What matters when?
Time-varying sparsity in expected returns*

Daniele Bianchi†  Matthias Büchner‡  Andrea Tamoni§

First draft: May 2019.  This draft: January 10, 2020

Abstract
We provide a measure of sparsity for expected returns within the context of classical factor models. Our measure is inversely related to the percentage of active predictors. Empirically, sparsity varies over time and displays an apparent countercyclical behavior. Proxies for financial conditions and for liquidity supply are key determinants of the variability in sparsity. Deteriorating financial conditions and illiquid times are associated with an increase in the number of characteristics that are useful to predict anomaly returns (i.e., the forecasting model becomes more dense). Looking at specific categories of characteristics, we find that variables classified as value, trading frictions and, in particular, profitability are robustly present throughout the sample. A strategy that exploits the dynamics of sparsity to time factors delivers substantial economic gain out-of-sample relative to a random walk, a simple rolling window shrinkage estimator, as well as relative to models based on preselected, well-known characteristics like size, momentum, book-to-market, investment and accruals.

Keywords: Sparsity, Returns Predictability, Cross-Section of Returns, Anomalies, Asset Pricing.

JEL codes: C38, C45, C53, E43, G12, G17.

*We are grateful to Dimitris Korobilis for helpful comments and sharing part of his code. We thank Giovanni Ricco, Lucrezia Reichlin as well as participants at the 3rd Annual Workshop on Financial Econometrics at Örebro University and the research seminar at Now-Casting Ltd London.
†School of Economics and Finance, Queen Mary University of London, Mile End Rd, London E1 4NS, UK. E-mail: d.bianchi@qmul.ac.uk  Web: whitesphd.com
‡Warwick Business School, University of Warwick, Scarman Road, Coventry CV4 7AL, UK. E-mail: matthias.buechner.16@mail.wbs.ac.uk  Web: http://mbuechner.com
§Department of Finance, Rutgers Business School, 1 Washington Park, Newark, NJ 07102. E-mail: andrea.tamoni.research@gmail.com  Web: https://andreatamoni.meltinbit.com
1 Introduction

A fundamental tenet of asset pricing is that investors should be compensated for their exposure to sources of systematic risk. Unfortunately, empirical researchers are left with tens—if not hundreds—of sensible explanatory variables that may provide useful information about the behavior of future expected returns. Of course, theoretical models offer guidance in identifying factors driving risk premia. However, it is also the case that theoretical models are too stylized to explicitly describe all sources of risk in the economy. As a result, empiricists face a trade-off: they could pre-select the candidate explanatory variables for returns by appealing to economic theories, existing empirical literature, and a variety of heuristic arguments, with the risk of omitting important predictors.

Alternatively, one could use the entire set of available potential predictors. Confronted with a large set of predictors, however, the out-of-sample performance of standard techniques, such as ordinary least squares or Bayesian inference with uninformative priors tends to deteriorate as the dimensionality of the data increases; a well-known fact that goes under the name of curse of dimensionality. Perhaps not surprisingly, regularization and variable selection techniques have become more relevant to prevent over-fitting and to solve ill-posed problems by inducing sparsity and/or shrinkage.

In this paper we propose a novel modelling approach that: (1) starts with a large set of predictors and, thus, avoids pre-selection; (2) permits for a sparse relation between returns and sources of risk (i.e. only few variables may be driving expected returns) and, hence, is coherent with economic theories suggesting the presence of only a handful of factors; (3) and, finally, it allows for the (potentially sparse) relation between returns and predictors to change through time both in terms of the identity of the risk factors as well as in terms of their number. Importantly sparsity is entirely data-driven and is not imposed a priori.

We apply our methodology to forecast portfolios returns on eight popular factors over the period July 1963 through June 2017: value, size, momentum, investment, profitability, idiosyncratic volatility, betting-against-beta, and accruals. We use a large set of more than 60 characteristics that have been documented to predict (individually) returns in the cross-section of US equities. Our empirical analysis shows that sparsity varies substantially over time with the amount of active predictors ranging from less than 30% during the late 90s, to around 60% during the recession of early 2000s. Our measure-
of sparsity features substantial business cycle fluctuations: sparsity is generally low around recession periods, such as during the late 2000s and the 2007-2009, and higher in expansions.

Financial conditions and proxies for liquidity supply are important determinants of the dynamics of sparsity. Indeed, we find a strong association of sparsity with the VIX (a proxy for the level of risk faced by imperfectly diversified liquidity providers), and the leverage component of the National Financial Condition Index (NFCI) held by the Chicago Fed. Overall, our findings are consistent with the fact that when liquidity is abundant investors exploit a large number of characteristics to time factors. In principle, the factors could over time become too crowded, making some characteristics irrelevant (so that the probability of inclusion decreases). However, in bad times, when capital becomes scarce, several characteristics return to be useful to predict factor returns.

The dynamics of aggregate sparsity mask interesting heterogeneity across and within categories. Indeed, we find that the fraction of significant firms’ characteristics is relatively stable for profitability and trading frictions categories. On the other hand, the investment and value categories display a decreasing, respectively increasing, probability of inclusion, with substantial low frequency movements transmitted from these two categories to aggregate sparsity.

Turning to individual characteristics we find that the most robust signal within the category of past returns is given by firms’ performance 12 to seven months prior to portfolio formation in line with the findings by Novy-Marx (2012). Within the value category, we find that a composite of accounting ratios that incorporates earnings, cash, sales, and book value, along with price is needed to capture value (Asness et al., 2000). Several variables within profitability are found to be useful predictors of anomaly returns: profit margin, return on assets, return on net operating assets, and return on cash stand out with probability of inclusion at least as high as 70%. On the other hand, despite a relatively stable probability of inclusion displayed by the trading frictions category, we observe substantial variation in the relevance (for return predictability) of the characteristics pertaining to this category: for example the idiosyncratic volatility is more important in the first part of the sample, whereas the level and variability of trading activity (as proxied by volume, see Chordia et al., 2001b) become more relevant in the latter part of the sample; and the bid-ask spread enters intermittently in line with a probability of inclusion of about 50%.

To evaluate the economic significance of time-varying sparsity for expected returns, we develop a-
We start with a benchmark trading strategy that exploits the historical (unconditional) mean to allocate capital across portfolios. We also consider two strategies that, following Lewellen (2015), exploit a limited number of preselected, well-established, characteristics: book-to-market, momentum, size, accruals, split-adjusted shares outstanding, return on asset, and growth in total assets. Finally, we consider as alternative benchmark the trading signal from a rolling-window shrinkage estimator, namely a ridge regression. We find that our strategy that exploits time-varying sparsity consistently outperforms the benchmarks throughout the sample, and it delivers an annualized Sharpe ratio of 0.34.

Overall, our results provide little evidence in support of the conventional wisdom that posits a stable relation between expected returns and a handful of characteristics. In particular, we show that both the number and the identity of significant risk factors substantially change over time, something that has not been shown before in the literature. In addition, our results suggest that linking time variation in sparsity to the level of risk faced by liquidity providers seems a fruitful avenue of research.

1.1 Related literature

Our paper builds on a recent literature that exploits a large number of characteristics to predict the cross section of stock returns. Using state-of-the-art machine learning techniques, these papers conclude that among the large collection of characteristics, only a handful is statistically useful to explain return variation. On the other hand, Kozak et al. (2019) find that a small number of principle components (i.e., a dense model) predict the cross section better than a small number of characteristics. Inspired by these papers we entertain the hypothesis that the forecasting model is sparse with only a few characteristics being relevant; however, we allow for this sparsity to evolve through time. In other words, our framework can accommodate both sparse and dense models.

Other papers have highlighted that the importance of characteristics for expected returns may vary over time. For instance, Mclean and Pontiff (2016) show that for 97 return predictors, predictability decreases by 58% post-publication. Chinco et al. (2019) propose a novel approach to measure the

---


2The difference in findings of the above mentioned papers can also be explained partially by different sample selection criteria, definition of characteristics, sample period, estimation method used, etc.
probability of accepting a *as-yet undiscovered* risk factor. They find that the anomaly “base rate” fluctuates substantially over time.

By using preselected characteristics (on the first 10 years of data), Freyberger, Neuhierl and Weber (2017) document that the *functional association* between expected returns and characteristics varies over time. We instead fix the functional form to be linear and perform selection at each time $t$. Hence, whereas Freyberger, Neuhierl and Weber (2017) emphasize the variation in the functional relation for a fixed set of characteristics, we emphasize the variation in sparsity (or density) of the forecasting model for a fixed functional form.

Differently from all these papers, we forecast characteristic-sorted portfolios rather than individual stocks. In this sense our paper is closely related to Haddad et al. (2019). Whereas Haddad et al. (2019) use book-to-market ratios to forecast the dominant principal components (PCs) from a large set of stock “anomaly” portfolios, we instead use a large set of characteristics (including book-to-market) to forecast specific anomalies closely followed by investors. Of course, one could apply the novel approach proposed by Haddad et al. (2019) (i.e., forecasting the dominant PCs driving variation in the realized factors returns) together with our methodology to measure sparsity.

Our paper is related to the work by Dangl and Halling (2012) and Johannes et al. (2014) who study the effect of drifting regression coefficients for market predictability. There are, however, important differences between these papers and our work. We are interested in understanding the amount of time-varying coefficients, its drivers and economic consequences, in a (relatively large) setting with more than 50 predictors. Furthermore, our interest is in detecting predictable variation of anomaly portfolios returns. Indeed we find that the type of predictors, as well as the dynamics of sparsity, depend on whether we aim to predict characteristic-sorted portfolio returns or the aggregate market. Similar to Johannes et al. (2014) we document the importance of accommodating stochastic volatility when studying time-variation of regression coefficients.

Methodologically, there has been a great deal of interest in shrinkage and regularization methods within the context of models with time-varying parameters, see Nakajima and West (2013), Belmonte et al. (2014), Kalli and Griffin (2014), Uribe and Lopes (2017), Bianchi and McAlinn (2018), and Bitto.

---

A large literature investigates the role of structural breaks for predictive relation, (see, e.g., Paye and Timmermann, 2006; Pettenuzzo and Timmermann, 2011).
and Frühwirth-Schnatter (2019), among others. In practice, however, the actual implementation of these methodologies is restricted to a handful of forecasting variables due to their computational complexity since the estimation is based on computationally expensive MCMC algorithms.

Instead, we follow Rockova and McAlinn (2017) and Koop and Korobilis (2018) and implement a fast and reliable approximation method for the estimation of high-dimensional time-varying parameter models. In particular, our framework is similar in spirit to Koop and Korobilis (2018) whereby a variational-based method is used to approximate the true high-dimensional posterior distribution of the parameters of interest. We extend their time-varying sparsity framework to a panel dynamic regression with dynamic portfolio fixed effects.

The remainder of the paper proceeds as follows. Section 2 provides additional motivation and preliminary evidence in favor of time-varying sparsity. Section 3 introduces our econometric specification and outlines the estimation algorithm. Section 4 presents our empirical analysis with particular focus on our sparsity measure and its determinants. Section 5 develops a trading strategy that exploits time-varying sparsity and compares it with various benchmark strategies. Section 6 studies how the time-variation in characteristics useful for forecasting changes when we target the aggregate market returns rather than anomaly-based portfolios. Section 7 contains robustness checks. Section 8 concludes.

2 Sparsity: Motivation and Preliminary Evidence

Understanding the dynamics of risk premia is one of the fundamental goals in empirical asset pricing. However, risk premia are notoriously difficult to measure since the time series variation of expected asset returns is often buried in the noise of realized returns. A low signal-to-noise ratio has been often addressed by searching for additional sources of conditional information, that is by using more predictors. As a result, the collection of candidate conditioning variables for risk premia measurement is getting bigger and bigger (see Harvey and Liu, 2019 and references therein).

A direct consequence of the proliferation of predictors, and conditioning information more generally, is the lack of consensus of what matters and when. Table 2 shows this case in point. Different papers argue in favor of different sets of characteristics to capture the time series variation of risk-
premia, and, as the table clearly shows, there is overall little agreement among studies. Part of the problem is due to the fact that the predictors’ effects are assumed constant over time. Indeed, when dealing with a large set of covariates, the discussion in empirical asset pricing has been typically framed within the context of static linear shrinkage estimators designed to help regularize otherwise standard OLS estimates.

The simplest regularization/shrinkage technique takes the name of ridge regression and it is obtained by adding a squared penalty to the least-squares estimates of the betas. Such shrinkage estimator can be conveniently recast as a prior on the regression betas $\beta \sim N(0, \lambda^{-1}I_p)$; this implies a posterior estimate of the form

$$\hat{\beta} = \left(X'X + \lambda I_p\right)^{-1} X'Y,$$

which corresponds to a standard ridge regression with penalty parameter $\lambda$. The dependence of all parameters $\beta_j, j = 1, \ldots, p$ on the unknown parameter $\lambda$ helps regularizing the traditional least-squares estimator. As a matter of fact, for $\lambda \to 0$ one obtains $\hat{\beta} = (X'X)^{-1} X'Y = \hat{\beta}_{OLS}$.

It is easy to see that the shrinkage parameter is inherently static. However, accounting for a changing environment is important in asset pricing. Structural breaks and regime changes (see, e.g., Paye and Timmermann, 2006; Lettau and Van Nieuwerburgh, 2007), as well as anomalies that raise and decay over time (see, e.g., Mclean and Pontiff, 2016) are quite common in applications of equity premium forecasting. Figure 1 further elaborates on this point. The figure reports the amount of shrinkage (relative to a standard OLS) implied by a ridge regression estimated on a rolling basis. The blue (red) line shows the results based on a rolling window of 60 (120) months. The dependent variable is a set of long-short portfolios based on common anomalies, and the independent variable is a large set of firms’ characteristics used in the existing literature to capture expected excess returns. A complete description of the data is provided in Section 4.1 below.

[Insert Figure 1 here]

Two facts emerge; first, shrinkage in the cross-section varies over time as indicated by large fluctuations in the rolling estimates of $\lambda$. Interestingly, shrinkage tend to increase in the early 90s and around the
great financial crisis of 2008/2009, although with different magnitudes. Second, the time variation in shrinkage mostly depends on the size of the window in the rolling estimates, with a shorter window implying more variability in the shrinkage estimates. To sum up, the evidence in Table 2 and Figure 1 justifies an encompassing framework that (1) allows to be agnostic in the set of predictors initially considered, and (2) dynamically selects those predictors that are relevant to best capture the time-series variation of risk premia.

3 Modeling time-varying sparsity

We model expected returns \( y_{i,t} \) on portfolio \( i = 1, \ldots, n \) as a linear function of a set of characteristics \( C_{i,t-1} \) which are all a priori considered relevant but possibly with varying degree. A canonical and relevant approach is to use a time-varying parameter model with stochastic volatility of the form

\[
y_{i,t} = x'_{i,t-1} \beta_t + \sigma_{i,t} \varepsilon_{i,t}, \quad i = 1, \ldots, n
\]

\[
\beta_t = \beta_{t-1} + \eta_t, \quad i = 1, \ldots, n
\]

\[
\log(\sigma^2_{i,t}) = \log(\sigma^2_{i,t-1}) + \xi_{i,t},
\]

where \( x'_{i,t-1} = (1, C'_{i,t-1}) \) is a \((p + 1)\)-dimensional vector of predictors including an intercept, \( \beta'_{t} = (\beta_{0,t}, \beta_{1,t}, \ldots, \beta_{p,t}) \) is the \((p + 1)\) vector of loadings, \( \varepsilon_{i,t} \sim N(0,1) \) is some observation noise, \( \eta_t \sim N(0,Q_t) \) with \( Q_t \) a \((p \times p)\) state equation covariance matrix, and \( \xi_{i,t} \sim N(0, \nu_{i,t}) \). Notice that the regression restricts the loading on a characteristic to be the same in the cross-section, consistent with the idea that it is only the spread in characteristics that drives the cross-sectional variation in the expected return of the portfolios. Our proposed approach deals with two main obstacles that are often faced in empirical asset pricing: a large number of predictors and time-varying parameters. We discuss them in turn.

When the dimension of predictors \( p \) is large, it is well known that the ordinary least squares (OLS) or maximum likelihood estimator (MLE) often do poorly. Although the OLS/MLE estimator has the smallest variance among all linear unbiased estimators, an estimator with slight bias but smaller variance could be preferable for forecasting, leading to a substantial decrease in prediction error.
Recent studies address the trade-off between bias and variance through regularization, shrinkage, or both, which encourages simpler models. However, a priori, a different subset of predictors could be relevant at different points in time (see, e.g., Giannone et al., 2017 and the references therein). In these cases, i.e., when one has no dogmatic prior on the size of the model space, which in itself could be time-varying, classic shrinkage/regularization methods are not suitable. Similarly, data compression techniques such as principal component analysis and factor models, while using all of the predictors available, reduce them to a small set of latent factors that are often hard to interpret.

In this paper, we follow the lead of Koop and Korobilis (2018) and add shrinkage through a dynamic version of the variable selection mixture prior of George and McCulloch (1993). Our dynamic version of the George and McCulloch (1993) mixture prior reads as follows:

\[
\beta_{j,t} \mid \gamma_{j,t} \sim (1 - \gamma_{j,t}) N \left(0, v_{j,0}^2 \right) + \gamma_{j,t} N \left(0, v_{j,1}^2 \right),
\]

where \( \gamma_{j,t} \sim \text{Bernoulli} \left( \pi_{j,0} \right), j = 1, \ldots, p \). Here \( v_{j,0}^2, v_{j,1}^2 \) are fixed prior variances with \( v_{j,0}^2 \to 0 \) and \( v_{j,1}^2 \to \infty \), and \( \pi_{j,0} \) is a fixed prior hyperparameter. Under (5), the parameter \( \gamma_{j,t} \) decides which mixture component applies as a prior distribution for the coefficient \( \beta_{j,t} \). In particular, if \( \gamma_{j,t} = 1 \), the prior of \( \beta_{j,t} \) is flat and uninformative. In this case the posterior estimates are left unrestricted and the coefficient evolves essentially as a random walk. On the other hand, if \( \gamma_{j,t} = 0 \) the prior \( \beta_{j,t} \) is approximately a point mass at zero and the posterior estimates will also be essentially zero, so that the effect of the \( j \)-th predictor is removed from the model at time \( t \). Therefore \( \pi_{j,0} \) is the prior probability of inclusion of a given characteristic in the model. We highlight the fact that our mixture prior is independent over time allowing for a high degree of flexibility. Notice that a complete discussion of what constitutes a “small” and “large” prior variance is given in George and McCulloch (1993). Our prior choices are highlighted in Section 4.

One comment is in order. The time-varying sparse predictive model highlighted in Eqs. (2)-(5) exploits the fact the relevant information set could be sparse, but the degree of sparsity can change over time. In particular, the methodology proposed allows to account for two sources of sparsity in a multivariate dynamic regression problem: the dynamic (or vertical) sparsity, that is the possibility that the subset of relevant predictors is time-varying (through the presence of intermittent zeros in-

9
the regression coefficients at a specific time \( t \), and the horizontal sparsity, i.e., the case when an individual predictor is not relevant at all times \( t \). Table 1 shows an illustration with \( p = 5, C_1, \ldots, C_5 \) characteristics over a period of \( T = 12 \) months.

**Table 1: Illustration of Sparsity in dynamic regressions**

This table shows an illustration of both dynamic and horizontal sparsity with \( p = 5, C_1, \ldots, C_5 \) characteristics over a period of \( T = 12 \) months. Dynamic (or vertical) sparsity represents a situation in which \( \beta_{j,t} = 0 \) for several \( j \)s at given time \( t \). Horizontal sparsity consists of a situation in which \( \beta_{j,t} = \beta_{j,\cdot} = 0 \) for some characteristic \( j \).

<table>
<thead>
<tr>
<th></th>
<th>jan</th>
<th>feb</th>
<th>mar</th>
<th>apr</th>
<th>may</th>
<th>jun</th>
<th>jul</th>
<th>aug</th>
<th>sep</th>
<th>oct</th>
<th>nov</th>
<th>dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>( C_3 )</td>
<td>( \beta_{1,1} )</td>
<td>( \beta_{1,2} )</td>
<td>( \beta_{1,3} )</td>
<td>( \beta_{1,4} )</td>
<td>( \beta_{1,5} )</td>
<td>( \beta_{1,6} )</td>
<td>( \beta_{1,7} )</td>
<td>( \beta_{1,8} )</td>
<td>( \beta_{1,9} )</td>
<td>( \beta_{1,10} )</td>
<td>( \beta_{1,11} )</td>
<td>( \beta_{1,12} )</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>( \beta_{3,1} )</td>
<td>( \beta_{3,2} )</td>
<td>( \beta_{3,3} )</td>
<td>( \beta_{3,4} )</td>
<td>( \beta_{3,5} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( \beta_{3,9} )</td>
<td>( \beta_{3,10} )</td>
<td>( \beta_{3,11} )</td>
<td>( \beta_{3,12} )</td>
</tr>
<tr>
<td>( C_4 )</td>
<td>0</td>
<td>0</td>
<td>( \beta_{4,3} )</td>
<td>( \beta_{4,4} )</td>
<td>( \beta_{4,5} )</td>
<td>( \beta_{4,6} )</td>
<td>( \beta_{4,7} )</td>
<td>( \beta_{4,8} )</td>
<td>( \beta_{4,9} )</td>
<td>( \beta_{4,10} )</td>
<td>( \beta_{4,11} )</td>
<td>0</td>
</tr>
<tr>
<td>( C_5 )</td>
<td>( \beta_{5,1} )</td>
<td>( \beta_{5,2} )</td>
<td>( \beta_{5,3} )</td>
<td>( \beta_{5,4} )</td>
<td>( \beta_{5,5} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>( \beta_{5,9} )</td>
<td>( \beta_{5,10} )</td>
<td>( \beta_{5,11} )</td>
<td>( \beta_{5,12} )</td>
</tr>
</tbody>
</table>

In Table 1, some characteristics, e.g., \( c_2 \), are never significant across time (horizontal sparsity), while some others enter and exit the model set through time, e.g., \( c_3 \) (vertical sparsity). Such dynamic variable selection represents the novel feature of our multi-variate time-varying parameters regression model (2)-(4). As far as the other model parameters are concerned, we adopt conventional conjugate prior specifications for the state variance parameters:

\[
q_{j,t}^{-1} \sim \text{Gamma}(c_0, d_0), \quad j = 1, \ldots, p
\]

\[
\nu_{i,t}^{-1} \sim \text{Gamma}(f_0, g_0), \quad i = 1, \ldots, n
\]

where \( c_0, d_0, f_0, g_0 \) are fixed hyper-parameters which will be discussed in Section 4. With the exception of the dynamic “spike-and-slab” prior, the likelihood and prior specification are standard and similar to common choices in time-varying parameter regressions (see, e.g., Belmonte et al., 2014; Chan et al., 2012; Dangl and Halling, 2012, Bianchi and McAlinn, 2018, among others). However, the large model space makes prohibitive the use of typical Markov Chain Monte Carlo (MCMC) methods for the posterior approximation. In this respect, we follow Wang et al. (2016) and Koop and Korobilis (2018) and implement a Variational Bayes (VB) approach which allows to approximate intractable and large-scale posterior distributions. A complete discussion of the estimation procedure is provided in the Appendix. The reliability of our estimation procedure is investigated through a simulation.
Before concluding, it is useful to interpret some aspects of our model through the lens of the recent literature on machine learning and asset pricing. Kelly et al. (2018) propose a new modeling approach for the cross-section of stock returns, dubbed Instrumented Principal Components Analysis, and document that characteristics are risk exposures. In light of their results, our modeling approach can be interpreted as anchoring the latent factor loadings in a classical $\beta$-representation with observable characteristics. Our assumption of a linear relation between expected returns and characteristics is inline with Kelly et al. (2018). Furthermore, Chen et al. (2019) find that in isolation firm characteristics have a close-to-linear effect on the SDF. Finally, our model can accommodate non-linear associations between characteristics and returns by allowing for non-linear transformations of raw characteristics. We explore this set-up and the effects of non-linearities in a robustness check in Section 7.

4 Empirical results

In our empirical exercise we adopt the following baseline prior specification: for the dynamics mixture prior in Eq. (5) we choose $v_{j,0} = 0.001$ and $v_{j,1} = 10$, whereas for the prior probability of inclusion we choose $\pi_{j,0} = 0.5$. These prior specifications are rather standard in the literature (see Koop and Korobilis, 2018). Notice that the dynamic nature of our spike-and-slab prior implies that the impact of the prior tend to vanish as the sample size increases. In Appendix B we provide a series of prior sensitivity tests which shows that the posterior estimates are robust with respect to different initial probabilities of inclusion. For the conditional variance of both state equations we choose an uninformative Gamma distribution, i.e., $q_{j,t}^{-1} \sim \text{Gamma}(100, 1)$ and $\nu_{j,t}^{-1} \sim \text{Gamma}(100, 1)$. Finally, the initial values of the betas are drawn from a multivariate Normal with mean 0 and variance $100I_p$.

4.1 Data

In our benchmark analysis, we study the dynamics of long-short value-weighted portfolio returns on eight factors over the period July 1963 through June 2017: value, size, momentum, profitability, investment, idiosyncratic volatility (Ang et al., 2006), betting-against-beta (Frazzini and Pedersen, 2014), and accruals. The choice of factors is motivated by the fact that they are used in the most-
popular academic multi-factor models (Hou et al., 2015; Fama and French, 2015) and are closely followed by investors (see, e.g., Arnott et al., 2016, 2019).

In our study we follow Freyberger et al. (2017) and employ 62 characteristics as potential predictors. We follow Hou et al. (2015) in the classification of characteristics, and group them into six categories: (1) past return based predictors such as momentum and short-term reversal; (2) investment-related characteristics such as the annual percentage change in total assets (investment) or the change in inventory over total assets; (3) profitability-related characteristics such as gross profitability over the book-value of equity or return on operating assets (ROA); (4) intangibles such as operating accruals and tangibility, (5) value-related characteristics such as the book-to-market ratio (BE/ME) and earnings-to-price; and (6) trading frictions such as the average daily bid-ask spread and standard unexplained volume. Additional details are provided in the Appendix A.

4.2 Aggregate sparsity

The novel aspect of our framework is that we can explicitly accommodate for time variation in the number of significant characteristics in the cross-section of expected returns, i.e., dynamic sparsity. In this section we first provide an estimate of dynamic sparsity at the aggregate level. Specifically, we measure the percentage of significant predictors at each time $t$, i.e., $\text{Active Predictors} (%) = \frac{\sum_{j=1}^{p} \gamma_{j,t}}{p}$. The left panel of Figure 2 reports the estimates of the active number of predictors (expressed in percentage of the total number $p$) for the period that goes from February 1965 to June 2017. The grey vertical bars represent the NBER-based recession indicators for the U.S. from the peak through the trough.

![Insert Figure 2 here](image)

Few interesting results emerge. First, sparsity varies substantially over time, with the amount of characteristics that are significantly different from zero ranging from less than 40% during early 80s and late 90s, to around 60% before the recession of early 2000s. In line with our initial motivation (see Figure 1), it is evident that a model that restricts the number of significant characteristics...
to be constant in the data-generating process could be severely misspecified. Second, it is also easy to spot low-frequency movements, particularly so in the post-1990s. Indeed, the number of characteristics trends up over time for about 10 years reaching its peak in the early 2000s. We then observe a large decrease in the number of predictors. The proportion of significant predictors shrink considerably towards the end of the sample to a low value of about 40%, suggesting a large number of characteristics become irrelevant. Third, sparsity displays also business cycle fluctuations. For instance, the predictive set becomes more dense during the early 2000s, a period marked by several events such as financial scandals, e.g., Worldcom and Enron, the 9/11 attacks, and a mild economic recession. On the other hand, the proportion of significant predictors tends to substantially decrease around and after the great financial crisis of 2008/2009. Under the working hypothesis that characteristics are covariances (see Kelly et al., 2018), our evidence suggests that risk price dynamics are time-varying and, intermittently, risk prices turn to zero.

The right panel of Figure 2 shows that the proportion of active predictors tends to positively correlate with stochastic volatility. In particular, our results show that an increase in idiosyncratic volatility—averaged across portfolios—is associated with a more dense model for expected returns. Interestingly, idiosyncratic volatility can be thought of as an in-sample measure of discrepancies between realized and expected returns. Therefore, a natural question that arises is whether a high number of predictors is associated with stronger predictability. To address this question, we estimate the model on an expanding window basis and collect the one-step-ahead out-of-sample forecasting errors. In particular, we start with an initial sample equal to 50% of the total observations, that is starting date is April 1991, then generate a one-month-ahead forecast and calculate the realized forecasting error. Such procedure is repeated by estimating recursively the model and collecting the forecasting error for the remaining 50% of the sample. The forecasts are generated for each of the long-short portfolios, which represent our variables of interest. Figure 3 reports the results.

[Insert Figure 3 here]

The left panel Panel B shows the squared forecast errors averaged across portfolios, a measure of predictive inaccuracy, against the percentage of the active predictors during the out-of-sample period.
for which forecasts are generated.\footnote{The forecasting error for a given portfolio $i$ at a given time $t$ is calculated as $\hat{\varepsilon}_{i,t}^2 = (y_{i,t} - x'_{i,t-1}\hat{\beta}_t)^2$. The prediction error is then averaged as $\hat{\varepsilon}_{i,t}^2 = \sum_{t=1}^{n} \hat{\varepsilon}_{i,t}^2$.}

Two interesting facts emerge. First, the amount of predictability tends to be inversely related to sparsity, that is, a higher number of significant predictors in-sample does not translate into better predictive performance out-of-sample. This is in line with the bias-variance trade-off, that is, more parsimonious models, although biased by construction, lead to a substantial decrease in the prediction errors.\footnote{The advantage of more parsimonious models for predictive systems has been highlighted by Friedman et al. (2001) who introduced the concept of Bet on Sparsity, that is simpler models should be preferable.} Second, the discrepancy between the model and the realizations tends to increase during recessions. Again, this is in line with the fact that the squared forecast error can be thought of as a proxy of forecasting uncertainty, which has been shown to be highly correlated with recession periods and the business cycle more generally (see, e.g., Jurado et al., 2015).

Although informative, performance measures based on point forecasts only give a partial assessment since they ignore information on the full probability distribution of returns. Ideally, one also wants to evaluate the accuracy of the density forecasts. To this end, we compute an additional measure of forecasting accuracy, called the log-predictive likelihood. Given that our model (2)-(4) is conditionally Gaussian, for a portfolio $i = 1, \ldots, n$ the log-predictive likelihood is calculated as the log of the probability that a realized return at a given month $t$, $y_{i,t}$, is generated by a normal distribution with conditional mean $x'_{i,t-1}\hat{\beta}_t$ and conditional volatility $\hat{\sigma}_{i,t}$, i.e., $\text{LogPL}_{i,t} = N(y_{i,t}|x'_{i,t-1}\hat{\beta}_t, \hat{\sigma}_{i,t}^2)$. Similar to the squared forecasting error, the LogPL is evaluated recursively out-of-sample based on one-step ahead forecasts starting from April 1991 till the end of the sample.

The right panel in Figure 3 displays the results. The evidence shows that there is a negative relationship between the fraction of active predictors and the log-predictive likelihood. That means that more predictors do not translate into a higher predictive accuracy in a distributional sense. This result is consistent with the out-of-sample squared forecasting errors (see left panel).

4.2.1 Dynamic sparsity and correlations

We now delve further into the relationship between dynamic sparsity and the time-series properties of both predictors and portfolio returns. In particular, we first investigate, for each long-short
portfolio, the amount of (cross-sectional) correlations within a group of predictors. The presence of highly correlated predictors could be a problem as in principle different characteristics could carry a similar explanatory power for the time-series variation of expected returns. This implies that when correlation is high one cannot easily disentangle and pick the “right” set of firms’ characteristics as, at least statistically, different covariates could be seen as perfect substitute from an information perspective. Figure 4 shows the explained variation from the first principal component (PC) for each group of firms’ characteristics. The groups are defined following Hou et al. (2015) and Freyberger et al. (2017) as outlined in Section 4.1.

Left panel reports the in-sample variance explained by the first PC; the y-axis reports the long-short portfolios and the x-axis the groups of predictors. The last row reports the explained variance for each group of firms’ characteristics but averaged across portfolios. Interestingly, the first PC does not explain a large fraction of time-series variation, with a relative explanatory power of 50% on average across portfolios.

Right panel confirms these results by showing the explained time-series variation of the first two PCs per groups of firms’ characteristics and for each long-short anomaly-based portfolio. With the only exception of Investment and Intangibles, the cumulative explained variation is always below 65%. Notice that Investment and Intangibles groups only contain a handful of characteristics, which explains the above average explanatory power of the two first PCs. In sum, there seems to be enough independent information in the cross-section of characteristics and for the portfolios considered; our approach aims at exploiting the (non-redundant) information in this large set of predictors for forecasting portfolio returns.

Although instructive, the evidence provided by Figure 4 does not directly compare our measure of time-varying sparsity — see Figure 2 — with the explained variation of principal components. To investigate such relationship, we calculate the amount of time-series variation explained by the first two principal components (PCs) of either the active predictors or portfolio returns, and compare it with our measure of sparsity. More specifically, at each time $t$ we calculate for each portfolio the first-
two PCs of either the portfolio returns or the firms’ characteristics. As far as the firms’ characteristics are concerned, we explicitly take into account the dynamics in the model space by calculating the PCs only on the firms’ characteristics that are selected at each time \( t \). The PCs are calculated based on an expanding window starting in June 1975.

Similar to Figure 4, the goal of this exercise is to investigate whether the time variation of sparsity is due to increasing correlation among predictors and/or portfolio returns. The time-series implications are far from trivial; for instance, it could well be the case that despite having a high number of active predictors, these are highly correlated and well summarized by just few PCs. If this is the case, then more active predictors do not imply that more information is needed to capture the dynamics of expected returns. The left panel of Figure 5 compares our measure of dynamic sparsity (solid blue line) with the time-series variation of active predictors at each time \( t \) explained by the first two principal components.

An interesting result emerges. There is a slightly negative and significant correlation, equal to \(-0.42\), between the fraction of active predictors and the explained variance of the PCs. When the number of active predictors increases, we find that the explained variation of the PCs decreases, meaning more PCs are needed to capture the same variability that would be achieved with fewer predictors. If the added predictors were simply redundant, we should expect a constant explained variance by the first two PCs, which is not the case. We conclude that an increase in the number of significant predictors is associated with relevant information for the dynamics of expected returns.

The right panel of Figure 5 compares our measure of dynamic sparsity (solid blue line) with the explained variation of the first two PCs of portfolio returns (dashed orange line). Interestingly, the factor structure in the portfolio returns seems to be almost completely independent to the percentage of active predictors over time.

### 4.2.2 The role of time-varying volatility

The concept of time-varying volatility and its asset-pricing implications are well understood. The literature on long-run risk (Bansal and Yaron, 2004) attributes a key role to consumption volatility as a-
driver of expected returns. More recently, Campbell et al. (2018) showed the asset pricing implications of volatility risk within the context of an Intertemporal CAPM (see Merton et al., 1973). Importantly for our analysis, not only aggregate but also idiosyncratic volatility is time-varying: Herskovic, Kelly, Lustig and Nieuwerburgh (2016) find that even after saturating the factor regression with up to ten principal components, the residual firm- and industry volatilities continue to display substantial time-series variation (and co-movement) as seen in raw return volatilities.

The statistical implications of accounting for stochastic volatility (SV henceforth) in the econometric modeling of expected returns are also far from trivial. In many cases, a data-generating process of economic and financial variables seems to have drifting coefficients and shocks of stochastic volatility. If that is the case, then a model with time-varying coefficients but constant volatility raises the question of whether the estimated time-variation in parameters is likely to be biased because a possible variation of the volatility in disturbances is ignored. Indeed, a large literature has shown that incorporating time-varying volatility could play a crucial role in estimating a multivariate time-varying parameter model (see Cogley and Sargent, 2005; Primiceri, 2005). Finally, time-varying volatility has important implications for asset allocation (see, e.g. Han, 2005; Gargano et al., 2017).

In this section we discuss the role of time-varying volatility in modeling the dynamics of sparsity in expected portfolio returns. To address this issue, we estimate a restricted version of the main model Eq. (2)-(4) whereby the observation error term $\varepsilon_{i,t}$ has constant volatility, i.e., $\varepsilon_{i,t} \sim N(0, \sigma^2_i)$. Left panel of Figure 6 shows the dynamics of our sparsity measure—the number of active predictors as a fraction of the total—estimated from the main model (light-blue line) and the restricted, constant-volatility, model (magenta line with markers).

Two interesting aspects emerge. First, despite a similar pattern of high vs. low sparsity periods, the amount of sparsity substantially changes between the two models. In particular, a simple visual inspection shows that when stochastic volatility is explicitly included, the number of active predictors is much lower on average across the sample. In fact, the sample average of sparsity with SV is around

---

7Clark (2011) shows that adding stochastic volatility to VARs materially improves the real-time accuracy of density forecasts of U.S. GDP growth, unemployment, inflation, and the federal funds rate.
15% higher than without SV. Second, the number of active predictors in a model without SV relative to a model with SV is increasing around specific periods, like economic recessions (grey shaded areas). The right panel of Figure 6 offers a possible explanation. The figure shows the SV estimate (blue line) vis-a-vis the difference in the fraction of active predictors from the two models, with and without SV. A positive value implies that the model without SV has a higher number of active predictors, i.e., lower sparsity. Interestingly, the figure shows that the gap (in terms of how sparse are expected returns) between the two models increases precisely when idiosyncratic volatility tends to be higher.

As a whole, Figure 6 provides evidence that there is an interplay between the dynamics of idiosyncratic volatility and sparsity of expected returns. By explicitly accounting for time-varying idiosyncratic risk, our modeling framework mitigates concerns about possible bias estimates in the conditional mean due to the plausible instability of volatility in the data generating process of realized long-short portfolio returns.

4.3 A discussion on the origins of dynamic sparsity

A possible explanation of the pattern highlighted in Figures 2-5, i.e., the forecasting model is more dense in bad times (Figure 2) when the predictive accuracy decreases (Figure 3, Panel B), goes as follows. In normal times, investors exploit a large number of characteristics to time factors. Over time, the factors could, in principle, become too crowded making some characteristics irrelevant (so that the probability of inclusion decreases). However, in bad times, when capital becomes scarce, several characteristics return to be useful to predict factor returns. At the same time, the high uncertainty surrounding this period makes returns highly volatile and difficult to predict.

To test this explanation we investigate the link between dynamic sparsity and the supply of capital. Specifically, we regress our measure of time-varying sparsity on several proxies for liquidity supply. In particular, we include the level of idiosyncratic volatility (measured as the cross-sectional standard deviation of individual stock returns during the prior month) as a proxy for the level of risk faced by imperfectly diversified liquidity providers; the TED spread (the three-month Eurodollar deposit rate minus the three-month Treasury Bill rate) as a proxy for funding costs of financial intermediaries (see, e.g., Garleanu and Pedersen, 2011); and the VIX who Nagel (2012) has shown to capture well the time variation in the risk premium from providing liquidity.
We also study the relation between time-varying sparsity and aggregate financial conditions as proxied by the NFCI. The NFCI provides a weekly estimate of US financial conditions in money markets, debt and equity markets, and the traditional and shadow banking systems. The index is a weighted average of 105 measures of financial activity, each expressed relative to their sample averages and scaled by their sample standard deviations. The NFCI is constructed to be zero mean so that when the index is positive financial conditions are tighter than average, and vice versa. In addition to the aggregate NFCI we also consider a set of sub-indexes which allow for a more granular characterization of the forces driving aggregate financial conditions. Similar to the NFCI, each sub-index is constructed to have zero mean and unit standard deviation. We used three different sub-indexes; the “Risk” sub-index captures volatility and funding risk in the financial sector; the “Credit” sub-index is composed by measures of aggregate credit conditions; the leverage sub-index consists of debt and equity measures. Increasing risk, tighter financial conditions and higher leverage are all consistent with an increase in the NFCI.\(^8\)

Table 3 shows the results. Column (1) shows a positive association between our probability of inclusion and financial conditions. The next columns investigate the subcomponents of NFCI. Columns (2) and (3) show that such positive correlation can be attributed to the two sub-components related to risk and non-financial leverage. On the other hand, columns (4) shows no relation with the credit sub-index of the NFCI. Columns (8) and (9) show that, in a multiple regression, non-financial leverage is more important than the overall NFCI index or its risk component; it also drives away the TED spread (column (10)), a commonly used proxy of funding conditions (see, e.g., Frazzini and Pedersen, 2014). Turning to proxies for liquidity supply we find that sparsity is strongly associated with idiosyncratic volatility and VIX (columns (6) and (7)). However, in a joint regression, only the VIX continue to stay significant at usual confidence levels (column (11)). Finally, column (12) shows that both the VIX and the non-financial leverage sub-index are significant drivers of time-varying sparsity.

Overall, we find strong evidence in favour of time-varying sparsity with its dynamics that significantly...

\(^8\)We use the non-financial leverage index which captures leverage in both household and non-financial businesses. We find no evidence in favor of the financial leverage sub-index of NFCI, in line with the fact that non-financial business leverage measures in the NFCI prove to be a reliable leading indicator of financial stress (see Brave and Butters, 2012).
correlate with proxies for financial conditions (especially for households and non-financial businesses) and aggregate risk aversion/liquidity supply.

4.4 Group-specific sparsity and individual characteristics

The aggregate dynamic sparsity reported in Figure 2 masks interesting heterogeneity across and within categories. We start unravelling the dynamics of sparsity at the category level. To this end, for a fixed category, we investigate the proportion of active predictors at each time $t$, that is the amount of sparsity. The classification follows Hou et al. (2015) and Freyberger et al. (2017) as outlined in Section 4.1. Panel (b) in Figure 2 shows the results.

Three facts emerge. First, few groups, namely trading frictions, value and intangibles, are mostly responsible for the time-variation of sparsity. Second, the contribution to sparsity of characteristics pertaining to the profitability group is sizable and quite stable throughout our sample. Third, characteristics related to past returns and, to a lesser extent, investment seem to play a marginal role.

The dynamics of the group-specific sparsity shown in Figure 2–Panel (b) does not inform us on the actual characteristics that drive the dynamics of expected returns over time. To address this issue, we now delve further into the dynamic sparsity of individual characteristics; in particular, we report the time-varying posterior probability of inclusion for each of the 62 firms’ characteristics at each time $t$. This level of granularity should give a sense of the heterogeneity of sparsity within each category of predictors. For the ease of exposition, Figure 7 shows each of the estimated posterior probabilities in the form of a heat map. Specifically, each panel shows a different category, and the labels on the $y$-axis of a subplot describe the characteristics within a category. A dark-blue color implies that a given characteristic is not included in expected returns at a given $t$ while bright yellow indicates that a given predictor is active across portfolio returns.

[Insert Figure 7 here]

Top-left panel shows the characteristics classified as trading frictions. We find idiosyncratic volatility (IdioVol), the bid-ask spread (spread), the level as well as the variability of trading activity (proxied...
by suv and std volume; see Chordia et al., 2001a) to be included in the dynamics of expected returns for a large fraction of the sample. In addition, total assets (at, see Gandhi and Lustig, 2015) is also significant for more than 90% of the sample period. Our results are in line with, e.g., Kelly et al. (2018) who find that book assets, unexplained volume, and idiosyncratic volatility are significant characteristics related to future returns. We also find size adjusted by industry (lme adj) to be more important than un-adjusted size, consistent with the idea that several anomalies are mostly an within-industry rather than across-industry effect. However, even after having been adjusted by industry, size turns out to be of secondary importance relative to other controls.\footnote{Differently from Freyberger et al. (2017) we do not find closeness to the 52 weeks high (REL, TO, HIGH) to be important.}

Top-right panel shows the results for the value characteristics. Consistent with other findings in the literature (see, e.g., Asness et al., 2000), we find that the within-industry component of book-to-market (beme adj) is a stronger signal than an unadjusted book-to-market, especially for the first half of the sample.\footnote{The near absence of book-to-market from the list of included value characteristics is surprising given its prominence in the empirical asset pricing literature. A possible explanation is that our sample includes financial stocks along with non-financials, and book-to-market ratios may be incomparable across these two groups.} Total assets to Size (a2me), the earning-to-price ratio (e2p) and the net payout ratio (nop, see Boudoukh et al., 2007) are also strong predictors throughout the sample until the recent crisis. In particular, the net payout ratio starts with a high probability in the first part of the sample, it becomes irrelevant for about 10 years, and then it returns significant again from early 90s till the great financial crisis of 2008/2009. Differently from this intermittent behavior, assets-to-market cap is robustly present throughout the sample period (with the sole exception of the aftermath of the financial crisis).

Characteristics pertaining to profitability (mid-left panel) are often present throughout the sample. In particular, we find few very robust predictors such as profit margin (PM), return on assets (ROA), return on net operating assets (RNA), return on cash (ROC) and return on equity (ROE): these characteristics stand out with a probability of inclusion close to 80% throughout the sample.

Variables in the “investment” category (mid-right panel) start out as very important (yellow areas are pervasive), but their probability of inclusion decreases throughout the sample. For instance, the average probability of inclusion for the change in PP&E and inventory over lagged AT (ΔP12A) and the change in shares outstanding (Δshrout) is high in the first part of the sample but it becomes...
essentially zero in the second part. There are however exceptions. The % change in inventories (ivc) becomes more important over the sample with its probability of inclusion raising from half insignificant to mostly significant when we move from the first to the second half of the sample. Finally, net operating assets (NOA) is present throughout the sample period with only minor interruptions around the 1996 and the 2016.

Within intangibles, variables like absolute value of operating accruals (AOA) and operating leverage (sum of costs of goods sold and selling, general, and administrative expenses to total assets, OL) are robustly present throughout the sample. On the other hand, variables like tangibility (tan) are only significant at times, in particular in the early part of the sample and before the mild recession of 2001.

Bottom-right panel shows the results for past returns. At the category level, the amount of active predictors is rather stable, with a slight tendency to increase over time, especially after the recession of early 2000s, see figure ?? . The stronger momentum signal is given by firms’ performance 12 to seven months prior to portfolio formation in line with the findings by Novy-Marx (2012). Interestingly, there seems to be some substitutability between echo-momentum and classical momentum: in the time window where momentum portfolios formed from 12 to 7 months do not matter, then classical momentum becomes relevant. Finally, short-term reversal (r2,1) does not play any role throughout the sample (differently, e.g., from Kelly et al., 2018).

4.4.1 Unconditional results

As a by-product, our modeling framework allows to investigate a second source of sparsity, namely horizontal sparsity. This is the unconditional probability that a given firm’s characteristic is not included in the dynamics of expected returns, that is the probability that $\beta_{j,t} = 0, \forall t$ for some characteristic $j$. In our setting, this can be calculated by looking at how frequently a given characteristic turns out to be significant in the set of predictors.

Figure 8 reports the estimates for each of the characteristics. The colors of the vertical bars identify the groups of predictors: light-blue are past returns, red investment, green profitability, black-
intangibles, blue and brown value and trading frictions, respectively.

As expected, the figure confirms the findings of Figure 7, that is, value and profitability constitute the most important categories with high probabilities of inclusion for many firms’ characteristics. Interestingly, within the value category, we find that a composite of accounting ratios that incorporates earnings, cash, sales, and book value of assets, along with price are needed to predict future returns. This seems in line with (Asness et al., 2000) who use a composite measure to proxy for value. Within past returns, only the performance from 12 to 7 months seem to matter.

5 Market timing and economic significance

Can time-varying sparsity help to better capture predictable variation in portfolio returns? The answer is positive and contributes to the debate about the benefit of timing factor returns. In particular, we employ strategies that rotate across the eight factor portfolios based on the level of the expected return. Namely, the trading rule is given by:

\[
\sum_{j=1}^{n} w_{j,t} x r_{j,t+1} \quad (6)
\]

\[
w_{j,t} \propto E_{t}[r_{j,t+1}] \quad (7)
\]

with \(n = 8\). The set of anomaly-based portfolios is the same as in the main empirical analysis, that is, long-short value-weighted returns on eight factors over the period July 1963 through June 2017: value, size, momentum, profitability, investment, idiosyncratic volatility (Ang et al., 2006), betting-against-beta (Frazzini and Pedersen, 2014), and accruals. We benchmark our strategy based on time-varying sparsity against four strategies based on alternative models to form expected returns, \(E_{t}[r_{j,t+1}]\). To start, we consider an unconditional strategy that invests based on the historical average return until time \(t\). The idea behind the second and third strategies is to challenge a data-rich environment like-
ours with a model that instead carefully preselects characteristics that have been found to robustly predict returns. To this end we follow Lewellen (2015), and employ panel regressions (with portfolio fixed effects) of returns on few, well-known firm characteristics. The second strategy exploits size, book-to-market, and past 12 months returns (skipping the most recent). The third strategy exploits accruals, return on asset, split-adjusted shares outstanding, and growth in total assets, in addition to size, book-to-market and past returns. The fourth strategy is based on the prediction obtained from a ridge regression. The last strategy is based on our model and intends to take advantage of the time-varying nature of sparsity in real-time.

To put the four benchmark strategies (the historical mean, the two strategies based on preselected characteristics, and the one based on ridge) on a fair ground, we calculate the forecast on a rolling window basis. In this way each strategy features a simple form of time-varying parameters and is given a fair chance to compete against our model that naturally allows for variation in betas. In particular, we start with an initial window of either 60 or 120 observations, generate a one-month ahead forecast and calculate the trading signal from each model as in Eq. (7). Then we add another observation and discard the first one, generate a forecast and calculate again the optimal weights, and so on. The sample starts in 1991 and ends on 2017, providing more than 25 years of data to evaluate the out-of-sample performance of the strategies.

Figure 9 presents the results. The left panel is based on a 5-year rolling window. The figure shows the total amount of money that one would have in the account in month $t$ if one followed either our time-varying sparsity strategy or any of the alternative ones.

[Insert Figure 9 here]

We observe that our strategy based on time-varying sparsity (red line with triangles) produces substantial gains relative to the alternative benchmarks. The ridge models perform at par with our time-varying sparse model until the early 2000s. However, by June 2017 the wealth generated by the sparsity-adjusted strategy is about twice as large as the wealth generated by a rolling window ridge regression. Interestingly, both models based on a small set of pre-selected predictors underperform a simple strategy based on the recursive mean.
The right panel of Figure 9 presents results obtained using a 10-year rolling window. Although the pecking order remains intact, the magnitude of the gaps possibly increases. For instance, a trading signal obtained from our time-varying sparsity model generates a final wealth which is three times higher than a simple rolling window ridge regression. Interestingly, similar to the case with window size of 60 observations, the gap in favour of our strategy substantially increases in the aftermath of the great financial crisis. This is possibly due to the fact that our modeling framework explicitly accounts for stochastic variations in idiosyncratic volatility, which arguably turned out to be relevant during market failures.

Although our framework is fully dynamic, and therefore should naturally be evaluated from a time-dependent perspective, it is instructive to also look at the unconditional portfolio measures. Table 4 reports the descriptive statistics for the performance generated by our time-varying sparse predictive regression against the recursive mean and the set of alternative strategies described above.

The top panel shows the results whereby the estimates of the competing models are based on a rolling window of 60 observations. The annualized Sharpe ratio of the benchmark strategy with three and seven characteristics are lower at -0.019 and 0.021 during our sample period, respectively. By contrast, the sparsity-adjusted strategy has an annualized Sharpe ratio that is more than twice as high at 0.301. This difference is due to simultaneously exploiting a larger set of predictors and dropping predictors that have weak signals. Consistent with the existing evidence, the unconditional mean performs similar to a simple ridge regression with an annualized Sharpe ratio of 0.210 (see, e.g., Campbell and Thompson, 2007 and Welch and Goyal, 2007).

Panel B confirms such pecking order by using a longer rolling window of 120 observations. In particular, our time-varying sparsity strategy obtains an annualized Sharpe ratio equal to 0.385, a number twice as large as the recursive mean (0.192) and around 68% higher than the rolling ridge regression (0.228), which turns out to be the closest among the competitors.

Overall we find that the time-varying sparsity trading strategy outperforms the benchmark trading strategies in a way that is economically meaningful.
What matters for the aggregate stock market?

A recent literature has documented market return predictability using firm-level variables aggregated across stocks, such as the share of equity issues, accruals or short interest (see e.g. Baker and Wurgler, 2000; Hirshleifer et al., 2009; Rapach et al., 2016). More recently, however, Engelberg et al. (2019) show that these results do not extend well to several other cross-sectional anomalies. In fact, after accounting for multiple testing, Engelberg et al. (2019) find weak evidence of market predictability out-of-sample using (aggregated) cross-sectional predictors.

These papers suggest that our analysis at the portfolio level, and in particular our measure of sparsity and the associated identity of predictors, can be different for the aggregate market. Thus, we next employ our framework to investigate the properties of time-varying sparsity at the market level.

Figures 10 and 11 report our measure of sparsity together with the probability of inclusion for the various characteristics when we predict the aggregate value-weighted US market return. Comparing Figure 10 to 2, we note that sparsity is higher at the market level with a probability of inclusion below 0.5 during most of the sample period. Not only the level but also the dynamics of sparsity is different: Figure 10 shows that, contrary to the portfolio level, the probability of inclusion decreases substantially during recessions. Finally, at the market level, the probability of inclusion displays substantial low-frequency movements, with a downward trend from the 1980 till the great recession of 2007-2009, and a subsequent upward trend. Perhaps unsurprisingly, the substantial time-variation in sparsity at the market level suggests that time-invariant models that force a predictor to be present throughout the sample period are doomed to detect a lack of out-of-sample predictability.

It is also interesting to compare Figure 11 to Figure 8. In line with our previous discussion, we find that, for the market, none of the predictors attains a probability of inclusion greater than 0.65; this stands in sharp contrast with the portfolio level analysis where several predictors have a probability of inclusion of 0.8 or greater. Furthermore, several characteristics that were important for the predictability of anomaly portfolios turn out to be irrelevant for the market. E.g., variables like return-on-assets (ROA), return-on-net-operating-assets (RNA), and return-on-cash (ROC) that were ranking among the strongest predictors within the profitability category are almost never included to predict the market. Similarly, within the trading friction category, total
asset (AT) moves from ranking top for the anomaly portfolios to be almost absent for the market; and, conversely, the probability of inclusion of turnover (LTURNOVER) raises from 0.2 for portfolios to about 0.55 for the market. In sum, our results show that, even without accounting for multiple testing and data-mining as in Engelberg et al. (2019), several cross-sectional predictors that are useful to forecast in the time-series anomaly returns, turn out to be irrelevant for the time-series predictability of the market.

Despite these differences between portfolios and the aggregate market, our framework does find evidence for some characteristics to be useful time-series predictors for both anomalies and market returns. E.g., the signal based on 12 to 7 months returns prior to the current month (R127), net operating assets (NOA), return on equity (ROE), operating leverage (OL), and the earnings-to-price (E2P) are found to have — within their respective categories — high probability of inclusion both at the portfolio and market levels.  

Thus, our analysis suggest that the dynamics of portfolios expected returns may not be totally disconnected from those of the expected return on the market, with several signals driving both anomaly-based portfolios and the aggregate market.

7 Robustness

In this Section we provide further evidence and discussion on the reliability of our modeling framework. In particular, we first test the robustness of the aggregate measure of sparsity with respect to the inclusion of non-linearities in the dynamics of expected returns. Second, we compare the dynamics of our model-implied sparsity against few typical shrinkage methods. Third, we report the results of an extensive simulation exercise tailored on the empirical analysis. Finally, we discuss the differences of our framework with respect to existing alternatives to model dynamic sparsity.

\[13\] The probability of inclusion of the signal based on 12 to 7 months returns decreases from 0.8 to about 0.5 when we move from the portfolio to the market level. Despite the drop, this signal continues to stay the strongest predictor within the past-return based characteristics.
7.1 Time-varying sparsity and non-linearities

So far, the firms’ characteristics enter linearly in the regression model. We are interested in understanding how our measure of sparsity is affected by the inclusion of non-linear effects (see, e.g., Freyberger et al., 2017 for empirical evidence of non-linearities).

To this end, we implement two different exercises to capture the potential nonlinearity of the SDF: first, we follow Freyberger et al. (2017) and we interact each of the 61 firm characteristics other than firm size with firm size for a total of 123 firm characteristics. This exercise is inspired by Chen et al. (2019) who finds strong evidence in favor of interaction between several characteristics. Second, we follow Feng et al. (2017) and we add as controls 62 squared terms, one for each portfolio’s characteristic. Notice both exercises are implemented ceteris paribus, meaning all of the other specifics of the estimates, e.g., sample, priors, etc., are exactly the same as in the main empirical analysis.

The left panel of Figure 12 reports the percentage of active predictors out of the original 62 firms’ characteristics when interactions with size are included as additional regressors (123 predictors in total).

The results are largely consistent with the empirical analysis; that is, sparsity is highly time-varying and tends to correlate with the business cycle, being generally low during recessions and increasing thereafter. In fact, the correlation between the two sparsity measures (with and without size interactions) is as high as 0.7.

The right panel shows the percentage of active predictors out of the original 62 firms’ characteristics when the square of each characteristic is included as additional regressor (124 predictors in total). Despite the divergence after the great financial crisis of 2008/2009, the correlation between our original sparsity measure and the one obtained by controlling for squared characteristics is still high at 0.8.

7.2 Comparison against rolling window estimates

In this section, we compare the results obtained when we employ our model with time-varying sparsity against the results one would obtain by employing alternative, common penalized regression
models. For the ease of exposition we considered three alternative penalized regression specifications which have been widely used in the literature, namely the ridge $L_2$ penalty, the Lasso $L_1$ penalization and the Elastic-Net which represents a combination of the two penalty strategies. To put the alternative methodologies on the same ground as our methodology, we adopt a Bayesian estimation framework; this is particularly convenient since the shrinkage estimates can be derived as a posterior-estimates under suitable continuous shrinkage priors. We have discussed the ridge prior in Section 2 already; as far as the lasso is concerned, Tibshirani (1996) notices that the lasso penalty is equivalent to the posterior mode estimate under a Laplace prior. Park and Casella (2008) build on this intuition and suggest that a consistent prior can be defined as a scale mixture of normals with an exponential mixing density. With such prior, the posterior mode estimates are similar to the estimates under the $L_1$ penalty term.\footnote{Notice that by integrating $\lambda_j^2$ out one obtains the original Tibshirani (1996) double-exponential prior.}

An alternative shrinkage estimator which combines both an $L_1$ and an $L_2$ penalty term is the so-called elastic net. Originally proposed by Zou and Hastie (2005), the elastic net mitigates the lasso concerns related to $p > T$ and the presence of group-wise correlated regressors. Li and Lin (2010) show that the shrinkage prior for the elastic net penalization can be expressed as a hierarchical mixture prior. Notice that providing an in-depth discussion of these methodologies and their posterior sampling procedures is beyond the scope of the paper. A detailed description of the estimation procedure as well as empirical finance applications for the ridge, the lasso and the elastic net in a Bayesian framework can be found in Bianchi and Tamoni (2020).

Figure 13 reports the estimation results. The top panels compare our time-varying sparsity measure with the rolling shrinkage estimates from a ridge regression and a lasso. Notice the scale of the shrinkage parameters from the penalized regressions is inverted to have the same interpretability of our measure of sparsity, i.e., \% active predictors out of total.

![Insert Figure 13 here]

The figure shows that our time-varying sparsity measure tend to diverge with the rolling window estimates both at the beginning of the sample, until the 90s, as well as in the aftermath of the great financial crisis. Conversely, there is a quite substantial coherence in the trend from early 90s to and
around the crisis of 2008/2009. In fact, rolling window estimates and our measure tend to give similar implications, that is ridge and lasso imply a steadily decreasing shrinkage from the early 1990s to 2001 and at the same time our measure of sparsity shows that the number of active firms’ characteristics is increasing over the same period.

Bottom panels show the estimates from an elastic-net regression model. More specifically, the left (right) panel shows the shrinkage from the L2 (L1) penalty term. The pattern described above is largely confirmed, although the estimates of the regularization parameters are slightly smoother than those obtained from the ridge and the lasso.

As a whole, Figure 13 confirms that our measure indeed capture some fundamental structure in the dynamic of expected returns. However, the divergence with respect to standard penalized linear regression estimates coupled with the main empirical results in Section 4 suggest that standard rolling window sparse estimates cannot fully recover such structure.

7.3 A discussion on alternative methodologies

Recently, there has been great interest in regularizing the coefficients within problems in which the parameters vary over time. Within the context of dynamic linear regression models, recent contributions have been proposed by Nakajima and West (2013), Belmonte et al. (2014), Kalli and Griffin (2014), Koop and Korobilis (2018), Bitto and Frühwirth-Schnatter (2019), and Kowal et al. (2019). The basic approach in Nakajima and West (2013) is to set a threshold under which a dynamic coefficient is automatically switched off, i.e., set to zero. Although convenient, such an approach does not scale up easily to a large set of predictors and leaves mainly unspecified the optimal choice of the significance threshold. Belmonte et al. (2014) and Kalli and Griffin (2014) take a different approach such that the dynamics of each regression coefficient $\beta_{j,t}$ is rewritten in terms of scaled latent states $\tilde{\beta}_{j,t} = \beta_{j,t}/\omega_j$, so that the shrinkage is modelled by assigning priors to $\omega_j$. In particular, while Belmonte et al. (2014) used the Laplace prior for shrinking, Bitto and Frühwirth-Schnatter (2019) used more general Normal-Gamma priors. Similarly, Kalli and Griffin (2014) used an autoregressive specification of the Normal-Gamma prior to shrink the dynamics of predictor-specific betas. Finally, Koop and Korobilis (2018) and Kowal et al. (2019) took a different perspective and generalized to a time-varying setting a discrete mixture prior and a continuous global-local prior, respectively.
These approaches share two main common features: first, they accommodate that only a subset of predictors could be significant at a given time $t$. Second, the selected subset could change over time.

All these approaches are suitable for univariate regressions, that is, the variable of interest is a scalar at time $t$. In this respect, they may be of limited usefulness for typical empirical asset pricing applications which often involved the cross-section of asset returns. On the other hand, our approach extends these existing approaches to a multivariate regression framework. More precisely, we model explicitly the possibility that the set of significant firms’ characteristics could change over time, while at the same time restricting the loading on a characteristic to be the same in the cross-section, consistent with the idea that it is only the spread in characteristics that drives the cross-sectional variation in the expected return of the portfolios. As a result, our proposed approach deals with few main obstacles that are often faced in empirical asset pricing: (1) a large number of predictors; and systematic risk exposures are (2) time-varying and (3) common drivers of the cross-section of asset returns.

One comment is in order. A major drawback of our approach is that it relies on a discrete mixture of normal, i.e., spike-and-slab, in order to induce sparsity in the dynamics of regression betas. Although convenient, such an approach suffers with one main problem, that is it can be difficult to set the hyper-parameters for the prior probability of inclusion and the variances of the spike and the slab distribution. We show in Appendix B that the posterior estimates of both our measure of sparsity and the stochastic volatility are reasonably stable under various choices of the prior probability of inclusion. Nevertheless, the choice of the prior variances of the Gaussian mixtures is also relevant (see, e.g., Giannone et al., 2017). While the choice of using non-informative priors put us on par with the existing literature, less conservative prior variance specifications could in principle affect the results. One possible solution could be to assume Exponential densities for both the spike and the slab, which leads to a discrete mixture of Laplace instead of a discrete mixture of Normals (see, e.g., Bernardo et al., 2011 for a discussion). Another viable alternative is to generalize a continuous global-local shrinkage prior such as the Horseshoe (see Carvalho et al., 2010 and Kowal et al., 2019) to a multivariate dynamic regression case. Leaving aside that our main goal is not to provide a definitive methodology to model time-varying sparsity, we think these are interesting.
avenues for future research.

7.4 Simulation study

In this section we test the reliability of our sparse dynamic panel regression through a simulation study. In particular, we evaluate the performance of our algorithm when it comes to track the time-varying parameters in a sparse setup using artificial data. In this respect, one should evaluate our estimation procedure along two key dimensions: first, the ability to track the dynamics of the parameters which are not-zero in the Data Generating Process (DGP), and, second, the ability of our estimation strategy to detect those parameters which are zero in the DGP.

Our Monte Carlo study involves generating data from a sparse time-varying parameter panel regression which restricts betas on a single characteristic to be the same in the cross-section, consistent with the theory (see, e.g., Kelly et al., 2018). In particular, the returns on the $n$ portfolios $y_t = (y_{1t}, \ldots, y_{nt})$ are related to the $p$ characteristics $x_{1t}, \ldots, x_{pt}$ as follows:

$$y_t = \beta_{i0t} + x_{1t}\beta_{1t} + x_{2t}\beta_{2t} + \ldots + x_{pt}\beta_{pt} + \epsilon_t \quad \text{with} \quad \epsilon_t \sim N(0, \Sigma)$$

Notice the slope parameters $\beta_{jt}$ are constant in the cross-section, whereas we assume pricing errors $\beta_{i0t}$ are different across portfolios. The simulation of the portfolio returns $y_t$ is based on two additional assumptions: (1) idiosyncratic errors are uncorrelated, i.e., $\text{corr}(\epsilon_{st}, \epsilon_{kt}) = 0$, and (2) characteristics are correlated with each other, i.e., for the portfolio $y_{it}$ we assume $\text{corr}(x_{ijt}, x_{int}) = \rho$ for $j \neq n$.

Sparsity in the regression coefficients is imposed as a mixture distribution whereby $\beta_{jt}$ is different from zero with a given probability $\pi$, i.e.,

$$\beta_{jt} = d_j \times \theta_{jt} \quad \text{where} \quad d_j = \begin{cases} 0 & \text{with probability } \pi \\ 1 & \text{with probability } 1 - \pi \end{cases}$$

that is, $d_j, j = 0, \ldots, p$ is a Bernoulli random variable that determines whether or not the coefficient $\beta_{jt}$ is zero or not. The dynamics in the regression coefficients for both the intercept and each of the
We set $\phi_1 = 0.99$ in order to generate a persistent but not explosive dynamic for the expected returns.

We consider $n = 3$ portfolios, each with $p = 10$ characteristics. We assume $\sigma_1^2 = 0.1, \sigma_2^2 = 0.2, \sigma_3^2 = 0.4$ and $\pi = 0.5$; the chosen value of $\pi$ implies that, on average, only half of the predictors are included in the dynamics of expected returns. The length of the time series simulated is $T = 600$ which is in line with the empirical implementation. Figure 14 reports the simulated values of the eight time-varying parameters and the respective estimates. Top panel shows the results when the regressors are assumed orthogonal to each other.

The average estimates track quite closely the dynamics of the simulated parameters $\beta_{1t}, \beta_{4t}, \beta_{6t}, \beta_{7t}$. Notice that such accuracy is far from trivial to obtain as the estimates are obtained through a filtering method and therefore are not smoothed. For the remaining four coefficients that were set to zero in the DGP, the estimates accurately track sparsity with the partial exception of the very initial period estimates which slightly different from zero, before eventually being shrunk to zero.

The bottom panel shows the estimates for a situation where regressors are highly correlated, that is $\rho = 0.95$. This is a setting in which it is notoriously difficult to capture sparsity even in a static setting (see, e.g., Su et al., 2017), let alone with time-varying parameters. The coefficient estimates of the coefficients $\beta_{4t}$ and $\beta_{6t}$ lie slightly below the true value, but some small bias has to be expected as our method is a time-varying version of classical penalized estimators which are known to be biased, especially in the presence of highly correlated regressors (see, e.g., Korobilis, 2013). As a matter of fact, top panel of Figure D.1 in Appendix D shows that with a moderate correlation of $\rho = 0.5$ across regressors, the estimates closely follow the dynamics of the “active” regressors while at the same time the regressors which are zeros are never significant, perhaps with the only exception of the initial period where estimates are slightly different from zero although by a very small amount.

The DGP implies time-varying portfolio-specific intercepts. In our context, these intercepts can be interpreted as time-varying pricing errors, so capturing their dynamics turns out to be crucial as

\[ \phi_j = \phi_0 + \phi_1 (\theta_{j,t-1} - \phi_0) + \tau_{j,t} \quad j = 0, \ldots, p \]
far as the asset-pricing implications of time-varying sparsity are concerned. Top panel of Figure 15 reports the results when the regressors are assumed orthogonal to each other.

![Insert Figure 15 here]

The estimates closely track the dynamics of the simulated intercepts. Consistent with bottom panel of Figure 14, the accuracy of the estimates tends to deteriorate in the presence of highly correlated regressors. Nevertheless, the dynamics of the intercepts is largely captured, although with a bias for $\beta_{03}$. The role of predictors’ correlation is somewhat confirmed by bottom panel of Figure D.1 in Appendix D, that shows the simulated intercepts against the estimates for moderately correlated regressors. The fitted values are now much closer to the true ones than with highly correlated regressors, that is, the bias is kept relatively small compared to the magnitude of the intercepts.

8 Conclusion

In this paper, we propose a framework to model expected returns within a dynamic regression model when the decision maker is confronted with a large number of risk factors. Our approach allows for the relation between returns and sources of risk to be sparse, i.e. with only few variables driving expected returns. More importantly, our framework permits this relation to change through time both in terms of the identity of the risk factors as well as in terms of their number.

We provide evidence that sparsity varies substantially over time with the amount of active predictors ranging from less than 30% during the early 90s, to around 60% before the mild recession of early 2000s. Delving further into the origins of such time variation we find a strong association of sparsity with the VIX (a proxy for the level of risk faced by imperfectly diversified liquidity providers) and non-financial leverage.

Finally, we develop a simple trading strategy to investigate the economic significance of time-varying sparsity in modeling expected returns. We find that our time-varying sparse predictive regression model delivers economic gains that are larger even relative to few established benchmarks, namely, the historical (unconditional) mean and the predictions from a three- and seven-factor models.
in addition to a simple, recursively estimated, ridge regression model.
References


36


38-


This table summarizes some of the recent evidence on the driving factors of expected returns. In particular, we report the main factors suggested by Freyberger et al. (2017), Kelly et al. (2018), DeMiguel et al. (2019), and Green et al. (2017).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>% chg. In shares outstanding</td>
<td>∆ShrOut</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log. chg. in split-adjusted shares</td>
<td>∆SO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% chg. In assets</td>
<td>Investment</td>
<td></td>
<td>agr</td>
<td></td>
</tr>
<tr>
<td>price-times shares outstanding</td>
<td>LME</td>
<td>mktcap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>last months volume to shares outstanding</td>
<td>LTurnover</td>
<td>turn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>standardized turnover</td>
<td></td>
<td>std turn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. scaled monthly volume in past 3 months</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short-term reversal (r2-1)</td>
<td>r2-1</td>
<td>strev</td>
<td>mom1m</td>
<td>mom1m</td>
</tr>
<tr>
<td>momentum (r12-2)</td>
<td>r12-2</td>
<td>mom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intermediate momentum (r12-7)</td>
<td>r12-7</td>
<td>intmom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>long-term reversal (r36-13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>momentum change: r6-1, r12-7</td>
<td></td>
<td></td>
<td>chmom</td>
<td></td>
</tr>
<tr>
<td>price to 52 week high price</td>
<td>rel-to-high</td>
<td>w52h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(size + long-term debt - total assets) to cash</td>
<td>roe</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>standardized unexplained volume</td>
<td>suv</td>
<td>suv</td>
<td></td>
<td>retvol</td>
</tr>
<tr>
<td>total assets</td>
<td>Total-vol</td>
<td></td>
<td>ret vol</td>
<td></td>
</tr>
<tr>
<td>correlations times ratio of vol</td>
<td></td>
<td></td>
<td>beta</td>
<td></td>
</tr>
<tr>
<td>idiosyncratic volatility from FF3</td>
<td>idiovol</td>
<td></td>
<td>retvol</td>
<td></td>
</tr>
<tr>
<td>(sales - cost of goods) to sales</td>
<td>pcm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>profit margin - mean pm - inf FF-48 industry</td>
<td>PM-adj</td>
<td>ato</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sales-to-assets</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unexpected quarterly earnings</td>
<td>sue</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gross profitability</td>
<td>gna</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>market beta using weekly returns over 3 years</td>
<td>betam</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>book-to-market ratio</td>
<td></td>
<td></td>
<td></td>
<td>cash</td>
</tr>
<tr>
<td>cash and cash equivalents divided by AT</td>
<td>zero-trade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zero trading days in recent month, turnover weighted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chg in number of analyst over t-3 to t (months)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sum of daily returns around earnings announcement increase in earnings relative to prior year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R&amp;D expense over market cap</td>
<td>rd-mve</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Time-varying sparsity and its determinants

The dependent variable is the probability of inclusion, see Figure 2. The regression includes aggregate financial conditions (proxied by the NFCI), and liquidity supply proxies. The NFCI is a composite financial conditions indicator that relies on information of 105 measures from money markets, debt and equity markets and the traditional and shadow banking systems. We also control for some of the NFCI sub-indices. Specifically in column (1) we employ the NFCI, whereas we employ the leverage sub-index of the NFCI in columns (2), the risk sub-index of the NFCI in column (3), and the credit sub-index of the NFCI in column (4). To control for liquidity supply we employ the TED (spread between three-month Eurodollar deposit rates and three-month Treasury Bill rates, see column (5)), the VIX (column (6)), and idiosyncratic volatility (measured as the cross-sectional standard deviation of individual stock returns, see column (7)). The sample period runs from January 1990 to June 2017.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
<th>(11)</th>
<th>(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFCI</td>
<td>0.023**</td>
<td></td>
<td></td>
<td></td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td></td>
<td></td>
<td></td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leverage</td>
<td>0.028***</td>
<td></td>
<td></td>
<td></td>
<td>0.027***</td>
<td>0.026***</td>
<td>0.026***</td>
<td>0.025***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td></td>
<td></td>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Risk</td>
<td>0.032***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit</td>
<td></td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TED</td>
<td></td>
<td>0.051***</td>
<td></td>
<td></td>
<td></td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.018)</td>
<td></td>
<td></td>
<td></td>
<td>(0.018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIX</td>
<td></td>
<td>0.002***</td>
<td></td>
<td></td>
<td></td>
<td>0.002**</td>
<td>0.002***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.001)</td>
<td></td>
<td></td>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IdioVol</td>
<td></td>
<td>1.433***</td>
<td></td>
<td></td>
<td></td>
<td>0.830</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.545)</td>
<td></td>
<td></td>
<td></td>
<td>(0.633)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N Obs</td>
<td>558</td>
<td>558</td>
<td>558</td>
<td>558</td>
<td>558</td>
<td>378</td>
<td>557</td>
<td>558</td>
<td>558</td>
<td>558</td>
<td>378</td>
<td>378</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>0.032</td>
<td>0.290</td>
<td>0.062</td>
<td>0.006</td>
<td>0.095</td>
<td>0.134</td>
<td>0.112</td>
<td>0.288</td>
<td>0.294</td>
<td>0.289</td>
<td>0.159</td>
<td>0.350</td>
</tr>
</tbody>
</table>
This table reports the unconditional descriptive statistics of our time-varying sparse predictive regression against fourth benchmarking methodologies. The benchmarks consist of an unconditional strategy that invests based on the historical average return until time $t$, two strategies relies on a panel regression (with portfolio fixed effects) that exploits either three firm characteristics only – namely size, book-to-market, and past 12 months returns, skipping the most recent – or seven characteristics – namely accruals, return on asset, split-adjusted shares outstanding, and log growth in total assets, in addition to size, book-to-market and past returns; and, finally, a strategy that exploits the prediction from a ridge regression. Panel A: shows the results where all alternative benchmarks are estimated based on a rolling window of 60 months. Panel B: reports the results where all benchmarks are estimated based on a rolling window of 120 months. The sample starts in 1991 and ends on 2017.

### Panel A: Rolling-window 60-months

<table>
<thead>
<tr>
<th></th>
<th>Recursive-Mean</th>
<th>TV-Sparsity</th>
<th>Three-Factor-Model</th>
<th>Seven-Factor-Model</th>
<th>Rolling Ridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp·Ret·</td>
<td>0.062·</td>
<td>0.256·</td>
<td>-0.005·</td>
<td>0.005·</td>
<td>0.150·</td>
</tr>
<tr>
<td>St·Dev·</td>
<td>0.374·</td>
<td>0.852·</td>
<td>0.287·</td>
<td>0.231·</td>
<td>0.717·</td>
</tr>
<tr>
<td>SR·</td>
<td>0.166·</td>
<td>0.301·</td>
<td>-0.019·</td>
<td>0.021·</td>
<td>0.210·</td>
</tr>
</tbody>
</table>

### Panel B: Rolling-window 120-months

<table>
<thead>
<tr>
<th></th>
<th>Recursive-Mean</th>
<th>TV-Sparsity</th>
<th>Three-Factor-Model</th>
<th>Seven-Factor-Model</th>
<th>Rolling Ridge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp·Ret·</td>
<td>0.052·</td>
<td>0.178·</td>
<td>0.016·</td>
<td>0.018·</td>
<td>0.065·</td>
</tr>
<tr>
<td>St·Dev·</td>
<td>0.271·</td>
<td>0.463·</td>
<td>0.281·</td>
<td>0.230·</td>
<td>0.286·</td>
</tr>
<tr>
<td>SR·</td>
<td>0.192·</td>
<td>0.385·</td>
<td>0.055·</td>
<td>0.080·</td>
<td>0.228·</td>
</tr>
</tbody>
</table>
**Figure 1: Rolling window shrinkage**

This figure shows the rolling window estimate of the shrinkage parameter from a ridge regression. The red (blue) line shows the shrinkage parameter for a window of 120 (60) months. The sample period runs from February 1965 to June 2017.

**Figure 2: Dynamic sparsity**

This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns. In particular, the left panel reports the percentage of characteristics for which the corresponding slope parameters are different from zero in a given month. The right panel decompose the percentage of active predictors by economic groups according to the classification provided in the main text. The sample period runs from February 1965 to June 2017.
**Figure 3: Dynamic sparsity and predictability**

This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns against two measures of out-of-sample (OOS) predictability. Left panel shows the average squared forecasting error whereas right panel shows the log-predictive likelihood, both averaged across portfolios. Predictability measures are calculated out-of-sample by an expanding window starting from April 1991 then adding one observation at a time. The sample period runs from February 1965 to June 2017.

(a) Dynamic sparsity vs. forecasting error  
(b) Dynamic sparsity vs. log-predictive likelihood

**Figure 4: Variance explained by PCAs across firms’ characteristics and portfolio returns**

This figure shows the percentage of the variance explained by the first (left panel) and the first two (right panel) principal components for each anomaly-based long-short portfolio (y-axis) and groups of predictors (x-axis). The last row on the y axis represent the average values across anomaly-based long-short portfolios. The sample period runs from February 1965 to June 2017.

(a) First PC  
(b) First two PCs
Figure 5: **Dynamic sparsity and principal components**

This figure shows the percentage of active characteristics at each time $t$ against the explained variation of the first two principal components (PCs) of the active predictors (top panel) and the portfolio returns (bottom panel). The principal components are recursively calculated on an expanding window basis. The sample period runs from February 1965 to June 2017.

![Figure 5: Dynamic sparsity and principal components](image)

(a) PCs of the active predictors  
(b) PCs of the portfolio returns

---

Figure 6: **The role of time-varying idiosyncratic volatility**

The left panel of this figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns with and without stochastic volatility explicitly considered. The right panel shows the difference in the proportion of active predictors between the two models together with the time-varying volatility from the main model specification. The sample period runs from February 1965 to June 2017.

![Figure 6: The role of time-varying idiosyncratic volatility](image)

(a) Dynamic sparsity with and w/o SV  
(b) Difference in Dynamic Sparsity and SV
Figure 7: Firms’ characteristics dynamics

This figure shows the inclusion of a given characteristic at each time $t$ in a forecasting regression of 1-month ahead factor returns. In particular, the figure reports the adjacency matrix. This adjacency matrix shows which predictors are included in the cross-section of expected returns each month. Darker areas imply that a given characteristic is included, whereas lighter areas imply that a given characteristic is not included. The sample period runs from February 1965 to June 2017.

(a) Trading Frictions
(b) Value
(c) Profitability
(d) Investment
(e) Intangibles
(f) Past returns
Figure 8: Unconditional probabilities of inclusion

This figure shows a measure of horizontal sparsity of a given characteristic throughout the entire sample period. The colors of the vertical bars identify the groups of predictors: light blue for past returns, red investment, green profitability, dark green intangibles, dark blue and brow value and trading frictions, respectively. The sample period runs from February 1965 to June 2017, monthly.
**Figure 9: Portfolio strategy**

This figure shows cumulative profits from “trading rules” using real time information. Each line plots the cumulative value of $r_{x_{j+1}} \times E_t[r_{x_{j+1}}]$. $E_t[r_{x_{j+1}}]$ are formed using one among: the unconditional historical average (blue line); a model that exploits three firm characteristics, namely size, book-to-market, and past 12 months returns (yellow line with square markers); a model based on accruals, return on asset, split-adjusted shares outstanding, and log growth in total assets, in addition to size, book-to-market and past returns (violet line with circles); a model that exploits the signal from a rolling window shrinkage, namely ridge, regression (dashed green line); and our model that exploits the time-varying nature of sparsity (red line with diamond circle markers). The left (right) panel shows the results based on a rolling of 60 (120) monthly observations.

![Figure 9: Portfolio strategy](image)

(a) Rolling window 60 months  
(b) Rolling window 120 months

**Figure 10: Time-varying sparsity for the aggregate stock market**

This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead aggregate stock market excess returns. In particular, the left panel reports the percentage of characteristics for which the corresponding slope parameters are different from zero in a given month. The sample period runs from February 1965 to June 2017. The right panel shows the dynamics of the stochastic volatility for the aggregate stock market.

![Figure 10: Time-varying sparsity for the aggregate stock market](image)

(a) Dynamic sparsity for the aggregate stock market  
(b) Dynamic sparsity and idiosyncratic vol
Figure 11: Unconditional probabilities of inclusion for the aggregate stock market

This figure shows a measure of horizontal sparsity of a given characteristic throughout the entire sample period for the aggregate stock market. The colors of the vertical bars identify the groups of predictors: light blue for past returns, red investment, green profitability, dark green intangibles, dark blue and brown value and trading frictions, respectively. The sample period runs from February 1965 to June 2017, monthly.
This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns and controlling for non-linearities in the input variables. In particular, top panel reports the percentage of characteristics for which the corresponding slope parameters are different from zero in a given month. In addition to the 62 firms’ characteristics we include 61 additional predictors which consist in the interaction of each characteristic with size (123 predictors in total). The bottom panel reports the same figure but now including the square of each characteristic instead of the interactions with size (124 predictors in total). The sample period runs from February 1965 to June 2017.
Figure 13: Time-varying sparsity vs. rolling window shrinkage estimates

This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns against standard rolling window estimates from standard shrinkage priors, namely ridge (see Hsiang, 1975), lasso (see Park and Casella, 2008), and elastic net (see Li and Lin, 2010). Rolling window estimates are based on 120 observations. The sample period runs from February 1965 to June 2017. Notice the scale of the shrinkage parameters from the penalized regressions is inverted to have the same interpretability of our measure of sparsity, i.e., % active predictors out of total.
Figure 14: Simulation results for the time-varying betas

This figure shows the true vs. estimated time-varying betas in the simulation study. The sample size is $T = 600$. Top panel shows the results for a set of orthogonal characteristics, i.e., $\rho = 0$. Bottom panel shows the results for a set of highly correlated characteristics, i.e., $\rho = 0.95$. The red line reports the true simulated betas and the blue line represents the estimates.

(a) Orthogonal Characteristics, $\rho = 0$

(b) Correlated Characteristics, $\rho = 0.95$
Figure 15: **Simulation results for the time-varying alphas**

This figure shows the true vs. estimated time-varying intercepts in the simulation simulation study. The sample size is $T = 600$. Top panel shows the results for a set of orthogonal characteristics, i.e., $\rho = 0$. Bottom panel shows the results for a set of highly correlated characteristics, i.e., $\rho = 0.95$. The red line reports the true simulated betas and the blue line represents the estimates.

(a) Orthogonal Characteristics, $\rho = 0$

(b) Correlated Characteristics, $\rho = 0.95$
Figure B.1: Sensitivity to different prior probabilities of inclusion

This figure shows the percentage of active characteristics at each time $t$ in a forecasting regression of 1-month ahead factor returns for the different prior inclusion probabilities. The base case (light-blue line) assumes $\pi_{j,0} = 0.5$. The two alternative prior specifications assume a more conservative, namely less sparse model, with $\pi_{j,0} = 0.8$ (magenta line with circles) and $\pi_{j,0} = 0.9$ (green line with squares). The sample period runs from February 1965 to June 2017.

(a) Active predictors (% of total)  
(b) Idiosyncratic volatility
Figure D.1: Simulation results with moderately correlated regressors

This figure shows the true vs. estimated time-varying parameters in the simulation study. The sample size is $T = 600$ and we assume regressors are highly correlated, that is $\rho = 0.5$. Top (bottom) panel shows the time-varying betas (alphas). The red line reports the true simulated betas and the blue line represents the estimates.
Appendix

A Data

Data comes from the Center for Research in Security Prices (CRSP) monthly stock file. We follow standard conventions and restrict the analysis to common stocks of firms incorporated in the United States trading on NYSE, Amex, or Nasdaq with market prices above USD $0.1.

All characteristics are winsorized monthly in those directions in which they could potentially have outliers. For example, a positive value variable for which 0 is a plausible value will only have winsorization at the 99th percentiles. For each characteristic, we sort all stocks into ten groups based on the NYSE breakpoints. The characteristics definition follows Freyberger et al. (2017) but for completeness we report here a brief description.

LME is the total market capitalization at the firm level at the end of the previous calendar month. LTurnover is the ratio of total monthly trading volume over shares outstanding at the end of the previous month. The bid-ask spread (spread) is the average daily bid-ask spread during the previous month. We also construct lagged returns over the previous month (r2,1), the previous 12 months leaving out the last month (r12,2), intermediate momentum (r12,7), and long-run returns from three years ago until last year (r36,13). We follow Frazzini and Pedersen (2014) in the definition of Beta (Beta), and idiosyncratic volatility (Idio vol) is the residual from a regression of daily returns on the three Fama and French factors in the previous month as in Ang et al., (2006).

Balance-sheet data are from the Standard and Poor’s Compustat database. We define book equity (BE) as total stockholders’ equity plus deferred taxes and investment tax credit (if available) minus the book value of preferred stock. Based on availability, we use the redemption value, liquidation value, or par value (in that order) for the book value of preferred stock. We prefer the shareholders’ equity number as reported by Compustat. If these data are not available, we calculate shareholders’ equity as the sum of common and preferred equity. If neither of the two are available, we define shareholders’ equity as the difference between total assets and total liabilities. The book-to-market (BM) ratio of year \( t \) is then the book equity for the fiscal year ending in calendar year \( t - 1 \) over the market equity as of December \( t - 1 \). We use the book-to-market ratio for estimation starting in June.
of year \( t \) until May of year \( t + 1 \) predicting returns from July of year \( t \) until June of year \( t \). We use the same timing convention for balance-sheet variables unless we specify it differently.

\( \text{AT} \) are total assets, \( \text{ATO} \) are sales scaled by net operating assets, and \( \text{C} \) is cash and short-term investments over total assets. \( \text{CTO} \) is capital turnover; \( \text{D2A} \) is depreciation and amortization over total assets, and \( \text{DPI2A} \) is the change in property, plant, and equipment. \( \text{E2P} \) is the earnings-to-price ratio. We define expenses to sales (FC2Y) as the sum of advertising expenses, research and development expenses, and selling, general, and administrative expenses over sales, and Free CF is net income and depreciation and amortization less the change in working capital and capex. Investment is the growth rate in total assets, \( \text{Lev} \) is the ratio of total debt to total debt and shareholders’ equity, and NOA are net operating assets to lagged total assets. We define operating accruals (OA) as in Sloan (1996), and operating leverage (OL) is the ratio of cost of goods sold and selling, general, and administrative expenses over total assets. We define the price-to-cost margin (PCM) as sales minus cost of goods sold over sales, the profit margin (PM) as operating income after depreciation to net sales, gross profitability \( \text{(Prof)} \) as gross profits over book value of equity, and \( \text{Q} \) is Tobin’s Q. Rel to high is the closeness to the 52-week high price and RNA is the return on net operating assets. The return on assets (ROA) is income before extraordinary items to total assets and the return on equity (ROE) is the ratio of income before extraordinary items to lagged book value of equity. \( \text{S2P} \) is the ratio of sales to market capitalization, \( \text{SGA2S} \) is the ratio of selling, general, and administrative expenses to net sales, spread is the monthly average bid-ask spread, and SUV is standardized unexplained volume.

### B Prior sensitivity

In this section we investigate the sensitivity of the model estimates to different prior inclusion probabilities \( \pi_{j,0} \). The base case scenario is assuming that the prior inclusion probability is distributed as a uniform \((0, 1)\), such that \( \pi_{j,0} = 0.5 \). Notice that our dynamic setting in principle makes the impact of the prior specification at time \( t = 1 \) less relevant as the sample size. Nevertheless, it is worth to investigate the robustness of the estimates to more conservative specifications. Figure B.1 shows the results and compares the posterior estimates under \( \pi_{j,0} = 0.8 \) (magenta line with circles) and \( \pi_{j,0} = 0.9 \).
The left panel shows the estimates of the time-varying sparsity measure. Except few nuances, the dynamics of the estimates is highly coherent. The correlation among the estimates is as high as 0.9, on average. The right panel shows the posterior estimates of the idiosyncratic stochastic volatility. The estimates are virtually equivalent and can not be distinguished each other for almost all of the sample period.

As a whole, Figure B.1 shows that our posterior estimates are not too sensitive to initial prior on the model size, that is on the initial inclusion probabilities. Of course, we are considering reasonable values of prior probabilities, meaning values that a prior assume the model is close to be non-sparse. On the other hand, one could assume a priori that the model is highly sparse, for e.g., \( \pi_{j,0} = 0.1 \). However, this is against the spirit of the empirical application as we want to be as much as possible explicitly agnostic about the degree of sparsity and let the data speak about the actual number of firms’ characteristics active at a given time \( t \).

C Estimation strategy

C.1 A Primer on Variational Bayes inference

Before we describe in detail our estimation strategy, we introduce how Variational Bayes (VB) methods can be used to approximate posterior distributions which are otherwise numerically intractable. VB methods have become increasingly popular to solve inferential problems in a data-rich environment, and in particular as a way to numerically approximate the posterior densities of quantities of interest in a big data framework. In this respect, VB generalise standard MCMC which are cumbersome, if not impossible, to implement when the posterior distribution is of too high dimension.\(^{15}\)

Consider some observable \( y \), a latent variable \( x \) and some parameter \( \theta \). In our setting, and as is common in the time-varying parameter literature, the latent state \( x \) represents the unobserved...

\(^{15}\)See, e.g., Ormerod and Wand, 2010 and Wand, 2017 for a survey of VB relating to machine learning and Hajargasht and Wozniak, 2018 for a survey on econometric applications of VB methods.
dynamic betas and sparsity indicators, while the parameters $\theta$ represent the static, structural, parameters such as the error variances.

Our object of interest is the joint posterior distribution of the latent states and parameters defined as $p(x, \theta | y)$. The underlying idea of VB is to find an approximate density $q(x, \theta)$ that is as close as possible to the target, intractable, distribution $p(x, \theta | y)$. “Closeness” is defined as the Kullback-Leibler divergence, i.e.,

$$
\text{KL} = \int q(x, \theta) \log \left( \frac{q(x, \theta)}{p(x, \theta | y)} \right) dx d\theta
$$

Note that the divergence metric is asymmetric, that is $\text{KL} \geq 0$ and equals to zero if and only if $q(x, \theta) = p(x, \theta | y)$. This is chosen for computational tractability, but can also be viewed as inducing approximations where areas of $q(.)$ are accurate, rather than areas of high $p(.)$.

The “mean-field” form of VB assumes that the joint, auxiliary, density $q(x, \theta)$ can be factorized as $q(x) = \prod_i q(x_i | \theta_i)$ and $q(\theta) = \prod_i q(\theta_i | x_i)$. In particular, it can be shown that, in general, the optimal choices for $q(x)$ and $q(\theta)$ are

$$
q(x) \propto \exp \left\{ \int q(\theta) \log p(x | y, \theta) d\theta \right\},
$$

$$
q(\theta) \propto \exp \left\{ \int q(x) \log p(\theta | y, x) dx \right\},
$$

In a nutshell, and similar to a standard Gibbs sampler, VB algorithms iterate over these two densities until convergence. In fact, similar to a Gibbs sampler, $q(x)$ and $q(\theta)$ involve the full conditional distributions. However, differently from a Gibbs sampler, VB algorithms do not repeatedly simulate from these full conditionals, making the approximation feasible in large dimensional settings, i.e., when the model space is practically too large to explore.

Our implementation relies on three assumptions: first, the complete likelihood of the data, i.e., $p(y | x, \theta)$ comes from the exponential family. Second, all priors need to be conjugate with respect to the complete likelihood. Third, we assume that the factorization $q(x, \theta) = q(x) q(\theta)$ holds. The first two assumptions are standard and not particularly restrictive; typical empirical linear factor models assume conditional Gaussian distributions, which lead to easily specified conjugate posterior.
updates. The third assumption is reasonable as far as posterior estimates of the latent states $x$ and the parameters $\theta$ are relatively low, that is, the joint density can be approximated as the product of the marginals. This is indeed the case in our setting.

C.2 Estimation strategy

With the exception of the dynamic “spike-and-slab” prior (5), the likelihood and prior specification are mainly standard and similar to common choices in time-varying parameter regressions (see, e.g., Belmonte et al., 2014; Chan et al., 2012; Dangl and Halling, 2012; Bianchi et al., 2017; Bianchi and McAlinn, 2018; among others). However, when dynamic sparsity is fully considered, the question arises as to how incorporate such prior into posterior computation. In addition, the large model space makes prohibitive the use of typical Markov Chain Monte Carlo (MCMC) methods for the posterior approximation.

We follow Wang et al. (2016) and Koop and Korobilis (2018) and implement an approximation strategy to estimate the posterior distributions of the variables of interest. First note that (3) can be viewed as a hierarchical prior for the regression coefficients $\beta_t$ of the form

$$
\beta_t | \beta_{t-1}, Q_t \sim N(\beta_{t-1}, Q_t)
$$

subject to the initial condition $\beta_0 \sim N(\beta_0, P_0)$. Without loss of generality, the covariance matrix of the state parameters $Q_t$ is assumed diagonal. As a result, the state variances are assumed to follow a prior of the form

$$
q_{j,t}^{-1} \sim \text{Gamma}(c_0, d_0), \quad j = 1, \ldots, p
$$

$$
\nu_{i,t}^{-1} \sim \text{Gamma}(f_0, g_0), \quad i = 1, \ldots, n
$$

where $c_0, d_0, f_0, g_0$ are fixed hyper-parameters which have been discussed in Section 4. We follow Wang et al. (2016) and re-write the dynamic mixture of normal (5) as a prior for the latent pseudo-
observation $z_{j,t} = 0$ which has a distribution of the form

$$z_{j,t} \sim N(\beta_{j,t}, v_{j,t}),$$

where we define $v_{j,t} = (1 - \gamma_{j,t})^2 v_{j,0}^2 + \gamma_{j,t}^2 v_{j,1}^2$. The resulting posterior is of the form

$$p(\beta_{1:T}, Q_{1:T}, V_{1:T}, \log \sigma_{1:T}^2, \nu_{1:T}|y^T, z^T) \propto \prod_{t=1}^T p(\beta_t|\beta_{t-1}, Q_t) p(y_t|\beta_t, \sigma_t^2) p(z_t|\beta_t, V_t) \times$$

$$\times p(\log \sigma_{1:T}^2|\log \sigma_{1:T}^2, \nu_t) p(\gamma_t) p(Q_t)p(\nu_t) \quad \text{(A.2)}$$

where $\beta^T = (\beta_1, \ldots, \beta_T)$ and $V_t = \text{diag}(v_{1,t}, \ldots, v_{p,t})$. The goal of a VB approach is to approximate the intractable joint posterior $p(\beta_{1:T}, Q_{1:T}, V_{1:T}, \log \sigma_{1:T}^2, \nu_{1:T})$ with a tractable distribution $q(\beta^T, Q^T, V^T, \log \sigma_{1:T}^2, \nu_{1:T})$. Applying a standard mean-field approximation once can factorize the tractable distribution as

$$q(\beta^T, Q^T, V^T, \log \sigma_{1:T}^2, \nu_{1:T}) = q(\beta_{1:T}) \prod_{t=1}^T \prod_{j=1}^p \prod_{i=1}^n q(v_{j,t}) q(\gamma_t) q(\log \sigma_{i,t}^2) q(\nu_{i,t}).$$

Notice that the factorization decomposes the parameters $q_t$, $v_t$ and $\gamma_t$ into components that are independent over $t$ and over $j$, in order to facilitate the computation. However, the dynamics of the betas is not factorized, that is, the time-series dependence between $\beta_t$ and $\beta_{t-1}$ remains intact. As a result, each of the elements in the factorization can be evaluated separately as follows: the marginal distribution $q(\beta^T) = \prod_{t=1}^T q(\beta_t|\beta_{t-1})$ can be evaluated by standard filtering recursions using the transformed state equation

$$\beta_t = \bar{F}_t \beta_{t-1} + \tilde{\eta}_t$$
where \( \bar{\eta} \sim N(0, \bar{Q}_t) \), with \( \bar{Q}_t = (Q_t^{-1} + V_t^{-1})^{-1} \) and \( \bar{F}_t = \bar{Q}_t Q_t^{-1} \). This can be derived as:

\[
q(\beta_t | \beta_{t-1}) \propto p(\beta_t | \beta_{t-1}, Q_t) p(z_t | \beta_t, V_t)
\]

\[
\propto \exp \left\{ -\frac{1}{2} (\beta_t - \beta_{t-1})' Q_t^{-1} (\beta_t - \beta_{t-1}) - \frac{1}{2} \beta_t' V_t^{-1} \beta_t \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} \beta_t' Q_t^{-1} \beta_t + \beta' Q_t^{-1} \beta_{t-1} - \frac{1}{2} \beta' V_t \beta_t \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} (\beta_t - \bar{F}_t \beta_{t-1})' \bar{Q}_t^{-1} (\beta_t - \bar{F}_t \beta_{t-1}) \right\}
\]

(A.3)

As a result, conditional on the other parameters of the model, \( q(\beta^T) \) can be evaluated through a standard Kalman filtering recursion. The functional form of \( q(v_{j,t}) \), the second ingredient in the factorization above, can be obtained using a standard mixture-of-normals specification, i.e.,

\[
v_{j,t} = \begin{cases} 
  v_0^2 & \text{if } \gamma_{j,t} = 0 \\
  v_1^2 & \text{if } \gamma_{j,t} = 1 
\end{cases}
\]

(A.4)

with \( q(\gamma_{j,t}) \propto \text{Bernoulli}(\pi_{j,t}) \). The inclusion probability at each time \( t \), \( \pi_{j,t} \) is defined as

\[
\pi_{j,t} = \frac{N(\beta_{j,t} | 0, v_{j,1}^2) \pi_{j,0}}{N(\beta_{j,t} | 0, v_{j,1}^2) \pi_{j,0} + N(\beta_{j,t} | 0, v_{j,0}^2)(1 - \pi_{j,0})}
\]

(A.5)

Thus, conditional on other model parameters, the form for \( q(v_{j,t}) \) allows for easy updating. Finally, the joint distribution of the state variance parameters can be obtained conditional on the other model parameters as follows:

\[
q(q_{j,t}) \propto \text{Gamma}(c_{j,t}, d_{j,t})
\]

where \( c_{j,t} = c_0 + .5 \cdot d_{j,t} = d_0 + D_{jj}/2 \) and \( D_{jj} \) is the \( j \)-th diagonal element of \( D = P_{tt} + \beta_{tt} \beta_{tt}' + \left( P_{t-1|t-1} + \beta_{t-1|t-1} \beta_{t-1|t-1}' \right) (I_p - \bar{F}_t)' \), with \( \beta_{tt} \) and \( P_{tt} \) the filtered mean and covariance of the parameter \( \beta_t \) at time \( t \).

As far as stochastic volatility is concerned, the non-negativity constraint imposed through a log-transformation \( \log \sigma_{i,t}^2 \) makes the system of equation non-linear, which obviously complicate things.
Notice that the state-space representation (2)-(4) implies that \( \log \varepsilon_{i,t}^2 \sim \log \chi^2(1) \) (see Kim et al., 1998 and Omori et al., 2007). We follow Koop and Korobilis (2018) and approximate such distribution with a normal distribution with the mean and variance that corresponds to the first and second moment of the \( \log \chi^2(1) \), i.e., \( N(-1.27, 4.93) \). Koop and Korobilis (2018) showed that such distribution represents a bad approximation of the true \( \log \chi^2(1) \) for small values of idiosyncratic volatility. However, we can help avoid this region of the parameter space by standardizing the firms’ characteristics so that they have unit variance. Our simulation study below suggest that such approximation is fairly accurate. The main advantage of this approximation is that it allows the system to be linear, that is standard conjugate updating for normal-inverse Gaussian distributions can be used to sample from \( q(\nu_{i,t}) \). Conditional on \( \nu_{i,t} \) the sampling of \( \log \sigma_{jt}^2 \) can be executed through a standard filtering recursion (see, e.g., Frühwirth-Schnatter, 1994).

### D Further simulation results

Top panel of Figure D.1 reports the simulated values of the eight time-varying parameters and the respective estimates when the predictors are assumed to be moderately correlated, i.e., \( \rho = 0.5 \).

[Insert Figure D.1 here]

The average estimates track closely the dynamics of the simulated parameters, including those parameters that have been set to zero in the data-generating process. As pointed out already in the main text, such accuracy is far from expected as the estimates are obtained through a filtering method which is forward-looking in nature. Bottom panel shows that also the estimates of the asset-specific alphas also match quite closely the true simulated values. As a whole, Figure D.1 shows that for the reasonable case of moderately correlated predictors our estimation strategy is able to capture quite closely the amount of sparsity and the dynamics of the significant parameters at the same time.