FORECAST ADJUSTMENT UNDER SHOCKS: SIMILARITY-BASED SOLUTIONS TO UNPRECEDENTED EVENTS

BY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2024

Urbana, Illinois

Doctoral Committee:

Assistant Professor Gökçe Dayanıklı (Committee Member)
Assistant Professor Daniel Eck (Chair and Advisor)
Associate Professor Jihyung Lee (Committee Member)
Professor Feng Liang (Committee Member)

Abstract

This work examines what to do under sudden ruptures in the data-generating process of time series.

Table of contents

List of Abbreviations

DGP	Data-Generating Processes
GARCH	Generalized Conditional heteroskedasticity

Chapter 1

Introduction

To be populated

Chapter 2

Post-Shock Volatility Forecasting Using Similarity-based Shock Aggregation

2.1 Introduction

Reacting to a seemingly unprecedented event might involve the question: what, if anything, does it resemble from the past? Such might be the case with event-driven investing strategies, where the identification of the event could arise via the news pages or corporate communications and hence contains a qualitative, narrative element [**?**]. Matching a current crisis to past events is a problem with unsurprising statistical angles: identification, sample size, weighting, risk, and robustness, among many others.

In the context of foreign exchange rate market structure, **?** speculate "[w]hether news is scheduled or non-scheduled its influence on exchange rates may be related to the state of the market at the time of the news arrival. News that arrives during periods of high uncertainty may have different effects on the exchange rate, than news that arrives in calmer periods." The authors also note that non-scheduled news may require more time for markets to digest, leading to greater heterogeneity (including but not limited to higher dispersion) of responses. We take inspiration from these arguments, developing a method suitable to the conditions that typically accompany news shocks. Central to our endeavor is the observation that news shocks have both a qualitative (the description of the event as well as future contingents surrounding it) and quantitative component.

In this work, we focus on the second central moment of a time series, the volatility. One of the most important aspects of a positively-valued time series $(P_t)_{t\in\mathbb{N}}$, especially financial time series, is the volatility of the return series $(r_t)_{t\in\mathbb{N}}$. A financial asset's price series may exhibit behavior that makes inapplicable and uninterpretable the traditional methods of time series analysis. In contrast, the return series is scale-free [**?**], easily-interpreted, and often at least weakly stationary. Our reasons for studying volatility are two-fold. Whereas once returns were thought to react to specific news events, now stock price movements are believed to be overwhelmingly noise-based (see **?** and references therein), and even if one could construct credible models for describing and forecasting price series and return series, that would not necessarily tell us much about the variability of such forecasts nor enlighten us about the evolution of the variability of $(P_t)_{t\in\mathbb{N}}$ and $(r_t)_{t\in\mathbb{N}}$ or under rapidly changing conditions.

No matter how a time series or its transformations are modeled, forecasting in the presence of news shocks requires a methodological framework that sensibly incorporates relevant information that has yet to manifest in market price or derivative quantities like volatility. In this setting, regime-change models (see **?** and citations therein) are of little use because under the assumption of a known exogenous shock, there is no need to estimate a regime-change time, nor is there data following the exogenous shock event to fit a model. Asymmetric GARCH models were an early attempt to account for the fact that negative returns typically beget greater volatility than positive returns [?]. Asymmetry and mixed-frequency are employed in ? to forecast under the presence of extreme shocks. Problematically, any asymmetric model will depend upon the observation of a signed return to provide the most updated volatility forecast, but under the circumstances posited herein, no such return has been observed.

Similar problems and more exist for Realized GARCH models [?] in our setting. These models incorporates observable measures of volatility, known as "realized measures", like implied volatility (IV). However, under the assumptions herein, no post-shock data is available, and even if it were, Realized GARCH does not promise to perform well, since Black-Scholes-implied volatility is a biased estimator of volatility [??], with the bias increasing in times of crises, when options may be far out-of-the-money. GARCH models have been shown slow to adapt to spikes in volatility [?].

The approach herein sidesteps the modeling shortcomings of Realized GARCH and other approaches which require observations of post-shock outcomes by substituting modeling assumptions. The method herein proceeds under the assumption that similar news events occasion volatility shocks arising from a common conditional shock distribution. The procedure proposed does not require post-shock information like returns or market-implied quantities from the time series under study. Hence, we also avoid questions about what realized measure to use and when as well as questions about the usefulness of high-frequency data, although these remain intriguing avenues for future work.

The primary methodological tool presented in this work is fixed effect estimation followed by a distance-based weighting procedure for pooling those estimates. The use of fixed effect estimation for the study of structural shocks has a pedigree in macroeconomic analysis (? as referenced in ?; see also discussion of deterministic exogenous events in ?). We employ fixed effect estimation on the basis of a well-established conceptual assumption that shocks of economic time series can be modeled as mixtures, in particular, mixtures of ordinary innovations and shocks due to rare events (see ? and references therein). In the forecasting literature, the term "intercept correction" has come to refer to a modeling technique in which nonzero errors are explicitly permitted [??]. They summarize the literature as distinguishing two families of intercept correction: so-called "discretionary" intercept corrections that attempt to account for future events, without hard-coding an ad-hoc adjustment into the model specification, and second, "automated" intercept corrections that attempt to adjust for persistent misspecification using past errors. ? use weighted subsets of a scalar series' own past to correct forecast errors by an additive term. ? introduce a similarity-based forecasting procedure for time-varying coefficients of a linear model. ? employ a form of intercept correction in order to adjust forecasts to the COVID-19 shock in the spring of 2020 based on the proportional misses of the same model applied to the Great Recession.

A researcher interested in forecast adjustment can choose between procedures that discretionary or automated, a variety of choices for the collection of data assembled to perform the correction, whether the data is internal (i.e. from the time series itself) or external, the parametric term to be corrected (e.g. intercept, coefficients), if any, as well as the correction function (i.e. the mapping from the data to the corrective term), including the weighting applied to the assembled data (e.g. Nearest-Neighbor, arithmetic mean, kernel methods).

Our procedure is a discretionary procedure for intercept correction that incorporates systematically data internal or external to the time series under study. The correction function, as we shall see, includes a convex combination of fixed effect estimates from a donor pool. In the causal inference literature, a donor pool is a set of units of observation that share distributional properties with the prime object of analysis. In ?, the authors build upon previous work in causal inference whereby a treatment effect can be estimated via comparison with a synthetic time series that represents the control. In their setting, the synthetic unit is constructed using a convex combination of the donors. The particular convex combination employed is a function of the distance between the time series under study and the donors. Donors that are closer to the time series under study are given greater weight. Conversely, donors that are unlike the time series under study receive little to no weight. ? adapt the notion of a donor pool as well as distance-based weighting for the purpose of prediction. Their one-step-ahead forecasts use distance-based-weighting to aggregate shock estimates from the donor pool's time series according to the similarity to the time series under study. Their approach does not take into account the ARCH effects commonly observed in time series, especially financial times series, leaving unaccounted for the variability in the variability of time series forecasts. Outside of **?**, we know of no prior work that both introduces a parametric specification for nonzero errors and introduces a procedure for weighting appropriately the nonzero errors of similar shocks occurring outside the time series under study. Likewise, we are not familiar with any prior work that attempts to account for anticipated nonzero errors using an explicit parametric adjustment, i.e., what we will call a "correction function".

The method proposed herein is supported by both numerical evidence as well as a real data example. The numerical evidence takes the form of simulations and analysis of the method's performance across a considerable grid of varying parameters, allowing us to evaluate reasonable hypotheses about the method but also uncover some intriguing nuances. The real data example, on the other hand, demonstrates the usefulness of the method by applying it to the United States (US) presidential election. We find that the method would have allowed a researcher to predict the volatility of a financial asset that immediately followed the 2016 US presidential. Our discussion of this example sheds light on some subtleties of the method as well as its robustness to various choices that must be made to construct and assemble donors and appropriate covariates [**?**].

2.2 Setting

In order to motivate our procedure, we provide a visual illustration. In Figure 2.1, we show how the aggregation of estimated shock-induced excess volatilities from donors in the donor pool works when the correction function is a specially-chosen convex combination of fixed effects from the donor pool. Our method assumes a credible, parsimonious parameterization in which the shock is an affine transformation of several key covariates. The key intuition behind this shock parameterization is that as the strength of the linear signal increases relative to the idiosyncratic error, the GARCH estimation of these effects increases in accuracy. From this, it follows that the aggregated shock estimate increases in accuracy.

2.2.1 A Primer on GARCH

We define the log return of an asset between $t-1$ and t as $r_t = \log(\frac{P_t}{P_{t-1}})$, where P_t denotes the price at time t. The class of ARIMA(p,d,q) models [**?**] provides a framework for modeling the autoregressive structure of r_t. These models assume a certain dependence structure between r_t and $(r_k)_{k\leq t}$, yet their errors — often called innovations in the financial time series context due to how they represent the impact of new information — are nevertheless assumed to be i.i.d. with mean zero and constant variance. The ARCH [**?**] and GARCH [**?**] models provide elegant alternatives to the homoskedasticity assumption. In fact, the GARCH framework in its most basic form disregards r_t and instead turns its interest to the series r_t^2 (once properly centered, i.e. after assuming a mean-model for returns).

To that end, let $a_t = r_t - \mu_t$, where μ_t is the (potentially time-varying) mean of the log return series r_t. We thus derive a mean-zero process $(a_t)_{t\in\mathbb{N}}$ with the property that $\mathbb{E}[a_t^2] = \mathrm{Var}[a_t]$. Under the assumption of time-invariant volatility, the series a_t^2 should exhibit no significant autocorrelation at any lag $\ell \geq 1$. This assumption motivates tests for ARCH effects, that is, tests for the clustering of volatility. These tests explore the alternative hypothesis that σ_t^2 is not only a time-varying parameter but furthermore a function of past squared residuals of the mean model. In particular, the ARCH(m) model is an autoregressive model in which σ_t^2 is a deterministic function of the past m values of r_t^2. The GARCH(m,s) framework take this one step further by modeling σ_t^2 as a linear combination of the past m values of r_t^2 and well as the past s values of σ_t^2. In functional form, a GARCH process (sometimes called a strong GARCH process [**?**, p. 19]) is given by

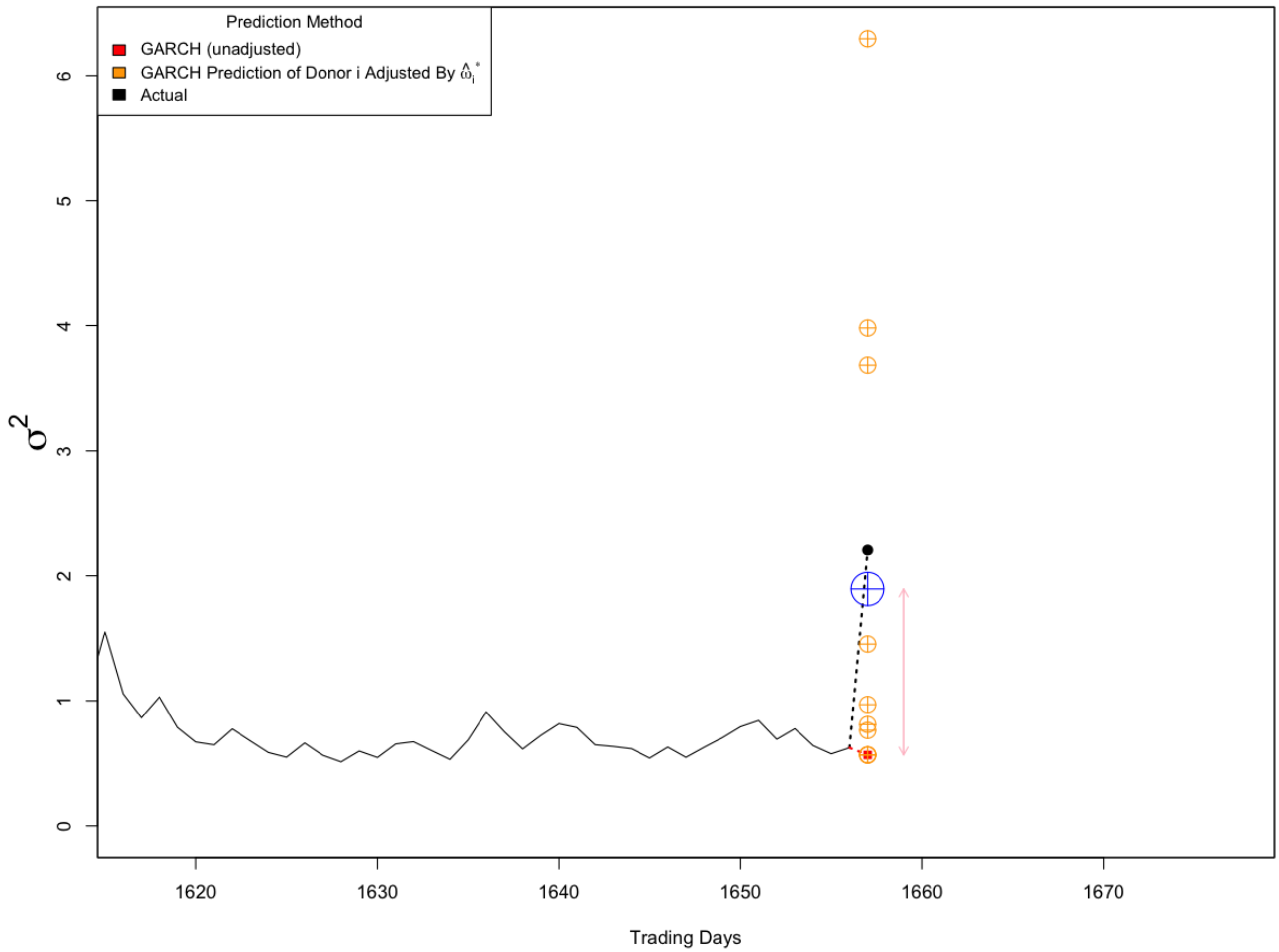


Figure 2.1: The simulated time series experiences a volatility shock between trading days 1,656 and 1,657. The GARCH prediction, in red, fails even to approach the volatility spike at T^*+1, as do several adjusted predictions, which are orange. In contrast, the GARCH forecast adjusted by $\hat{\omega}^* = \sum_{i=2}^{n+1} \pi_i \hat{\omega}_i^*$, a convex combination of the estimated shocks in the donor pool, achieves directional correctness as well as a smaller absolute loss in its prediction. The pink vertical line serves to indicate the adjustment of size $\hat{\omega}^*$ that allows the blue bullseye to approach more closely the ground truth.

$$
\begin{aligned}
&\sigma_t^2 = \omega + \sum_{k=1}^{m} \alpha_k a_{t-k}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \\
&a_t = \sigma_t \epsilon_t \\
&\epsilon_t \overset{iid}{\sim} E[\epsilon_t] = 0, Var[\epsilon_t] = 1 \\
&\forall k, j, \alpha_k, \beta_j \geq 0 \\
&\forall t, \omega, \sigma_t > 0 \ .
\end{aligned}
$$

Assuming further that σ_t^2 depends on a vector of exogenous covariates $\mathbf{x}_t$, we have a GARCH-X(m, s). The volatility equation then becomes

$$\sigma_t^2 = \omega + \sum_{k=1}^{m} \alpha_k a_{t-k}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 + \gamma^T \mathbf{x}_t \ . \tag{2.1}$$

2.2.2 Model setup

We will suppose that a researcher has multivariate time series data $\mathbf{y}_{i,t} = (r_{i,t}, \mathbf{x}_{i,t})$, $t = 1, \ldots, T_i$, $i = 1, \ldots, n+1$, where $r_{i,t}$ is scalar and $\mathbf{x}_{i,t}$ is a vector of covariates such that $\mathbf{x}_{i,t}|\mathcal{F}_{i,t-1}$ is deterministic. Suppose that the analyst is interested in forecasting the volatility of $r_{1,t}$, the first time series in the collection, which we will denote *the time series under study*. We require that each time series $\mathbf{y}_{i,t}$ is subject to an observed news event following $T_i^* \leq T_i + 1$ and before witnessing $T_i^* + 1$. We are implicitly leveraging the fact that financial assets are heavily-traded during market hours, yet only thinly traded (if traded at all) outside market hours. In contrast, the arrival of market-moving news does not obey any such restrictions. In light of the foregoing, we can denote our collection of GARCH-X volatility equations of interest using the following notation

$$\sigma_{i,t}^2 = \omega_i + \sum_{k=1}^{m_i} \alpha_{i,k} a_{i,t-k}^2 + \sum_{j=1}^{s_i} \beta_{i,j} \sigma_{i,t-j}^2 + \gamma_i^T \mathbf{x}_{i,t} \ .$$

Let $I(\cdot)$ be an indicator function. Let T_i denote the time length of the time series i for $i = 1, \ldots, n+1$, and let T_i^* denote the largest time index prior to the arrival of the news shock, with $T_i^* < T_i$, to ensure that there is at least one post-shock realization for each series i. Let $\delta, \mathbf{x}_{i,t} \in \mathbb{R}^p$. Let $\mathcal{F}_i$ with a single-variable subscript denote a univariate, time-invariant σ-algebra, and $\mathcal{F}_{i,t}$ denote the canonical product filtration for donor i at time t. Let $D_{i,t}^{return} = I(t \in \{T_i^* + 1, ..., T_i^* + L_{i,return}\})$ and $D_{i,t}^{vol} = I(t \in \{T_i^* + 1, ..., T_i^* + L_{i,vol}\})$, and let $L_{i,return}, L_{i,vol}$ denote the lengths of the level and volatility shocks, respectively. For $t = 1, \ldots, T_i$ and $i = 1, \ldots, n+1$, the model $\mathcal{M}_1$ is defined as

$$\mathcal{M}_1 : \begin{array}{l} \sigma_{i,t}^2 = \omega_i + \sum_{k=1}^{m_i} \alpha_{i,k} a_{i,t-k}^2 + \sum_{j=1}^{s_i} \beta_{i,j} \sigma_{i,t-j}^2 + \gamma_i^T \mathbf{x}_{i,t} + \omega_{i,t}^*, \\ a_{i,t} = \sigma_{i,t}((1 - D_{i,t}^{return})\epsilon_{i,t} + D_{i,t}^{return}\epsilon_i^*), \\ \omega_{i,t}^* = D_{i,t}^{vol}[\mu_{\omega^*} + \delta^T \mathbf{x}_{i,T_i^*+1} + u_{i,t}], \end{array}$$

with time-invariant error structure

$$\begin{aligned} &\epsilon_{i,t} \overset{iid}{\sim} \mathcal{F}_\epsilon \text{ with } \mathrm{E}_{\mathcal{F}_\epsilon}(\epsilon) = 0, \mathrm{Var}_{\mathcal{F}_\epsilon}(\epsilon) = 1, \\ &\epsilon_{i,t}^* \overset{iid}{\sim} \mathcal{F}_{\epsilon^*} \text{ with } \mathrm{E}_{\mathcal{F}_{\epsilon^*}}(\epsilon) = \mu_{\epsilon^*}, \mathrm{Var}_{\mathcal{F}_{\epsilon^*}}(\epsilon^*) = \sigma_{\epsilon^*}^2, \\ &u_{i,t} \overset{iid}{\sim} \mathcal{F}_u \text{ with } \mathrm{Var}_{\mathcal{F}_u}(u) = \sigma_u^2, \\ &\epsilon_{i,t} \perp\!\!\!\perp \epsilon_{i,t}^* \perp\!\!\!\perp u_{i,t}. \end{aligned}$$

Let $\mathcal{M}_0$ denote the subclass of $\mathcal{M}_1$ models such that $\delta \equiv 0$. Note that $\mathcal{M}_0$ assumes that nonzero shocks have no dependence on the covariates and are i.i.d. with $\mathbb{E}[\omega_{i,t}^*] = \mu_{\omega^*}$, where the lack of indices i or t on μ_{ω^*} indicates that it is shared across donors and is time-invariant. Models $\mathcal{M}_1$ and $\mathcal{M}_0$ parameterize changing dynamics after a shock is observed (when $t \geq T_i^* + 1$ for each series i). It is important to note that our modeling setup will additionally suppose that $\mathbf{x}_{i,t}$ will be observed before $\sigma_{i,t}^2$ is observed.

Note also the divergences from **?**. As already established, the model presented here is intended to account for ARCH effects. Additionally, here we also diverge by modeling the shocks $\omega^*_{i,t}$ as time-varying quantities obeying a mixture distribution. In particular, prior to and after the shock, the $\omega^*_{i,t}$ are uniformly zero. During the shock, i.e. from $T^*_i + 1$ through $T^*_i + L^{vol}_i$, the $\omega^*_{i,t}$ are distributed i.i.d, with a core signal component that depends solely on the unknown parameter δ and the observable vector $\mathbf{x}_{i,T^*_i+1}$. Wherever necessary, to distinguish the two distributions that make up the shock distribution, we use the term *nonzero shocks* to refer to the shocks that are not uniformly zero.

The nonzero shocks each include an idiosyncratic noise term that varies across time and donors. This feature allows the nonzero shocks to exhibit some variation as the shock progresses, even within a single donor. The noise term could be modeled with additional complexity, including serial correlation, or could exhibit deterministic or stochastic decay throughout the shock period, proportional to a deterministic or stochastic function of time, or could peak somewhere in the middle of the shock period. Each of these ideas might correspond to some particular, realistic feature of the world.

The model $\mathcal{M}_1$ is defined by a volatility equation and mean equation, as is any GARCH model. The decision to model the volatility shock $\omega^*_{i,t}$ as an additive random effect is consistent with our setup. However, the decision to model the level effect $\epsilon^*_{i,t}$ as a temporary rupture in the otherwise i.i.d. sequence of innovations $\epsilon_{i,t}$ may not be intuitive. One way of arguing for this choice is that, in a discrete time series model, if we assume the arrival of news in the time between T^*_i and $T^*_i + 1$, we do not have an easy way to express a conditional distribution of the innovation ϵ_{i,T^*_i+1} given the arrival of information. Using $\epsilon^*_{i,t}$ thus breaks this impasse. This justification also explains why we do not parameterize the level shock at $T^*_i + 1$ as a sum of two shocks, ϵ_{i,T^*_i+1} and ϵ^*_{i,T^*_i+1}, which would represent the level shock as generated by two independent sources of stochasticity. While we want to model the level shocks at $T^*_i + 1$ as potentially large in absolute value, we also want to retain the property of a unitary source of noise.

2.3 Methodology for Similarity-based Parameter Correction

We now introduce how a researcher can carry out predictions via distance-based weighting. The first step is to gather the p covariates that will parametrize a change in dynamics to model $\mathcal{M}_1$ and will represent the p-dimensional space in which weights are generated. This set of covariates parameterizes the shock to $\sigma^2_{i,t}$ of magnitude $\omega^*_{i,t}$ in model $\mathcal{M}_1$ above. Let $\mathbf{V}_t \in \mathbb{R}^{p\times n}$ be formed with columns $\mathbf{x}_{i,t}$, $i = 2, \ldots, n+1$. Without loss or generality, let the indices of each $\mathbf{x}_{i,t}$ be shifted and arranged such that $\mathbf{V}_{T^*+1}$ has as columns $\mathbf{x}_{2,T^*_2+1}, ..., \mathbf{x}_{n+1,T^*_{n+1}+1}$. Formally, in vector notation, the donor shocks are given by $\vec{\omega}_t = \vec{\mu}_{\omega^*} + \delta^T \mathbf{V}_t + \vec{u}_t$, where were have suppressed the $D^{vol}_{i,t}$ notation for simplicity. Ideally, $\mathbf{V}_t$ will display 'balance' in that p covariates exist for each of the n donors. In practice, missing values, corrupted values, or unacceptably extreme or noisy estimates may necessitate some sort of matrix completion, a problem that we do not tackle in this work.

2.3.1 Forecasting

We now turn to the next section, where $\mathbf{V}_t$ is employed in a procedure to arrive at a forecast adjustment. For illustration, we present two one-step-ahead forecasts for the time series under study. The first is the unadjusted GARCH forecast. The second is the adjusted forecast, which differs by the additive term $\hat{\omega}^*$ that is computed via our distance-based weighting procedure. These forecasts are:

$$\hat{\sigma}^2_{unadjusted,T^*_1+1} = \hat{\mathbb{E}}[\sigma^2_{1,T^*_1+1}|\mathcal{F}_{T^*_1}] \quad = \quad \hat{\omega}_1 + \sum_{k=1}^{m_1} \hat{\alpha}_{1,k} a^2_{1,T^*_1+1-k} + \sum_{j=1}^{s_1} \hat{\beta}_{1,j} \sigma^2_{1,T^*_1+1-j} + \hat{\gamma}^T_1 \mathbf{x}_{1,T^*_1+1},$$

$$\hat{\sigma}^2_{adjusted,T_1^*+1} = \hat{\mathbb{E}}[\sigma^2_{1,T_1^*+1}|\mathcal{F}_{T_1^*}] + \hat{\omega}^* \quad = \quad \hat{\omega}_1 + \sum_{k=1}^{m_1} \hat{\alpha}_{1,k} a^2_{1,T_1^*+1-k} + \sum_{j=1}^{s_1} \hat{\beta}_{1,j} \sigma^2_{1,T_1^*+1-j} + \hat{\gamma}_1^T \mathbf{x}_{1,T_1^*+1} + \hat{\omega}^* \ .$$

Note that GARCH models can be parameterized as ARMA models on the squares of the scalar time series $a^2_{i,t}$ [**??**, p. 18, p. 46], assuming that $a_{i,t}$ satisfies fourth-order stationarity. This fact matters for forecasting because the h-step-ahead forecasting function for a GARCH model is, just like for an ARMA model, the conditional expectation function, $\mathbb{E}[\sigma^2_{i,T_i^*+h}|\mathcal{F}_{T_i^*}]$, or practically speaking, the estimate thereof, $\hat{\mathbb{E}}[\sigma^2_{i,T_i^*+h}|\mathcal{F}_{T_i^*}]$ [**?**]. Here we have presented one-step-ahead forecasts for a GARCH-X(m,s). For $h = 2, 3, 4, ...$, the conditional expectation is computed recursively, as is standard for iterative autoregressive forecasts. Note that h-step-ahead predictions will require, of course, some reasonable idea about the length of time for which the adjustment quantity $\hat{\omega}^*$ should be included. Additionally, if $\mathbf{x}_{1,t}$ is used as an exogenous regressor in order to estimate γ_1 prior to the shock period, then strictly speaking, $\mathbf{x}_{1,t+h}|\mathcal{F}_t$ must be deterministic — at least for the relevant (t,h) pair for which a forecast is sought. In practice, it may be acceptable for $\mathbf{x}_{1,t+h}|\mathcal{F}_t$ to be merely well-estimated. Thus for simplicity of exposition, we will focus on the one-step-ahead forecast unless stated otherwise.

2.3.2 Excess Volatility Estimators

The problem of aggregating estimated donor shocks begins with the data constraints. Let us first introduce useful notation. Let $\hat{\omega}^*_{i,*}$ denote the shock estimate for donor i that is obtained via fixed effect estimation over time points $T_i^*+1, ..., T_i^*+L_{i,vol}$. That estimation procedure is justified by the assumption that at each nonzero shock point, the shocks will differ but will be equal in distribution.

Taking the estimated shocks as a given, we observe the pair $(\{\hat{\omega}^*_{i,*}\}_{i=2}^{n+1}, \{\mathbf{v}_i\}_{i=2}^{n+1})$. Let $\Delta^{n-1} = \{\pi \in \mathbb{R}^n : \sum_{i=1}^n \pi_i = 1, \pi_i \geq 0, i = 1, ..., n\}$. We wish to recover weights $\{\pi_i\}_{i=2}^{n+1} \in \Delta^{n-1}$ leading to favorable forecasting properties. These weights are used to define and compute an aggregate shock estimate

$$\hat{\omega}^* = \sum_{i=2}^{n+1} \pi_i \hat{\omega}^*_{i,*}, \tag{2.2}$$

which will be taken to be our forecast adjustment term. Since the weights $\{\pi_i\}_{i=2}^{n+1}$ are computed using $\mathcal{F}_{T_i^*}$, the set $\{\pi_i\}_{i=2}^{n+1}$ is deterministic and observed, *modulo* any stochastic ingredient in the numerical methods employed to approximate $\mathbf{x}_{1,T_1^*+1}$ using a convex combination of donor covariates. We say more about the properties of the shocks $\omega^*_{i,T_i^*+1}, ..., \omega^*_{i,T_i^*+L_i^{vol}}$ in section 2.4.

Following **???**, let $\|\cdot\|_{\mathbf{S}}$ denote any semi-norm on $\mathbb{R}^p$, and define

$$\{\pi\}_{i=2}^{n+1} = \arg\min_{\pi} \|\mathbf{x}_{1,T_1^*+1} - \mathbf{V}_{T^*+1}\pi\|_{\mathbf{S}}.$$

In the forecast combination literature, it is of interest whether the weights employed to aggregate forecasts strive toward and meet various optimality criteria [**??**]. In our endeavor, there are at least two senses of optimal weights that one might be interested in. First, we can think of optimal weights as a set $\{\pi_i\}_{i=2}^{n+1}$ such that $\omega_1 = \sum_{i=2}^{n+1} \pi_i \hat{\omega}_{i,*}$, i.e., ω_1 is recovered perfectly, as it belongs to convex hull of the estimated shocks. However, ω_1 is never revealed to the practitioner, and hence there is no way of verifying the extent to which this condition is satisfied.

A more promising aim is finding weights such that $\mathbf{v}_1 = \sum_{i=2}^{n+1} \pi_i \mathbf{v}_{i,T_1^*+1}$, meaning that the covariates of the time series under study lies within the convex hull of the donor covariates. This condition underwrites asymptotic results in **?**, and the intuition there extends to this work: if the shock is parameterized by an affine function of covariates, then finding a linear combination that recreates the shock should serve us well. Because the method proposed uses a point in Δ^{n-1}, it is important to head-off possible confusion. What we are proposing is not a forecast combination method. What we are aggregating and weighting (not combining) are subcomponents of forecasts, not forecasts themselves. Moreover, from a broader perspective, forecast combination is an inapt term for what is being proposed here. First, the donor time series do not provide forecasts, nor would forecasts be needed for random variables that have already been realized. Second and more fundamentally, the theoretical underpinnings of forecast combination, while diverse [**?**], are distinct from the setting presumed in this work.

Supposing that we can define optimal weights, then uniqueness may still be a concern. **?** discuss sufficient conditions for uniqueness as well as the implications of non-uniqueness. **?** invoke the Carathéodory Theorem to argue for the sparseness of the weight vector. We make additional comments as well. $(\mathbb{R}^n, \|\cdot\|)$ is a Chebyshev space, and hence for any element x and any convex set $C \subset \mathbb{R}^n$, there exists a unique element $y \in C$ that minimizes $\|x - y\|$. However, the pre-image of y with respect to a particular operator and constraint set might not be unique. Let p', n' denote the number of linearly independent rows of $\mathbf{V}_t$ and linearly independent columns of $\mathbf{V}_t$, respectively. Let $\mathrm{col}(\cdot)$ denote the column space of a matrix, and let $\mathrm{Conv}(\cdot)$ denote the convex hull of a set of column vectors. The following table is useful for categorization of when a perfect fit and uniqueness prevail.

	$\mathbf{v}_1 \in \mathrm{Conv}(\mathrm{col}(\mathbf{V}_t))$	$\mathbf{v}_1 \notin \mathrm{Conv}(\mathrm{col}(\mathbf{V}_t))$
$p' \geq n'$	Perfect fit; minimizer unique	Fit not perfect; minimizer unique
$p' < n'$	Perfect fit; minimizer not necessarily unique, Carathéodory Theorem applies	Fit not perfect; minimizer not necessarily unique, Carathéodory Theorem applies

It should be noted, however, that even when the minimizer is unique, finding that unique minimizer may require numerical methods. In the event that one's preferred numerical optimization routine fails to arrive at a single solution on the simplex in repeated runs of the algorithm, some sort of forecast combination might be advised. We discuss forecast combination later in this paper in the context of robustifying against misspecification.

On the questions of convex geometry and convex optimization, comparison with least-squares estimation is also illustrative. Consider least-squares estimation for the p-vector δ in the shocks $\omega^*_{i,t}$:

$$\hat{\omega}^*_{OLS} = \arg\min_{\delta} \|\hat{\omega}^* - \delta^T \mathbf{V}_{T^*+1}\|_2^2,$$

where $\mathbf{V}_{T^*+1}$ could include an adjoined row $\mathbf{1}_n$, in order to estimate the locator parameter of the shock. One problem is that this optimization problem is an optimization problem over p-vectors δ — over linear combinations of the covariates, whereas what we seek is an n-vector — a linear combination of donors. Additionally, there is no guarantee that $\hat{\omega}^*_{OLS} = \hat{\delta}^T_{OLS}\mathbf{x}_{1,T^*+1}$ would perform poorly, but given the small number of donors supposed in our setting, it is risky. Last, and perhaps most challenging for the least-squares estimate, $\hat{\delta}^T_{OLS}$ faces the usual requirement of more donors than linearly independent covariates, and that may not often be easy to satisfy in our setting.

2.3.3 Ground Truth Estimators

The time-varying parameter σ_t^2 is a quantity for which even identifying an observable effect in the real world is far more challenging. In this work, we use a common estimator of the variance called realized volatility (RV), one which has the virtue of being "model-free" in the sense that it requires no modeling assumptions [**?**]. The realized variance itself can be decomposed into the sum of a continuous component and a jump component, with the latter being less predictable and less persistent [**?**], cited in **?**, two factors that further motivate the method employed herein.

Suppose we examine K units of of time, where each unit is divided into m intervals of length $\frac{1}{m}$. We adapt the notation of **?**. Let $p_t = \log P_t$, and let $\tilde{r}(t, \frac{1}{m}) = p_t - p_{t-\frac{1}{m}}$. Suppressing the i index, we estimate the variance the log return series using Realized Volatility of the K consecutive trading days that conclude with day t, denoted

$RV_t^{K,m}$, using

$$RV_t^{K,m} = \frac{1}{K}\sum_{v=1}^{Km} \tilde{r}^2(v/m, 1/m),$$

where the K trading days have been chopped into Km equally-sized blocks.

Assuming that the K units $\tilde{r}(t,1) = p_t - p_{t-1}$ are such that $\tilde{r}(t,1) \overset{iid}{\sim} N(\mu, \delta^2)$, it is easily verified that $\mathbb{E}[RV^{K,m}] = \frac{\mu^2}{m} + \delta^2$, which is a biased but consistent estimator of the variance. We will proceed using $m = 77$, corresponding to the 6.5-hour trading day chopped into 5-minute blocks, with the first block omitted in order to ignore unusual trading behavior at the start of the day. Other sensible choices for m can replace our specification if desired.

2.3.4 Loss Functions

We are interested in point forecasts for $\sigma^2_{1,T_1^*+h}|\mathcal{F}_{T_1^*}$, $h = 1, 2, ...$, the h-step ahead conditional variance for the time series under study. Let L^h with the subscripted pair {prediction method, ground truth estimator}, denote the loss function for an h-step-ahead forecast using a given prediction function and ground truth estimator. For example, the h-step-ahead MSE when forecasting from time t using our method and using Realized Volatility as the ground truth is

$$\text{MSE}^h_{\text{adjusted prediction, RV}} = (\hat{\sigma}^2_{\text{adjusted prediction},t+h} - \hat{\sigma}^2_{RV,t+h})^2 \ .$$

Also of interest in absolute percentage error for an h-step-ahead forecast, defined as

$$\text{APE}^h_{\text{method, ground truth}} = \frac{|\hat{\sigma}^2_{\text{method},t+h} - \hat{\sigma}^2_{\text{ground truth},t+h}|}{\hat{\sigma}^2_{\text{ground truth},t+h}} \ .$$

Finally, we introduce the QL (quasi-likelihood) Loss [**?**]:

$$\text{QL}^h_{\text{method, ground truth}} = \frac{\hat{\sigma}^2_{\text{ground truth},t+h}}{\hat{\sigma}^2_{\text{method},t+h}} - \log\frac{\hat{\sigma}^2_{\text{ground truth},t+h}}{\hat{\sigma}^2_{\text{method},t+h}} - 1 \ .$$

What distinguishes QL Loss is that it is multiplicative rather than additive. This has benefits, both practical and theoretical. As **?** explain, the technical properties of the QL Loss allow researchers to compare forecasts across heterogeneous time series, whereas additive loss functions like MSE unfairly penalize forecasts made under market turbulence. For this reason and others, we proceed to evaluate the method, both in simulations and real data examples, using the QL loss.

2.4 Properties of Volatility Shocks and Shock Estimators

In this section we will provide estimation properties for our shock estimator (2.2) and our adjusted forecast. These results will be with $m_i = s_i = 1$, for all $i = 1, \ldots, n+1$ in model $\mathcal{M}_1$ unless otherwise stated, i.e. a GARCH(1,1) model with additional parameterizations for shocks. Note that the dual level-volatility shock in $\mathcal{M}_1$ has a marginal effect on the conditional variance $\sigma^2_{i,t}$ that reflects the geometric decay of innovations in autoregressive models. As usual, assume $\alpha + \beta < 1$. Furthermore, assume that both the volatility shock and the level shock are of length one only, and consider a circumstance with no exogenous covariate $\mathbf{x}_{i,t}$, except in the conditional shock distribution. Assume also that $r \geq 2$, which is necessary in order to isolate the effects of the level shock $\epsilon^*_{i,t}$. Then

$$\begin{aligned} \sigma^2_{i,T^*_i+r+1}|\mathcal{F}_{T^*_i+r} &= \omega_i + \alpha_i a^2_{T^*_i+r} + \beta_i \sigma^2_{i,T^*_i+r} \\ &= \omega_i + \alpha_i(\sigma_{i,T^*_i+r}\epsilon_{T^*_i+r})^2 + \beta_i \sigma^2_{i,T^*_i+r} \\ &= \omega_i + \sigma^2_{i,T^*_i+r}(\alpha_i(\epsilon_{T^*_i+r})^2 + \beta_i) \ . \end{aligned} \tag{2.3}$$

In Equation (2.3), observe that ω^*_{i,T^*_i+1} and ϵ^*_{i,T^*_i+1} each appear at most once, through the term $\sigma^2_{T^*_i+r}$. This might lead one to suspect geometric decay of the shocks ω^*_{i,T^*_i+1} and ϵ^*_{i,T^*_i+1}. Such a suspicion is easier to justify by examining the conditional expectation of the variance, $\mathbb{E}[\sigma^2_{i,T^*_i+r+1}|\mathcal{F}_{T^*_i+r}]$, which also happens to be the principal forecasting tool for a GARCH model [?]. Indeed, if we assume unit variance for all $\epsilon_{i,t}$ except, of course, $\epsilon^*_{i,t}$, then we have

$$\begin{aligned} \mathbb{E}[\sigma^2_{i,T^*_i+r+1}|\mathcal{F}_{T^*_i+r}] &= \mathbb{E}[\omega_i + \alpha a^2_{T^*_i+r} + \beta\sigma^2_{i,T^*_i+r}|\mathcal{F}_{T^*_i+r}] \\ &= \omega_i + \mathbb{E}[\alpha(\sigma_{i,T^*_i+r}\epsilon_{T^*_i+r})^2|\mathcal{F}_{T^*_i+r}] + \beta\sigma^2_{i,T^*_i+r} \\ &= \omega_i + \alpha\sigma^2_{i,T^*_i+r} + \beta\sigma^2_{i,T^*_i+r} \qquad \text{(Due to the unit variance assumption)} \\ &= \omega_i + \sigma^2_{i,T^*_i+r}(\alpha+\beta) \ . \end{aligned}$$

By repeated substitution, in conditional expectation, the shock belongs to the equivalence class $\mathcal{O}((\alpha+\beta)^r)$. We generalize this observation in the following proposition.

Proposition 1. *Let $a_{i,t}$ be a mean-zero time series obeying a GARCH(1,1) specification with unit-variance errors, all prior to the arrival of a volatility shock of length $L^{vol}_i \geq 1$ and level shock of length $L^{return}_i \geq 1$ at some time $T^*_i + 1$. Then for any i, $1 \leq i \leq n+1$, and r such that $r \geq max\{L^{return}_i, L^{return}_i\} + 1$,*

$$\mathbb{E}[\sigma^2_{i,T^*_i+r+1}|\mathcal{F}_{T^*_i+r}] = \omega_i + (\alpha+\beta)\sigma^2_{i,T^*_i+r}.$$

In other words, for a GARCH(1,1), once two time points removed from the longest shock length, the volatility shock and level shock can be subsumed into one. However, prior to being two time points removed, there is no such guarantee. For example, one can take $r = 1$ and level shock of length at least 1 to see that

$$\begin{aligned} \mathbb{E}[\sigma^2_{i,T^*_i+2}|\mathcal{F}_{T^*_i+1}] &= \mathbb{E}[\omega_i + \alpha a^2_{T^*_i+1} + \beta\sigma^2_{i,T^*_i+1}|\mathcal{F}_{T^*_i+1}] \\ &= \omega_i + \mathbb{E}[\alpha(\sigma_{i,T^*_i+r}\epsilon^*_{T^*+1})^2|\mathcal{F}_{T^*_i+1}] + \beta\sigma^2_{i,T^*_i+1} \\ &= \omega_i + \alpha\sigma^2_{i,T^*_i+1}(\mu^2_{\epsilon^*} + \sigma^2_{\epsilon^*}) + \beta\sigma^2_{i,T^*_i+1} \\ &= \omega_i + \sigma^2_{i,T^*_i+1}(\alpha(\mu^2_{\epsilon^*} + \sigma^2_{\epsilon^*}) + \beta), \end{aligned}$$

where $(\alpha(\mu^2_{\epsilon^*} + \sigma^2_{\epsilon^*}) + \beta)$ may be greater than 1, permitting explosive behavior, at least in the short term. After both shocks have been exhausted, their influence disappears quickly. This short-memory effect has implications for the method being developed herein. First, there may be different risks associated with overestimating and underestimating level shock and volatility shock lengths. Therefore, estimation of effects among the donors should err on the side of underestimating, not overestimating, the length of the max shock, since overestimation of the shock length brings with it the risk of underestimating $\omega^*_{i,t}$. Second, a practitioner of the method needs some idea of how long the the respective shocks in the time series under study might last. There are couple of obvious strategies: take all the donors, and over all the donor shock lengths, take the minimum. Alternatively, one could take the maximum.

2.4.1 Consistency of the shock estimators

The estimators $\hat{\omega}^*_{i,t}$ in (2.2) are central to our forecast adjustment strategy. Here we show that the quantities $\hat{\omega}^*_{i,t}$ possess an important asymptotic property that bolsters that overall reliability of the method proposed herein.

Proposition 2. *Assume*

1. *For each i, $1 \leq i \leq n$, $\{a_{i,t}\}_{t=0,...,T_i}$ obeys a GARCH-X(m, s), as laid out in Equation (2.1), with volatility shocks found in $\mathcal{M}_1$, where T_i is the length of the ith series.*
2. *For each $i, \{\omega^*_{i,t}\}_{t=0,...,T_i}$ is potentially non-zero at $\{T^*_i + 1, ..., T^*_i + L^{vol}_i\}$, $\omega^*_{i,T^*_i+1} \equiv ... \equiv \omega^*_{i,T^*_i+L^{vol}_i}$, and zero otherwise, where the arrival of T^*_i is governed by a time-invariant distribution on $\{a_{i,t}\}_{t=0,...,T_i-1}$, and both the arrival and conclusion of the shock is observable by the researcher.*
3. *The conditions in Assumption 0 of **?** hold.*

Then for any $i, 1 \leq i \leq n+1$, and for any $r, 1 \leq r \leq L^{vol}_i$, $\hat{\omega}^*_{i,T^*_i+r} \xrightarrow{p} \omega^*_{i,T^*_i+r}$ as $t \to \infty$. Additionally, $\hat{\omega}^*_{i,*} \xrightarrow{d} \omega^*_{i,T^*_i+r}$ as $t \to \infty$, and if for all $i, 1 \leq i \leq n+1$, $u_{i,t} \equiv 0$ on $\{T^*_i+1, ..., T^*_i+L^{vol}_i\}$, then $\hat{\omega}^*_{i,T^*_i+r} \xrightarrow{p} \omega^*_{i,T^*_i+r}$.

Lemma 1. *Under assumption 2, for each $i, i = 1, ..., n+1$, $\{\omega^*_{i,t}\}_{t=0,...,T_i}$ is a strictly stationary series.*

2.4.2 Consistency of the Conditional Forecast Function

Having established the consistency of the estimators $\hat{\omega}^*_{i,T^*_i+r}$, we extend that result to prove asymptotic properties of the conditional forecast function itself.

Proposition 3. *Assume*

1. *All conditions listed in Proposition 2.*
2. *There exist weights $\{\pi_i\}_{i=2}^{n+1}$ such that $\boldsymbol{v}_{1,T^*_1} = \sum_{i=2}^{n+1} \pi_i \boldsymbol{v}_{i,T^*_i}$.*

Then for any r, $1 \leq r \leq L^{vol}_1$, $\hat{\sigma}^2_{adjusted,T^*_1+r} \xrightarrow{d} \sigma^2_{1,T^*_1+r}$ as $t \to \infty$ in the donor pool, and if for all $i, 1 \leq i \leq n+1$, $u_{i,t} \equiv 0$ on $\{T^*_i+1, ..., T^*_i+L^{vol}_i\}$, then $\hat{\sigma}^2_{adjusted,T^*_1+r} \xrightarrow{p} \sigma^2_{1,T^*_1+r}$.

2.4.3 Asymptotic Loss

We now evaluate the loss and risk of our method under two scenarios: first, under arbitrary distribution of σ^2_{t+1}, and then second, under the assumption that the data-generating process is correctly specified. We proceed with the same notation for adjusted and unadjusted forecasts introduced in Section 2.3.1. For a 1-step-ahead forecast of σ^2_{t+1} where $t = T^*$, consider the difference

$$\begin{aligned}
&QL(\hat{\sigma}^2_{unadjusted,t+1}, \sigma^2_{t+1}) - QL(\hat{\sigma}^2_{adjusted,t+1}, \sigma^2_{t+1}) \\
&= (\frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{unadjusted,t+1}} - \log \frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{unadjusted,t+1}} - 1) - (\frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{adjusted,t+1}} - \log \frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{adjusted,t+1}} - 1) \\
&= \frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{unadjusted,t+1}} - \frac{\sigma^2_{t+1}}{\hat{\sigma}^2_{adjusted,t+1}} + \log \frac{\hat{\sigma}^2_{unadjusted,t+1}}{\hat{\sigma}^2_{adjusted,t+1}} \\
&= \frac{\sigma^2_{t+1}(\hat{\sigma}^2_{adjusted,t+1} - \hat{\sigma}^2_{unadjusted,t+1})}{\hat{\sigma}^2_{adjusted,t+1}\hat{\sigma}^2_{unadjusted,t+1}} + \log \frac{\hat{\sigma}^2_{unadjusted,t+1}}{\hat{\sigma}^2_{adjusted,t+1}} \ . \qquad (2.4)
\end{aligned}$$

For simplicity, we work with a GARCH(1,1) that experiences a volatility shock at a single time point for which we would like to provide a point forecast. Then (2.4) can be expressed as

$$\frac{\sigma^2_{t+1}\hat{\omega}^*_{t+1}}{\hat{\sigma}^2_{adjusted,t+1}\hat{\sigma}^2_{unadjusted,t+1}} + \log\frac{\hat{\omega}+\hat{\alpha}a^2_t+\hat{\beta}\sigma^2_t}{\hat{\omega}+\hat{\alpha}a^2_t+\hat{\beta}\sigma^2_t+\hat{\omega}^*_{t+1}}.$$

It is easily verified that as $\hat{\omega}^*_{t+1} \to 0^+$, the difference in the losses goes to zero. On the other hand, as $\hat{\omega}^*_{t+1}$ becomes large, the difference in the losses turns negative, with the lesson being that $\hat{\omega}^*_{t+1}$ must be in appropriate proportion to σ^2_{t+1} in order for the adjusted forecast to outperform the unadjusted forecast. This explains why it is so important to avoid using a naive adjustment estimator, $\overline{\omega^*}$, the arithmetic mean of the estimated shocks. We conclude this section with a broader result.

Proposition 4. *Assume the conditions in Propositions 2 and 3. Then as $t \to \infty$ in the donor pool,*

$$QL(\hat{\sigma}^2_{unadjusted,T^*+1},\sigma^2_{T^*+1}) - QL(\hat{\sigma}^2_{adjusted,T^*+1},\sigma^2_{T^*+1}) \xrightarrow{d} \frac{\omega^*_{1,T^*+1}}{\sigma^2_{1,T^*+1}-\omega^*_{1,T^*+1}} + \log\frac{\sigma^2_{1,T^*+1}-\omega^*_{1,T^*+1}}{\sigma^2_{1,T^*+1}} \geq 0,$$

and if for all $i, 1 \leq i \leq n+1$, $u_{i,t} \equiv 0$ on $\{T^*_i+1,...,T^*_i+L^{vol}_i\}$, the convergence is in probability. Hence, a correctly specified $\mathcal{M}_1$ model will outperform the unadjusted forecast asymptotically.

2.5 Numerical Examples

In this section, we demonstrate the effectiveness of the proposed method using Monte Carlo simulations. All simulations will use $\mathcal{M}_1$ volatility models and $\mathcal{M}_0$ models on the returns, i.e. $D^{return}_{i,t} \equiv 0$ for each i and each t. We first explain the parameters to be varied, the parameters that remain fixed, and the behavior we expect to observe. For our purposes, a parameter is defined to be fixed whenever it does not vary across any of the simulations performed, and conversely, a parameter is varying whenever it varies across at least two simulations performed. The overarching story told by the simulations is that our method performs well as the magnitude of the signal in the shock grows relative to the variance of the idiosyncratic noise term of the shock.

2.5.1 Fixed Parameters

Each simulation, $i = 1,...n+1$ is a GARCH(1,1) process of length T_i chosen randomly from the set of integers $\{756,...,2520\}$, corresponding to approximately 3-10 years of daily financial data. We fixed the intercept of the processes at the *garchx* package default of $\omega = .2$ [?]. We also use the values $\alpha = .1, \beta = .82$, corresponding to a GARCH process with greater influence from past values of $\sigma^2_{i,t}$ than past values of $a^2_{i,t}$.

2.5.2 Varying Parameters

We vary exactly five parameters: $\mu_V, \sigma_V, \mu_\delta, \mu_{\omega^*}, \sigma_u$, each of which is critical to evaluating the method's responsiveness to changing conditions in the shock distribution. μ_V, σ_V govern the elements of the covariates. μ_δ interacts with $\mathbf{V}_t$ via the dot-product operation, of course. μ_{ω^*} is a location parameter for the volatility shock, and σ_u is the standard deviation of the volatility shock.

We add an important note about the parameter μ_δ, which governs a vector δ of length p with elements that increase monotonically in proportion to $1,...,p$, after which they are scaled by the factor $\frac{2\cdot\mu_\delta}{p(p+1)}$, so that a randomly selected entry of the vector has mean μ_δ. The heterogeneity of the entries in δ is critical to simulating the plausible heterogeneity that will exist in the random effects structure. Absent a heterogeneous vector δ, the shock would not vary systematically with regard to each variable in the covariate vector. Such a setup would fail to approximate the realities of financial markets.

2.5.3 Evaluative Framework and Monte Carlo Results

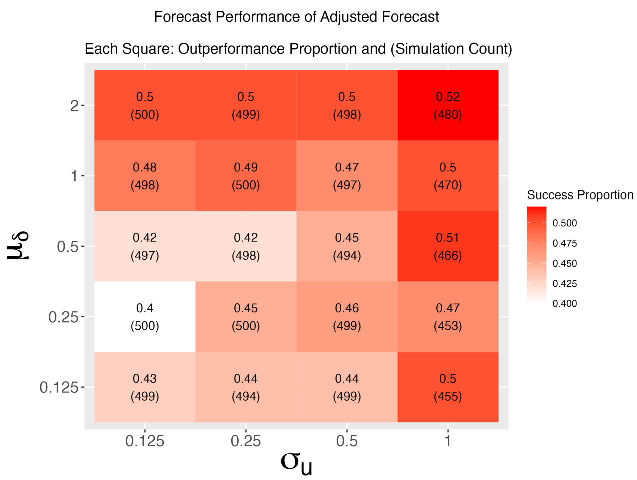
Consistent with our evaluative framework declared in 2.3.4, we compare adjusted and unadjusted forecasts using QL Loss, calculating the fraction of the simulations in which the adjusted forecast yields a QL Loss no smaller than that of the unadjusted forecast. The simulations fall into three broad categories, which we will denote **Signal and Noise** (displayed in Figure 2.2), **Interaction between Signal and Volatility Profile Means** (displayed in Figure 2.3), and lastly, **Interaction between Shock Intercept and Shock Noise** (displayed in Figure 2.4).

The first category plots two parameters of the shock equation, μ_δ and σ_u, in grids while increasing the mean of the covariate vectors $\mathbf{x}$ over three plots. Recall that δ is a parameter shared between donors, whereas each vector $\mathbf{x}_{i,t}$ in $\mathbf{V}_{i,t}$ is i.i.d. multivariate normal. In Figure 2.2a, we see very little variation across the grid. However, in Figures 2.2b and 2.2c, where μ_v is increased, we observe an expected phenomenon: for any fixed column of the plots, the performance of the adjusted forecast improves with increasing μ_δ. A somewhat subtler phenomenon is also observed: for a fixed row of the grids in Figures 2.2b and 2.2c, for $\mu_\delta \geq .5$, increasing σ_u leads to a decline in performance. However, this is not observed for smaller levels of μ_δ, where the signal is too weak.

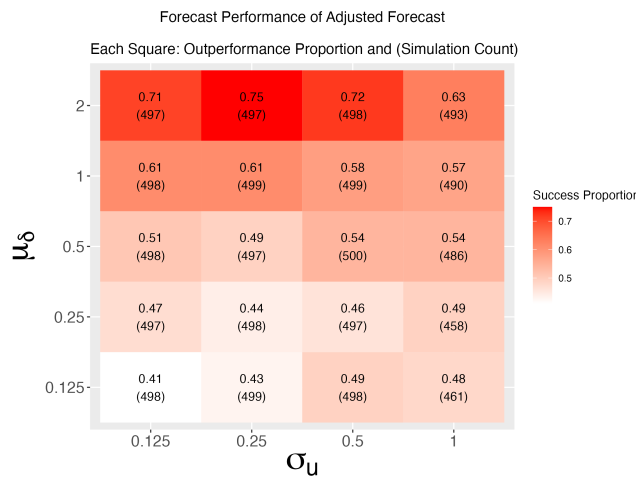
The second category of simulations examines μ_δ and μ_v, while increasing the noise σ_u. The row-wise and column-wise increases in performance that we witnessed in the first category largely hold in the second category of simulations, with one similar disclaimer: the relationship is strained under large levels of noise, like $\sigma_u = 1$.

Lastly, the third set of simulations takes a look at the role of the intercept of the conditional shock distribution, μ_{ω^*}, in particular, its interaction with σ_u. We can think of μ_{ω^*} as being the sole component of $\mathbf{x}$ that has no variance, i.e. because its corresponding entry in the vector $\mathbf{x}$ is the scalar 1. For larger values of μ_{ω^*}, the outperformance of the adjusted forecast declines with increases in σ_u, but this effect is absent for smaller values of μ_{ω^*}

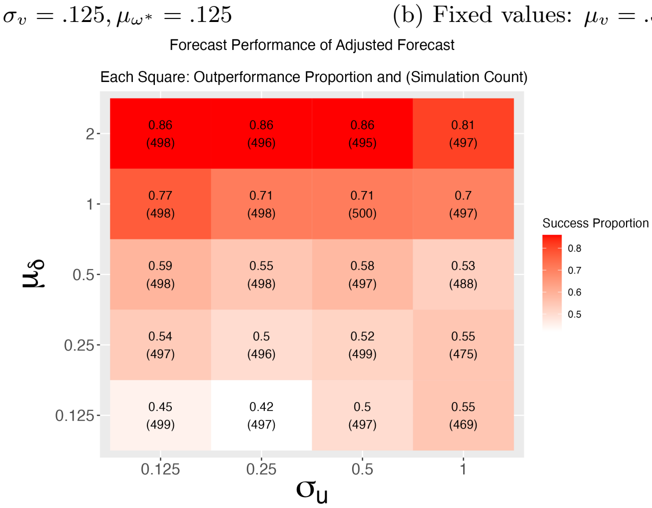
Signal and Noise



(a) Fixed values: $\mu_v = .125, \sigma_v = .125, \mu_{\omega^*} = .125$



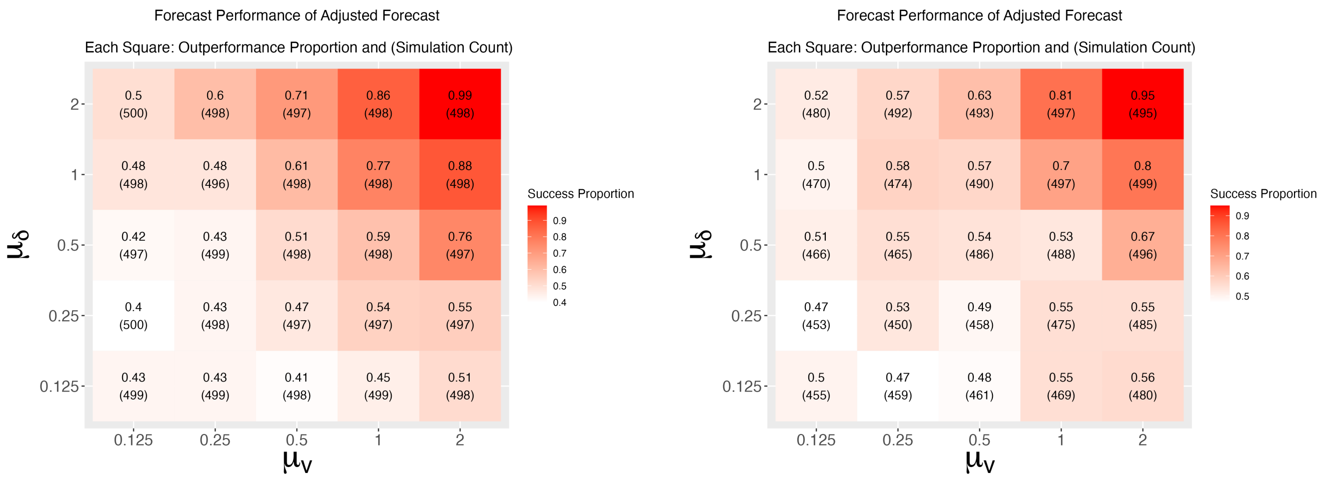
(b) Fixed values: $\mu_v = .5, \sigma_v = .125, \mu_{\omega^*} = .125$



(c) Fixed values: $\mu_v = 1, \sigma_v = .125, \mu_{\omega^*} = .125$

Figure 2.2: In the progression from 2.2a to 2.2b to 2.2c, we see that increasing μ_δ leads to improved performance of the adjusted forecast, and this effect is intensified as μ_v increases. However, it is not consistently true that an increasing noise undermines the performance of the adjusted forecast. Rather, that phenomenon requires both large quantites of μ_δ and μ_v.

Interaction between Signal and Volatility Profile Means

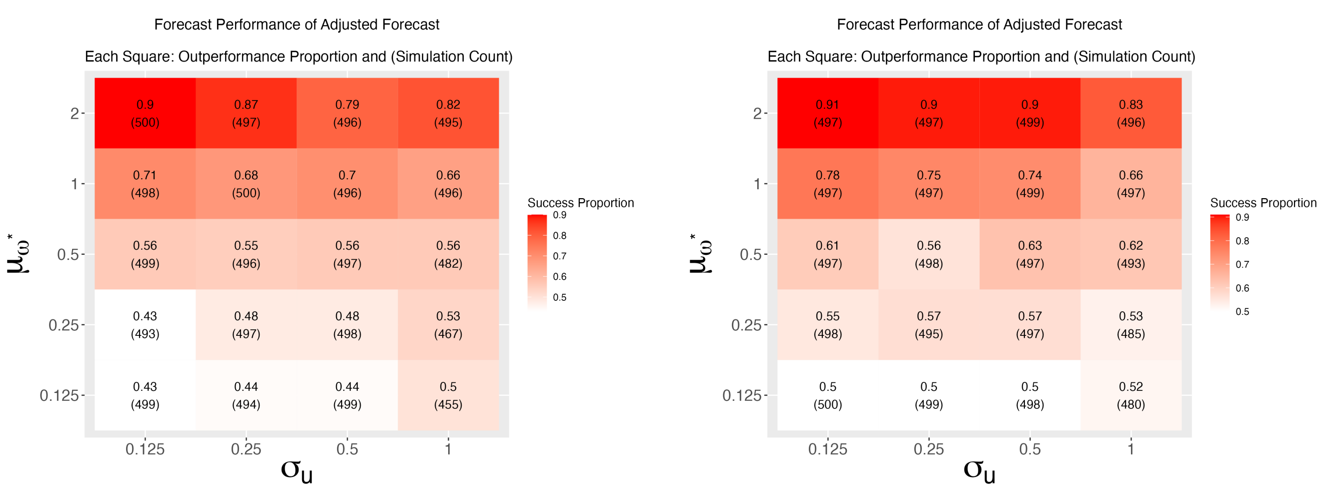


(a) Fixed values: $\sigma_v = .125, \mu_{\omega^*} = .125, \sigma_u = .125$

(b) Fixed values: $\sigma_v = .125, \mu_{\omega^*} = .125, \sigma_u = 1$

Figure 2.3: We compare the interaction of μ_δ and μ_v at two different levels of shock noise. In the low-noise regime, the peak performance of the adjusted forecast is higher, and the ascent is faster along both dimensions. However, in the high-noise regime, the adjusted forecast performs well even at low levels of μ_δ and μ_v.

Interaction between Shock Intercept and Shock Noise



(a) Fixed values: $\mu_v = .125, \sigma_v = .125, \delta = .125$

(b) Fixed values: $\mu_v = .125, \sigma_v = .125, \delta = 2$

Figure 2.4: The interaction between the intercept μ_{ω^*} and σ_u suggests that the parameters behave as expected. However, a larger value of δ in 2.4b attenuates the effect of increasing noise.

2.6 Forecasting after the results of the 2016 US presidential election

We demonstrate the applicability of our methodology using a real data set at the intersection of financial trading and electoral politics. The goal of this analysis will be to provide a forecast for the volatility of IYG[1] (an ETF composed of American financial majors JPMorgan, Bank of America, etcetera) on Wednesday November 9th, 2016 using only the information available at the close of US financial markets at 4pm ET on November 8th, 2016, either after Donald Trump won the 2016 US presidential election or in a scenario in which Trump would win the day. Our specifications for this analysis are outlined below:

1. **Model choice** We assume a GARCH(1,1) for the daily log return series of IYG in each donor. As argued in **?**, a GARCH(1,1) is rarely dominated by more heavily-parameterized GARCH specifications. It thus provides a defensible choice when motivation or time for choosing another model is lacking. For the time series under study and the donor series alike, we fit a GARCH(1,1) on almost four years of market data prior to the shock.

2. **Covariate Choice** We choose covariates that could plausibly satisfy the model assumptions spelled out earlier, that is, risk-related and macroeconomic covariates that could plausibly be weighted and summed in a shock distribution. We thus choose the log return Crude Oil (CL=F), the VIX (VIX) and the log return of the VIX, the log returns of the three-month, five-year, ten-year, and thirty-year US Treasuries. Finally, we include the squares of the demeaned log return of IYG for the three trading days preceding the shocks, in an effort to capture the 'local' ARCH effects prior to the shocks. For each variable in the volatility profile, we compute the sample mean and sample standard deviation across the $n + 1$ events, allowing us to scale the variables to have zero mean and unit variance. Hence, no single variable can dominate the distance-based weighting procedure.

3. **Donor pool construction** We choose the three most recent US presidential elections prior to the 2016 election as well as several Brexit-related events in the first half of 2016. The inclusion of the three US presidential elections is straightforward and easily justified. Those three elections are the only presidential elections since the advent of the ETF IYG, and it is not unreasonable that three elections held between 2004 and 2012 would share a conditional shock distribution with the time series under study in 2016. We exclude the midterm congressional elections in the US (i.e. those held in even years not divisible by four), which generate far lower voter turnout and feature no national races. The inclusion of Brexit-related donors is less straightforward but perhaps more interesting for our method. We first explain the three donors and then justify their inclusion. The first Brexit-related donor is former UK Prime Minister David Cameron's announcement, on February 20th, 2016, that a referendum would occur to decide the UK's future in the EU [**?**]. The second is a poll concluded on June 13th, 2016 in the UK reflecting higher support for Brexit than had been thought [**?**]. Last is the Brexit vote itself, which occurred on June 23rd, 2016, with polls closing that evening and markets reacting the next morning when the final vote count became known [**?**]. Even though the movement to withdraw from the EU was a long-running, complex phenomenon that occurred outside of the United States, nevertheless the Leave campaign possessed a very similar right-wing populism to that of the Trump campaign [**?**]. Furthermore, Brexit preceded the 2016 US presidential election by less than five months and was, according to several news outlets, seen as a harbinger of Trump's victory [**??**]. This view was made clear by Anne Applebaum who stated: "I do realize that it's facile to talk about the impact on a U.S. election which is still many months away, that it's too simple to say "first Brexit, then Trump." But there is a way in which this election has to be seen, at the very least, as a possible harbinger of the future. This referendum campaign, as I wrote a few days ago, was not fought on the issues that are normally central to British elections. Identity politics trumped economics; arguments about "independence" and "sovereignty" defeated arguments about British influence

[1] It has been noted that GARCH effects are more attenuated in aggregated returns [**?**], which suggests against using the S&P 500 or similar indices as an example.

and importance. The advice of once-trusted institutions was ignored. Elected leaders were swept aside. If that kind of transformation can take place in the U.K., then it can happen in the United States, too. We have been warned" [?]. However, it is important to note that the Leave decision foretelling a Trump victory was not a consensus view [??]. In our forecasting setting, these three Brexit donors will be an imperfect donors, as will the preceding US presidential elections which feature different candidates and campaigns.

4. **Choice of estimator for volatility** We use the sum of squared five-minute log returns of IYG on November 9th, 2016, otherwise known as the Realized Volatility (RV) estimator of volatility [?], as our proxy. We exclude the first five minutes of the trading day, resulting in a sum of 77 squared five-minute returns generated between 9:35am and 4pm.

5. **Data Sources** All daily market data is provided via the YahooFinance API available in the quantmod package in R [?]. In order to calculate log equity returns, we use Adjusted Close of IYG. The realized volatility is computed using high-frequency quote data available from Wharton Research Data Services (WRDS) [?].

We now discuss results displayed in the three subplots in Figure 2.5 in order from left to right. On the left, we see that distanced-based weighting places the largest weights on the 2004 and 2012 elections and Brexit-relate donors, with only modest weight on the 2008 election. Assuming an approximately correct specification of the covariates, this is interpreted to mean that the 2016 US election had a general climate of risk and tension less extreme than 2008 and more similar to conditions of the other donors. The three Brexit donors also have large, similar weights, which is not entirely surprising, as they were temporally proximate both to one another as well as the 2016 US presidential election. In the middle plot, we notice that the fixed effect estimate for the 2008 election is considerable, while the 2012 election and three Brexit donor estimates are modest by comparison, with only 2004 registering a near-zero fixed effect estimate. On the right, we observe in black the series $\hat{\sigma}^2_t$ fitted by a GARCH(1,1) for the time series under study. We also observe four colored points, each listed in the legend: three predictions and the ground truth. We include the prediction derived by adjusting the GARCH(1,1) prediction by the arithmetic mean of the fixed effect estimates. As is evident, our method comes remarkably close the ground truth. The prediction is not only directionally correct, i.e. we predict a volatility spike where there is one; the prediction also outperforms the unadjusted prediction. Remarkably, the arithmetic-mean based prediction here demonstrates the inherent risk in failing to weight each donor appropriately. The 2008 election receives far more weight than is called for, as simple averaging ignores the radically different conditions on the evening of those two events.

Naturally, one might ask how sensitive this prediction is to at least two kinds of model specification: donor pool specification and covariate specification. There are two responses to these concerns. First, although the practitioner lacks a priori knowledge of the adequacy of the donors with respect to the time series under study, it is possible to gauge the diversity of the donor information by examining the singular values of the volatility profile. Second, we follow **?** in a executing a multiverse analysis. In the supplement we carry out leave-one-out analyses on both the donor set and the covariate set. Our multiverse analysis concludes that our adjustment approach beats the unadjusted forecast in every leave-one-out configuration.

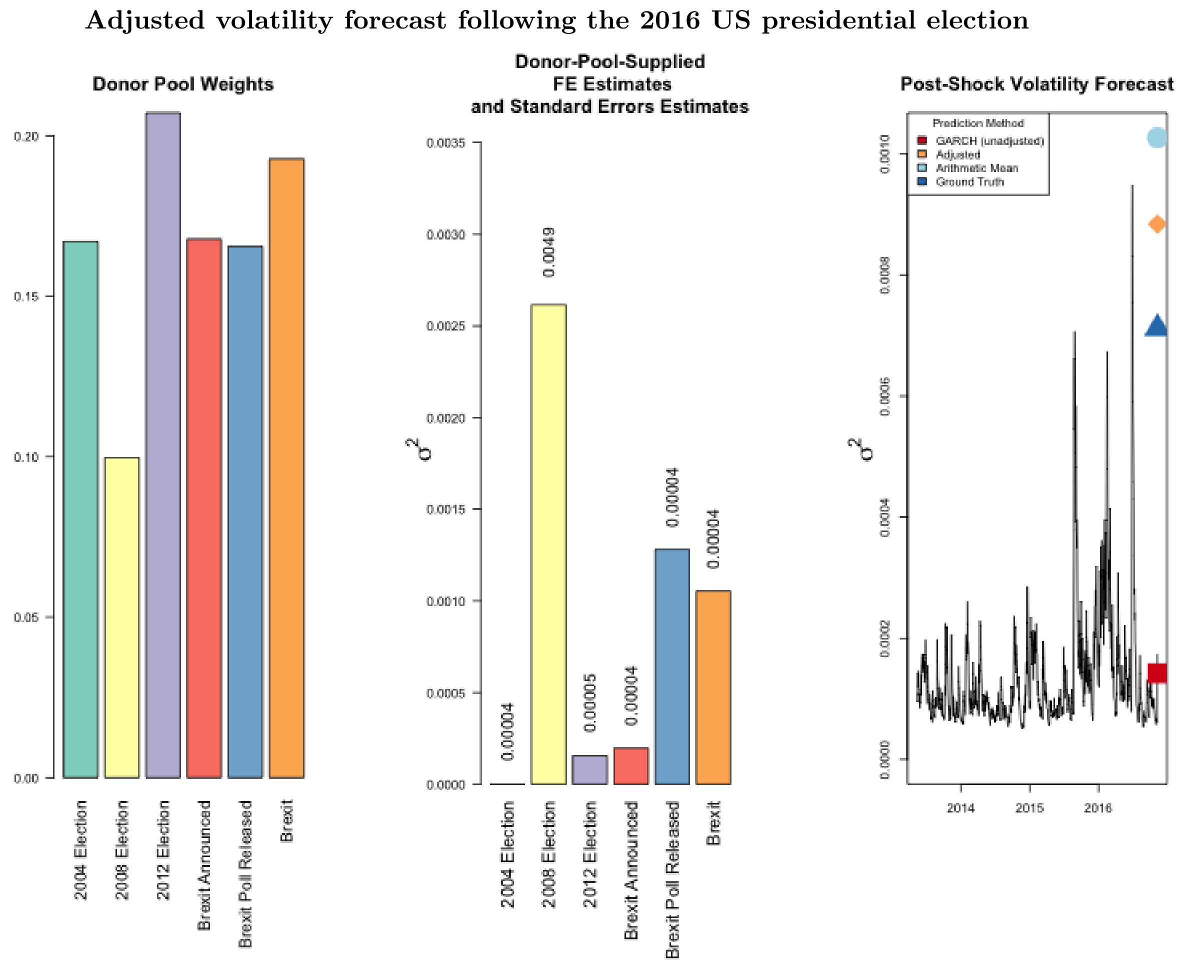


Figure 2.5: Donor pool weights (left panel), individual shock effects (center panel), and forecasts (right panel).

2.7 Discussion

This present work applies and innovates techniques in similarity-based forecasting and information aggregation in order to provide better GARCH forecasts under news shocks. It extends **?** principally by substituting a GARCH model for an AR(1), i.e. by modeling both the mean and volatility of a univariate time series. An interesting insight related to **?** is that the GARCH model, under weak assumptions, can be represented as an ARMA on the squared residuals a_t^2. Hence, a shock $\omega^*_{i,t}$ to the volatility at time t is an identical shock to a_t^2. However, this use-case becomes less attractive in situations where the sign of the time series under study following the shock is uncertain. Insofar as this work has made advances that accommodate heteroskedastic time series, the benefits may redound most amply to applications like Value-At-Risk (VaR) and Expected Shortfall, where $\hat{\sigma}_t^2$ is an input.

This work as well as **?** itself can be viewed as a generalization of **?** where the rare event probability λ is assumed to be 1 and the error associated with the rare event is estimated using external series. In our setting, if the shock to the time series under study were known and yet so underdescribed, i.e. so lacking in qualitative context, then **?** would be recommended. However, in our setting the rare event is not only rare but also contains specific qualitative features that are best accounted for via data aggregation.

The method under development does not strictly require knowledge of the length of the shocks in the donor pool, but correctly sizing up those shock lengths is helpful to proper estimation of the shocks in the donor pool. An important question remains: even if the donor pool shock lengths are assumed to be known, how do we advise the operator to forecast the time series under study? In other words, for how long is the adjustment estimator ω^* valid, applicable, and reliable? One idea suggested by our work is obvious: why not aggregate the shock lengths from the donors as well and round that quantity or take the floor or ceiling of any non-integer value? This is worth pursuing. However, it may be that estimating the persistence of a volatility shock induced by news is an endeavor deserving of its own study, where aggregation methods might naturally arise as helpful tools.

There is also a broader discussion to be had regarding the degree of model heterogeneity permitted in fitting the donors' series. To return to the real data example above, it could be that the 2008 election could be improved as a donor by fitting a more highly-parameterized GARCH, so as to accommodate the unusual market conditions in 2008. We leave this broad topic for future work.

2.7.1 Comparison with KNN

? (cited in **?**) note intercept corrections find their theoretical basis in the occurrence of structural breaks, whereas Nearest-Neighbor methods, being nonparametric, are theoretically more adept at accounting for nonlinearity. The present work examines news shocks, which are more closely related to structural breaks (though though one could easily imagine more exotic conditional shock distributions, including nonlinearity in the covariates). Nevertheless, there are deeper observations to be made about KNN. The method presented here is unlike KNN in that we are not trying to learn a function, first and foremost. We are trying to estimate a parameter. Also, KNN encounters the curse of dimensionality. In contrast, large p is not a problem in Synthetic Control and related methods, since the intermediary object estimated is the vector π with $n-1$ degrees of freedom. For KNN, a high-dimensional space, i.e. large p, corresponding to many covariates, is a difficult space in which to work with distances [**?**]. In contrast, large p is not a problem in and of itself for Synthetic Control — in fact, asymptotic results exist for p [**?**].

As is pointed out in **?**, KNN regression performs well when K is chosen small enough that one can simply average the points $\{y_i\}_{i=1}^{N_{train}}$ in the neighborhood around each element in $\{y_i\}_{i=1}^{N_{test}}$ to get good predictions. As we have noted above, the arithmetic mean-based estimator of ω_1^*, denoted $\overline{\omega^*}$, corresponds to KNN when $K = n$, the number of donors. Fundamentally, the idea that n would be small enough and the donors homogeneous enough that one could simply average the $\hat{\omega}_{i,*}$ is at odds with the assumed variation in the shocks in our modeling setup, and the real data example attests to what can do wrong when that variation is not accounted for.

Additionally, in KNN regression, the hyperparameter K must be learned. In similarity-based parameter correction, the number of donors is not learned. A donor pool is curated, and then careful rules of thumb can be applied to determine whether a given donor should be included or excluded. This brings us to a deeper point about the distinction between similarity-based methods in the style of **?** and KNN. In KNN, the exogenous variables are taken as a given and nearness to the object to be predicted depends on the distance function chosen. In contrast, in **?**, the determination of nearness begins with a qualitative step, i.e. curating the donor units between which we will calculate distances and from which we will ultimately derive weights.

2.7.2 Donor Pool Construction

Should we gather as many donors as possible and pick them quantitatively? It would be counter to the method proposed to assemble a vast number of donors, lacking careful scrutiny of the qualitative fit, and let the optimization simply pick the donors, via weighting. What makes a donor good is not merely its quantitative fit but its qualitative fit as well. **?** make a similar point about large donor pools.

What matters is that the donors chosen are properly situated in the p-dimensional predictor space, so as to allow proper estimation of the weights. That being said there may be subjectivity in selecting donors, and care will be needed to make good selections. In this work we included Brexit and the previous three US presidential elections as donors used for forecasting volatility following the 2016 US presidential election. The resulting adjusted forecast with this donor pool performed very well. In our supplement we show that adjusted forecasts perform well under a variety of donor pool choices and covariate choices.

2.7.3 The Nature and Estimation of Volatility Shocks

Not all of the volatility of an asset return may be related to news [**?**]. This explains our inclusion of an idiosyncratic noise term in the shock specification. However, this point also gestures in the direction of possible unexplained variation in the shocks. **?** find that for predicting 1-minute returns, highly transitory firm-specific news is useful. The authors conclude that news about fundamentals is predictive.

It would be a pyrrhic victory for our method if the volatility profile indeed underlies real-world shocks but the volatility profile is radically high-dimensional or the signal in the shocks is overwhelmed by the noise term. Even unbiased predictions can be unhelpful if they are high in variability. To that point, in Section 2.8.4 discuss the potential benefits of forecast combination. Another possibility is high-frequency data and the use of linear models like HAR [**?**]. HAR would not only increase the sample size available for discovering the autoregressive structure of a series' realized volatility. It would also open the door to high-dimensional regression methods, shrinkage estimators, and more.

2.7.4 Parameter Stability

There is an important question about the stability of the GARCH parameters under the presence of a shock, and on parameter instability as it pertains to forecasting, we refer readers to **??**. There are at least two reasons that we do not explore parameter instability or methods to adjust for it. First, the marginal effect of coefficient changes at the shock time would, under the assumptions in this work, be swamped by the random effect. Second, the estimation of post-shock parameter values would require at least several — better yet, dozens — of post-shock data points in the time series under study, whereas this work assumes access to zero post-shock data points. However, it is possible that similarity-based estimators for the GARCH coefficients could be produced, for example, by adapting the methods of **?**.

2.8 Supplement

Here we provide additional useful information about the foregoing results and analyses. We begin with several sections that re-examine the results of our real data example through a multiverse perspective. We also explore forecast combination as a tool for real data analysis by evaluating the mean and median of the full set of forecasts generated by dropping one covariate and one donor at a time. Finally, we provide proofs for the numerous formal results found earlier in this work.

2.8.1 Leave-one-out: How we analyzed a multiverse of 63 predictions

Given the option of redoing our analysis, with eight covariates and six donors, a natural question from a place of skepticism is, how stable are the results — how contingent are they on a particular specification? To answer these questions, in Table 2.1 we generate all 63 predictions yielded the decision of leave out any one of the donors (or none) and any one of the covariates (or none). This analysis includes the prediction presented above in Figure 2.5, which can be viewed as the null model, at least in the sense that it provides and baseline for comparison.

2.8.2 Sensitivity to Covariates Chosen

Any covariate that appears near to the bottom of Table 2.1 more often is, in a greedy sense, a more important covariate for this prediction task, since by dropping it, a larger loss results. Similarly, any covariate that appears more often in forecasts below the null are covariates that lead to a worse forecast than the null. The covariate that thus stands out in positive way is the squared, demeaned log return of IYG, of which we use the three days preceding the date T_i^*+1. This suggests that the behavior of the observable time series itself (and its transformation) may be a useful proxy for risk conditions, a result wholly consistent with the core consistency result for Synthetic Control found in **?**.

2.8.3 Sensitivity to Donors Chosen

The most glaring result visible in Table 2.1 regarding donor selection is two-fold: on one hand, omitting the donor that is the Brexit Poll of June 13th, 2016 leads to the worst QL Loss quantities, far worse than having the full set of donors. It could very well be that the poll concluded that day attenuates the contribution of the Brexit vote itself, and without accounting for the full variety of Brexit-related events in 2016, a poor prediction results. Hence, the inclusion of all three Brexit donors is justified.

On the other hand, the 2004 election appears most often at the top of the table, suggesting that is an unhelpful donor. This is unsurprising if one considers that the fixed effect estimate for 2004 is nearly zero. That means that by dropping the 2004 election, other donors must account for the approximately 17% weight it had in the null model. The donors that fill that void then have the opportunity to vault the prediction much closer to the ground truth. This discussion hints at a conjecture one might make: that the removal of an unjustified donor will improve forecast performance in expectation, under certain conditions. More fundamentally, however, we should ask why it is that 2004 is such a poor donor, beyond the fact that its fixed effect is small. It could be that the 2004 election delivered such little surprise or is so far distant, temporally, that its inclusion in our donor pool is not fully justified. The potential lesson to practitioners of the method is that even when there appears to be a default donor pool available, some degree of curation and vetting may be beneficial.

2.8.4 Can we combine forecasts to outperform the individual forecasts?

We consider a very simple forecast combination procedure – in fact, we consider two: the average forecast and the median forecast. While forecast combination can be justified in any number of ways, here we invoke forecast combination as a way to robustify forecasts against misspecification. We combine in two simple ways, using the mean of all 63 forecasts and the median of all 63 forecasts. The forecast losses are all specifications – including the average and median forecasts – are presented in Table 2.1.

2.8.5 Formal Results

Proof of Proposition 1 We claim

$$\begin{aligned}
\mathbb{E}[\sigma^2_{i,T_i^*+r+1}|\mathcal{F}_{T_i^*+r}] &= \mathbb{E}[\omega_i + \alpha a^2_{T_i^*+r} + \beta\sigma^2_{i,T_i^*+r}|\mathcal{F}_{T_i^*+r}] && (2.5)\\
&= \omega_i + \mathbb{E}[\alpha(\sigma_{i,T^*+r}\epsilon_{T_i^*+r})^2|\mathcal{F}_{T_i^*+r}] + \beta\sigma^2_{i,T_i^*+r} && (2.6)\\
&= \omega_i + \alpha\sigma^2_{i,T_i^*+r} + \beta\sigma^2_{i,T_i^*+r} && (2.7)\\
&= \omega_i + (\alpha+\beta)\sigma^2_{i,T_i^*+r}\ . && (2.8)
\end{aligned}$$

The volatility equation of a GARCH(1,1) dictates that for any r, the one-step-ahead volatility is given by the expression inside the expectation in (2.5). By the mean-model assumption of a GARCH(1,1), we have $a_{i,t} = \sigma_{i,t}\epsilon_{i,t}$, and hence by substituting $\sigma_{i,t}\epsilon_{i,t}$ for $a_{i,t}$, we arrive at equation (2.6) above. Using the unit-variance assumption on $\epsilon_{T_i^*+r}$, we can compute explicitly the expectation, yielding (2.7). Finally, by rearranging terms, we arrive at equation (2.8). □

Proof of Lemma 1 Since the shock is assumed to arrive uniformly at random for each i, $1 \leq i \leq n+1$, and last for a discrete number of indices, the sequence $\{\omega^*_{i,t}\}_{t=0,\ldots,T_i}$ is governed by a distribution $F_{\{\omega^*_{i,t}\}_{t=0,\ldots,T_i}}$ that is invariant to shifts in time and thus satisfies strong stationarity. □

Proof of Proposition 2.2 The result for each $\hat{\omega}^*_{i,*}$ follows from the consistency proof of the QMLE in GARCH-X models, as established by **?**. As for a convex combination of $\{\hat{\omega}^*_{i,*}\}_{i=2}^{n+1}$, consider the following argument, where we have suppressed the time-indexation:

$$\begin{aligned}
\mathbf{x}_1 &= \sum_{i=2}^{n+1} \pi_i \mathbf{x}_i && \text{(by an assumption of the proposition)}\\
\implies \delta^T \mathbf{x}_1 &= \delta^T \sum_{i=2}^{n+1} \pi_i \mathbf{x}_i = \sum_{i=2}^{n+1} \pi_i \delta^T \mathbf{x}_i && \text{(by left-multiplication by } \delta^T)\\
\implies \delta^T \mathbf{x}_1 &= \sum_{i=2}^{n+1} \pi_i (\omega^*_i - \mu_{\omega^*} - u_{i,t}) && \text{(by substitution and definition of } \omega^*_i)\\
\implies \mu_{\omega^*} + \delta^T \mathbf{x}_1 + \sum_{i=2}^{n+1} \pi_i u_{i,t} &= \sum_{i=2}^{n+1} \pi_i \omega^*_i && \text{(by moving terms to the left-hand side)}
\end{aligned}$$

Now, since each $\hat{\omega}^*_i$ converges in probability to ω^*_i, by Slutsky, any finite linear combination of $\hat{\omega}^*_i$ converges in probability to the corresponding finite linear combination of ω^*_i. Hence, $\sum_{i=2}^{n+1} \pi_i \hat{\omega}^*_i$ converges in probability to the right-hand side of the final line. By setting $u_{i,t}$ to zero for all shock points and all i, we then have $\omega^*_1 = \mu_{\omega^*} + \delta^T \mathbf{x}_1$ on the left-hand side. Without taking the $u_{i,t}$ to be zero at the shock points, however, we nevertheless have convergence in distribution. The proof is as follows. For any collection of $n+1$ i.i.d variables, any convex combination of any n of them is equal in distribution to the one left out. Therefore, the left-hand side of the final line is equal in distribution to ω_1, as desired. □

Proof of Proposition 3 Without loss of generality, we prove the case for $r = 1$. The general case follows from iterative one-step-ahead forecasts. Recall the conditional expectation of the variance for the GARCH-X(m,s) model:

$$\mathbb{E}[\sigma^2_{adjusted,t+1}|\mathcal{F}_t] = \omega_i + \sum_{k=1}^{m_i} \alpha_{1,k} a^2_{1,t+1-k} + \sum_{j=1}^{s_1} \beta_{1,j} \sigma^2_{1,t+1-j} + \gamma_1^T \mathbf{x}_{1,t+1} + \omega^*_{1,*} \ .$$

By replacing parameters with their estimates, we arrive at the prediction

$$\hat{\sigma}^2_{adjusted,t+1}|\mathcal{F}_t = \hat{\omega}_1 + \sum_{k=1}^{m_1} \hat{\alpha}_{i,k} a^2_{i,t+1-k} + \sum_{j=1}^{s_1} \hat{\beta}_{i,j} \hat{\sigma}^2_{1,t+1-j} + \hat{\gamma}_1^T \mathbf{x}_{1,t+1} + \hat{\omega}^*_{1,*} \ ,$$

which converges to

$$\sigma^2_{adjusted,t+1}|\mathcal{F}_t = \omega_i + \sum_{k=1}^{m_i} \alpha_{1,k} a^2_{i,t+1-k} + \sum_{j=1}^{s_1} \beta_{i,j} \sigma^2_{1,t+1-j} + \gamma_1^T \mathbf{x}_{1,t+1} + \omega^*_{1,*}$$

in distribution or probability (depending upon how $\hat{\omega}^*_*$ converges) as $t \to \infty$ in the donor pool by a simple application of Slutsky's Theorem. $\square$

Proof of Proposition 4 First, note that the function $g : (-\infty, \omega^*_i) \to \mathbb{R}$ given by $g(x) = \frac{x}{\sigma^2_{T^*+1} - x} + \log \frac{\sigma^2_{T^*+1} - x}{\sigma^2_{T^*+1}}$ is nonnegative, convex, obtains a minimum at $x = 0$, and being continuous, preserves both convergence in distribution and consistency. The conclusion follows from facts that the model is correctly specified and the convergence of $\hat{\sigma}^2_{adjusted,T^*+1}$ is guaranteed by Proposition (3). $\square$

Table 2.1: Adjusted volatility forecasts for the 2016 US presidential election ranked by QL loss. This table contains all forecasts made by applying simultaneous leave-one-out approaches made to the covariates and the donors forming the donor pool. The forecast from our analysis in the main text is included with a yellow highlight. Mean and median of all forecasts are included, and they are highlighted in gray. The unadjusted forecast is included. It is highlighted in red.

QL Loss of the Adjusted Forecast	Omitted Covariate	Omitted Donor
0.0001	Log Return of ˆIRX	None
0.0007	NA (Average Adjusted Forecast)	NA (Average Adjusted Forecast)
0.0007	Log Return of CL=F	2004-11-02
0.0008	Log Return of ˆIRX	2004-11-02
0.0034	Log Return of ˆTNX	None
0.0060	Log Return of ˆTNX	2004-11-02
0.0061	Log Return of ˆVIX	2016-02-19
0.0077	RawˆVIX	2004-11-02
0.0079	None	2004-11-02
0.0080	Log Return of CL=F	2008-11-04
0.0099	Log Return of ˆVIX	2004-11-02
0.0099	NA (Median Adjusted Forecast)	NA (Median Adjusted Forecast)
0.0115	Log Return of CL=F	2016-02-19
0.0127	None	2016-02-19
0.0137	RawˆVIX	2016-02-19
0.0151	Log Return of ˆTNX	2016-02-19
0.0160	Log Return of ˆTYX	2016-02-19
0.0193	RawˆVIX	2016-06-23
0.0194	Log Return of ˆVIX	2016-06-23
0.0216	Log Return of ˆIRX	2016-06-23
0.0219	None	None
0.0222	Log Return of ˆFVX	2016-02-19
0.0232	Log Return of ˆVIX	2008-11-04
0.0237	Log Return of ˆTNX	2016-06-23
0.0240	Log Return of ˆIRX	2016-02-19
0.0245	Log Return of ˆTNX	2008-11-04
0.0251	None	2016-06-23
0.0280	Log Return of CL=F	2016-06-23
0.0315	Log Return of ˆFVX	2008-11-04
0.0322	None	2008-11-04
0.0360	Log Return of ˆTYX	2008-11-04
0.0393	Log Return of ˆVIX	2012-11-06
0.0396	RawˆVIX	2008-11-04
0.0473	Log Return of ˆIRX	2008-11-04
0.0484	IYG	2004-11-02
0.0499	Log Return of ˆFVX	2016-06-23
0.0542	Log Return of ˆTYX	2016-06-23
0.0578	IYG	2012-11-06
0.0595	Log Return of ˆVIX	None
0.0596	RawˆVIX	None
0.0598	Log Return of ˆTYX	None
0.0598	IYG	None
0.0598	Log Return of ˆFVX	None
0.0604	Log Return of ˆTYX	2004-11-02
0.0627	Log Return of ˆFVX	2004-11-02
0.0651	Log Return of CL=F	None
0.0692	Log Return of ˆVIX	2016-06-13
0.0758	Log Return of ˆIRX	2012-11-06
0.0922	IYG	2008-11-04
0.0924	Log Return of ˆTNX	2012-11-06
0.0990	None	2012-11-06
0.1038	RawˆVIX	2012-11-06
0.1117	Log Return of ˆIRX	2016-06-13
0.1147	Log Return of CL=F	2012-11-06
0.1241	RawˆVIX	2016-06-13
0.1356	IYG	2016-06-23
0.1359	Log Return of ˆTYX	2012-11-06
0.1480	Log Return of ˆFVX	2012-11-06
0.1621	None	2016-06-13
0.1630	Log Return of ˆTNX	2016-06-13
0.1928	Log Return of CL=F	2016-06-13
0.2102	Log Return of ˆFVX	2016-06-13
0.2435	Log Return of ˆTYX	2016-06-13
0.3013	IYG	2016-02-19
1.0261	IYG	2016-06-13
2.3529	All	All

Chapter 3

Forecast Adjustment Under Shocks: A Unification

3.1 Introduction

In time series and panel data, sometimes the salient challenge is not predicting when an event will occur or begin but what its key properties will be. In the familiar case of scalar time series, those properties can include the time series' post-event direction, magnitude of change, moments, duration of the event, or correlation structure, all over an arbitrary horizon or perhaps multiple horizons. This is not to say that predicting the arrival of an event is easy. In some cases, it may be difficult or impossible. The event might not even be conceivable to many before it happens, as was the case with the COVID-19 pandemic. On the other end of the spectrum are events whose arrival need not be predicted because they are scheduled in advance, e.g. scheduled news, like releases of economic data or election results. In those cases, conditional forecasts allow a statistician to proceed as if the event has arrived. Consider the analysis of Mike Wilson, Morgan Stanley's CIO and Chief US Equity Strategist, in October 2024 regarding the upcoming 2024 US Presidential US [**?**]:

> While some argue a Trump win would be a headwind for growth and equity markets, due to tariff risks and slowing immigration, we think there's an additional element from the 2016 experience that is worth considering—rising animal spirits. More specifically, in 2016 Trump's pro-business approach led to the largest 3-month positive impact on small business confidence in the past 40 years. It also translated into a spike in individual investor sentiment. It appears to me that markets may be trying to front-run a repeat of this outcome as Trump's win in 2016 came as a surprise to pundits and markets alike.

Wilson here assesses how market participants are incorporating current information as well as past forecast failures, including the potential overweighting of more recent failures. His assessment implies that the shock conditional on a Trump victory may disappoint observers, as it has already been pulled forward into current forecasts. In short, this time around may be different. Of course, it may be that participants have merely pulled *too much* forward by correctly assessing the conditional impact of a Trump victory but overestimating its likelihood, or vice-versa. Regardless, what Wilson is engaging in is what **?** identify as a crucial reaction to uncertain territory: what, if anything, does it resemble from the past?

Herein we attempt to unify a range of conceptual approaches to forecasting amid shocks that have developed across the broad ecosystem of the econometric and forecasting literatures. We attempt to locate our main contribution, Similarity-based Parameter Correction, at the intersection of several important strands of thought. We in fact delineate a specific type of SPC called post-shock forecasting, which is induced by nuanced and interesting choices that fall under the general SPC framework. In particular, this current manuscript focuses on model adjustment amid news shocks that undermine the reliability of the model at hand. Forecasting under shocks raises unavoidable questions: should the forecast model be abandoned in favor of a discretionary or ad-hoc or one-off adjustment? Does the does the discretion of a forecaster rule out a quantitative method for making the adjustment? What is the ultimate purpose of the adjustment, and how it is to be used? For how long is the adjustment necessary or reliable?

Forecast model adjustment received its earliest attention under the auspices of "intercept correction" [**????**]. Of fundamental importance is the distinction between discretionary and automated intercepts corrections. **?** define scalar intercept corrections to be automated when they follow the simple rule of adding an estimation or prediction residual e_t to subsequent (possibly but not necessarily all) forecasts $\hat{f}_{t+1}, \hat{f}_{t+2}, [...]$. "*[S]etting the model back on track*" is the informal label they provide for that procedure.

The term intercept correction can be construed more broadly to encompass the correction of terms besides the intercept, and justification for it can be found beyond the original goal of setting forecasts back on track. In later work, **?** present seven additional interpretations for intercept correction, and **??** themselves note the possibility of correcting a coefficient in a forecast model specification. **?** develop a similarity-based procedure for correcting the parameter β using past information of the time series itself. For a review of more recent work in similarity-based forecasting, we refer the reader to **?** and references therein.

What if the break in the DGP is brought about by a very particular kind of event, a news shock? What if we could predict well the intercept shift that occurs between T^* and T^*+1? In **?**, the authors list six necessary conditions for reacting to what we in this current paper call a news shock. Of greatest interest to us are two of the six:

- "the forecasting model already embodies that source of information", and
- "there is an operational method for selecting an appropriate model".

For intercept shifts that occur between T^* and T^*+1, forecasting procedures have been explored in **??**, where the AR(1)-X and GARCH(m, s) cases, respectively, are treated. Both works target additive parameters in scalar time series, predicting the additive parameter in the time series under study by aggregating information from other time series. The authors leave several stones unturned, including a more general, dare say comprehensive treatment of forecasting under shocks.

As the literature has developed, a current practitioner of forecast adjustment can now choose (i) procedures that are discretionary or automated, (ii) adjusting using internal data (i.e. from the time series itself) or external data, (iii) the term to be corrected (e.g. intercept, coefficients), if any, (iv) as well as the specification of the correction function (i.e. the mapping from the donor unit data to the correction term in the time series under study), including the aggregation mechanism applied to the assembled data (e.g. Nearest-Neighbor, arithmetic mean, kernel methods). We discuss this in depth in Section 3.5.

If intercept correction is the correction of a forecast by adding the residual at time T^*+1 to forecasts at some subset of $\{T^*+2, T^*+3, ...\}$, in this manuscript, we are doing something rather close to that by adding the predicted correction term to each forecast in some some subset of $\{T_1^*+1, T_1^*+2, ...\}$, where T_1^*+1 is first integer index for the time series under study following the shock. The particular procedure presented, post-shock forecasting, is a discretionary procedure for intercept correction that integrates data internal or external to the time series under study in a systematic manner. It is a discretionary procedure in the sense that substantive choices must be made at each instance of use. The correction function uses the notion of similarity at multiple levels: first, to identify similarity between breaks in different time series, and second, in borrowing from the causal inference literature to aggregate information from those similar circumstances. Parametric specifications for location shifts in the DGP can be found in **?**, where locations shifts are modeled as induced by additional observable variables, as well as in **?**, where the arrival of the news shock is observable, even if its quantitative instantiation is not. Beyond **??**, we are not

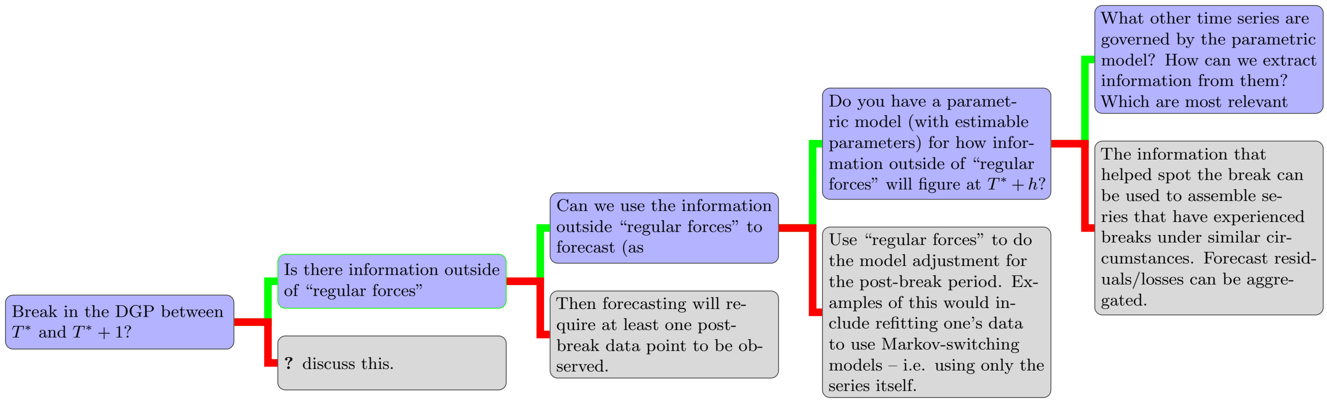


Figure 3.1: Forecast Model Adjustment: A Decision Tree

aware of prior work that proposes a method for aggregating similar shocks occurring outside the time series under study in order to correct the time series under study.

The structure of this manuscript is as follows. We provide a deeper and more critical review of several literatures, including intercept correction, model-evaluation, and similarity-based forecasting. We then provide a canonical setting in which we attempt to focus our work. This setting will be presented at a level of generality that showcases the broad applicability of our method. We then introduce the method from a global perspective, abstracting away from familiar applications. We then show several specific applications, including novel applications that cannot be found elsewhere in the literature. We close with a discussion, including possible extensions.

3.2 Forecasting Amid Shocks

[**?**, p. 177] characterize breaks in the DGP (and hence the misspecification of the default forecasting method) as the "most obvious and best understood rationale" for intercept correction. In this section, we wish to locate the task of forecasting amid shocks within the broader literature on forecasting under breaks in the DGP.

Some terminological notes are in order before we proceed, however. Errors are a staple of stochastic modeling. They go by the synonyms "noise", "shocks", or "innovations" as well. The meaning of these terms may not differ as much in their mathematical representations in symbols but in how they are interpreted (but see **?** on the ambiguity in the concept of an information set $\mathcal{F}_t$). Whereas in the psychometric and social-scientific fields, errors may represent unmeasured or latent capacities of a unit, in other fields like the natural sciences, errors may be used to account for variability in the instruments used to collect data. "Innovations" or "shocks", in econometric literature, are generally understood to represent the arrival of previously unknown (but not necessarily uncontemplated) information. Shocks are called structural when they are can be linked back to some key feature of an scientific (usually economic) theory; otherwise, they are idiosyncratic. Additionally, there exist other subcategories of shocks that crosscut the distinction just made. For example, a news shock is a shock in which the information material to the market is delivered via newsmedia and moves markets no earlier than the news release itself. Shocks can also be permanent or transitory, supply-related or demand-related, all-at-once or have a sequential, 'slow drip', or sporadic quality to them. The taxonomy is vast. We use the term "shock" as a special case of a break in the DGP, even though, strictly speaking, some shocks will neither result from a break in the DGP nor herald the emergence of a new regime. The news shocks examined in this manuscript, however, should be considered breaks in the DGP, no matter whether they are transient or permanent. Additionally, in this work, we will be focused on all-at-once news shocks that occur strictly between two discrete time points, or information available at a "fractional lag" [**?**], so that a practitioner is faced with the question of what action is most advisable for the following time point, given the news shock.

3.2.1 The Role of Outside Information

When is more information better? That is a central question with regard to forecasting. In **?** (cited in **?**), the authors note that when forecasts fail, it often has to do with location (i.e. intercept shifts). Therefore, using additional information is not likely to help unless it can address potential location shifts. We should clarify what we mean by “additional information”. Do we mean any information not included in the default forecast function? Could that include something as straightforward as additional lags of the time series being forecast? Perhaps a more useful distinction to make is that between information internal to the time series being forecast and information external to the time series being forecast. Of course, the usefulness of this distinction will hinge, in part, on whether we can cleanly partition information into internal and external buckets. A cross-cutting distinction originates in **?**, which discusses necessary conditions for forecasting under breaks as well as a taxonomy of information available for doing so. The authors make a fundamental distinction between “regular forces”, which includes the typical standbys of economic theory, and information potentially relevant to location shifts.

We wish now to sharply distinguish between essential and inessential concepts. For example, whether information is relevant to location shifts or not is a worthy and complex question that depends upon numerous factors, including but not necessarily limited to the DGP. We prefer not to conflate that distinction with the internality-externality distinction, or whether the information belongs to the traditional sort of information used in econometric modeling, though certain tendencies and general patterns may emerge. We make clear what is essential in Table 3.1.

Table 3.1: A 2x2 Schema of Forecast Information, With Examples

	Conventional Econometric Models	Outside Conventional Econometric Models
internal	Lags of the series itself; past shocks	Polynomial expansion of the feature space and other transformations without solid theoretical motivation
external	Macro variables like interest rates, commodity prices; weather-related variables	Google Trends, high-frequency data like prediction markets; past shocks under similar conditions

3.2.2 The Meaning and Use of Similarity

This manuscript aims to contribute at the intersection of post-break forecasting and similarity-based forecasting. In particular, we use the former to motivate the latter. The notion of similarity appears in various statistical contexts, including matching, synthetic control, nearest-neighbor methods, not to mention the massive area of approximation theory. Similarity-based reasoning can also be given a philosophical treatment when viewed as a special case of analogical reasoning [**?**], from which we can glean a useful general framework in which reasoning by analogy involves a source domain and a target domain. Our target domain, simply put, contains the thing we’d like to know, while the source domain is our reservoir of information from which to draw.

Our objective as statisticians is to make forecast corrections by exploiting information from previous time series that have undergone shocks under similar circumstances. As a provisional definition, we say that a break in the DGP of a time series is similar to a break in another just in case the causes of the breaks share key, material qualities. Note here that we have not said anything about the time series themselves (the details of their DGPs or their moments) being similar, nor the entities that they represent (e.g. the firms, financial instruments, governments, etcetera) being similar. Why that is so will become clear in Section 2.2.

Having provided only a provision definition, we would like to introduce useful distinctions regard the concept of similarity. First, quantitative ways of determining similarity naturally tend toward the idea that similarity is reducible to determining what serves as a best approximation. Those approaches include matching a target object to "source" objects along one or more variables, where unique matches or exact matches may be likely or may be statistically certain, likely, unlikely, or impossible, subject to the setting. In metric spaces, the notion of a metric that quantifies nearness is crucial to fundamental concepts of proximity. This suggests a set of optimization problems such that for any target object Q, the search for similarity means a search for an approximator or a collection of approximators A such that $\delta(A, Q)$ is smallest, where δ is an appropriate metric.

The notion of approximating a target does not necessarily mean that the target need be similar to the approximator. This suggests a role for asymmetric distance functions δ (which fail to meet the methematical requirements of a metric) for exploring differences between donors and weighting their contributions accordingly. Consider an asymmetric Jaccard Index $\delta_{\mathcal{J}}(\cdot, \cdot)$ that maps from any two finite-cardinality sets A, B to the interval $[0, 1]$ [**?**], penalizing A more for its incongruence with B than B is penalized for its incongruence with A. Asymmetry also plays an important role in weighting schemes far beyond traditional forecasting, including Transformer mechanisms inside the field of deep learning. The asymmetry inherent in query, key, and value matrices allow hierarchical relations in language to be represented mathematically [**?**, p. 364].

So far, we have discussed only binary measures of similarity. The distance-based weighting scheme found in **?** and repurposed in **?** can be thought of as an n-ary similarity measure, where the n weights means little in isolation but can tell a greater story when viewed alongside the approximation they produce. However, an n-ary similarity measure will not always reduce to the task of constructing an approximator to the target using n original objects. We might instead insist on a notion of similarity such that n objects are said to be similar, as a collection, to the target if and only if each of the n matches the target perfectly.

What about qualitative ways of determining similarity? In **??**, donors are identified by matching the qualitative aspects of the shock in the time series under study to shocks previously witnessed in other series. Only after the donor pool has been identified based on qualitative facts about the shocks is there a role for quantitative methods. The role of the covariate vector in their models is two-fold: the covariates appear in the shock distribution, and the covariates also help determine which donors are most important. It is to a generalization of this procedure that we now turn.

3.3 Setting

We will suppose that a researcher has a multivariate time series data $\mathbf{y}_{i,t} = (y_{i,t}, \mathbf{x}_{i,t})$, $t = 1, \ldots, T_i$, $i = 1, \ldots, n+1$, where $y_{i,t}$ is scalar and $\mathbf{x}_{i,t}$ is a vector of covariates such that $\mathbf{x}_{i,t}|\mathcal{F}_{i,t-1}$ is deterministic and observed. Suppose that the analyst is interested in forecasting $y_{1,t}$, the first time series in the collection, which we will denote *the time series under study*. We require that each time series $\mathbf{y}_{i,t}$ be subject to an observed news event following T_i^* and before witnessing $T_i^* + 1 \leq T_i^*$. The reasons for this represent the heart and soul of the framework proposed. We are not proposing a method to correct for unanticipated shocks that have been reflected in the observed data. Instead, we are proposing a method to correct for shocks for which the qualitative kernel of information is known and the observed quantitative instantiation of which is not yet known.

We assume the following general data-generating process

$$\mathcal{M}_1 \colon \begin{aligned} y_{i,t} &= F(\mathcal{F}_{i,t-1}) + \alpha_{i,t} + \epsilon_{i,t} \\ \alpha_{i,t} &= \mathbf{x}_{i,t}^T \lambda_{i,t} \\ \mathbf{x}_{i,t}^T &= (1, x_{i,t}^1, ..., x_{i,t}^p) \\ \lambda_{i,t}^T &= (u_{i,t}, \lambda_t^1, ..., \lambda_t^p), \end{aligned} \tag{3.1}$$

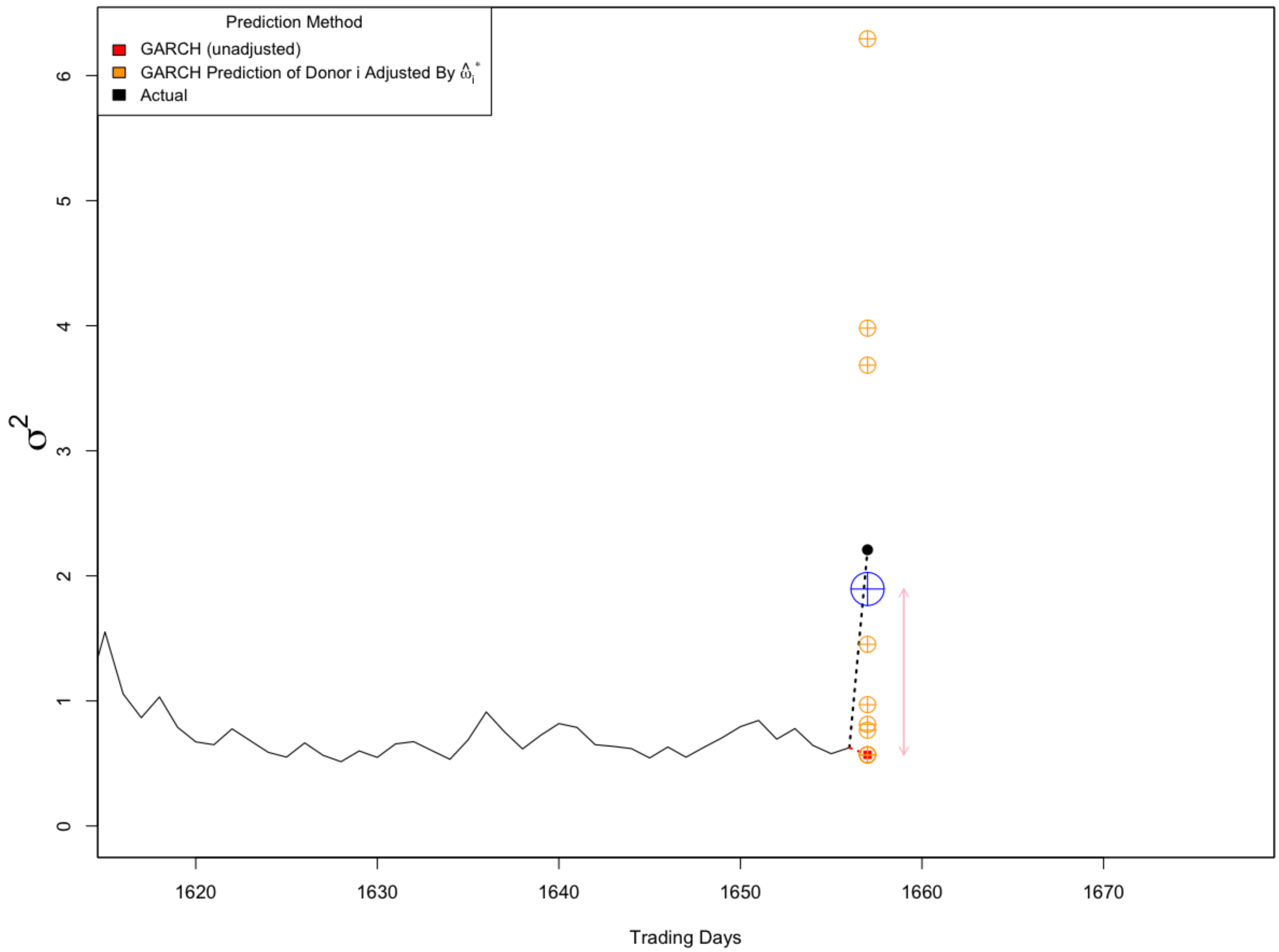


Figure 3.2: The simulated time series experiences a volatility shock between trading days 1,656 and 1,657. The GARCH prediction, in red, fails even to approach the volatility spike at $T^* + 1$, as do several adjusted predictions, which are orange. In contrast, the GARCH forecast adjusted by $\hat{\omega}^* = \sum_{i=2}^{n+1} \pi_i \hat{\omega}_i^*$, a convex combination of the estimated shocks in the donor pool, achieves directional correctness as well as a smaller absolute loss in its prediction. The pink vertical line serves to indicate the adjustment of size $\hat{\omega}^*$ that allows the blue bullseye to approach more closely the ground truth.

with time-varying and observable covariate vector $\mathbf{x}_{i,t}^T$, time-varying and unknown parameter vector

$$\lambda_{i,t} \sim \mathcal{F}_{\lambda} \text{ with } \mathbb{E}_{\mathcal{F}_{\lambda}}(\lambda) = \mu_{\lambda_t}, \mathrm{Var}_{\mathcal{F}_{\lambda}}(\lambda) = \Sigma_{\lambda_t},$$

and time-invariant error structure

$$\epsilon_{i,t} \overset{iid}{\sim} \mathcal{F}_{\epsilon} \text{ with } \mathbb{E}_{\mathcal{F}_{\epsilon}}(\epsilon) = 0, \mathrm{Var}_{\mathcal{F}_{\epsilon}}(\epsilon) = \sigma^2_{\epsilon},$$

where F maps objects from the past into future objects of $y_{i,t}$, and $\epsilon_{i,t}$ uncorrelated across time and donors. Note that the dot product $\alpha_{i,t} = \mathbf{x}_{i,t}^T \lambda_{i,t}$ includes the term $u_{i,t}$ that is not shared among donors. It encodes the known possibility that the shocks will vary across donors both in ways that can predicted based on pre-shock covariates as well as in ways that cannot be predicted. That the parameter vector is time-varying allows us to capture that most time points are without news shocks (in which case $\alpha_{i,t}$ is of neglible effect), but conditional upon information arriving between T_1^* and $T_1^* + 1$, we may know with near certainty that λ_{1,T_1^*+h} will be nonzero in norm for some $h > 0$. Note that the specification (3.1) above is sufficiently capacious as to include data-generating processes of the form

$$y_{i,t} = \log Y_{i,t} = \log a_{i,t} y_{i,t-k} e^{\alpha_{i,t} + \epsilon_{i,t}} = \log a_{i,t} + \log y_{i,t-k} + \alpha_{i,t} + \epsilon_{i,t}$$

with $k \geq 1$ and $a_{i,t}, y_{i,t}$ supported on $\mathbb{R}^+$.

The term $\mathbf{x}_{i,t}^T \lambda_{i,t}$ will be deemed the *correction term* for donor i at time t, and the function

$$\xi \colon \mathcal{F}_{2,T_2^*+h} \times \ldots \times \mathcal{F}_{n+1,T_{n+1}^*+h} \to \mathcal{F}_{1,T_1^*+h}$$

that estimates or predicts α_{1,T_1^*+h} will be deemed the *correction function*. The correction function maps donor information — be it observables or inferences — to the prediction function for the time series under study.

So far we have reviewed a long and rich literature about model adjustment, and we have also introduced a data-generating process upon which the rest of this manuscript will be based. We now proceed to introduce a particular framework of solutions to the circumstances assumed in Section 2.2. What we are doing is distinct in at least two ways. First, we are incorporating outside information faster, and second, we are proposing a principled, systematic way to do it that defies that curse of dimensionality. If one wanted to forecast using linear models and high-dimensional information set, then some kind of method for reducing model complexity would be required. We avoid that.

3.4 SPC Forecasting Methodology and Correction Functions

We now present two one-step-ahead forecasts for the time series under study. The first is the unadjusted forecast. The second is the adjusted forecast, which differs by the predicted correction term. The forecasts are:

$$\begin{aligned} &\text{Forecast 1: } \hat{y}_{unadjusted,T_1^*+1} = \hat{\mathbb{E}}[\mathcal{F}_{T_1^*}] \\ &\text{Forecast 2: } \hat{y}_{adjusted,T_1^*+1} = \hat{\mathbb{E}}[\mathcal{F}_{T_1^*}] + \hat{\alpha}_{T_1^*+1} \ . \end{aligned}$$

The problem of aggregating estimated shocks or residuals begins with the data constraints. Let us first introduce useful notation. Let $\hat{\alpha}_{i,*}$ denote estimated correction term for donor i. Taking the estimated shocks or residuals as a given, we essentially observe the tuple of information $(\{\hat{\alpha}_{i,*}\}_{i=2}^{n+1}, \mathcal{F}_{2,T_2^*} \times \ldots \times \mathcal{F}_{n+1,T_{n+1}^*}\})$. One approach to aggregation comes from **?**, where weights belonging to the simplex, $\{\pi_i\}_{i=2}^{n+1} \in \Delta^{n-1}$, are chosen via a distance-based weighting procedure, resulting in the correction function

$$\hat{\alpha} = \sum_{i=2}^{n+1} \pi_i \hat{\alpha}_{i,*},$$

which is just one way among many to build a correction function. Another interesting correction function can be found in **?**, where percentage residuals of past forecasts are used. For yet another example, consider a correction function that aggregates the survey panel of forecasts by the Survey of Professional Forecasts [**?**]. An obvious advantage of human forecasters is that they are able to integrate information revealed arbitrarily close to the scheduled release of the forecast, even if human forecaster's lack of systematicity or closed-form forecasts, for example, leads to non-optimal forecast performance in general. By calculating forecast differences between the low-frequency information forecast, $\tau_{i,t} = \hat{y}_{unadjusted,t} - \hat{y}_{i,t}$, we prepare ourselves for late-breaking events – e.g. shocks – for which we might want a correction function

$$\xi \colon \boldsymbol{\tau} \to \mathcal{F}_{1,T_1^*}$$

that weights the i-th forecaster according to that forecaster's past performance or according to the similarity of the forecaster's stated information set to the information available for the default forecasting model.

3.4.1 Diagnostics for Similarity-Based Parameter Correction

The assumptions we make on the DGP for shocks to both donors and the time series under study allow us to evaluate our prediction tool even before we predict. In this subsection, we will elaborate on a permutation method proposed in **?**. The idea is simply to use each of the n donors as an object to predict, while the remaining $n-1$ serve as actual donors. Permutation tests for inference were proposed as early as **?** in the context of synthetic control. Although we will not embark upon a formal exploration of the permutation method for predictive model evaluation, we will note briefly its motivation, chiefly that the donors share a DGP and the model particulars are invariant to permutation. To the extent that those assumptions hold in practice, model performance can be gauged by methods as simple as evaluating the losses of the 'pseudo-time-series-under-study' to the more sophisticated, like assessing the correlation between the goodness of approximation and the losses. One can also assess the plausibility of the $\{\hat{\alpha}_i\}$ being uncorrelated, which is central to MSE-minimization, as we will see in Section 3.6.1.

3.5 Model Adjustment Using Similarity-Based Parameter Correction: A Global Overview

In this section, we introduce and discuss a general framework for model adjustment that both generalizes and is motivated by the circumstances laid out in Section 2.2 by boiling down SPC to its essential five elements. The essential elements of similarity-based parameter correction are

1. **Object-to-predict** Most fundamentally, the method requires a random object (indexed over time and possibly space, as well) that obeys a specification with additive errors, or, at the very least, a specification that can be transformed to have additive errors. This requirement is suitably weak, so as to include models that are not linear or not linear in each of their parameters. It also includes multidimensional objects as well as objects from non-Euclidean probability spaces, like some function spaces.

   Notice that we did not say *parametric* specification. The reason for this is that, given any forecast function f from an information set $\mathcal{F}_t$, and given any appropriate loss function L, we can define a residual $e_t = y_t \odot \hat{y}_t$, where $\odot$ is subtraction in the simple case of mean squared-error. The residuals (or transformations of those residuals) of those n models can be weighted as part of a correction function. This fact is especially useful when the forecast function f is nonparametric, black-box, or stochastic with respect to $\mathcal{F}_t$.

2. **Common Model Family on the Shocks** The method requires that residuals be governed by a model that is shared across all units. This ensures that in the estimation of news shocks in the donor pool, the estimators will enjoy similar properties that will produce a good aggregated shock estimator. This condition is satisfied by the parametric shock distributions found in **??**.

3. **Reliable and Shared Model-Fitting Procedure** There must exist a reliable model-fitting procedure for the $n+1$ units, one that will allow the assumption of a shared family for the residuals to be reflected in the residuals. Here, reliable could mean any number of several things. It means that the estimation procedure must produce a credible description of each data-generating process or a good prediction function, so as to aid in estimating residual(s) in each unit. Reliable may mean the estimators have low variance, and it may also mean that the estimation procedure is robust, to some degree, to misspecification bias. However, in theory, the lack of these properties is not harmful to the method unless the lack of these properties harms the correction term's estimation. When we use fixed effect estimation (under ordinary assumptions), we can construct confidence intervals for the fixed effect estimates, and then assuming independence, we can get confidence intervals for convex combinations of fixed effect estimates.

4. **Reliable Correction Term Estimation** Fourth, there must exist a reliable procedure for modeling and estimating the correction term in each unit. Again, here we care about low variance as well as robustness to misspecification. The very simplest correction term is the residual itself. Alternatively, the correction term could be an inner product that depends upon external covariates, as is found in **??**.

   This might not always be straightforward. Some models like GARCH, for example, might deliver very noisy estimates for indicator variables that occur just once.

5. **Reliable Correction Function Estimation** There must exist a correction function (presumably based on the correction term) that maps data from the donor pool to the *predicted* correction term in the time series under study based on some notion of similarity. In some cases, there may exist a posited DGP that the correction term estimates. In other cases, there may be no posited DGP. Similarly, if there exists a posited DGP for the correction term, it may depend upon data internal to the time series under study, or it may depend on external data, e.g. outside data could be aggregated (e.g. Nearest-Neighbor, arithmetic mean, kernel methods) to estimate the correction function.

3.5.1 The Method of Multiple Shocks

Sometimes, we might have reason to believe that a forecast requires adjustment due to not one but multiple shocks. **?** explored a "dual" shock in their analysis of oil company equity price shocks due to the March 2020 COVID lockdowns. Here we discuss two ways of navigating that interesting challenge.

Method 1: For each source of the k shocks, assemble n_j donors and p_j covariates, with $j = 2, ..., n_j + 1$, and compute weights $\{\pi_{i,j}\}_{i=2}^{n_j+1}$. With these k weight sets, a set of k aggregated shock estimators can be constructed and averaged (or aggregated in some other way). Ultimately, each weight set results in a shock adjustment terms, and the sum of those estimated terms becomes the final adjustment term, as is done in **?**. One potential downside of this approach is that is has no in-built way to control the size of the final adjustment term, which is somewhat in tension with original idea of taking convex combinations of estimated shocks, and therefore it can be seen as a relaxation of the weight constraint.

Method 2: Each of the k shocks can be represented using an application of distance-based weighting, where the vector δ and the vector $\mathbf{x}_{i,T_i^*+1}$ can encode a partition of shocks by using zeros in appropriate entries. Under this method, the sum of the weights remains 1, the estimation space for the weights is increased to kp. Unlike **Method**

1, this approach remains committed to a weight constraint. An obvious concern is that a space of dimension kp may be too large for weight estimation.

In practice, the best approach may depend on k, p as well as the task at hand. Additionally, an assumption of independence or at least uncorrelatedness among the k donor pools may helpful here. The foregoing, we hope, provides a basis upon which future work can build.

3.6 Formal Properties and Model-Specific Considerations

In this section, we discuss particular models for which our approach has been implemented as well as others for which it has yet to be implemented. For those for which an implementation exists, we introduce the model in a more general light, while also commenting on the model-specific considerations.

3.6.1 ARIMA

Recall Specification (3.1), which models a news shock as an affine function of covariates:

$$y_{i,t} = F(\mathcal{F}_{i,t-1}) + \alpha_{i,t} + \epsilon_{i,t}$$

We begin our recounting of model-specific cases by recalling **?**, in which it was established that for a family of AR(1)-X-distributed scalar time series with a common shock distribution, forecast risk can be reduced via a similarity-based adjustment procedure. An AR(p) is an easy model to work with, in part because it can be consistently estimated via OLS, and an AR(p) can approximate an ARMA(p,q) with arbitrarily-small error through via a truncation of an AR(∞) representation. Setting aside the estimation method, consider a one-step-ahead forecast for an AR(p) at time t with an adjustment estimator, $\hat{\alpha}_t$. For ease of exposition, we suppress subscripts for donors:

$$\hat{y}_t = f(\mathcal{F}_{t-1}) = \hat{\mu} + \sum_{k=1}^{p} \hat{\rho}_k y_{t-k} + \hat{\alpha}_t \tag{3.2}$$

which can be understood as an AR(p) with time-varying coefficients. The mean squared-error of the forecast above, i.e.

$$\begin{aligned}
\mathbb{E}[(y_t - \hat{y}_t)^2] &= \mathbb{E}[(y_t - \hat{\mu} + \sum_{k=1}^{p} \hat{\rho}_k y_{t-k} + \hat{\alpha}_t)^2] && (3.3)\\
&= \mathbb{E}[(\mu + \sum_{k=1}^{p} \rho_k y_{t-k} + \alpha_t + \epsilon_t - (\hat{\mu} + \sum_{k=1}^{p} \hat{\rho}_k y_{t-k} + \hat{\alpha}_t))^2] && (3.4)\\
&= \mathbb{E}[((\mu - \hat{\mu}) + \sum_{k=1}^{p} (\rho_i - \hat{\rho}_k) y_{t-k} + (\alpha_t - \hat{\alpha}_t) + \epsilon_t)^2] && (3.5)
\end{aligned}$$

admits of the typical bias-variance decomposition. If it is known that $\mu = 0$, (3.5) becomes

$$\mathbb{E}[(\sum_{k=1}^{p} (\rho_k - \hat{\rho}_k) y_{t-k} + (\alpha_t - \hat{\alpha}_t) + \epsilon_t)^2]$$

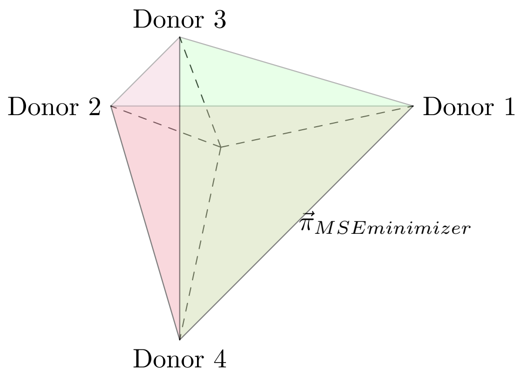


Figure 3.3: The 3-Simplex, Δ^3, where hypothetical minimizer is a convex combination of Donors 1 and 4.

$$
\begin{aligned}
=&\text{Bias}^2[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}+(\alpha_t-\hat{\alpha}_t)]+\text{Var}[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}+(\alpha_t-\hat{\alpha}_t)]+\sigma^2_\epsilon && (\epsilon_t \text{ idiosyncratic})\\
=&\big[\mathbb{E}[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}+(\alpha_t-\hat{\alpha}_t)]\big]^2+\text{Var}[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}+(\alpha_t-\hat{\alpha}_t)]+\sigma^2_\epsilon && (\text{Definition of Bias})\\
=&\big[\sum_{k=1}^{p}(\rho_k-\mathbb{E}[\hat{\rho}_k])y_{t-k}+(\alpha_t-\mathbb{E}[\hat{\alpha}_t])\big]^2+\text{Var}[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}]+\text{Var}[\hat{\alpha}_t]+\sigma^2_\epsilon\\
& && (\text{if donor shocks are uncorrelated with time series under study})\\
=&\big[\sum_{k=1}^{p}(\rho_k-\mathbb{E}[\hat{\rho}_k])y_{t-k}+(\alpha_t-\mathbb{E}[\hat{\alpha}_t])\big]^2+\text{Var}[\sum_{k=1}^{p}(\rho_k-\hat{\rho}_k)y_{t-k}]+\sum_{i=2}^{n+1}\pi_i^2\text{Var}[\hat{\alpha}_{i,t}]+2\sum_{i\neq j}\pi_i\pi_j\text{Cov}[\hat{\alpha}_{i,t},\hat{\alpha}_{j,t}]+\sigma^2_\epsilon
\end{aligned} \tag{3.6}
$$

Assume we have no control over the accuracy of the estimates $\{\hat{\rho}\}_{k=1}^{p}$ nor any control over the estimates $\{\hat{\alpha}_{i,t}\}_{i=2}^{n+1}$. Then our only control over the MSE comes comes from the weight vector $\vec{\pi}$. **?** conditions under which $\vec{\pi}$ is unique, but when uniqueness fails, then a practitioner has meaningful choices to make regarding forecast risk. If we insist that $\hat{\alpha}_t$ be an unbiased predictor, then $\vec{\pi}$ may not be sparse, and hence the covariance terms in 3.6 may be large. However, if some bias can be tolerated, the variance and covariance terms may be small. All of this is to say that control of the MSE may require tight control of the variance in the prediction function, which in turn will gesture in the direction of a sparse $\hat{\alpha}_t$ and sparse estimation of the autoregressive structure.

How does the MSE of the adjusted forecast compare with the unadjusted forecast? Since the unadjusted forecast MSE would lack the final two sums in expression 3.6, the adjusted model therefore has lower MSE if and only if the reduction in the squared bias is greater than $\sum_{i=2}^{n+1}\pi_i^2\text{Var}[\hat{\alpha}_{i,t}]+2\sum_{i\neq j}\pi_i\pi_j\text{Cov}[\hat{\alpha}_{i,t},\hat{\alpha}_{j,t}]$.

The finite-sample prediction performance of our method, as discussed above, will hinge at least partly on questions of donor independence and uncorrelatedness, as well as the extent to which bias can be minimized through a suitable convex combination of the donors. As for asymptotic performance, those things still matter. We will now offer three theorems that, in their totality, bolster the credibility of the forecast adjustment method on offer here.

Proposition 5. *Assume*

1. *For each i, $1\leq i\leq n+1$, let $\{y_t\}_{i=1}^{T_i}$ follow an AR(p)-X as laid out in Section 3.6.1.*
2. *Assume for each i, $1\leq i\leq n+1$, the shocks $\alpha_{i,t}$ are uncorrelated across donors.*
3. *Assume for each i, $1\leq i\leq n+1$, the shocks $\alpha_{i,t}$ are uncorrelated with α_{i,T_i^*+1}*

Then the tuple of estimators $(\hat{\rho}_{i,1},...,\hat{\rho}_{i,p},\hat{\alpha}_{i,T^*+1})$ is consistent as $t\to\infty$.

Proposition 6. *Assume*

1. *All conditions listed in Proposition 5.*
2. *There exist weights* $\{\pi_i\}_{i=2}^{n+1} \in \Delta^{n-1}$ *such that* $\boldsymbol{v}_{1,T_1^*} = \sum_{i=2}^{n+1} \pi_i \boldsymbol{v}_{i,T_i^*}$.
3. *For all* i, *the* $\{u_{i,t}\}$ *are equal in distribution.*

Then the aggregated estimator $\hat{\alpha_{T_1^*+1}}$ *converges in distribution to* $\alpha_{T_1^*+1}$ *as* $t \to \infty$. *Furthermore, if the* $\{u_{i,t}\}$ *are constant with probability 1, the convergence is in probability.*

Proposition 7. *Let* $\{\hat{y}_{1,T_1^*+r}\}_{r=1}^{h}$ *denote the vector of adjusted predictions (adjusted through* h *steps ahead) in the time series under study. Assume all conditions listed in Propositions 5 and 6. Then* $\{\hat{y}_{1,T_1^*+r}\}_{r=1}^{h} \xrightarrow{d} \{y_{1,T_1^*+r}\}_{r=1}^{h}$. *Furthermore, if the* $\{u_{i,t}\}$ *are constant with probability 1, the convergence is in probability.*

When forecasting financial time series using the method above, the question of prices versus returns (or in the language of economics or engineering, levels versus flows) is relevant. Consider, for example, the persistence of a shock. A shock to an asset's return series at time t implies a persistent shock to the asset's price series at time t and beyond, as one can verify by inverting the discrete-differencing operation on a time series. However, if one forecasts the price series of the asset directly, that persistence beyond the first instantiation of the shock must be modeled explicitly. This suggests interesting opportunities in the analysis and forecasting of shock persistence for situations in which level-forecasting is appropriate, e.g. for stationary time series.

It must be noted here that adding a past residual to future forecasts assumes implicitly that whatever caused the residual has some persistence. Hence, intercept correction might be thought more appropriate for time series of levels, or at least times series of rates that are not necessarily mean-reverting, like inflation. Of course, if one has a strong grasp of how long the intercept correction will be necessary, then all of that is moot, but in practice, knowledge of that persistence may be harder to come by than foreknowledge of the need for intercept correction.

We now turn to the limitations of linear time series model class like ARIMA. The first is that linearity is a nontrivial restriction, and for the purpose of modeling financial time series, applications of SPC to TAR and SETAR would be welcome. Additionally, the efficient market hypotheses implies that market prices reflect all publicly-available information, which in turn implies that ARIMA models should be of no predictive value for market returns. This critique, in a technical sense, alleges that market returns are random walks centered at zero, and for that reason, the coefficients of an ARMA approximation to a financial return series are uniformly zero, and that any incipient deviation from that equilibrium would not go unnoticed by market participants, who would exploit those deviations and push the coefficients back to zero, leaving only idiosyncratic noise. Such concerns about forecasting in markets with heavy feedback loops motivates our next section.

3.6.2 GARCH

The methods of **?** are adapted in **?** to forecast volatility. This is in response to concerns we raised above about the usefulness of ARIMA models for forecasting financial returns as well as concerns about heteroskedasticity. Financial time series are known to have both time-varying volatility as well as volatility that *clusters*, i.e. with volatile trading periods that bunch together. A GARCH model, under weak conditions, is an ARMA model on the squared residuals of a time series, and empirical evidence strongly suggests that those ARMA (i.e. GARCH) coefficients are not zero, in general. The predictability of either those squared returns or the predictability of σ_t^2 (in this context, the two tasks reduce to one) is bounded, however, by a function of the kurtosis of a GARCH process [**?**]. That the GARCH model accommodates excess kurtosis in both the unconditional and conditional returns of a financial asset is a feature of the model, and perhaps for that reason, it has been noted that GARCH models do a much better job at modeling the evolution of volatility (a descriptive or inferential objective) as opposed to forecasting volatility.

In this way, **?** bridged a gap between the descriptive and predictive roles of GARCH. Fixed effects (correction terms) are estimated in the donor pool, and those fixed effects are aggregated via corrected function. A forecast adjustment is then made in the time series under study. Consider how GARCH fits into our Specification (3.1). First consider, for an arbitrary multivariate time series indexed by i, the volatility equation of a GARCH-X model:

$$\sigma_{i,t}^2 = \omega_i + \sum_{k=1}^{m_i} \alpha_{i,k} a_{i,t-k}^2 + \sum_{j=1}^{s_i} \beta_{i,j} \sigma_{i,t-j}^2 + \gamma_i^T \mathbf{x}_{i,t} \ . \tag{3.7}$$

which corresponds to our Specification (3.1)

$$y_{i,t} = F(\mathcal{F}_{i,t-1}) + \alpha_{i,t} + \epsilon_{i,t}$$

The autoregressive component of (3.7) is clear, as is the fixed effect. The most likely question a newcomer might have is where the idiosyncratic error can be found in (3.7). This can be answered by noting that the GARCH model is a deterministic specification for the variance function; alternatively, it is well know that under weak assumptions, the GARCH model can be rearranged and interpreted as an ARMA.

3.6.3 HAR

The Heterogeneous Autoregressive (HAR) model for realized volatility [**?**] uses ordinary least-squares (OLS), WLS, or some other appropriate linear model to fit a predictive regression model. It explores the influence of previous volatility on current volatility, of course, but uses a specification motivated by theory: that a heterogeneous collection of market participants interacts at different frequencies. Therefore, the volatility can be well-modeled as a linear combination of the average volatility over the previous day, previous week, and previous month (usually taken to be 22 days). Formally, HAR assumes that the log price of an asset is a governed by a diffusion process (a model beyond our scope herein), and HAR estimates

$$RV_t = \beta_0 + \beta_{\text{1-day}} RV_{t-1} + \beta_{\text{5-day}} \overline{\{RV_{t-1}, ..., RV_{t-5}\}} + \beta_{\text{22-day}} \overline{\{RV_{t-1}, ..., RV_{t-22}\}} + \epsilon_t$$

Recently, **?** find that overnight news improves the predictive ability of HAR regressions with a logarithmic transformation on the realized volatility estimate. The authors distinguish between earnings-related and non-earnings-related news occurring after hours, finding that predictive performance is improved for one-day-ahead forecasts but no more. Consider now how HAR fits into Specification (3.1):

$$\mathcal{M}_1 \colon \quad \begin{aligned} & RV_{i,t} = \beta_0 + \beta_{\text{1-day}} RV_{i,t-1} + \beta_{\text{5-day}} \overline{\{RV_{i,t-1}, ..., RV_{i,t-5}\}} + \beta_{\text{22-day}} \overline{\{RV_{i,t-1}, ..., RV_{i,t-22}\}} + \alpha_{i,t} + \epsilon_t \\ & \alpha_{i,t} = \mathbf{x}_{i,t}^T \lambda_{i,t} \\ & \mathbf{x}_{i,t}^T = (1, x_{i,t}^1, ..., x_{i,t}^p) \\ & \lambda_{i,t}^T = (u_{i,t}, \lambda_t^1, ..., \lambda_t^p), \end{aligned} \tag{3.8}$$

with time-invariant error structure

$$\epsilon_{i,t} \overset{iid}{\sim} \mathcal{F}_\epsilon \text{ with } \mathrm{E}_{\mathcal{F}_\epsilon}(\epsilon) = 0, \mathrm{Var}_{\mathcal{F}_\epsilon}(\epsilon) = 1,$$

$$
\begin{aligned}
&\epsilon^*_{i,t} \overset{iid}{\sim} \mathcal{F}_{\epsilon^*} \text{ with } \mathrm{E}_{\mathcal{F}_{\epsilon^*}}(\epsilon) = \mu_{\epsilon^*}, \mathrm{Var}_{\mathcal{F}_{\epsilon^*}}(\epsilon^*) = \sigma^2_{\epsilon^*},\\
&u_{i,t} \overset{iid}{\sim} \mathcal{F}_u \text{ with } \mathrm{Var}_{\mathcal{F}_u}(u) = \sigma^2_u,\\
&\epsilon_{i,t} \perp\!\!\!\perp \epsilon^*_{i,t} \perp\!\!\!\perp u_{i,t}.
\end{aligned}
$$

We should emphasize that the regressions under discussion here are predictive regressions, with no suggestion that the model is fundamentally driven by the linear specification that is estimated. However, if it were, we would especially desire answers to the following

1. Are the relevance and novelty filters used by **?** appropriate, and does it represent a qualitative decision? Does it represent an instance of similarity-based matching?
2. Could we relax the requirement on the relevance score and instead let a numerical algorithm determine which stories and relevant and how relevant?
3. Could we relax the requirement on the novelty score?
4. Bias-variance tradeoff: are we able to reduce the bias or the variance or both via the method of **?**?
5. What is the 'frontier' that separates using a similarity-based approach like SPC as opposed to an approach using count variables?
6. What are advantages and disadvantages of using counts versus indicator variables versus alternatives? Can we confirm the theory that count variable approaches reduce variance while, as is well-established, SPC reduces bias?
7. Does SPC do better when the shock is a nonlinear function of the covariates compared to a method that uses regression terms for each covariate? Does SPC do better under those circumstances compared to count variable-based approaches?
8. What DGP on overnight firm-specific news would most justify the approach in **?**?
9. Is earning-related and non-earnings-related the appropriate disaggregation to use?
10. Is it appropriate to use counts at each time point, or would indicator variables be superior, or rather, under what conditions would indicator variables be superior?

3.6.3.1 Real Data Application of Shock-Adjusted HAR Forecasting

In the autumn of 2008, as Great Financial Crisis (GFC) gathered steam, the US Federal Reserve reduced its overnight lending rate to banks to between 0 and 25 basis points [**?**]. Though the US economy (and the globe as a whole) returned to growth by 2009, the US Fed kept interest rates there for more than half a decade. Only in 2015 did rates begin their gradual and well-telegraphed ascent. Between 2015 and 2018, the Fed would raise rates at nine meetings, culminating in the December 2018 hike [**?**]. In this real data example, we use the eight hikes preceding December 2018 as donors in order to predict the volatility induced December 2018 rate hike.

1. **Model choice** We proceed with a log-HAR model, that is, a HAR model with the realized volatility under the natural log transformation. The logarithmic transformation for RV was suggested as early as **?** and is justified by the variance-stablizing effect of the log.
2. **Covariate Choice** We choose covariates that could plausibly satisfy the model assumptions spelled out earlier, that is, risk-related and macroeconomic covariates that could plausibly be weighted and summed in a shock distribution. We thus choose from four sets of covariates:

- The most recent daily changes in VIX, IRX (the 13-week Treasury Bill rate), XAU (the spot price of an ounce of gold bullion)
- The most recent closing level of the VIX
- The change in the volume SPY on the most recent trading day
- The credit spread between AAA-debt bonds and BAA-rated debt

3. **Donor pool construction** We choose as donors the eight rates hikes occurring in the span 2015 to 2018 (inclusive), excepting our time series under study. We should note here that in each of these eight donors, the rate hike was a fully observed event, whereas in the time series under study, we proceed under the assumption that a rate hike will occur. This represents a more refined and realistic application of the method presented herein, where the shock in the time series under study has not yet occurred, but we nevertheless proceed under the assumption that it will.

4. **Choice of estimator for volatility** We use the sum of squared five-minute log returns of SPY on December 18th, 2018, otherwise known as the Realized Volatility (RV) estimator of volatility [**?**], as our proxy.

5. **Data Sources** All daily market data is provided via the YahooFinance API available in the quantmod package in R [**?**] and the R package highfrequency **?**. The realized volatility estimates of SPY is provided exclusively by the highfrequency R package. The credit spread is calculated using data Federal Reserve Economic Data (FRED) [**?**].

In Figure 3.4, we observe that our distance-based weighting procedure identifies the December 2015 and March 2018 hikes as the only two relevant donors with respect to the December 2018 meeting. The fixed effect estimates in the middle panel range from large and positive to modestly negative, in contrast to the estimates in 2.1. The GARCH model, under the implementation in the garchx package bounds estimates below by zero. In the panel on the right, we observe the unadjusted and adjusted predictions, as well as our estimate of the ground truth. Although the adjusted prediction does miss the volatility spike by a large margin, it appears that this is nearly the best the assembled donors could do, given the donor pool we have to work with. In that sense, this is a prediction task where the time series under study has covariates far outside the support of the donor data, which poses a serious and relevant challenge for our method, which cannot extrapolate.

3.6.4 "Exponential Breaks"

In **?**, the authors discuss intercept shifts that decay deterministically. Those shifts are parameterized by a scalar η and a decay rate ψ. We adapt their model for our own purposes:

$$y_{i,t} = \alpha_i + \eta[1 - e^{-\psi[t-T_i^*]}]\mathbf{1}_{t\geq T_i^*+1} + \epsilon_{i,t}, \epsilon_{i,t} \overset{iid}{\sim} [0, \sigma^2] \ . \tag{3.9}$$

Such a specification captures the dynamics of a shock that is realized over an infinite horizon but which can be practically approached as a permanent shock. This poses a novel case for our method, in that the model is parametric, and yet without estimability of all shock parameters, aggregation may not be straightforward or easy. Consider, for any i, $2 \leq i \leq n+1$, the estimates $\hat{\alpha}_{i,OLS}$ obtained via regression on the first $T_i^* + 1$ points as well as $\hat{\eta}_{OLS}$ obtained by estimating a fixed effect at $t = T_i$. Consider also a subsequent estimator, an estimator for ψ, that is derivable from inverting the model specification and plugging in the estimates $\hat{\alpha}_{i,OLS}, \hat{\eta}_{OLS}$:

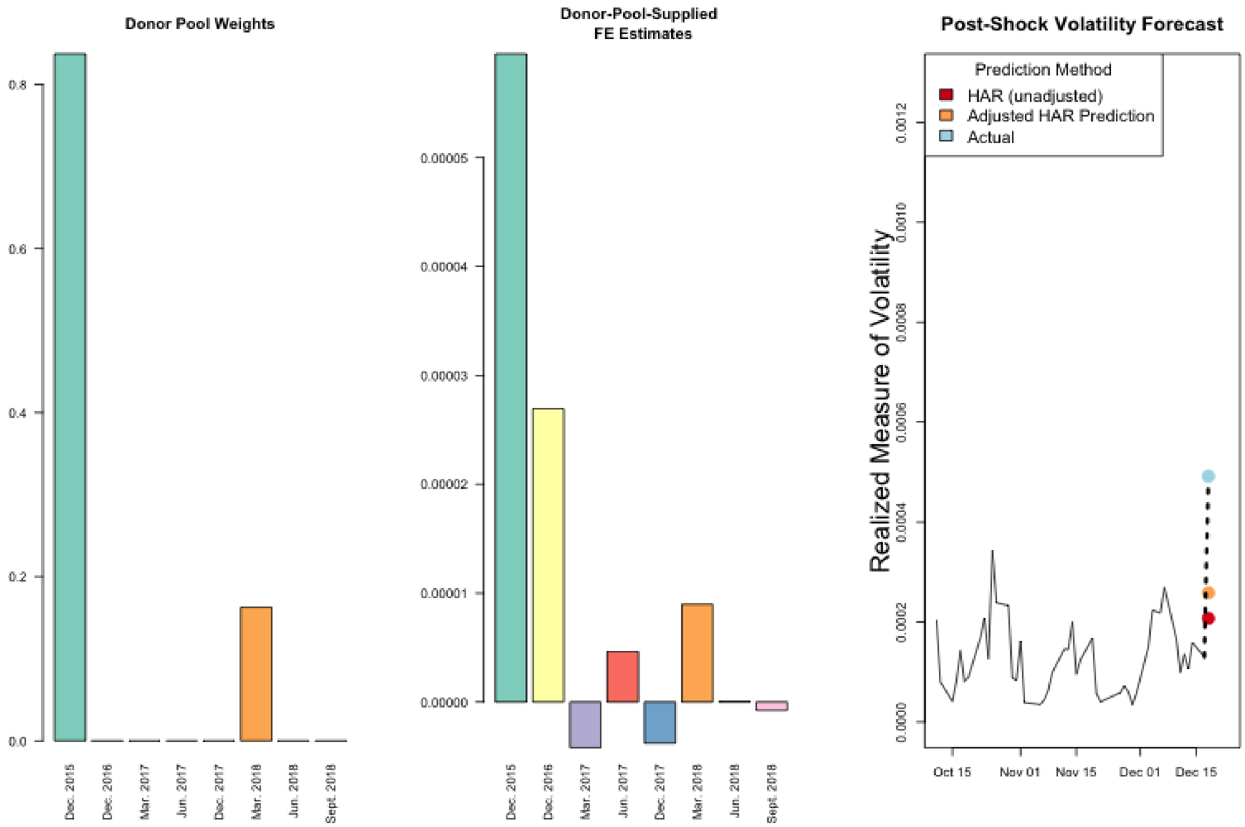


Figure 3.4: Over the course of the 2010s, the US FOMC began a rate-hiking cycle. Left pane: weights of the donor pool of FOMC meetings; right-pane: fixed effect estimates; right pane: the unadjusted forecast, adjusted forecast, and the estimate of the volatility provided to us by realized volatility.

$$\hat{\psi}_{i,plugin} = \frac{1}{T_i - T_i^*} \sum_{t \geq T_i^*+1}^{T_i} -log\big[\big(1 - \frac{y_{i,t} - \hat{\alpha}_i}{\hat{\eta}}\big)\big]/[t - T_i^*] \ . \tag{3.10}$$

(3.10) is, to put it lightly, a complicated estimator with potentially dubious properties. To begin, (3.10) is defined only when the term inside the logarithm is positive for each $t \geq T_i^* + 2$. We could get around this by dropping any summands for which that term is nonpositive and instead calculate an average over only the remaining subset of summands. Additionally, variance estimates for $\hat{\psi}_{i,plugin}$ may require resampling methods or the delta method. Estimation error from $\hat{\alpha}$ and $\hat{\eta}$ would naturally propagate to $\hat{\psi}_{i,plugin}$, and the expected estimation error would depend upon the size of the pre-shock sample.

? report that MSFE is lower by using intercept correction instead of non-linear methods to estimate ψ. In our simulations in the next section, we are going to employ non-linear estimation.

Now consider a practitioner in need of a conditional forecast for $T_1^* + h, h \geq 1$. Could aggregating the prediction residuals $r_{i,T_i^*+h} = y_{T_i^*+h} - \hat{y}_{i,T_i^*+h}, h \geq 1$ lead us to a superior predictions in the time series under study?

Consider also these complicating scenarios

1. We can estimate λ using a cross-sectional regression, and that may be better estimator than doing it one-by-one in each donor.

2. Here we have taken λ to be fixed and shared across donors. If λ is a random, however, then estimating it becomes harder.

3.6.4.1 Simulation Study

We now investigate the exponential decay model introduced just above. We model the decay parameter ψ as a dot product of donor-specific and donor-invariant quantities.

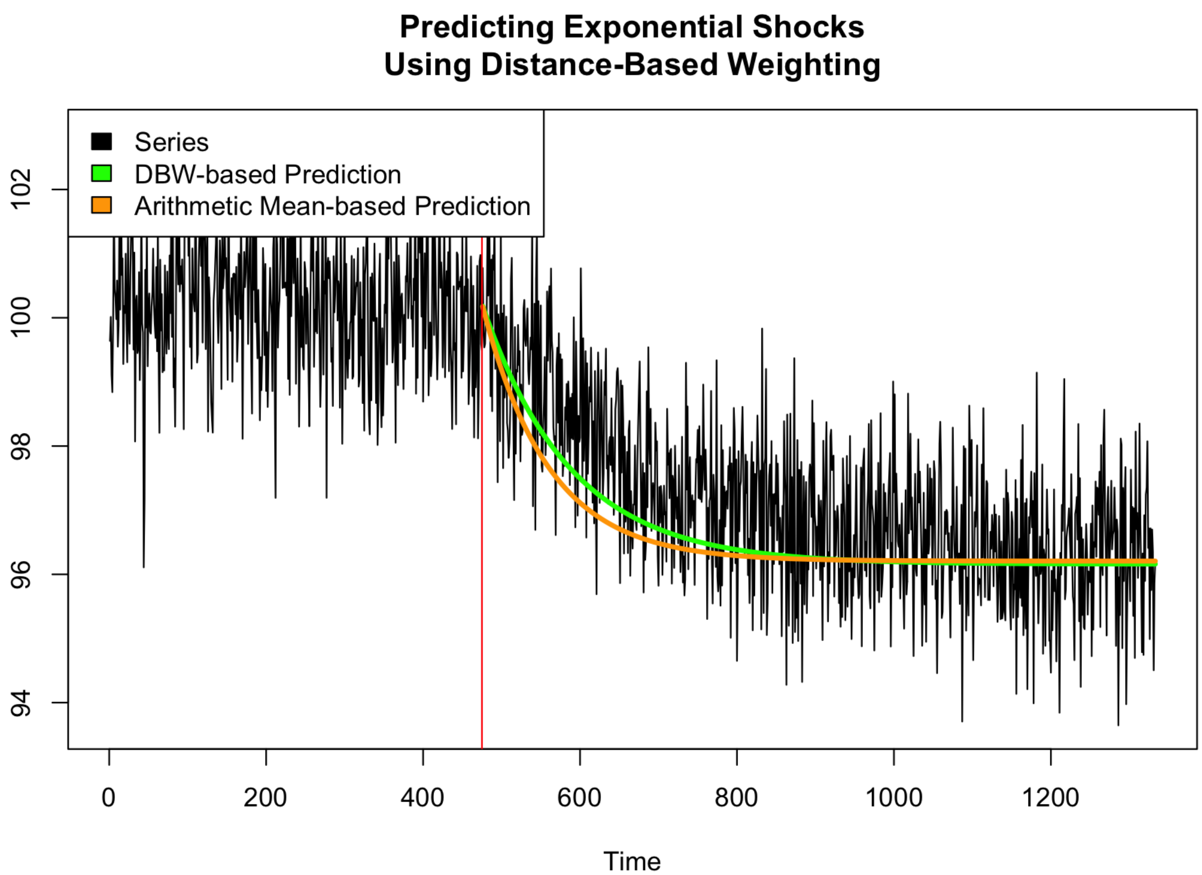
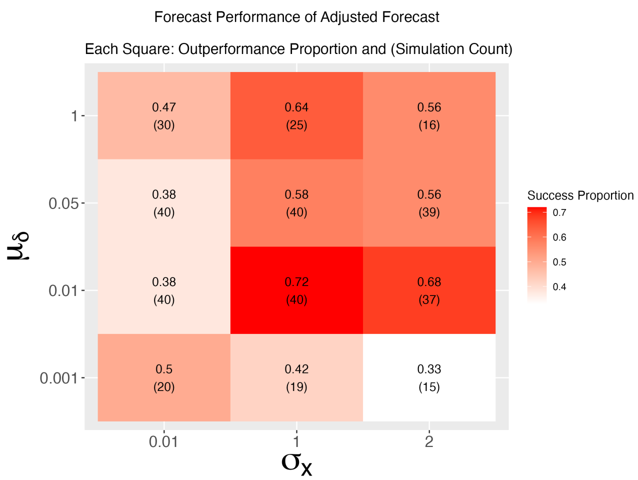
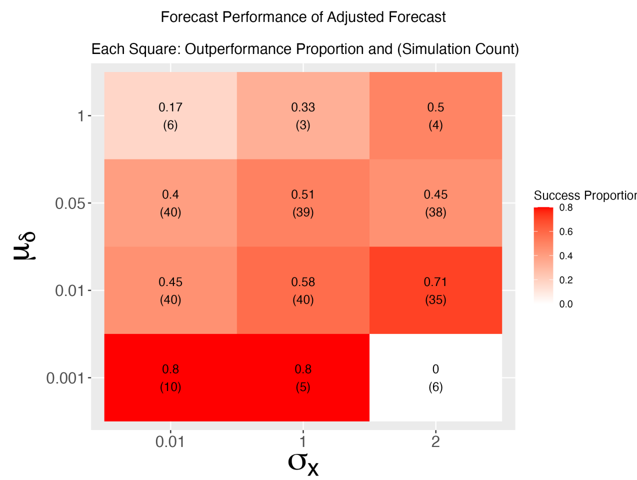


Figure 3.5: We simulate a random walk centered at $\mu = 100$ that is subject to a shock at approximately time index 440. The shock of size -4 is not realized over a single index. Instead, the shock is governed by a decay exponential decay parameter ψ_1, as are exponential shocks in $n = 10$ donor series. We estimate ψ_1 using a convex combination of the estimated decay parameters $\{\psi_i\}_{i=2}^{n+1}$, resulting in a distance-based weighting prediction of the shock. We also illustrate the arithmetic-mean-based prediction of the shock using the color orange. This prediction is based on an overestimate of ψ_1 and hence results in residuals that undershoot the time series under study.

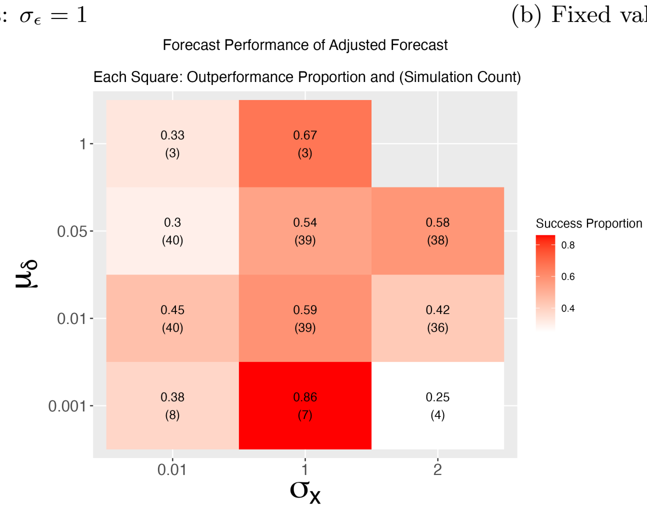
Covariate Diversity and Mean



(a) Fixed values: $\sigma_\epsilon = 1$



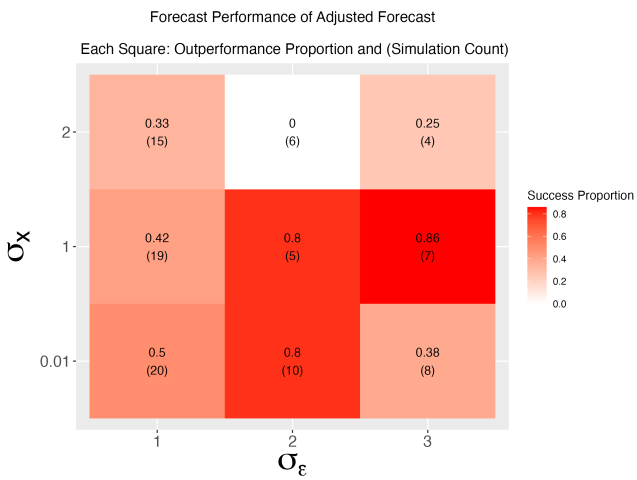
(b) Fixed values: $\sigma_\epsilon = 2$



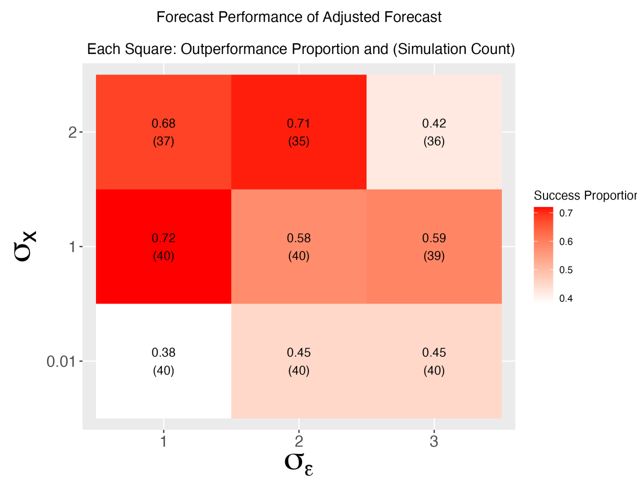
(c) Fixed values: $\sigma_\epsilon = 3$

Figure 3.6: In the progression from 3.6a to 3.6b to 3.6c, we see that increasing μ_δ leads to improved performance of the adjusted forecast, and this effect is intensified as μ_v increases. However, it is not consistently true that an increasing noise undermines the performance of the adjusted forecast. Rather, that phenomenon requires both large quantites of μ_δ and μ_v.

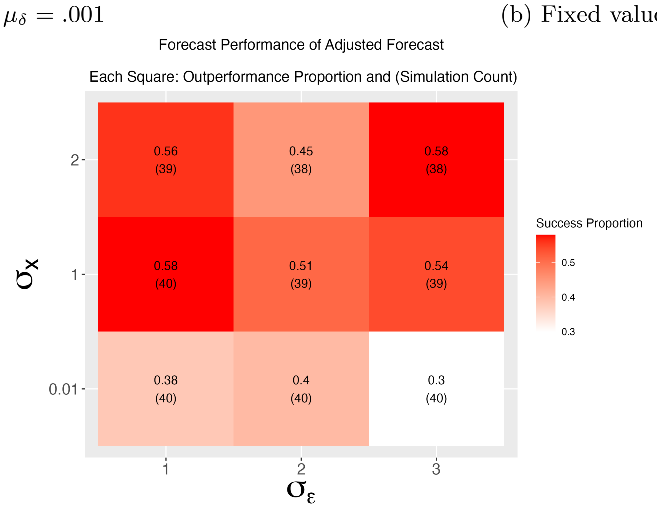
Idiosyncratic Noise and Covariate Diversity



(a) Fixed values: $\mu_\delta = .001$



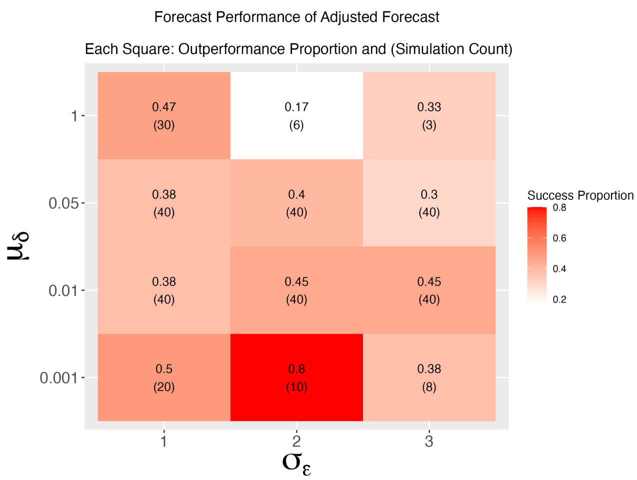
(b) Fixed values: $\mu_\delta = .01$



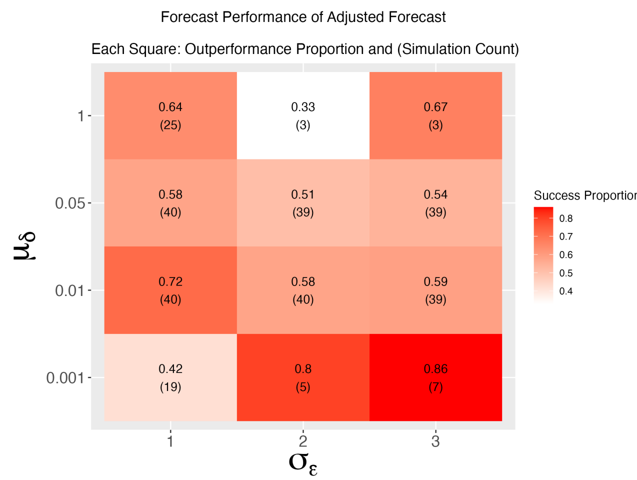
(c) Fixed values: $\mu_\delta = .05$

Figure 3.7: In the progression from 3.7a to 3.7b to 3.7c, we see that increasing μ_δ leads to improved performance of the adjusted forecast, and this effect is intensified as μ_v increases. However, it is not consistently true that an increasing noise undermines the performance of the adjusted forecast. Rather, that phenomenon requires both large quantites of μ_δ and μ_v.

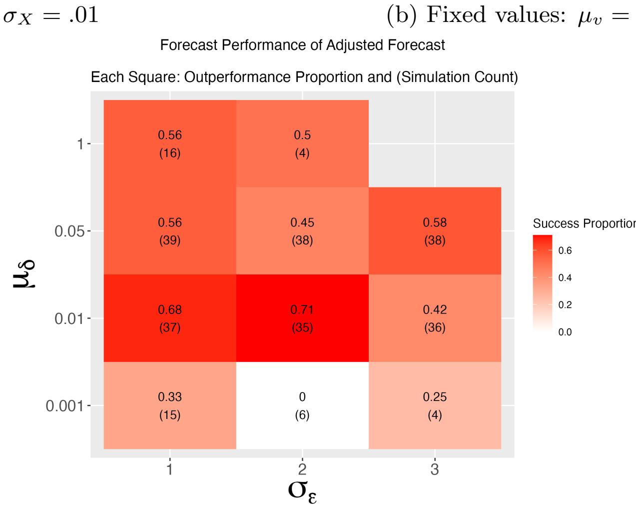
Idiosyncratic Noise and Covariate Mean



(a) Fixed values: $\sigma_X = .01$



(b) Fixed values: $\mu_v = 1, \sigma_v = .125, \mu_{\omega^*} = .125$



(c) Fixed values: $\mu_v = 2, \sigma_v = .125, \mu_{\omega^*} = .125$

Figure 3.8: In the progression from 3.8a to 3.8b to 3.8c, we see that increasing μ_δ leads to improved performance of the adjusted forecast, and this effect is intensified as μ_v increases. However, it is not consistently true that an increasing noise undermines the performance of the adjusted forecast. Rather, that phenomenon requires both large quantities of μ_δ and μ_v.

3.7 Discussion

The forecast horizon — does it matter? If so, how so? **?** has a one-period horizon. [**?**, p. 203] discuss how long to keep the forecast adjustment in place. For a corrected "slope parameter", the effect of h is not so clear.

3.7.1 Asking questions about the covariate vector and conditional shock distribution

We pose two questions.

What if a given entry of the covariate vector was helpful at only one of those roles? For example, what if an entry was particularly useful in yielding good weights but was not important for the shock distribution — perhaps because the corresponding shock parameter was zero.

Conversely, suppose a given entry was known to be part of the shock distribution, but the inclusion of that covariate in distance-based weighting was often not worth it, perhaps because it was hard to approximate with using donors, and the corresponding shock parameter was small in magnitude with high probability.

3.7.2 Shock propagation and estimation accuracy

Consider **?**, in which a real world example shows how volatility can be forecasted following surprising election outcomes. There, the estimation of fixed effects is done around elections in a donor pool. That means that for each past election or poll, etcetera, the method must accurately estimate the surprise-induced volatility boost. That means picking a window around such an effect.

proposals

1. For a given data-generating process, simulate the accuracy and variability of shock estimates using post-shock measurement periods of difference lengths.

2. Develop formal results to quantify (estimate, bound, etc) the risk of using different post-estimation measurement lengths. This matters because it might be that a donor has experienced a shock in the last m days, and if m is too small, using that donor may be risky.

3.7.3 Aggregating residuals as opposed to aggregating fixed effect estimates?

Cases:

1. Fixed effect is estimated with accuracy

2. Fixed effect is estimated with considerable noise. Hence, the idiosyncratic error is estimated with considerable noise.

3. idiosyncratic error large

4. idiosyncratic error small

3.8 Extensions

- Forecast for series besides the mean or conditional mean If you have a family of models estimated used a Gaussian loss function, and the post-shock residuals are not conditionally Gaussian (e.g. they could be conditionally Cauchy), then why even start with the default forecasting function? Why not throw it out entirely? If only the distribution of the innovations changes, then would that even matter for a forecast that is purely concerned with the conditional mean? Yet another potential motivation springs from the way we evaluate model performance. In **?**, the authors discuss six dichotomies regarding forecast evaluation as a way of advising caution about simplistic notions of performance evaluation. These ideas matter for the current manuscript for at least one reason. We must entertain the possibility that breaks in the DGP will invalidate the standard forecast evaluation framework under which we had been working. If, for example, we have a family of models estimated used a Gaussian loss function, and the post-shock residuals are not conditionally Gaussian (e.g. they could be conditionally Cauchy), then why even start with the default forecasting function? Why not throw it out entirely? If only the distribution of the innovations changes, then would that even matter for a forecast that is purely concerned with the conditional mean?

- Bias-variance decomposition for ARIMA, as introduced in (3.6.1)
- Shrinkage estimation of autoregressive structure on all donors, then use our method
- TAR
- SETAR
- Function forecasts
- Binary Outcome Forecasts
- Density Forecasts
- Quantile Forecasts

3.8.1 Forecast Combination

what we are talking about here is not forecast combination, but there may be, nevertheless, a role for forecast combination: combining the forecasts generated by small differences in covariate and/or donor choice, as is done in **?**.

Consider the leave-one-out method used on pairs of donors and covariates in **?**. By averaging over $(n+1)\cdot(p+1)$ predictions, we average over $(n+1)\cdot(p+1)$ convex combinations (where, when a donor is ommitted, its weight can be imputed as zero).

3.8.2 Limitations

3.8.2.1 How important is a shared DGP?

In particular, can we relax this assumption and proceed with weaker assumptions on the donors and time series under study?

3.8.2.2 Relaxing the shared family assumption

The assumption of a shared family can be relaxed in certain circumstances. For example, if a certain donor is not like the time series under study with respect to the covariates, then the lack of a shared family will not matter much.

3.9 Supplement

3.9.1 Formal Results

Proof of Proposition 5 We claim

The result for each $\hat{\alpha}$ follows from the consistency proof of the MLE in AR-X models.

$\square$

Proof of Proposition 6 Here we suppress time-indexation for each of exposition.

$$\begin{aligned}
\mathbf{x}_1 &= \sum_{i=2}^{n+1} \pi_i \mathbf{x}_i && \text{(by an assumption of the proposition)} \\
\implies \lambda^T \mathbf{x}_1 &= \lambda^T \sum_{i=2}^{n+1} \pi_i \mathbf{x}_i = \sum_{i=2}^{n+1} \pi_i \lambda^T \mathbf{x}_i && \text{(by left-multiplication by } \lambda^T\text{)} \\
\implies \lambda^T \mathbf{x}_1 &= \sum_{i=2}^{n+1} \pi_i \alpha_i && \text{(by substitution and definition of } \alpha_i\text{)} \\
\implies \alpha_1 &= \sum_{i=2}^{n+1} \pi_i \alpha_i && \text{(by the definition of } \alpha_1\text{)}
\end{aligned}$$

If, on the right-hand side, we replace the α_i with $\hat{\alpha}_i$, then it will converge in probability to the expression on the right-hand side, by a simple application of Slutsky. Hence, $\hat{\alpha}_1$ is a consistent estimator. $\square$

Proof of Proposition 7

Recall from Proposition 5 that for any i, $2 \leq i \leq n+1$, the tuple $(\hat{\rho}_1, ..., \hat{\rho}_p, \hat{\alpha}_{i,T^*+1})$ is consistent at $t \to \infty$. Next consider the prediction function for an AR(p)-X, and let $t = T_1^*$:

$$\mathbb{E}[y_{adjusted,t+1} | \mathcal{F}_t] = y_{t+1} + \alpha_{t+1} = f(\mathcal{F}_{t-1}) + \alpha_{t+1} = \mu + \sum_{k=1}^{p} \rho_k y_{t-k} + \alpha_{t+1} \ .$$

By replacing parameters with their estimates, we arrive at the prediction

$$\mathbb{E}[\hat{y}_{adjusted,t+1} | \mathcal{F}_t] = \hat{y}_{t+1} + \hat{\alpha}_{t+1} = \hat{f}(\mathcal{F}_{t-1}) + \hat{\alpha}_{t+1} = \hat{\mu} + \sum_{k=1}^{p} \hat{\rho}_k y_{t-k} + \hat{\alpha}_{t+1} \ .$$

which converges to

$$\mu + \sum_{k=1}^{p} \rho_k y_{t-k} + \alpha_{t+1} \ .$$

in distribution or probability (depending upon how $\hat{\alpha}_{t+1}$ converges) as $t \to \infty$ in the donor pool by a simple application of Slutsky's Theorem.

$\square$

Chapter 4

Conclusion

TO BE POPULATED

Appendix A

An appendix