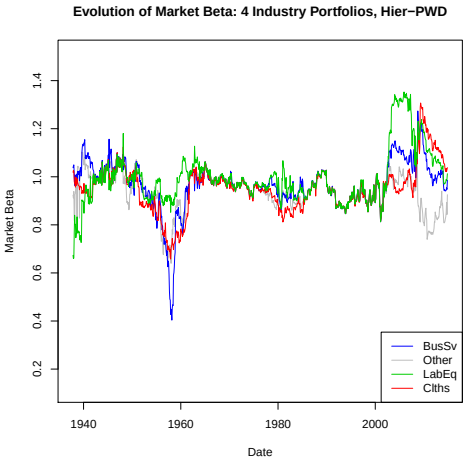

Fig 5: Estimated $\hat{\beta}^m$ from `Hier-PWD` for four industries over time

ment and clothes ("Lab Eq" and "Clths") have the lowest amount of non-stationarity (highest $\alpha_j^\star$ values). It is unsurprising that the "Other" industry has high non-stationarity since its risk profile and industry mix is most likely to change over time, while an industry like "Clothes" has a more stable risk profile over time.

Figure 4 provides further examination of the role of our PWD weighting on β^m, the sensitivity of industry portfolio returns to the overall market over time. Specifically, we plot the estimated of β_j^m over time for the Business Services industry as estimated by the `Stationary`, `Stat-Hier`, `Sep-PWD` and `Hier-PWD` models. Our PWD approaches suggest that β_j^m for the Business Services industry is far less stable over time than implied by the `Stationary` and `Stat-Hier` models. For example, after the burst of the technology stock market bubble in the early 2000's, our PWD approach inferred a sharp rise in β_j^m which is indicative of heightened sensitivity of returns to overall market movements, while the stationary models made no such adjustment.

Figure 5 compares the evolution of β_j^m estimated by our `Hier-PWD` approach for the four industries that represented the highest and lowest degrees of non-stationarity in Figure 3. Interestingly, there were two time periods in which β_j^m's sharply diverged from 1.0: the period preceding 1960 and immediately following 2000. These fluctuations would not be detectable by a stationary model that uses all historical data to estimate β_j^m.

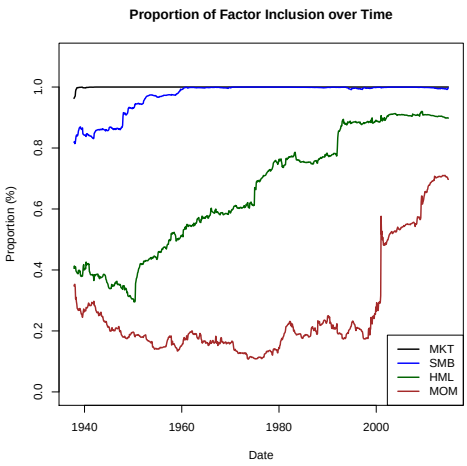

Fig 6: Inclusion probability for each factor, averaged over portfolios

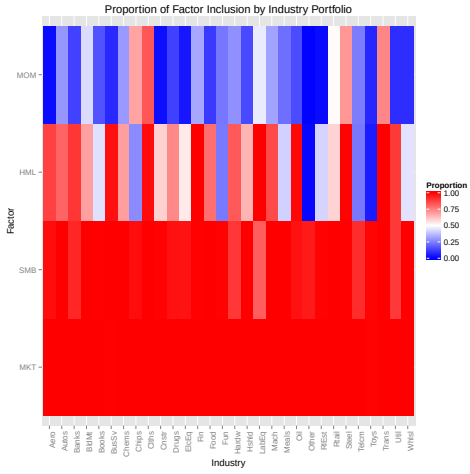

Fig 7: Inclusion probability for each factor by industry portfolio

4.2. *The Evolution of Factor Weightings in Bayesian Model Averaging.* The Bayesian model averaging variant of our PWD approach provides additional insight into the importance of the different predictor factors over time and across industries. As outlined in Section 2.4, our PWD-BMA calculates posterior model probabilities (equation 2.24) for the 16 possible models that can be formed by the inclusion/exclusion of our four factors. We calculate the *posterior probability of inclusion* for each factor as the sum of the posterior model probabilities over the subset of models that included that factor. These inclusion probabilities are calculated for each portfolio j and for each time point t (using only data up to that time point).

In Figure 6 we plot the evolution of the inclusion probability, averaged over the thirty portfolios, for each of the four factors: MKT, SMB, HML and MOM. In Figure 7, we give the inclusion probability for each of the four factors averaged over time separately for each of the thirty portfolios.

We see in Figure 6 that the MKT and SMB factors have inclusion probabilities near to 1.0 for almost the entire time series. The HML and MOM factors initially have much lower inclusion probabilities for most of the time series, with the momentum factor being particularly interesting. For almost 60 years, MOM's inclusion probability vacillated between 15% and 25% with an inclusion probability of 17% in November 1997. The MOM inclusion probability abruptly increased to 58% by January 2001 and then further increased to 70% by November 2014. It is probably not coincidence that

Carhart (1997), which first introduced the momentum factor, immediately preceded a sharp rise in the importance of MOM after 60 years of relative unimportance. We observe a similar phenomenon with the HML factor, which had a step function-like increase from 78% in December 1991 to 89% in January 1993, the month before Fama and French (1993) was published.

Of the 16 regression models considered, the one with the highest posterior probability over the time period from November 1937 to May 1960 was a two factor model including only MKT and SMB factors. From June 1960 to January 2001, the three factor model had the highest posterior probability. Thereafter, the four factor model had the highest posterior probability. We see in Figure 7 the considerable heterogeneity across industries in the inclusion probabilities of the MOM and HML factors. For example, the "Other", "Aerospace" and "Real Estate" industries have MOM inclusion probabilities of under 2%, while the "Steel", "Transportation" and "Clothes" industries have MOM inclusion probabilities above 70%.

These results could impact how financial practitioners may want to go about hedging their positions. For example, our PWD approach does suggest that MOM factor is more important than it has ever been in explaining cross-sectional heterogeneity in returns across stock portfolios.

5. Summary and Discussion. As an alternative to standard times series models, we have developed a power-weighted densities (PWD) approach where observations in the distant past are down-weighted in the likelihood function relative to more recent observations (2.2). Our general approach provides an effective way to allow for non-stationarity in time series while still giving the practitioner control over the choice of the underlying data model, which could be useful in a wide variety of applications. In this paper, we focused on a specific exponentially-decreasing weighting scheme (2.3) though other weighting schemes could be considered. For example, the most popular way of allowing for non-stationarity in finance, rolling window estimation, is another special case of our PWD approach.

Our PWD approach is a simpler alternative for allowing non-stationarity compared to dynamic linear state space methods (West and Harrison, 1998) that explicitly model the evolution of an underlying state vector. Our approach has the greatest benefit when the goal is forward-looking prediction, which is relevant in our application: prediction of future prices given the concurrent movement of market factors is often the primary goal in the financial markets. With this emphasis on prediction, we have focused heavily on the posterior distribution for the parameters $\boldsymbol{\theta}_{\mathrm{T}}$ at the terminal time point T, instead of inferring the entire evolution of an underlying state vector $\boldsymbol{\theta}_{1:T}$

as is done in state space models.

Our simulation evaluation (Section 3) suggests that our PWD approach performs well in terms of both predictive accuracy and computational cost across different data settings and should be considered in situations where the practitioner suspects the underlying process generating the data evolves over time.

In Section 2.3, we developed the specific methodology for our PWD approach for a hierarchical linear regression model, which was needed for our application to industry portfolios in Section 4. In that application, our PWD approach showed superior predictive performance over models that assume stationary parameters, as well as alternative non-stationary methods such as dynamic linear models and rolling windows. In Section 2.4, we developed a PWD variant of Bayesian Model Averaging which yielded the best predictions in our application, and also gave interesting insights into the evolution in the importance of market factors over time.

6. Acknowledgements. Thanks to the Wharton Research Computing team for their HPC resources and to Joshua Magarick, Tengyuan Liang, Robert Stine and Nathan Stein for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement A: Discussion of "Improving Market Factor Estimation with Power Weighted Densities"
(). We show the conjugacy for exponential families under our PWD approach and the Kullback-Leibler optimality of the general PWD setup. We provide additional results for computational cost and simulations comparing additional PWD variants to competing models. An adaptive PWD variant which switches between linear and exponentially decaying weights is also explored.

References.

Aiolfi, M. and Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics* **135** 31–53.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics* **64** 423–458.

Berry, D. A. and Stangl, D. K. (1996). Bayesian methods in health-related research. *Bayesian Biostatistics* 3–66.

Berry, S. M., Carlin, B. P., Lee, J. J. and Muller, P. (2010). *Bayesian adaptive methods for clinical trials* **38**. CRC press.

Brian, N. (2010). Bayesian analysis using power priors with application to pediatric quality of care. *Journal of Biometrics & Biostatistics*.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance* **52** 57–82.

Carter, C. K. and Kohn, R. (1994). On Gibbs Sampling for State Space Models. *Biometrika* **81** 541-553.

Chen, Y. and Singpurwalla, N. D. (1994). A non-Gaussian Kalman filter model for tracking software reliability. *Statistica sinica* **4** 535–48.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics* **106** 157–181.

Dawid, A. P. (1992). Prequential data analysis. *Lecture Notes-Monograph Series* 113–126.

Fama, E. F. and French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of financial economics* **25** 23–49.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics* **33** 3–56.

Figlewski, S. (1984). Hedging performance and basis risk in stock index futures. *The Journal of Finance* **39** 657–669.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* 501–514.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian data analysis*. CRC press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **6** 721-741.

Grunwald, G. K., Raftery, A. E. and Guttorp, P. (1993). Time series of continuous proportions. *Journal of the Royal Statistical Society. Series B (Methodological)* 103–116.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67** 1047–1056.

Hoeting, J. A., Raftery, A. E. and Madigan, D. (2002). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics* **11** 485–507.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science* 46–60.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association* **90** 773–795.

Lewellen, J. and Nagel, S. (2006). The conditional CAPM does not explain asset-pricing anomalies. *Journal of Financial Economics* **82** 289–314.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89** 1535–1546.

Paez, M. S. and Gamerman, D. (2013). Hierarchical Dynamic Models. *The SAGE Handbook of Multilevel Modeling* 335.

Petkova, R. and Zhang, L. (2005). Is value riskier than growth? *Journal of Financial Economics* **78** 187–202.

Petris, G., Petrone, S. and Campagnoli, P. (2009). *Dynamic linear models with R*. Springer.

Raftery, A. E. and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* **98** 931–938.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2009). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial*

Studies hhp063.

Shephard, N. (1994). Local scale models: State space alternative to integrated GARCH processes. *Journal of Econometrics* **60** 181–202.

Shiller, R. (2014). Online Data. http://www.econ.yale.edu/ shiller/data.htm.

Smith, J. (1979). A generalization of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society. Series B (Methodological)* 375–387.

Smith, J. (1981). The multiparameter steady model. *Journal of the Royal Statistical Society. Series B (Methodological)* 256–260.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* **23** 405–430.

Tan, S., Machin, D., Tai, B., Foo, K. and Tan, E. (2002). A Bayesian re-assessment of two Phase II trials of gemcitabine in metastatic nasopharyngeal cancer. *British journal of cancer* **86** 843–850.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* **21** 1455–1508.

West, M. and Harrison, J. (1998). Bayesian Forecasting and Dynamic Models (2nd edn). *Journal of the Operational Research Society* **49** 179–179.

Wilmott, P. (1995). *The mathematics of financial derivatives: a student introduction.* Cambridge University Press.

Daniel McCarthy
Department of Statistics
The Wharton School
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104
Email: danielmc@wharton.upenn.edu

Shane T. Jensen
Department of Statistics
The Wharton School
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104
Email: stjensen@wharton.upenn.edu