ANNALS OF ECONOMICS AND FINANCE **18-2**, 291–322 (2017)

Market Timing under Limited Information: An Empirical Investigation in US Treasury Market

Guoshi Tong*

Hanqing Advanced Institute of Economics and Finance
Renmin University of China, Beijing, China
E-mail: gstong@ruc.edu.cn

I examine the welfare value of bond return forecasts in timing the market under a limited data trading environment. Using monthly US data, I estimate the utility benefit of each return forecast and test its significance through a structural approach of forecast evaluation. I find that predictor based market timing with finite historical data creates occasional but large portfolio loss. The benchmark welfare level under no-predictability view is hard to beat by parametric or non-parametric strategy. However, a Bayesian shrinkage strategy with no-predictability prior leads to significant welfare gain at certain range of prior confidence.

Key Words: Bond return predictability; Limited information; Structural forecast evaluation.
JEL Classification Numbers: C12, E47, G11, G17.

1. INTRODUCTION

A vast literature, such as Fama and Bliss (1987), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009), has documented that expected returns in the US Treasury bond market vary over time and are predictable by the shape of the yield curve and macroeconomic fundamentals. The variation in the expected return is also economically large. Taking the Cochrane Piazzesi factor, the measured conditional expected annual-excess-return of a 5-year bond varies over time with a standard deviation of around 2.5%, while its unconditional mean is less than 1%.

This essay examines whether such return predictability can be exploited by a bond investor facing finite history of data to improve trading decision.

* Supported by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China, 16XNF009.

1529-7373/2017
All rights of reproduction in any form reserved.

In theory, a high magnitude of predictability calls for aggressive market timing, that is, the optimal portfolio weight should depend on the level of return predictor. In practice, additional obstacles exist. First, investors only have limited information on the relevant predictive relation. For example, although they understand that the return predictor at hand is correlated with future return, the true conditional distribution of future return given current predictor value is unknown. The return forecasting process typically relies on either a parametric or non-parametric return model, which is inevitably mis-specified.[1] On top of that, either a parametric or non-parametric return model needs to be estimated. As investors only face a limited data history, estimation uncertainty becomes a concern. Secondly, given certain estimated and mis-specified return forecasting model, there is an extra portfolio decision based on it. The errors in estimation or model specification would be further transformed during some portfolio optimization processes and their ultimate impact to investor welfare would be unknown.

Given the above-mentioned concerns, the objective of this essay is to quantify the portfolio value of bond return forecasts under a limited information constraint. To this end, I consider a finite data bond trading scenario. I assume that, under this scenario, a CRRA bond investor has access to a finite history of data on bond return predictor values and subsequent realized bond excess returns. These historical data are used by investors to infer the relevant predictive relation. I then adopt a decision theory approach to view each allocation strategy that exploits bond return predictability as a function of historical data, or in other words, an estimator. That portfolio estimator hence maps any observed data in investor's information set towards a bond portfolio weight scheme. The finite sample properties of this portfolio estimator thus indicates the welfare value of return predictor under limited data.

Methodologically, the contribution of my approach is to propose a conceptually new means of assessing return predictors. Traditional works focus on raw predictability, i.e., whether the return predictor helps to reduce mean squared forecast error (MSFE). In contrast, I shift attention to the welfare value of each predictor, an object of ultimate interest to investors. I argue that although the traditional approach helps to understand the underlying data generating process, it does not provide sufficient information to guide trading decisions. For instance, one predictor could forecast the correct sign of excess return in each period while completely missing the magnitude. In that case, the traditional criteria of MSFE would not favor that signal despite the fact that it is obviously useful for trading. Unlike

[1]This can be due to inadequately modeled dynamics, incorrect functional form or any combination of these.

MSFE, the utility metric accounts for the consequences any forecast error would have on the portfolio. Under this metric, a small forecast error is more valuable when the sign forecast is correct and the magnitude of forecast is large. Secondly, traditional works test on the population level of mean squared forecast error assuming the true value of slope coefficients in predictive regression are known. Yet, my work recognizes the fact that return forecasts and the subsequent portfolio decisions are both conducted with limited data. Hence, the focus here is on the forecast errors when relevant parameters are estimated and how they translate into portfolio risks. Note that the predictive relation, even if it does exist in the population level, needs to be specified and estimated precisely enough within a finite sample in order to achieve welfare gain.

Using monthly US data, I estimate and test the significance to welfare gain of a list of bond return predictors from either yield curve, macro fundamental or technical analysis in the aforementioned limited data framework. First, I find that linear parametric strategies, driven by any of the predictors considered above, create large losses occasionally, despite positive gains in most states. As a result, the corresponding welfare estimates are inferior to a simple benchmark rule which ignores predictability. Second, nonparametric policies, advocated by Brandt (1999) to account for potential nonlinearity in the predictive relation, exhibit similar performance instability and fail to beat the benchmark. Third, such performance instability is not specific to any business cycle or market volatility regime and hence, it is hard to forecast in advance. Fourth, shrinkage strategies, suggested in Connor (1997) and Brandt (2009) and implemented through Bayesian predictive regressions with no-predictability prior, lead to significant welfare improvement when the degree of prior confidence is high.

The above findings suggest that errors in forecasting model estimation and specification can indeed create large welfare loss. Specifically, when predictive relations are subject to un-modeled and hence un-expected instability, forecast errors would be magnified under a market timing decision that is made based on the historical data observed. In addition, the market timing policies, either parametric or non-parametric, are more complex than the benchmark rule, which ignores predictors. Their estimations are hence more sensitive to the realization of historical data observed. This additional estimation uncertainty also translates into extra volatility in realized returns and hence lowers welfare. The shrinkage policy represents an attempt to reduce the portfolio risks due to forecasting model estimation and mis-specification. By taming down estimated return forecast, bond predictability is only partially exploited under this policy. Our results indicate that at a high shrinkage, the information value gained from incorporating predictors outweighs the welfare loss due to mis-specification and estimation.

In assessing strategy performance, my estimation of utility expectation, given a single path of realized US data, rests upon a series of "pseudo" repeated experiments generated by an out-of-sample portfolio construction exercise.[2] In particular, at the end of each month, our investor is asked to make an allocation decision based only on the most recent data. Meanwhile, a rolling window scheme is imposed, so that the same size of historical data or information set is available at each portfolio decision experiment. Thus, the resulting average realized utility serves as a consistent estimator of the (unconditional) welfare measure averaged over time.

In testing the significance of expected-utility difference, I build the inference procedures upon those established in the forecast evaluation literature, Diebold and Mariano (1995), West (1996), Giacomini and White (2006). This stream of works has traditionally evaluated point forecasts by statistical measures of accuracy, such as the mean squared forecast error or predictive likelihood. However, in our context, the evaluation object is a portfolio estimator. Thus, forecast evaluation is addressed in a structural way with a portfolio optimization process embedded and the evaluation metric modified to be the expected utility.

Speaking to the bond return forecasting literature, my work is most closely related to Thornton and Valente (2012) but complements their analysis of out-of-sample bond return predictability in two dimensions. First, in examining the economic significance of predictability, Thornton and Valente (2012) fix a mean variance rule and then evaluate the information value of yield curve in shaping bond portfolios. In contrast, I argue that the portfolio value of a return predictor also depends on the way it is exploited. Hence, my work focuses on a joint evaluation, where for each return predictor, a bunch of policy functions such as parametric or non-parametric or Bayes rules are included in measuring welfare value. The return predictors examined also go beyond those based on the yield curve and contain macro and technical analysis driven factors. Secondly, while out-of-sample analysis is conducted in Thornton and Valente (2012), inference is still on a population level statement of the forecast errors, assuming the values of parameters are known. Nonetheless, my work relies on a formal estimation and inference procedure that validates the finite sample properties of return predictors against estimation and mis-specification risks. Since investors only face limited data in reality, this finite sample approach seems to be more relevant for portfolio management.

The rest of the essay is structured as follows. 2 lays out the limited data bond investment framework. There, I also describe the utility metric along with the associated estimation and inference procedure. 3 conducts em-

[2]While investors only have access to limited data, econometricians could use the whole time series (future data) in performance evaluation.

pirical analyses in the US Treasury bond market and discusses the results. Robustness checks are illustrated in 4, and the last chapter concludes.

2. INVESTMENT FRAMEWORK

This lays out the investment decision framework. I consider a single period bond allocation problem in which the excess return of long term bond is predictable. However, the true state/forecasting variable as well as its joint distribution with return are unknown and have to be estimated with a finite history of data. I describe respectively the allocation rule, its performance measure, the estimation of performance as well as the relevant inference procedure.

2.1. Bond allocation rule

Consider a Treasury bond investor who allocates his current wealth W_t between a short term *1*-year discount bond and a longer term n-year one. The investment horizon is τ so the position is held until $t + \tau$ and then liquidated. With τ equal to a year, the *1*-year bond matures at face value and is risk-less in nominal terms. While, the n-year bond will be sold as an *(n-1)*-year bond whose price is unknown beforehand. The long-term bond's log return $r^{(n)}_{t+\tau}$ is therefore random. Its expected value in excess of the log risk-less rate $r^f_t = r^{(1)}_{t+\tau}$ is referred to as bond risk premium $E_t[r^{(n)}_{t+\tau} - r^f_t]$.

The investor's preference admits an expected utility representation with a CRRA function defined over the terminal wealth $W_{t+\tau}$. At time t, the investor puts a fraction α_t of wealth into the n-year bond based on the conditional density of return and his risk tolerance $1/\gamma$. The allocation decision is thus a directional / market-timing bet that collects risk premium and does not involve any cross-sectional arbitrage.

In making allocation decision, the true conditional density is unknown, but a finite history (sample realization) of return $\vec{r}_t = \{r^{(n)}_\tau, \ldots, r^{(n)}_t\}$ and state variables $\vec{z}_t = \{z_0, \ldots, z_t\}$ is available. Following bond literature, z_t include both yield curve and macroeconomic fundamental data which form the basis of various return forecasting factors. This historical data, of length t and denoted as $\phi_t = \{\vec{r}_t, \vec{z}_t\}$, may be used to estimate the desirable forecasting model and the corresponding portfolio policy. Thus, the allocation choice α_t in this finite history setting is data dependent, and the rule $\alpha(.)$ can be viewed as a generic estimator formally defined as follows:

Definition 2.1. An allocation rule, or portfolio strategy, $\alpha(.)$ is a mapping from realizations of historical data in the estimation window to the set of allocation positions.

$$\alpha(\phi_t) \quad : \Phi_t \to \mathcal{A}, \tag{1}$$

where Φ_t is the range of historical (sample) data ϕ_t and $\mathcal{A} = (-\infty, +\infty)$ is the admissible set of portfolio weight on long term bond.

2.2. Measurement of performance

The performance of each *allocation rule* $\alpha(.)$ is assessed based on the expected utility it generates. Given a history of ϕ_t, the realized utility is derived as

$$U(\alpha(\phi_t), r_{t+\tau}^{(n)}) = \frac{(\alpha(\phi_t)e^{r_{t+\tau}^{(n)}} + (1-\alpha(\phi_t))e^{r_t^f})^{1-\gamma}}{1-\gamma}, \tag{2}$$

where dependence on r_t^f is suppressed since it is observed at t and thus is an element of ϕ_t. This realized utility is a random variable as both ϕ_t and $r_{t+\tau}^{(n)}$ are random. Accordingly, it should be averaged across realizations of both historical data and future return, as suggested in the following (unconditional) notion of performance measure:

Definition 2.2. An unconditional welfare measure of the allocation rule $\alpha(.)$ is the unconditional expectation of realized utility:

$$EU[\alpha(.)] = E_{f_{\phi_t, r_{t+\tau}^{(n)}}}[U(\alpha(\phi_t), r_{t+\tau}^{(n)})], \tag{3}$$

where $f_{\phi_t, r_{t+\tau}^{(n)}}$ is the joint density of historical data and forecasting period return.

By integrating over historical data, the above metric explicitly accounts for the effect of estimation uncertainty on portfolio performance. Meanwhile, it reflects mis-specification risk as modeled forecasting relations do not necessarily coincide with the true one. These two ingredients help us to focus on the practical, or limited data usefulness of any portfolio strategy. Besides, this unconditional welfare measure can be further modified to examine potential heterogeneity in allocation rule performance. In particular, we will consider utility expectations that are conditional on certain regime of the business cycle or market volatility measured by an economic state variable s_t,

$$EU[\alpha(.)|s_t = s] = E_{f_{\phi_t, r_{t+\tau}^{(n)}|s_t=s}}[U(\alpha(\phi_t), r_{t+\tau}^{(n)})], \tag{4}$$

where $f_{\phi_t, r^{(n)}_{t+\tau}|s_t=s}$ is the joint density of historical data and future return conditional on current economic state being s. The regime s, for example, can be a low / high unemployment episode, $s_{LowUnemp}$ / $s_{HighUnemp}$, if the contemporary unemployment rate is below / above its average level, or a high / low turbulence episode, s_{hvol} / s_{lvol}, if past year's realized bond market volatility is greater / less than its mean.

2.3. Estimation of welfare metric

Our welfare measure, either unconditional or conditional, is a frequentist notion of average realized utility achieved over repeated samples drawn from the true distribution. However, in estimating this quantity, only one single path of data is available. To overcome this issue, I rely on a sequence of "pseudo" repeated experiments generated by an out-of-sample portfolio construction exercise. Specifically, let T denote the total number of observations available to the econometrician and t be the number of observations accessible to the investor (in his portfolio estimation window). Thus, $m = T - t - \tau + 1$ would represent the number of out-of-sample periods. At each time j, $t \leq j < t + m$, our investor is asked to make allocation decision based only on the historical data within $[j - t + 1, j]$, i.e., rolling window scheme. The rationale of using rolling window rather than the expanding one is that: in quantifying limited data value, our allocation rule $\alpha(\phi_t)$ and performance measure are set to be history size specific. Accordingly, the length of data available to investor at each portfolio decision experiment needs to stay the same.

Based on the above argument, an estimator of the (unconditional) welfare can be the out-of-sample average of realized utility, expressed as:

$$\widehat{EU}[\alpha(.)] = \frac{1}{m} \sum_{j=t}^{T-\tau} U(\alpha(\phi_j), r^{(n)}_{j+\tau}), \tag{5}$$

where ϕ_j stands for the sample data between $[j - t + 1, j]$. Assume that the whole time series follows certain mixing property and denote that the population level of utility expectation at period j to be $EU_j[\alpha(.)]$, $\widehat{EU}[\alpha(.)] - \frac{1}{m}\sum_{j=t}^{T-\tau} EU_j[\alpha(.)]$ will converge almost surely to zero as m goes to infinity. (See the strong law of large numbers for mixing process in White (1984) Corollary 3.48 p.49)[3] Here, stationarity is not required, hence the whole time series can be characterized by structural shifts at unknown date. This assumption of data heterogeneity is more realistic than the assumption of stationarity. It further justifies the use of rolling window scheme since local approximation may be less biased in cases of instability.

[3]When data is stationary, unconditional expected utility $EU_j[\alpha(.)] = EU[\alpha(.)]$ is the same across time, so $\widehat{EU}[\alpha(.)]$ will converge to the constant $EU[\alpha(.)]$.

The conditional notion of welfare can be estimated in a similar way, except that averaging is now only over portfolio exercises with the same economic regime. Denote s_j to be the level / regime of certain economic state at allocation time j, the performance of $\alpha(.)$ conditional on regime s is then estimated as:

$$\widehat{EU}[\alpha(.)|s] = \frac{1}{m_s} \sum_{j=t}^{T-\tau} U(\alpha(\phi_j), r_{j+\tau}^{(n)}) I(s_j = s), \tag{6}$$

with $I(.)$ being the indicator function and m_s being the number of observations with regime s. Finally, all welfare estimates, either unconditional or conditional, are translated into certainty equivalent returns, $\widehat{CE}[(\alpha(.))] = U^{-1}(\widehat{EU}[\alpha(.)])$, for ease of exposition.

2.4. Inference on utility benefit

Our performance estimate of any allocation rule $\alpha(.)$ is benchmarked against that of a simple strategy which ignores predictability. In particular, the benchmark strategy, denoted by $\alpha^0(.)$, does not have access to past values of return predictors $\vec{z}_t$ and can only map past realization of returns into portfolio position: $\alpha^0(\phi_t \backslash \vec{z}_t) = \alpha^0(\vec{r}_t)$. The difference in estimated welfare between an allocation rule that uses the predictors and a benchmark that discards them, $\widehat{EU}[\alpha(.)] - \widehat{EU}[\alpha^0(.)]$, thus reflects the portfolio value of relevant return predictors. However, $\widehat{EU}$ is only a point estimate of the true expected utility. Hence, to account for the sampling variability in $\widehat{EU}$, a formal inference procedure is needed.

Unconditional inference

For unconditional inference, the null hypothesis we are interested in is that, on average, market timing does not generate any expected utility difference relative to the benchmark:

$$H_0 \ : \ E[U(\alpha(\phi_j), r_{j+\tau}^{(n)}) - U(\alpha^0(\vec{r}_j), r_{j+\tau}^{(n)})] = 0, \ \forall j = t, ..., T-\tau. \tag{7}$$

Note that, expectation here is taken with respect to all possible sample paths of the entire stochastic process $\{r_{t+\tau}, z_t\}_{t=0}^{T-\tau}$.

The alternative to H_0 is specified in a global way, as distribution of sample data and return are non-identical over time. Denote $\Delta U_{j,j+\tau} = U(\alpha(\phi_j), r_{j+\tau}^{(n)}) - U(\alpha^0(\vec{r}_j), r_{j+\tau}^{(n)})$ and let $\Delta \overline{U}_{t,m} = \frac{1}{m} \sum_{j=t}^{T-\tau} \Delta U_{j,j+\tau}$,

$$H_A : E[|\Delta \overline{U}_{t,m}|] \geq \delta > 0, \text{ for small } \delta \text{ and all } m \text{ sufficiently large.} \tag{8}$$

The testing procedure borrows from those developed in the forecast evaluation literature (e.g. Diebold and Mariano (1995), West (1996), Clark and McCracken (2001), Giacomini and White (2006)). This stream of research has traditionally focused on equal forecast accuracy between two competing forecasts, in which the objects of interest are typically quadratic loss (squared error), directional accuracy, or predictive log-likelihood. However, in our framework, the primary purpose of return forecast is to make allocation decision. Accordingly, forecast evaluation is addressed in a structural way with the portfolio optimization process embedded. The relevant loss function is the negative of realized utility and can no longer be expressed as a function of forecast errors.

While conceptually distinct, the asymptotic results established in existing literature can still be applied. The test is based on the following Wald-type statistic:

$$\mathrm{T}_{t,m} = m(\Delta\overline{U}_{t,m})\hat{\Omega}_m^{-1}(\Delta\overline{U}_{t,m}), \tag{9}$$

where $\hat{\Omega}_m$ is a suitable HAC estimator of the asymptotic variance $\Omega_m = var[\sqrt{m}\Delta\overline{U}_{t,m}]$.

A level α test rejects the null of equal performance whenever $\mathrm{T}_{t,m} > \chi^2_{1,1-\alpha}$, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of χ^2_1 distribution. The underlying justification of such test follows central limit theorem for mixing process stated in Wooldridge and White (1988) and other standard asymptotic arguments in White (1984) and Giacomini and White (2006).

Conditional inference

Whereas the above analysis focused on the unconditional value of market-timing, conditional inference tests for expected utility difference conditional on a particular economic regime. The null hypothesis considered now is:

$$H_0^c \;:\; E[\Delta U_{j,j+\tau}|s_j = s] = 0,\ \forall j = t, ..., T-\tau, \tag{10}$$

with s being certain business cycle or market volatility regime. As mentioned above, I consider current economy being at a low / high unemployment episode, $s_{LowUnemp}$ / $s_{HighUnemp}$, if the contemporary unemployment rate is below / above its average level, and at a high / low turbulence regime, s_{hvol} / s_{lvol}, if past year's realized bond market volatility is greater / less than its mean. Those conditioning instruments will help us to examine whether relative performance of market timing is uniform across the business cycle or turbulence dependent.

Fixing a regime s, the testing procedure relies on the same Wald-type statistic as in the unconditional test except that it uses only samples with $s_j = s$. As before, under certain regularity conditions on the mixing coefficients (c.f., White (1984); Giacomini and White (2006)), such test has correct size and is consistent against the alternative of $H_A^c : E[\ |\Delta\overline{U}_{t,m_c}|\ \mid s] \geq \delta > 0$, where $\overline{U}_{t,m_c}$ is now the average of utility realizations conditional on the state s.

3. EMPIRICAL ANALYSIS

Using framework developed in the previous, I now look into the performance of bond market timing empirically. I first describe the data used, the return predictors considered and the types of portfolio policies entertained. I then present the empirical findings and discuss their implications.

3.1. Data and utility assumption

I use monthly data on US Treasury bond and macroeconomic fundamentals. Bond prices are obtained from Fama-Bliss data set in Center for Research in Security Prices (CRSP) and contain observations of zero coupon (discount) bonds with maturity one to five years. Macro fundamental data consists of a balanced panel of 135 economic series. Such data set is originally collected in Stock and Watson (2002) and Stock and Watson (2005), later expanded by Ludvigson and Ng (2009, 2011), McCracken and Ng (2016), and available on FRED database. The spanning period considered for both yield curve and macroeconomics data starts from Jan 1964 and ends at Dec 2014.

Regarding primitive on the CRRA preference, it is common practice in the portfolio allocation literature to consider relative risk aversion γ ranging from 5 to 10, but a higher value of $\gamma = 20$ are also entertained when gauging the effect of varying γ (See for instance, Barberis (2000)).[4] Following this tradition, we pick $\gamma = 10$ for most of our portfolio allocation exercises and then change this risk aversion level at 5, 15 and 20 as robustness checks. While the main conclusions are robust to each γ, my analysis shows that a low value of risk aversion at 5 would induces highly levered positions for certain predictors considered and lead to ex-post bankruptcy at some states.

3.2. Implementing allocation rules

Recall that, since an allocation rule is defined as a function of sample data ϕ_t, the size of investor's information set need to be pre-specified. I assume that investors always face a historical data of 15 years length, i.e. ϕ_t include 180 monthly observations. Then based on this 15 year rolling window scheme, I describe on how to construct various return forecasting factors and how to estimate the associated policies using available sample on bond prices and macroeconomic series.

[4]Decision theory literature and experimental economists have shown some evidence that individual's risk aversion level when making lottery choices should not exceed a number of 5. We point out here that a portfolio manager operating in the financial market may have a different risk appetite. In fact, according to Figure 1 in van Binsbergen et al. (2012) which estimates the cross sectional distribution of US mutual fund managers' risk appetite, the density of risk aversion peaks at 10 to 25 and is skewed to the right.

3.2.1. Return predictors construction

I construct major return predictors identified in the bond forecasting literature. Those factors are either directly observed or themselves estimated through historical data. I classify them into three categories: those based on yield/forward curve; on macro fundamental; or on bond market technical analysis.

Yield/forward curve driven factors:

The first factor I consider is the Fama-Bliss (FB) forward spread. I calculate log forward rate at time t for loans between $t+n-1$ and $t+n$ as: $f_t^{(n)} = p_t^{(n-1)} - p_t^{(n)}, n = 1, ..., 5$, where $p_t^{(n)}$ is the log price at time t of the n-year discount bond. I then record the n-year forward spread to be $FB_t^{(n)} = f_t^{(n)} - f_t^{(1)}$. As documented in Fama and Bliss (1987), this factor forecast annual excess return of the n-year bond, which we label as $rx_{t+1}^{(n)} = r_{t+1}^{(n)} - r_t^f = p_{t+1}^{(n-1)} - p_t^{(n)} + p_t^{(1)}$.

The second predictor I study is the Cochrane-Piazzesi (CP) factor. While the $FB_t^{(n)}$ predictor is maturity dependent, Cochrane and Piazzesi (2005) suggests that a single factor summarizes bond premium across maturity. This single return-forecasting factor is estimated through a (first stage) predictive regression of average excess return on the whole forward curve. Specifically, let $\overline{rx}_{t+1} = \frac{1}{4}\sum_{n=2}^{5} rx_{t+1}^{(n)}$ be the average (across maturity) annual excess return, CP factor is formed as the fitted value from:

$$\overline{rx}_{t+1} = \gamma_0 + \gamma_1 f_t^{(1)} + \gamma_2 f_t^{(2)} + \ldots + \gamma_5 f_t^{(5)} + \overline{\epsilon}_{t+1}. \qquad (11)$$

The regression uses only data on bond prices within the information set ϕ_t, and the (estimated) CP factor is denoted by: $\widehat{CP}_t = \hat{\gamma}_0 + \hat{\gamma}_1 f_t^{(1)} + \hat{\gamma}_2 f_t^{(2)} + \ldots + \hat{\gamma}_5 f_t^{(5)}$.

The third predictor I account for is the cycle factor (cf) proposed in Cieslak and Povala (2015). The construction of this factor rests on a decomposition of log yields, $y_t^{(n)} = -\frac{1}{n} p_t^{(n)}$, into persistent component η_t and shorter-lived fluctuations $c_t^{(n)}$ (cycles). η_t relates to the long run inflation expectation and is proxied by discounted moving average of realized core CPI, while $c_t^{(n)}$, the transitory part, is counted by residual. Following the authors' suggestion, I regress log yields of different maturity on the contemporary level of long-run inflation expectation proxy: $y_t^{(n)} = b_0^{(n)} + b_\eta^{(n)} \eta_t + \epsilon_\eta$, and obtain cycle as the fitted residual $c_t^{(n)} = y_t^{(n)} - \hat{b}_0^{(n)} - \hat{b}_\eta^{(n)} \eta_t$. I then project the average excess return onto the cross-sectional composition of these cycles to form a (single) return-forecasting factor. In particular, I estimate

$$\overline{rx}_{t+1} = \theta_0 + \theta_1 c_t^{(1)} + \theta_2 \overline{c}_t + \overline{\epsilon}_{t+1}, \quad \text{where} \quad \overline{c}_t = \frac{1}{4}\sum_{n=2}^{5} c_t^{(n)}, \tag{12}$$

and record the fitted linear combination as cycle factor $\widehat{cf}_t = \hat{\theta}_0 + \hat{\theta}_1 c_t^{(1)} + \hat{\theta}_2 \overline{c}_t$.

Macro and technical analysis driven factor:

Macroeconomic fundamentals also predict bond excess return. I follow Ludvigson and Ng (2009) and Ludvigson and Ng (2011) to estimate one such factor. I first extract J principal components, $\hat{f}_t = (\hat{f}_{t,1}, ..., \hat{f}_{t,J})$ from the set of 135 macroeconomic series, where $J << 135$. Extraction relies on asymptotic PCA, and the number J is determined by the information criteria developed in Bai and Ng (2002). I then perform best subset selection among different subsets of $\{\hat{f}_{t,1}^3, \{\hat{f}_{t,j}, \hat{f}_{t,j}^2; j = 1, ..., J\}\}$ using the BIC criteria.[5] Given a preferred subset $\widehat{F}_t$, I estimate the Ludvigson and Ng factor by running

$$\overline{rx}_{t+1} = \delta_0 + \delta_1' \widehat{F}_t + \overline{\epsilon}_{t+1}, \tag{13}$$

and it follows that $\widehat{LN}_t = \hat{\delta}_0 + \hat{\delta}_1' \widehat{F}_t$.

The technical analysis driven factor I consider is implemented in a similar way except that, principal components are extracted from a set of technical indicators. Following Goh et al. (2013), I build those indicators by comparing two (short and long) moving averages of forward spread. Let $MA_t^{n,j} = \frac{1}{j}\sum_{k=0}^{j-1} f_{t-j/12}^{(n)}, j \in \{s, l\}$ be the s (short) or l (long) months moving average of n-year forward spread, $I(MA_t^{n,s} > MA_t^{n,l})$ would then define one such signal. Combining $n = 2, 3, 4, 5, s = 3, 6, 9$, and $l = 18, 24, 30, 36$ gives us a total of 48 signals. The return-forecasting factor $\widehat{TA}$ is the fitted value of predictive regression on the selected subset of the extracted principal components from the 48 signals.

As a finally remark, all the estimations use only past 15 years of data. Hence, both J and the composition of subset $\widehat{F}_t$ in $\widehat{LN}$ and $\widehat{TA}$ may vary over time as ϕ_t gets updated monthly.

3.2.2. Policy function estimation

Given return predictors $z \in \{FB, \widehat{CP}, \widehat{cf}, \widehat{LN}, \widehat{TA}\}$, I now turn to the question of how to transform the information contained into portfolio

[5]The motivation is that pervasive components in $\hat{f}$ (those with large eigenvalues) is not necessarily the ones most relevant for prediction.

decisions. I lay out two empirical procedures to estimate the portfolio weights. The first one assumes and estimates a conditional log-normal distribution for the return generating process, and then solve for an optimal policy function under the estimated distribution. While, the second one, based on Brandt (1999), bypasses the need to specify a statistical model on return and directly estimates the portfolio weight in a non-parametric GMM framework.

Linear parametric rules:

This is the standard plug-in strategy with portfolio weight determined by the estimated mean excess return divided by its estimated variance, scaled down by the risk aversion level. I assume that bond excess returns $rx_{t+1}^{(n)}$ is log-normally distributed conditional on z and model the return generating process through a predictive regression expressed as $rx_{t+1}^{(n)} = \beta^{(n)} z_t + u_{t+1}$, with homoskedastic error term $u_{t+1} \sim N(0, \sigma^2)$. Using conditional distribution estimated from data ϕ_t, the approximate (up to a log linearization) optimal portfolio weight under this estimated distribution is solved to be:

$$\alpha(\phi_t) = \frac{\hat{\beta}^{(n)} z_t + \hat{\sigma}^2/2}{\gamma \hat{\sigma}^2}. \tag{14}$$

Note that this allocation rule is linear in terms of the current value of return predictor, a consequence of the log-normal assumption. The benchmark strategy, which ignores predictors, will be a special case of the parametric strategy with $z_t \equiv 1$.

Non-linear non-parametric rules:

This strategy allows for a non-linear response to the value of the predictor. Following Brandt (1999), the optimal portfolio weights given a return predictor are now estimated directly through investor's conditional Euler equations. In particular, denote α_t to be the choice variable on portfolio weight at time t, the first order conditions that characterize the portfolio optimization problem can be expressed as:

$$E_t \left[(\alpha_t e^{rx_{t+1}^{(n)} + r_{f,t}} + (1 - \alpha_t) e^{r_{f,t}})^{-\gamma} (e^{rx_{t+1}^{(n)} + r_{f,t}} - e^{r_{f,t}}) \mid z_t = z \right] = 0. \tag{15}$$

These FOCs serve as a set of moment conditions (for each z), and then the method of moments estimator is applied separately in each value of the predictor. Collectively, this yields a point-wise, or non-parametric estimate of the allocation rule $\alpha(z)$.

Operationally, to replace the conditional expectation (point-wise on z) with a proper empirical counterpart, I use sample analog with each observation weighted according to the similarity of its predictor level with the current value z. I adopt a normal kernel density $g(\frac{z_j - z}{h_t})$ on each observation $\{rx_{j+1}^{(n)}, z_j\}$, where h_t is a data dependent bandwidth. By standard practice, I set $h_t = 1.06\hat{\sigma}_z t^{-0.2}$, with $\hat{\sigma}_z$ being the standard deviation estimate of $\{z_j\}_{j=1}^{t}$ within the estimation window and t the window length. The empirical moment condition at time t, denoted by $Q_t(\alpha_t)$, with $z_t = z$ and given choice variable α_t, is expressed as

$$Q_t(\alpha_t) = \frac{\sum_{j=1}^{t-\tau}\left((\alpha_t e^{rx_{j+1}^{(n)} + r_{f,t}} + (1-\alpha_t)e^{r_{f,t}})^{-\gamma}(e^{rx_{j+1}^{(n)} + r_{f,t}} - e^{r_{f,t}})exp(-\frac{(z_j - z)^2}{2h_t^2})\right)}{\sum_{j=1}^{t-\tau} exp(-\frac{(z_j - z)^2}{2h_t^2})}, \quad (16)$$

with the numerator normalizing the weights to sum up to one. The optimal portfolio weight at time t, conditioning on $z_t = z$, is estimated through,

$$\alpha(\phi_t) = \underset{\alpha_t}{\operatorname{argmin}}(Q_t(\alpha_t))^2. \quad (17)$$

Note that, the above procedure has not relied on any statistical model of the return process or any (parametric) functional form of the portfolio policy. Thus, the resulting estimator is less biased and robust to policy function mis-specification. However, non-parametric estimation comes at the cost of loosing observations. As the effective sample size is decreased due to kernel weighting, variance of estimated portfolio weight would be increased (relative to a correctly specified parametric estimator).[6] Such efficiency loss is particularly severe when the current value of predictor, z_t, falls into a sparse region of $\{z_j\}_{j=1}^{t-\tau}$, and thus lacks similar observations. Technically, the numerator in moment estimate $\sum_{j=1}^{t-\tau} exp(-\frac{(z_j - z_t)^2}{2h_t^2})$ at sparse sate would be too close to zero, and the portfolio weight estimate at this point/state would be poor. To partially address this concern, I consider trimming z_t when $\sum_{j=1}^{t-\tau} exp(-\frac{(z_j - z_t)^2}{2h_t^2})$ is below certain threshold. As an example, we use the 10% quantile of density estimates computed at all other observations $\{\sum_{j=1}^{t-\tau} exp(-\frac{(z_j - z_i)^2}{2h_t^2})\}_{i=1}^{t-\tau}$. When triggered, trimming will switch the portfolio advice to the benchmark one. In this way, estimation error at sparse states are controlled. Yet, the remaining additional risk at other states would still be disliked by a risk averse investor. Hence, in a limited data environment, it is not clear à prior whether non-parametric strategies will dominate the parametric ones.

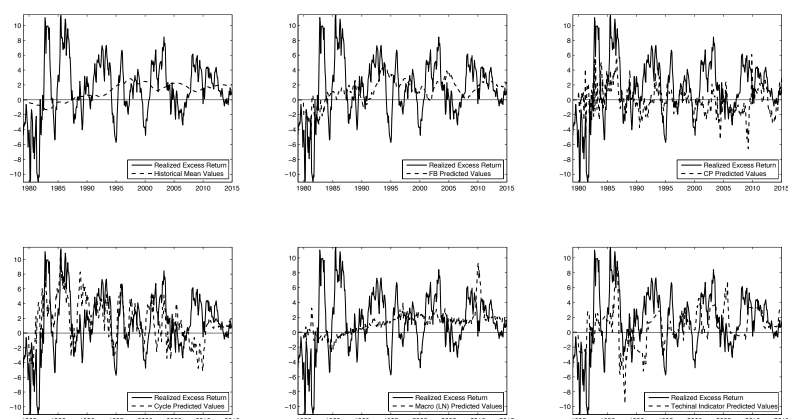
[6]see Brandt (1999) for the expression of standard error on the estimated portfolio weights and the relevant discussion on the estimator's asymptotic properties.

3.3. Empirical findings

3.3.1. Statistical accuracy

I start by looking at the statistical accuracy of the above mentioned predictors in the out-of-sample environment. As a benchmark, I first use the historical mean, which complies with a no-predictability belief. I then switch to the constructed factors $z \in \{FB, \widehat{CP}, \widehat{cf}, \widehat{LN}, \widehat{TA}\}$. I plot in Figure 1 the 15 years rolling window return forecasts (dashed), averaged across maturities: $\sum_{n=2}^{5} \widehat{rx}_{t+12}^{(n)}$, in comparison to the actual values (solid). Each panel corresponds to a particular predictor. As shown in the graph, the benchmark forecast based on the historical mean (top left) turns out too flat relative to the realized ones which fluctuates heavily. In contrast, the forward spread FB (top right) picks up some of the variations in realized return. The forward curve factor $\widehat{CP}$ (middle left) further improves the forecast precision in many episodes. Yet, it breaks down in certain periods such as early 80s and the 07-10 crisis. Similarly, the cycle factor $\widehat{cf}$ (middle right) captures lots of the return spikes, especially from mid 80s to early 00s. But, as $\widehat{CP}$, it fails severely during the early 80s and the crisis period. Macro based predictor, $\widehat{LN}$ (bottom left), is relatively more persistent, but it catches a lower frequency return trend. Finally, technical factor, $\widehat{TA}$ (bottom right), forecasts the correct sign in most instances, while as FB, it does not perform too badly during the crisis.

FIG. 1. Out-of-Sample Bond Return Forecasts with Different Predictors Solid lines represents the realized annual bond excess return (averaged over 2 to 5-year maturity bonds) and dashed lines denote the predicted value based on historical mean (Hist, top left); forward spread (FB, top middle); forward rates (CP, top right); cycle factor (Cycle, bottom left); macro factor (LN, bottom middle) and technical analysis factor (TA, bottom right).



To provide a more quantitative assessment, I resort to the metric of out-of-sample R^2_{OS}, and employ the Clark and West (2007) MSFE-adjusted test to gauge the significance of R^2_{OS}. The null hypothesis is that, predictor does not reduce expected squared forecast error, i.e., $H_0 : E[R^2_{OS}] \leq 0$, and is against a one-sided alternative that it does, i.e., $H_A : E[R^2_{OS}] > 0$. I examine both the whole sample period (ended at Dec 2014) and the pre-crisis one (ended Dec 2007). According to the p-values reported in Table 1, we find that, except for FB at 5 year maturity and $\widehat{LN}$ for whole sample period, all nulls are rejected at 10% confidence level. We therefore conclude: (1) bond excess returns are not characterized by random walk and (2) predictors considered, $z \in \{FB,\ \widehat{CP},\ \widehat{cf},\ \widehat{LN},\ \widehat{TA}\}$, are generally still valid in terms of statistical accuracy under this out-of-sample environment.[7]

3.3.2. Portfolio value – unconditional evaluation

I now turn to the welfare value of above predictors in making bond allocation / market timing decisions. I depict in Figure 2 the rolling window portfolio choices of a CRRA investor using either parametric or non-parametric rules with 15 years of data. Within each panel, a particular predictor is considered and the resulting portfolio weights on the long term bond are plotted against that of the benchmark strategy. For illustration, I visualize only the case of risk aversion $\gamma = 10$ and maturity of long term bond $n = 5$. We see that parametric rule weights based on alternative predictors are generally quite similar to the non-parametric ones, suggesting that linear policy is a reasonable approximation. However, when value of predictor falls into the sparse region, non-parametric estimates will get too noisy and are thus trimmed. Besides, the magnitude of parametric and non-parametric (trimmed) weights range from about -200% to 300% (on the 5-year risky bond) for FB; -400% to 400% for CP; -600% to 600% for cf; -200% to 500% for LN; and -700% to 700% for TA, which looks quite extreme. But given that bond risk premiums are small and predictive regression R^2s are high (at least relative to the case of equity return prediction), it is indeed intriguing to take a bit leverage in collecting those premiums.

[7]I am aware of the multiple and simultaneous hypothesis testing issue that with a total number of $5 \times 4 = 20$ hypothesis tested in the same time, the likelihood of witnessing a rare event and therefore the family-wise type I error rate increase (See Dunn (1961) and Holm (1979)). I conduct Bonferroni correction and Holm-Bonferroni method as conservative ways to control for family wise error rate. With whole sample data, we can no longer reject the null hypothesis of no-predictability. But with pre-crisis data, the cycle factor survives these corrections since the associated p-values are lower than the significance level $\alpha = 0.05$ divided by 20.

TABLE 1.

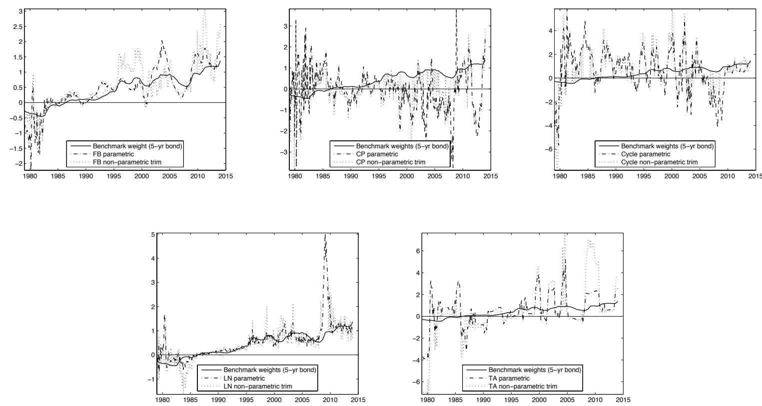
Statistical Accuracy of Bond Return Forecast in Rolling Window Scheme

Entries are the out-of-sample R^2_{OS}s and the p-values on its significance

A. Whole sample: 1964/01 : 2014/12					
Maturity	FB	CP	cf	LN	TA
2-years	0.068	-0.078	0.170	-0.087	0.020
	(0.025)	(0.016)	(0.005)	(0.404)	(0.007)
3-years	0.084	-0.052	0.146	-0.043	0.083
	(0.013)	(0.015)	(0.005)	(0.195)	(0.007)
4-years	0.093	-0.029	0.164	-0.040	0.119
	(0.015)	(0.012)	(0.004)	(0.228)	(0.006)
5-years	0.005	-0.016	0.158	-0.033	0.149
	(0.144)	(0.014)	(0.004)	(0.263)	(0.005)
B. Pre-crisis period: 1964/01 : 2007/12					
Maturity	FB	CP	cf	LN	TA
2-years	0.083	0.040	0.299	-0.018	0.036
	(0.016)	(0.006)	(0.002)	(0.082)	(0.006)
3-years	0.098	0.075	0.289	0.012	0.103
	(0.009)	(0.006)	(0.002)	(0.055)	(0.006)
4-years	0.0992	0.100	0.308	0.034	0.143
	(0.014)	(0.004)	(0.001)	(0.041)	(0.005)
5-years	-0.001	0.103	0.308	0.038	0.173
	(0.182)	(0.005)	(0.001)	(0.038)	(0.004)

I then measure the performance of these allocation rules through the unconditional expected utilities they achieve. I report in Table 2, for each predictor in turn, the point estimates of certainty equivalent gross returns (CERs) of parametric and non-parametric strategy, as well as the unconditional inference results on their welfare benefits relative to the benchmark. From the rows *para CER* and *non-para CER*, we observe that estimated CERs of predictors based timing strategies are frequently lower than that of the no-predictor benchmark (row *Bench CER*). Taking LN factor as an example, the resulting CER estimates for either policies under different bond maturities n range from about 1.035 to 1.053, while those of the benchmark strategy are no lower than 1.061. Likewise, cf, CP, TA based market timing all deliver below benchmark welfare estimates. In fact, the cycle factor, cf, which has the highest out-of-sample R^2, is generating the lowest estimated CERs for each maturity (column) n. The forward spread FB, interacted by linear parametric policy, stands out as an exception for which the CER estimates reach 1.062 and 1.065 when $n = 2, 3$. How-

FIG. 2. Out-of-Sample Bond Portfolio Weights under Different Rules Solid lines denote the estimated portfolio weight based on predominant mean, which serves as the benchmark. Dash-dot and dotted lines denote the portfolio weights generated from, respectively, the parametric and nonparametric rules.



ever, according to the p-values of the unconditional tests (row *p-value (P)* in panel FB), those welfare improvement are not statistically significant. These findings tells us that, despite the fact of non-random walk / return predictability, it is indeed difficult to transform the information contained in identified predictors into expected utility gains at least by the above policies.[8]

As an additional check, I contrast the performance of each parametric timing strategy against the corresponding non-parametric one with same underlying predictor. In terms of CERs estimates, we find evidence are mixed, favoring non-parametric (and trimmed) policies when equipped with CP, LN, or TA, but leaning to parametric ones when using FB or cf. While unreported analysis suggest that non of the welfare differences are statistically significant by conventional standard.[9]

3.3.3. Portfolio value – conditional evaluation

Thus far, the focus has been on the unconditional notion of welfare measure. But as mentioned above, relative performance of market timing can

---

[8]I focus on the economic value of each single return predictor, as combining multiple predictors in multivariate predictive regressions do not lead to superior allocation decisions.

[9]I also repeat all the unconditional performance evaluation analysis using only precrisis data (ended Dec 2007). Results are quantitatively the same, and hence, the above findings are not purely driven by the financial crisis happened in 2008.

TABLE 2.

Unconditional Evaluation of Bond Allocation Rules

	Maturity			
	2-year	3-year	4-year	5-year
Bench CER	1.061	1.062	1.064	1.062
	FB			
Para CER	1.062	1.065	1.064	1.048
p-value (P)	(0.721)	(0.524)	(0.875)	(0.404)
Non-para CER	1.061	1.057	1.049	1.054
p-value (NP)	(0.896)	(0.459)	(0.217)	(0.167)
	CP			
Para CER	0.971	0.968	0.951	0.917
p-value (P)	(0.166)	(0.179)	(0.192)	(0.239)
Non-para CER	1.005	1.000	1.026	1.035
p-value (NP)	(0.245)	(0.227)	(0.151)	(0.122)
	cf			
Para CER	0.729	0.428	0.322	0.257
p-value (P)	(0.296)	(0.305)	(0.302)	(0.304)
Non-para CER	0.297	0.018	0.020	0.040
p-value (NP)	(0.306)	(0.299)	(0.307)	(0.307)
	LN			
Parametric CER	1.035	1.053	1.053	1.048
p-value (P)	(0.405)	(0.572)	(0.484)	(0.432)
Non-para CER	1.046	1.056	1.056	1.055
p-value (NP)	(0.388)	(0.453)	(0.291)	(0.215)
	TA			
Para CER	0.864	0.730	0.666	0.579
p-value (P)	(0.295)	(0.304)	(0.304)	(0.303)
Non-para CER	1.036	0.957	0.944	0.963
p-value (NP)	(0.212)	(0.224)	(0.182)	(0.138)

be heterogeneous across different economic regimes. Accordingly, I conduct welfare estimates and utility benefit inferences conditional on each particular regime. In particular, I let economy be at low / high unemployment state if the contemporary unemployment rate is below / above its average, (which amounts to 6.3% for our whole sample), and a high / low turbulence regime when past year's realized bond market volatility is greater / less than its mean.[10]

[10]In principal, we could have a more refined definition of economic regime such as three or four stage regimes, but this would reduce the number of data for conditional tests and hence decrease power.

I present in Table 3 the estimated CERs of predictors- based timing strategies, along with the associated inference results, conditioning on low unemployment regime (left panel) and high unemployment regime (right panel). We notice that, while forward spread FB and CP factors appear to generate higher CERs in high unemployment than low unemployment periods, the rest of the factors create higher expected utility estimates in low rather than high unemployment episodes. However, conditioning on either regime, the relative performance of timing against the benchmark is generally not significant. Exceptions are the macro LN and technical analysis factor TA driven parametric policies when conditioning on the low unemployment rate regime. The associated p-values of the structural tests on welfare difference are below 10% when $n = 2, 3$, and the rejections are favoring marketing timing for LN and benchmark for TA. This suggests that one could use LN based parametric timing while avoid TA based one using 2-year or 3-year bond when the current economic state is a low unemployment one.

Table 4 examines whether relative performance of bond market timing are specific to market turbulence regime. Following [**?**], I measure such turbulence at annual frequency through the realized or integrated daily return volatility between time $t - 252$ and t, i.e., $\sum_{i=t-252}^{t}(r_{i,d}^{n})^2$, where $r_{i,d}^{n}$ is the daily return of a n-yr bond. Using this volatility measure, I report separately the portfolio evaluation results conditioning on the state of higher (left panel) and lower (right panel) than mean volatility. Based on the relevant CER estimates, we find that most timing strategies more profitable in low volatility than high state. One exception is the non-parametric policy coupled with CP factor, which has higher CER estimates in turbulent state. Conditioning on either volatility regime, none of the relative performance against the benchmark is significant according to the p-values of conditional tests (rows p-value (P) and p-value (NP)).

3.3.4. Discussion on the failure of timing

To better understand the performance of timing strategies, I plot in Figure 3, for each predictor in turn, the rolling window realized returns of parametric (dash-dotted) and non-parametric (dotted) policies in comparison to the benchmark one (solid). I observe that, although for many periods market timing generates extra profits (dash-dotted/dotted curve above the solid line), there are certain episodes in which they lead to huge losses. These episodes tend to be characterized by two features. First, the realized returns in those episodes deviate significantly from the forecasted

TABLE 3.

Conditional Evaluation of Bond Allocation Rules: Unemployment rate

	Low Unemp				High Unemp			
	Maturity				Maturity			
	2	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Bench CER	1.063	1.065	1.066	1.064	1.058	1.059	1.061	1.059
	FB				FB			
Par CER	1.059	1.063	1.066	1.066	1.063	1.065	1.061	1.032
p-val(P)	(0.526)	(0.729)	(0.969)	(0.626)	(0.268)	(0.385)	(0.940)	(0.352)
Non-par CER	1.059	1.049	1.049	1.053	1.061	1.064	1.048	1.054
p-val(NP)	(0.192)	(0.269)	(0.246)	(0.158)	(0.513)	(0.188)	(0.496)	(0.539)
	CP				CP			
Par CER	0.926	0.925	0.908	0.868	1.051	1.041	1.026	1.019
p-val (P)	(0.172)	(0.203)	(0.235)	(0.274)	(0.544)	(0.424)	(0.361)	(0.356)
Non-par CER	0.969	0.964	1.004	1.018	1.056	1.052	1.052	1.053
p-val (NP)	(0.241)	(0.235)	(0.178)	(0.142)	(0.682)	(0.462)	(0.459)	(0.517)
	cf				cf			
Par CER	0.994	0.982	0.938	0.828	0.676	0.395	0.297	0.237
p-val (P)	(0.146)	(0.189)	(0.232)	(0.283)	(0.300)	(0.302)	(0.298)	(0.301)
Non-par CER	1.046	1.045	1.039	1.031	0.274	0.017	0.018	0.037
p-val (NP)	(0.283)	(0.305)	(0.337)	(0.355)	(0.304)	(0.296)	(0.302)	(0.304)
	LN				LN			
Par CER	1.071	1.071	1.071	1.067	1.008	1.038	1.037	1.031
p-val (P)	(0.017)	(0.025)	(0.066)	(0.173)	(0.345)	(0.450)	(0.399)	(0.393)
Non-par CER	1.070	1.070	1.069	1.065	1.026	1.044	1.044	1.045
p-val (NP)	(0.113)	(0.185)	(0.378)	(0.721)	(0.289)	(0.267)	(0.197)	(0.174)
	TA				TA			
Par CER	1.030	1.038	1.046	1.056	0.808	0.676	0.615	0.535
p-val (P)	(0.057)	(0.091)	(0.117)	(0.359)	(0.307)	(0.303)	(0.302)	(0.300)
Non-par CER	1.047	1.049	1.055	1.061	1.024	0.909	0.892	0.914
p-val (NP)	(0.168)	(0.212)	(0.301)	(0.755)	(0.360)	(0.241)	(0.186)	(0.129)

ones. Second, the predictive regressions within the estimation windows preceding those episodes provide good fit, which induce high leverages in portfolio decisions. A typical example is the cycle factor, cf, driven strategy. As the in-sample (within estimation window) fit by this predictor is very good, leverage ratio in portfolio decision increases. While such high leverage effectively captures lots of the risk premiums when forecasted correctly, it also enormously magnify the portfolio loss in presence of forecast instability or breaks. On the other hand, the no-predictability benchmark, although mis-specified, does not create that amount of downside risk and

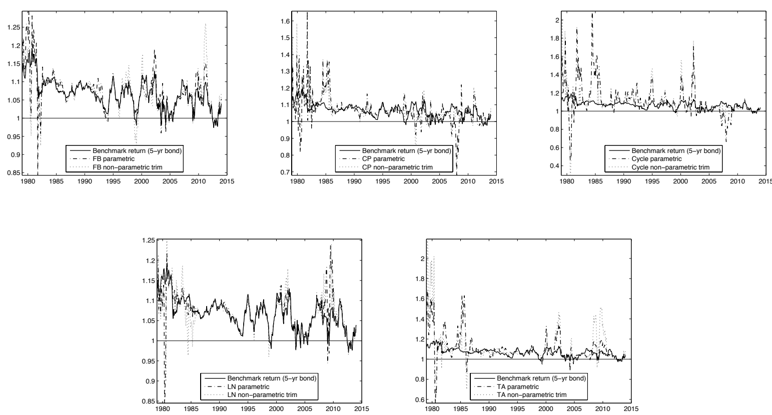
TABLE 4.

Conditional Evaluation of Bond Allocation Rules: Realized Volatility

Condition on	High volatility				Low volatility			
	Maturity				Maturity			
	2	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Bench CER	1.061	1.065	1.068	1.069	1.061	1.060	1.059	1.056
	FB				FB			
Par CER	1.063	1.068	1.067	1.040	1.061	1.061	1.062	1.057
p-val (P)	(0.495)	(0.349)	(0.964)	(0.361)	(0.930)	(0.818)	(0.728)	(0.604)
Non-par CER	1.061	1.066	1.049	1.057	1.060	1.048	1.049	1.050
p-val (NP)	(0.994)	(0.698)	(0.314)	(0.204)	(0.818)	(0.314)	(0.407)	(0.414)
	CP				CP			
Par CER	0.952	0.942	0.917	0.873	0.994	1.003	1.005	1.006
p-val (P)	(0.246)	(0.222)	(0.204)	(0.232)	(0.092)	(0.085)	(0.123)	(0.221)
Non-par CER	1.062	1.060	1.061	1.063	0.968	0.963	1.000	1.012
p-val (NP)	(0.902)	(0.645)	(0.541)	(0.526)	(0.223)	(0.219)	(0.151)	(0.115)
	cf				cf			
Par CER	0.677	0.396	0.298	0.238	1.036	1.031	1.008	0.961
p-val (P)	(0.292)	(0.300)	(0.295)	(0.298)	(0.439)	(0.460)	(0.382)	(0.326)
Non-par CER	0.275	0.017	0.018	0.037	1.062	1.057	1.047	1.037
p-val (NP)	(0.301)	(0.293)	(0.300)	(0.301)	(0.944)	(0.892)	(0.677)	(0.590)
	LN				LN			
Par CER	1.014	1.045	1.046	1.040	1.062	1.062	1.061	1.056
p-val (P)	(0.392)	(0.522)	(0.452)	(0.416)	(0.753)	(0.452)	(0.624)	(0.748)
Non-par CER	1.034	1.055	1.058	1.061	1.059	1.057	1.053	1.049
p-val (NP)	(0.404)	(0.518)	(0.418)	(0.357)	(0.777)	(0.658)	(0.422)	(0.305)
	TA				TA			
Par CER	0.808	0.677	0.617	0.536	1.057	1.062	1.064	1.066
p-val (P)	(0.291)	(0.298)	(0.299)	(0.298)	(0.770)	(0.853)	(0.648)	(0.299)
Non-par CER	1.014	0.907	0.892	0.915	1.064	1.063	1.063	1.062
p-val (NP)	(0.173)	(0.212)	(0.168)	(0.115)	(0.674)	(0.728)	(0.664)	(0.477)

hence is not dominated by the timing strategy. Meanwhile, market timing in a limited data environment also renders allocation decision *more* sensitive to data realizations. While benchmark strategy only needs the unconditional mean and volatility estimates, timing strategy requires the estimation of both return predictor and a parametric or non-parametric policy. Such complexity, of additional parameters (for parametric policy) and different estimation procedure (for non-parametric policy), leads to higher estimation uncertainty and eventually translates into extra volatility in realized returns. The same argument also applies to explain why the less mis-specified non-parametric policy does not outperform the parametric one in a limited data environment. Finally, the negative effects of both model complexity and forecast instability will intertwine, and our findings actually suggest that, the resulting portfolio losses are not dominated by the benefits of incorporating predictors, at least for the parametric and non-parametric policies.

FIG. 3. Out-of-Sample Bond Portfolio Returns under Different Rules
Solid lines denote the gross return of the portfolio based on predominant mean, which serves as the benchmark. Dash-dot and dotted lines denote the gross returns generated from, respectively, the parametric and nonparametric portfolio rules in conjunction with the corresponding predictor.



3.3.5. Shrinkage policy

Given above findings, it is natural to ask whether we can reduce the loss due to model mis-specification and estimation in exploiting predictability. Here, I make one such attempt by compromising between our benchmark strategy and the dogmatic market timing. In particular, I adopt a shrinkage strategy, suggested in Connor (1997) and Brandt (2009) and implemented through a Bayesian predictive regression with informative prior on the slope coefficients. By setting the prior to be no-predictability (slope equal to zero), I effectively tame down the estimated return forecast towards the unconditional mean and only partially exploit the predictability. With confidence on prior expressed in terms of expected R^2 in the predictive regression, the shrinked return forecast $\hat{r}^s_{t+1}$ can then be derived as

$$\hat{r}^s_{t+1} = \left[1 - \frac{t}{t+\frac{1}{\rho}}\right]\hat{\bar{r}} + \left[\frac{t}{t+\frac{1}{\rho}}\right]\hat{\beta}_{ols}z_t, \text{ for } \rho = E^{prior}\left[\frac{R^2}{1-R^2}\right], \quad (18)$$

where $\hat{\bar{r}}$ is the estimated unconditional mean return, $\hat{\beta}_{ols}z_t$ is the original forecast, t is the sample size and $\frac{t}{t+\frac{1}{\rho}}$ is the shrinkage factor.[11] Such representation can be thought of as an intermediate view between the benchmark and predictive regression, with shrinkage factor as the relative weight. In specifying this weight, I examine the whole range of prior confidence, or equivalently the shrinkage factor, to see whether this strategy has potential to out-perform the benchmark.

I report in Table 5, for each bond maturity in turn, the estimated certainty equivalent returns of various shrinkage strategies. I consider shrinkage factors ranging from 0.1 to 0.9 with an increment of 0.1.[12] As the degree of shrinkage gets larger (shrinkage factor smaller), the estimated CERs increase first and then drop. Such pattern holds for all predictors. This suggests that, when return forecasts are tamed down, the reduction in estimation and specification risks will initially benefit the welfare despite of a distorted return forecast. But gradually, risk reductions would be limited and forecast distortion is too big. When compared against the benchmark, we find that at certain (high) level of shrinkage (low value of factor) and especially for those based on TA predictor, the welfare benefit is statistically significant at conventional levels.[13] For example, with a 5-year maturity bond, the TA driven strategy with 50% of shrinkage generates almost 80 basis point gain in certainty equivalent which is significant at 5% confidence level. This indicates that, eventually the investor can partially exploit the return predictability without being completely offset by the associated estimation and mis-specification risks. The results become stronger when we conduct the same analysis using pre-crisis data only. In addition, and as will be shown in the robustness section below, the range of desirable shrinkage that results in utility improvement is not risk aversion

[11] For multivariate predictive regression, each of the slope coefficient will be shrinked according to its marginal degree of predictability, i.e $\left[\frac{t}{t+\frac{1}{\rho_j}}\right]$, where $\rho_j = E\left[\frac{R^2_j}{1-R^2}\right]$ and R^2_j the marginal coefficient of determination by variable j. See Connor (1997) and Brandt (2009) for more detail.

[12] A value of 0 or 1 corresponds to the two extreme cases of benchmark and dogmatic market timing.

[13] However, with a total number of 45 hypothesis tested simultaneously, the rejection of null does not survive the Bonferroni correction to control for family-wise type I error.

specific. One explanation is that shrinkage here only measures the conservativeness on forecasted return distribution. It is not directly affected by the risk averse level, which will be accounted for in the portfolio decision stage.

TABLE 5.

Unconditional Evaluation of Shrinkage Strategies

	FB	CP	cf	LN	TA	FB	CP	cf	LN	TA
	Maturity n=2					Maturity n=3				
shrink	Bench CER: 1.0609					Bench CER: 1.0623				
0.1	1.0613	1.0615	1.0629	1.0623	1.0620	1.0630‡	1.0632	1.0647‡	1.0634	1.0640†
0.2	1.0617	1.0618	1.0642	1.0640†	1.0627	1.0636	1.0636	1.0662	1.0649‡	1.0652‡
0.3	1.0620	1.0615	1.0645	1.0653‡	1.0629	1.0641	1.0635	1.0664	1.0660‡	1.0660
0.4	1.0622	1.0607	1.0635	1.0661	1.0623	1.0644	1.0628	1.0646	1.0666	1.0658
0.5	1.0624	1.0591	1.0597	1.0659	1.0601	1.0647	1.0615	1.0579	1.0667	1.0641
0.6	1.0625	1.0567	1.0498	1.0642	1.0542	1.0648	1.0595	1.0371	1.0661	1.0587
0.7	1.0625	1.0534	1.0235	1.0595	1.0374	1.0648	1.0565	0.9720	1.0643	1.0433
0.8	1.0625	1.0490	0.9586	1.0490	0.9895	1.0647	1.0527	0.8249	1.0609	0.9999
0.9	1.0624	1.0436	0.8416	1.0286	0.8851	1.0645	1.0479	0.6302	1.0554	0.9046
	FB	CP	cf	LN	TA	FB	CP	cf	LN	TA
	Maturity n=4					Maturity n=5				
shrink	Bench CER: 1.0635					Bench CER: 1.0621				
0.1	1.0644‡	1.0646	1.0665†	1.0652†	1.0657†	1.0624	1.0630	1.0653†	1.0639†	1.0647*
0.2	1.0652	1.0650	1.0685‡	1.0669†	1.0675†	1.0625	1.0633	1.0675‡	1.0655†	1.0669*
0.3	1.0657	1.0650	1.0692	1.0681‡	1.0689†	1.0625	1.0629	1.0682	1.0667‡	1.067*
0.4	1.0661	1.0642	1.0675	1.0689‡	1.0695‡	1.0623	1.0619	1.0665	1.0675†	1.0698†
0.5	1.0663	1.0627	1.0606	1.0691	1.0689	1.0620	1.0599	1.0597	1.0677	1.0699†
0.6	1.0663	1.0602	1.0383	1.0688	1.0662	1.0615	1.0568	1.0394	1.0672	1.0680
0.7	1.0662	1.0566	0.9674	1.0676	1.0588	1.0608	1.0523	0.9795	1.0658	1.0614
0.8	1.0658	1.0518	0.8093	1.0653	1.0399	1.0598	1.0461	0.8395	1.0634	1.0424
0.9	1.0653	1.0457	0.6054	1.0616	0.9964	1.0585	1.0383	0.6401	1.0593	0.9943

Note: $*$, † and ‡, denote significance at 1%, 5% and 10% level against benchmark.

3.3.6. Estimation window averaging

As an additional check, I investigate an alternative way to mitigate forecast instability. In previous sections, our portfolio strategies have relied on the use of all available data in the information set ϕ_t. But in presence of potential break in forecast relation *within* this estimation window, it may or may not be optimal to use the whole 15 years length. The reason is that, when the size of break is small, adding pre-break data may reduce forecast error variance. However, when the size is big, it is the effect of bias that dominates. In addition, the estimation of time and size of a break with limited data is usually subject to considerable uncertainty. To alleviate these concerns, I borrow the idea in Pesaran and Timmermann (2007), which combines return distribution forecasts based on predictive regressions with estimation window of different length. In particular, rather than selecting a single estimation window, I pool three return forecasts based respectively on 5 years; 10 years; and 15 years data with equal weight. More specifically, I generate one set of simulated returns from all three estimated distributional forecasts (by a particular predictor) and then solve for the portfolio decision numerically. I also poor over different predictors as an additional model averaging check. Table 6 documents the evaluation results of such window averaging strategies. Comparing with Table 2, pooling over windows improves the estimated CERs for most of the predictors with the exception of FB and TA at short term maturities ($n = 2, 3$). But relative to the benchmark, it still fails to outperform. Finally, pooling over predictors does not help either (sub-panel "Model average").

TABLE 6.

Estimation Window Averaging

	Maturity			
	2-year	3-year	4-year	5-year
Bench CER	1.061	1.062	1.064	1.062
	FB			
CER	1.058	1.060	1.059	1.045
	(0.543)	(0.584)	(0.503)	(0.299)
	CP			
CER	1.024	1.027	1.024	1.004
	(0.051)	(0.048)	(0.052)	(0.092)
	cf			
CER	1.010	0.912	0.850	0.854
	(0.204)	(0.275)	(0.279)	(0.239)
	LN			
CER	1.052	1.061	1.061	1.059
	(0.565)	(0.795)	(0.599)	(0.476)
	TA			
CER	0.738	0.854	0.931	0.900
	(0.305)	(0.304)	(0.304)	(0.303)
	Model average			
CER	1.059	1.061	1.064	1.062
	(0.745)	(0.819)	(0.923)	(0.971)

4. ROBUSTNESS CHECK

This part of the paper checks the robustness of our baseline evaluation results in section 3.3 with respect to: (1) different level of relative risk aversion γ; and (2) different size of estimation window. The first exercise involves the consideration of parametric and non-parametric strategies at two alternative risk aversion coefficients $\gamma = 5$ and $\gamma = 15$. As shown in Table 7, in both cases, the results are qualitatively similar to the ones in Table 2, so that none of the timing strategies differs from the benchmark significantly. And noteworthy, when risk aversion is low, i.e., $\gamma = 5$, some of the entries on CER estimates, especially for cf and TA, are close to zero. This is due to the fact that leverage ratio of less risk averse investor would go up sharply. Then under forecasting model instability, such high leverage leads to ex-post bankruptcy at some states and kills the corresponding strategy.[14]

The second test focuses on the shrinkage strategies with risk aversion ranging from $\gamma = 5, 10, 15$ to 20. I report only the estimated CERs and their significance against the benchmark when the maturity of long term bond equals 5 years. As illustrated in Table 8, entries in different panels, which correspond to different γ s, exhibit similar pattern. Just as our baseline results, when shrinkage factor gets smaller gradually, all estimated CERs initially go up and then slightly drop. In addition, the range of shrinkage that leads to a significant utility benefit remains the same across γ, especially for TA based strategies.[15] This indicates that the role of shrinkage is not specific to any particular choice of risk aversion.

The third robustness test considers changing the size of information set. Specifically, I set the length of limited data available to investor at 20 years. According to Table 9, evaluation results are still qualitatively similar to Table 2, so that a lot of competing forecasts fail to beat the benchmark. Exceptions are the technical indicator TA based parametric timing strategy operated with a 5-year bond and LN based parametric strategies on longer maturity bonds. But the significance of welfare improvement merely cross the 10% threshold. To summarize, our conclusion that it is generally hard to exploit bond return predictability in a limited data environment is not sensitive to the level of risk aversion and size of information set.

---

[14]Since CRRA utility is not defined on negative payoff, we truncate the loss at a gross return of 0.01 (close to bankruptcy). This explains why some of the entries on CER estimates, especially in the left panel when $\gamma = 5$, are close to zero.

[15]The range of shrinkage in which cycle factor LN based strategies outperform the benchmark are also approximately the same.

TABLE 7.

Parametric and Non-parametric Strategies at Different Risk Aversion

Risk aversion	$\gamma = 5$				$\gamma = 15$			
	Maturity				Maturity			
	2	**3**	**4**	**5**	**2**	**3**	**4**	**5**
Bench CER	1.073	1.074	1.076	1.073	1.055	1.056	1.057	1.056
	FB				FB			
Par CER	1.077	1.080	1.066	0.803	1.055	1.058	1.059	1.054
p-val(P)	(0.643)	(0.530)	(0.647)	(0.310)	(0.796)	(0.543)	(0.617)	(0.612)
Non-par CER	1.073	1.075	1.012	1.046	1.054	1.050	1.051	1.053
p-val(NP)	(0.969)	(0.926)	(0.295)	(0.130)	(0.678)	(0.389)	(0.259)	(0.258)
	CP				CP			
Par CER	0.490	0.398	0.240	0.045	1.003	1.002	0.995	0.981
p-val (P)	(0.297)	(0.303)	(0.306)	(0.307)	(0.156)	(0.167)	(0.181)	(0.226)
Non-par CER	0.847	0.793	0.959	1.002	1.025	1.023	1.036	1.040
p-val (NP)	(0.280)	(0.279)	(0.195)	(0.141)	(0.233)	(0.212)	(0.142)	(0.116)
	cf				cf			
Par CER	0.038	0.034	0.034	0.030	0.920	0.775	0.716	0.677
p-val (P)	(0.299)	(0.294)	(0.292)	(0.194)	(0.265)	(0.299)	(0.297)	(0.294)
Non-par CER	0.038	0.034	0.034	0.034	0.634	0.407	0.358	0.445
p-val (NP)	(0.299)	(0.294)	(0.294)	(0.294)	(0.304)	(0.306)	(0.305)	(0.301)
	LN				LN			
Par CER	0.776	1.012	1.014	0.980	1.046	1.054	1.054	1.051
p-val (P)	(0.313)	(0.384)	(0.366)	(0.340)	(0.521)	(0.780)	(0.602)	(0.516)
Non-par CER	1.062	1.040	1.042	1.047	1.050	1.055	1.054	1.053
p-val (NP)	(0.497)	(0.309)	(0.222)	(0.173)	(0.527)	(0.681)	(0.422)	(0.291)
	TA				TA			
Par CER	0.045	0.045	0.023	0.032	0.980	0.926	0.895	0.849
p-val (P)	(0.307)	(0.306)	(0.297)	(0.288)	(0.275)	(0.296)	(0.299)	(0.299)
Non-par CER	0.216	0.045	0.241	0.086	1.044	1.010	1.003	0.993
p-val (NP)	(0.306)	(0.307)	(0.275)	(0.306)	(0.262)	(0.198)	(0.164)	(0.173)

5. CONCLUDING REMARKS

In this essay, I adopt a hypothesis testing approach to assess the portfolio value of a variety of identified bond return forecasts in timing the bond market. I emphasize on the practical usefulness of return predictors in a limited data environment where forecast relations can merely be estimated. I consider allocation rules that vary from not only the return predictors but also the policy functions. I evaluate their performances relative to that of a simple no-predictability benchmark strategy on both unconditional and

TABLE 8.

Shrinkage Strategies at Different Risk Aversion

	FB	CP	cf	LN	TA	FB	CP	cf	LN	TA
	$\gamma = 5$					$\gamma = 10$				
shrink	Bench CER: 1.073					Bench CER: 1.062				
0.1	1.073	1.075	$1.082^{\dagger}$	$1.077^{\dagger}$	1.079^{*}	1.062	1.063	$1.065^{\dagger}$	$1.064^{\dagger}$	1.065^{*}
0.2	1.073	1.076	$1.088^{\ddagger}$	$1.080^{\dagger}$	1.084^{*}	1.062	1.063	$1.067^{\ddagger}$	$1.066^{\dagger}$	1.067^{*}
0.3	1.072	1.076	1.090	$1.083^{\ddagger}$	1.088^{*}	1.062	1.063	1.068	$1.067^{\ddagger}$	1.069^{*}
0.4	1.071	1.074	1.087	$1.084^{\ddagger}$	$1.090^{\dagger}$	1.062	1.062	1.066	$1.067^{\ddagger}$	$1.070^{\dagger}$
0.5	1.069	1.070	1.067	1.085	$1.090^{\ddagger}$	1.062	1.060	1.060	1.068	$1.070^{\dagger}$
0.6	1.066	1.064	0.962	1.083	1.084	1.062	1.057	1.039	1.067	1.068
0.7	1.061	1.054	0.310	1.078	1.056	1.061	1.052	0.979	1.066	1.061
0.8	1.054	1.038	0.037	1.068	0.913	1.060	1.046	0.839	1.063	1.042
0.9	1.043	1.015	0.034	1.045	0.343	1.059	1.038	0.640	1.059	0.994
	FB	CP	cf	LN	TA	FB	CP	cf	LN	TA
	$\gamma = 15$					$\gamma = 20$				
shrink	Bench CER: 1.056					Bench CER: 1.052				
0.1	1.057	1.057	$1.058^{\ddagger}$	$1.058^{\ddagger}$	1.058^{*}	1.052	1.052	1.053	$1.053^{\ddagger}$	1.053^{*}
0.2	1.057	1.057	1.059	$1.059^{\ddagger}$	1.059^{*}	1.053	1.052	1.053	$1.054^{\ddagger}$	1.054^{*}
0.3	1.057	1.057	1.059	$1.059^{\ddagger}$	1.060^{*}	1.053	1.052	1.053	$1.054^{\ddagger}$	1.055^{*}
0.4	1.057	1.056	1.058	1.060	1.061^{*}	1.053	1.051	1.052	1.055	1.055^{*}
0.5	1.057	1.054	1.054	1.060	$1.061^{\dagger}$	1.053	1.050	1.050	1.055	$1.056^{\dagger}$
0.6	1.057	1.052	1.044	1.060	1.060	1.053	1.049	1.043	1.055	1.055
0.7	1.057	1.050	1.018	1.059	1.057	1.053	1.047	1.029	1.055	1.054
0.8	1.056	1.046	0.962	1.058	1.050	1.053	1.044	0.998	1.054	1.050
0.9	1.056	1.041	0.873	1.056	1.031	1.053	1.041	0.947	1.053	1.041

Note: $*$, † and ‡, denote significance at 1%, 5% and 10% level against benchmark.

conditional bases. While the unconditional assessments ask whether return predictor is valuable on average, the conditional ones allow for performance heterogeneity and gauged their relative performance conditional on different economic regimes. The estimation of performance measure relied on an out-of-sample portfolio construction exercise and the inference procedure built on the forecast evaluation literature in a structural way.

Empirically, using monthly US data, I find that major return predictors identified based on either yield curve, macro-fundamental or technical analysis indicators, coupled with parametric or non-parametric strategies, fail to outperform the benchmark rule. This suggest that welfare loss due to estimation uncertainty and forecasting model instability is not dominated by the benefit of incorporating return predictors. Conditional tests indi-

TABLE 9.

A Longer Length of Information Set / Limited Data

	Maturity			
	2-year	3-year	4-year	5-year
Bench CER	1.055	1.056	1.057	1.056
	FB			
Par CER	1.058	1.061	1.064	1.061
p-val (P)	(0.401)	(0.309)	(0.224)	(0.102)
Non-par CER	1.058	1.057	1.057	1.056
p-val (NP)	(0.376)	(0.646)	(0.995)	(0.929)
	CP			
Par CER	1.025	1.027	1.025	1.024
p-val (P)	(0.125)	(0.138)	(0.173)	(0.220)
Non-par CER	1.050	1.052	1.054	1.053
p-val (NP)	(0.460)	(0.426)	(0.532)	(0.609)
	cf			
Par CER	1.038	1.030	1.014	0.956
p-val (P)	(0.458)	(0.399)	(0.362)	(0.320)
Non-par CER	1.059	1.058	1.054	1.046
p-val (NP)	(0.582)	(0.865)	(0.838)	(0.643)
	LN			
Par CER	1.061	1.063	1.061	1.060
p-val (P)	(0.185)	(0.090)	(0.083)	(0.051)
Non-par CER	1.056	1.056	1.051	1.048
p-val (NP)	(0.869)	(0.892)	(0.519)	(0.391)
	TA			
Par CER	1.056	1.062	1.065	1.068
p-val (P)	(0.817)	(0.446)	(0.261)	(0.079)
Non-par CER	1.053	1.054	1.054	1.052
p-val (NP)	(0.801)	(0.715)	(0.641)	(0.678)

cate that the failure of market timing is not specific to any economic regime of unemployment level and market turbulence state. On the other hand, a shrinkage strategy implemented through Bayesian predictive regression, combined with random walk prior, manage to beat the benchmark at certain range of prior confidence. The main results are shown to be robust to investor's risk aversion, to the size of information set, and are not completely driven by the outlier of the 2008 financial crisis.

REFERENCES

Bai, Jushan, and Serena Ng, 2002. Determining the number of factors in approximate factor models. *Econometrica* **70(1)**, 191–221.

Barberis, Nicholas, 2000. Investing for the long run when returns are predictable. *Journal of Finance* **55(1)**, 225–264.

Brandt, Michael, 1999. Estimating portfolio and consumption choice: A conditional euler equations approach. *Journal of Finance* **54(5)**, 1609–1645.

Brandt, Michael, 2009. Portfolio choice problems. *in: Handbook of Financial Econometrics.* **1**, 269–336.

Cieslak, Anna, and Pavol Povala, 2015. Expected returns in treasury bonds. *Review of Financial Studies.*

Clark, Todd, and Michael McCracken, 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105(1)**, 85–110.

Clark, Todd, and Kenneth West, 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* **138(1)**, 291–311.

Cochrane, John, and Monika Piazzesi, 2005. Bond risk premia. *American Economic Review* **95(1)**, 138–160.

Connor, Gregory, 1997. Sensible return forecasting for portfolio management. *Financial Analyst Journal*, 44–51.

Diebold, Francis, and Roberto Mariano, 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13(3)**.

Dunn, Olive, 1961, Multiple comparisons among means. *Journal of American Statistical Association* **56(293)**, 52–64.

Fama, Eugene, and Robert Bliss, 1987. The information in long-maturity forward rates. *American Economic Review*, 680–692.

Giacomini, Raffaella, and Halbert White, 2006. Tests of conditional predictive ability. *Econometrica* **74(6)**, 1545–1578.

Goh, Jeremy, Fuwei Jiang, Jun Tu, and Guofu Zhou, 2013. Forecasting bond risk premia using technical indicators. *Working Paper.*

Holm, Sture, 1979. A simple sequentially rejective multiple test procedure. *Scandinavia Journal of Statistics*, 65–70.

Ludvigson, Sydney, and Serena Ng, 2009. Macro factors in bond risk premia. *Review of Financical Studies.*

Ludvigson, Sydney, and Serena Ng, 2011. A factor analysis of bond risk premia. *in: Handbook of Empirical Economics and Finance.*, 313.

McCracken, Michael, and Serena Ng, 2016. Fred-md: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics.*

Pesaran, Hashem, and Allan Timmermann, 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137(1)**, 134–161.

Stock, James, and Mark Watson, 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* **20(2)**, 147–162.

Stock, James, and Mark Watson, 2005. Implications of dynamic factor models for var analysis. *Working paper.*

Thornton, Daniel, and Giorgio Valente, 2012. Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies* **25(10)**, 3141–3168.

van Binsbergen, Jules, Michael Brandt, and Ralph SJ Koijen, 2012. Decentralized decision making in investment management. *in: The Oxford Handbook of Quantitative Asset Management.*.

Viceira, Luis, 2012. Bond risk, bond return volatility, and the term structure of interest rates. *International Journal of Forecasting* **28(1)**, 97–117.

West, Kenneth, 1996. Asymptotic inference about predictive ability. *Econometrica*, 1067–1084.

White, Halbert, 1984. *Asymptotic Theory for Econometricians*. Academic press New York.

Wooldridge, Jeffrey, and Halbert White, 1988. Some invariance principles and central limit theorems for dependent heterogeneous processes. *Econometric Theory* **4(02)**, 210–230.