Giacomini, Raffaella; Lu, Jason; Smetanina, Katja

Working Paper

Perceived shocks and impulse responses

cemmap working paper, No. CWP21/24

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Giacomini, Raffaella; Lu, Jason; Smetanina, Katja (2024) : Perceived shocks and impulse responses, cemmap working paper, No. CWP21/24, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2024.2124

This Version is available at:
https://hdl.handle.net/10419/306646

cemmap
centre for microdata methods and practice

Perceived shocks and impulse responses

Raffaella Giacomini
Jason Lu
Katja Smetanina

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP21/24

UKRI Economic and Social Research Council

Perceived shocks and impulse responses*

Raffaella Giacomini‡, Jason Lu§, Katja Smetanina¶

November 22, 2024

Abstract

This paper develops a novel approach that leverages the information contained in expectations datasets to derive empirical measures of beliefs regarding economic shocks and their dynamic effects. Utilizing a panel of expectation *revisions* for a single variable across multiple horizons, we implement a time-varying factor model to nonparametrically estimate the latent shocks and their associated impulse responses at every point in time. The method is designed to accommodate small sample sizes and relies on weak assumptions, requiring no explicit modeling of expectations or assumptions about agents' forecasting models, information sets, or rationality. Our empirical application to consensus inflation expectations identifies a single perceived shock that closely aligns with observed inflation surprises. The time-varying impulse responses indicate a significant decline in the perceived persistence of this shock, suggesting that inflation expectations have become more "anchored" over time.

JEL classification: C38, C14, E37, E65

Keywords: Beliefs; Time-varying factor models; Nonparametric estimation; Principal Components Analysis; Heteroskedasticity; Small samples.

*We thank: Gadi Barlevy, Martin Ellison, Guido Lorenzoni, Morten Ravn, Esther Ruiz, Dacheng Xiu, Andrei Zeleneev and seminar participants at several seminars and conferences for useful comments and suggestions.

‡University College London, Department of Economics. Email: r.giacomini@ucl.ac.uk

§International Monetary Fund. Email: jlu2@imf.org

¶University of Chicago Booth School of Business. Email: E.Smetanina@chicagobooth.edu

1 Introduction

Data on expectations are increasingly recognized as crucial for understanding economic dynamics and the formation of beliefs. This paper contributes to the literature by demonstrating how the horizon dimension in expectations datasets, combined with a focus on expectation *revisions*, can be leveraged to derive empirical measures of key economic quantities: the shocks perceived by agents and their corresponding dynamic effects (the impulse responses).

The necessary data are a balanced panel of expectation revisions for a single variable across multiple future horizons and over time. These revisions can be obtained from expectations datasets that encompass both time-series and horizon dimensions, which are widely available in economics. Examples of such datasets include: surveys of expectations (e.g., Blue Chip Analysts, Survey of Professional Forecasters, I/B/E/S Earning Forecasts, Survey of Firms' Inflation Expectations, University of Michigan Survey of Consumers), market-based expectations (e.g., Treasury Inflation-Protected Securities break-even inflation, inflation swap contracts, currency futures, implied volatility from option contracts), and combinations of surveys and/or market-based expectations (e.g., Cleveland Fed inflation expectations).[1]

At the core of this paper lies a simple yet powerful idea: fitting a factor model with time-varying loadings to the panel of expectation revisions allows us to recover two latent components at any given time. First, we extract a low-dimensional vector of independent common factors - representing the shocks that drive agents' revisions across all horizons. Second, we extract the corresponding loadings - representing agents' beliefs about the dynamic effects of each shock, which can vary over time. A key challenge we face is that existing nonparametric econometric methods are inadequate in our context, where the small horizon dimension typical of the data leads to small-sample bias. We favor nonparametric methods for their ability to avoid additional structure and assumptions, and we propose a novel econometric approach specifically designed to address these issues.

[1] A revision is defined as the difference between the expectation of a variable at a future horizon (e.g., inflation in May) generated in the current period (e.g., April) and the expectation of the same variable (inflation in May) produced in the preceding period (e.g., March).

Fitting a factor model to expectation revisions may initially appear arbitrary; however, we demonstrate that such a factor structure emerges naturally from various theories of expectation formation. For instance, when a rational agent utilizes a Structural Vector Moving Average (SVMA) model to generate expectations (see, e.g., Plagborg-Møller, 2019), the expectation revisions for a given variable across multiple horizons exhibit a factor structure without idiosyncratic errors, with factors and loadings corresponding to the structural shocks and structural impulse responses, respectively.[2] Similarly, a dynamic factor model for the target variable inherently suggests a factor structure for the revisions. Furthermore, certain theories of information rigidities, such as the noisy information model in Coibion and Gorodnichenko (2015), also imply a factor structure for expectation revisions without idiosyncratic errors, where the loadings represent the model-implied impulse responses and the factor is the agent's perceived shock after filtering out the noise.

To illustrate one possible use of our method, we apply it to extract historical perceived shocks and impulse responses related to inflation. We construct a term structure of consensus expectation revisions across various horizons by integrating two data sources: the Blue Chip Analysts for short- and medium-term horizons and the Cleveland Fed expectations for long-term horizons. Our key findings are threefold. First, a single perceived shock drives inflation revisions across all horizons and is highly correlated with inflation surprises, particularly those from Stock and Watson (2007)'s model. However, the extracted impulse responses diverge significantly from this model's predictions. Second, the perceived impulse responses exhibit time-varying shapes, showing a noticeable secular decline in perceived shock persistence. Finally, we illustrate that our method can inform important policy questions, such as whether shifts in long-term expectations signal beliefs in enduring effects of the shock - indicative of deanchoring - or merely reflect the perception of a large shock. Subject to the caveats of end-of-sample nonparametric estimation, our results suggest that the large changes in long-term expectations observed in 2022 primarily resulted from a large perceived shock rather than deanchoring.

[2]Note the potential to extract multiple shocks from expectations of only one of the variables in the system.

A key feature of our approach is its foundation on weak assumptions. By focusing on revisions, we avoid the need to assume how agents form expectations, relying solely on expectations data to extract beliefs about shocks and their dynamic effects. This stands in contrast to the extensive literature on testing belief accuracy and rationality, which utilizes ex-post data to construct forecast errors and imposes assumptions regarding belief formation (e.g., the seminal paper by Coibion and Gorodnichenko, 2015 and the literature it spurred). Similarly, studies on 'belief distortions' and their impact on aggregate fluctuations (e.g., Bianchi et al., 2022, Enders et al., 2021) also depend on forecast errors, necessitating assumptions about information sets and the connection between expectations and fundamentals. Furthermore, we adopt a non-parametric approach to estimate both shocks (via Principal Components Analysis, or PCA) and time-varying impulse responses (allowing for smooth and flexible patterns of temporal variation). Our methodology is robust to unmodeled serial correlation in shocks and idiosyncratic errors, and it is specifically tailored to address the small horizon dimension that is typical of expectations datasets.[3]

In terms of impulse responses, our method for deriving empirical measures of beliefs regarding the dynamic effects of shocks is, to our knowledge, a novel contribution to the literature. The insights gained from perceived impulse responses can be valuable across various contexts, and engage with multiple strands of literature. First, beliefs regarding the long-term effects of shocks are critical for central banks, and our method can yield new insights into the anchoring of long-term inflation expectations (e.g., Carvalho et al., 2023), such as those provided by our empirical application. Second, our framework enables the documentation of new stylized facts related to the perceived persistence of shocks and the shape of impulse response functions, which can be instrumental in testing theories that link beliefs about shock persistence to aggregate fluctuations (e.g. Blanchard et al., 2013). Third, the flexible nonpara-

[3]The core premise of this paper - fitting a time-varying factor model to expectation revisions across horizons - could equally be approached through a parametric lens. For instance, one might consider the Bayesian approach to factor models with time-varying loadings and stochastic volatility by Del Negro and Otrok (2008). However, it is important to note that a parametric strategy requires several crucial modeling decisions. These include modeling the serial correlation of both the factors and idiosyncratic errors, specifying the nature of time variation, selecting prior distributions if a Bayesian approach is employed, and addressing various identification and normalization challenges.

metric nature of our approach challenges the conventional dichotomy of permanent versus transitory shocks, allowing for a richer variety of impulse response shapes and persistence patterns over time. For example, our empirical findings reveal impulse response functions for inflation that are time-varying and differ markedly from those predicted by the model in Stock and Watson (2007).

Our method also contributes to the literature on extracting empirical measures of shocks by providing a rigorous econometric framework for shock extraction via PCA that capitalizes on the horizon dimension of expectations data, even though it is small. Our method not only corrects the small-sample biases of PCA methods, but also facilitates the selection of the appropriate number of shocks. Furthermore, our allowance for time-varying loadings results in the extraction of shocks that differ from those derived under constant loadings assumptions. A seminal paper in this literature, Gürkaynak et al. (2005), extracts the first principal component from changes in different measures of market-based expectations around monetary policy announcements. Applied to this type of data, our approach can be seen as providing a rigorous method that exploits the horizon dimension of the data.

A natural question to ask is whether the shocks extracted from expectations data using the method proposed in this paper can be given a structural interpretation, e.g., be considered a policy shock. This presents a challenge similar to that encountered in the literature on narrative measures of shocks - the need to disentangle policy shocks from any other change in information and behavior that may occur between the consecutive time periods underlying the revisions. Potential contamination of shocks can exist even when revisions are computed within a narrow time window around specific events such as monetary policy announcements (see, e.g., Miranda-Agrippino and Ricco, 2021). While it is difficult to provide a general solution to this problem, the method offers potential new avenues for identifying structural shocks. First, the method allows for the recovery of multiple shocks that are independent of each other, a key requirement for structural shocks. Second, the method recovers not only the shocks but also their associated impulse responses, which, by design, have different shapes (i.e., they are not collinear). The shapes of the impulse responses may provide valuable information for assigning structural interpretations to the extracted shocks.

For instance, if the analysis involves expectations for both price and quantity, and the method identifies two shocks - one driving price and quantity in the same direction and the other driving them in opposite directions - then the former can be interpreted as a demand shock and the latter as a supply shock.

Our methodological contribution is the development of a procedure for PCA in factor models with time-varying loadings and a small cross-sectional dimension. Existing approaches (Motta et al., 2011 and Su and Wang, 2017) assume a large cross section, but it is well-documented that PCA performs poorly in small samples due to heteroskedasticity in idiosyncratic errors (e.g., Bai and Wang, 2016). We address the problem by considering a finite-sample approach to PCA that accommodates heteroskedasticity from the matrix completion literature in statistics (Zhang et al., 2022) and adapting it to the context of time-varying loadings. We refer to this method as time-varying Heteroskedastic PCA (tvHPCA). As with all factor models, the extracted shocks and impulse responses are not separately identified. We address this challenge by normalizing either the shocks or the impulse responses, depending on which of the two objects is the primary interest of the analysis.

The intuition behind tvHPCA is simple: PCA under heteroskedasticity in small samples leads to discrepancies between the estimated eigenvectors of the sample covariance matrix and the underlying factors. Under the assumption of cross-sectionally uncorrelated idiosyncratic errors, the problem is with the diagonal of the sample covariance matrix. The solution is to iteratively substitute this diagonal with that of the low-rank approximation of the sample covariance matrix. Our tvHPCA method is easy to implement by applying the algorithm in Zhang et al. (2022) to a nonparametrically estimated local (in time) covariance matrix of the expectation revisions.

The tvHPCA method is based on three key assumptions. First, as noted in the previous paragraph, we assume that idiosyncratic errors in the factor model for expectation revisions are uncorrelated across horizons. Although some theories suggest a factor structure without these errors, practical issues - such as variations in survey participation and data merging - can introduce measurement error. In such cases, errors are plausibly uncorrelated.[4]

[4]Strong correlations would be captured by common factors, and if weak correlation is suspected,

The second key assumption is the continuity of time-varying impulse responses and a stable number of shocks over time, which facilitate the interpretation of shocks. Our algorithm ensures continuity and determines the number of shocks using a local version of the method by Onatski (2010), which also serves as validation of the stability assumption. Our modeling of time variation is common in the nonparametric literature and is based on the assumption of local stationarity. This assumes that expectation revisions are approximately stationary over short time intervals, thus ruling out immediate changes due to, e.g., new information or policy shocks. Using two-sided kernels for nonparametric estimation potentially raises concerns related to the Lucas critique, as shocks may change agents' behavior and affect local estimates. Our assumption of local stationarity accommodates shock-induced behavioral changes, provided these changes manifest slowly over time, with short bandwidths helping to guard against contamination.[5] While the horizon dimension of expectations datasets is small, the typically large time dimension allows us to estimate time variation in impulse responses nonparametrically. If the time dimension is also small, the method can still be applied by assuming time-invariant impulse responses, reducing it to the algorithm of Zhang et al. (2022).

The third key assumption is a standard one in the matrix completion literature, referred to as "incoherence", which is similar to the assumption of strong factors. We investigate the implications of this assumption through simulations and identify the worst-case scenario - calibrated to our empirical application - as one characterized by a low signal-to-noise ratio (a measure that can be calculated in applications) alongside a large proportion of zero impulse responses and rapidly decaying remaining responses. In this context, although our ability to recover shocks and the average bias of the estimated impulse responses remain largely unaffected, we do observe an increase in bias for certain impulse response estimates.

increasing the number of horizons by merging datasets can help, as PCA is robust to weakly correlated idiosyncratic errors in large cross-sections.

[5]While one-sided kernels could mitigate concerns about possible violations of local stationarity, they introduce boundary biases. An informal diagnostic check for potential violations of local stationarity could involve comparing results from two one-sided kernels, similar to techniques used in regression discontinuity analysis. However, a formal investigation of this approach is beyond the scope of this paper.

One of the contributions of Zhang et al. (2022) is the demonstration that HPCA not only performs well in simulations but also exhibits certain theoretical optimality properties. However, these theoretical results depend on the assumptions of time-invariant loadings and serially independent factors and idiosyncratic errors, which appear overly restrictive in our context. Through a series of simulations, we show that such stringent assumptions are in fact not necessary for the tvHPCA method to perform well in realistic small-sample scenarios, aside from some performance deterioration at the sample boundaries (a common issue in nonparametric estimation).

In certain applications, obtaining confidence intervals for the extracted impulse responses may be of interest, and we outline a procedure for doing so using the bootstrap. However, an important caveat is that inference for PCA is known to perform poorly in small samples (see, e.g., Maldonado and Ruiz, 2021 for a discussion on confidence intervals for factors). Our simulations indicate that bootstrap confidence intervals for impulse responses work well under serially uncorrelated shocks and idiosyncratic errors, though they tend to slightly undercover. The presence of serial correlation in the shocks exacerbates this undercoverage issue. To our knowledge, the literature lacks a proposal that ensures reliable finite-sample coverage of bootstrap confidence intervals for factors and loadings under general conditions. Therefore, until a satisfactory solution is identified, caution is warranted when interpreting confidence intervals for perceived impulse responses to serially correlated shocks.

The paper is organized as follows. Section 2 outlines the methodology, presenting the factor model idea, its link to theories of expectation formation, key assumptions, how to choose the number of shocks, estimation via the tvHPCA algorithm, and bootstrap inference. Section 3 discusses the simulation results, while Section 4 presents the empirical application. Section 5 concludes, and the appendix addresses bandwidth selection for nonparametric estimation.

2 Methodology

2.1 Set-up and notation

The information required by our method is a (balanced) panel of expectation revisions for one variable across a term structure of different horizons and over time. To fix notation, denote by t the frequency at which the expectations are measured. At each time $t = 0, \ldots, T$ we assume we have expectations of a target variable (denoted simply by Y_h) for a term structure of horizons $h = 1, \ldots, H$. We allow for flexibility in terms of the frequencies at which the expectations are produced and in the definition of the target variable. The simplest case is when the expectations and the target variable are based on the same frequency (e.g., monthly expectations of monthly inflation) and Y_h is the variable h months ahead, $Y_h = Y_{t+h}$. However, we can also accommodate mixed frequencies (e.g., monthly expectations of quarterly inflation), nowcasting (e.g., the first horizon is current-quarter inflation and t is a month within the quarter) and unequally spaced horizons (e.g., $h = 1, \ldots, H$ could denote 1-, 2- and 8-quarters ahead inflation). Also, Y_h could be measured differently at different horizons (e.g., one horizon could be one-quarter ahead inflation and another horizon could be 5-year 5-year inflation).

Let $\widehat{Y}_{h|t}$ and $\widehat{Y}_{h|t-1}$ denote the expectations of the same target variable Y_h made at times t and $t-1$, respectively. We denote the expectation revision at time t for the target variable at the h-th horizon as $X_{ht} = \Delta\widehat{Y}_{h|t} = \widehat{Y}_{h|t} - \widehat{Y}_{h|t-1}$.[6] Our panel data is thus given by expectation revisions X_{ht} for a set of horizons $h = 1, ..., H$ and over time periods $t = 1, ..., T$.

Throughout the paper we use the following notation: for a vector v and matrix M, we denote by v' and M' their transposes and $\|\cdot\|$ denotes the matrix spectral norm.

[6] Some care must be applied to the construction of revisions in the mixed frequency case. E.g., if t is April and the first point in the term structure is current quarter inflation, the expectation at $t-1$ is the expectation for 1-quarter ahead inflation made in March.

2.2 The idea: a factor model of expectation revisions

We model the expectation revisions across horizons h and over time t as a time-varying factor model:

$$X_{ht} = \lambda_{ht}' F_t + e_{ht}, \tag{1}$$

for $t = 1, ..., T$, $h = 1, ..., H$, where $F_t = (F_{1t}, F_{2t}, \ldots, F_{rt})'$ is a vector of $r < H$ independent latent factors, λ_{ht} is a vector of factor loadings for the h-th horizon at time t and e_{ht} is an idiosyncratic error with variance $\sigma^2_{h,t}$. The model thus assumes that there are a few common, latent factors that drive most of the comovements in expectation revisions across horizons, with loadings (as well as error variances) that are allowed to vary across horizons and over time.

At each time t, we interpret the independent latent factors F_t as the "perceived shocks" and the corresponding loading λ_{ht} for each horizon as the "perceived impulse response" at that horizon, that is, the effect of the corresponding shock on the target variable at the h-th horizon. A perceived impulse response function at time t is the plot of λ_{ht} as a function of h. Intuitively, the perceived shocks represent the drivers of the expectation revisions at time t that were unanticipated at time $t-1$. As discussed in the introduction, giving the shocks a structural interpretation generally requires additional assumptions. However, we note that the factors are independent of each other, so they satisfy one of the requirements for structural shocks. In addition, in some applications it is possible that the information contained in the extracted impulse responses can help give a structural interpretation to the extracted shocks.

Factor models have been extensively applied in economics and finance, see e.g. Chamberlain (1983), Diebold et al. (2005), Stock and Watson (2016), and their econometric properties have been studied by, e.g. Bai (2003), Stock and Watson (2006), Bai et al. (2008), Choi (2012), Bai and Ng (2019). A key challenge in our context is that the number of horizons H, i.e., the cross-sectional dimension in the factor model in (1), is typically small in the datasets of expectations available to economists. A main issue that arises in this context is the plausible presence of heteroskedasticity in the idiosyncratic errors e_{ht}. Under additional assumptions (see e.g. Bai, 2003, Bai, 2009), consistent estimates of both factors and loadings can

be achieved if H is large. A recent literature in statistics has however highlighted that, when H is finite, heteroskedastic errors can lead to inconsistent estimates of the factors and loadings, see e.g. Florescu and Perkins (2016), Zhang et al. (2022). In this paper we adapt the solution proposed in this literature to the general case of time-varying factor models.

2.3 Relationship with theories of expectation formation

This section shows that the factor structure for expectation revisions is compatible with some alternative theories of expectation formation. For simplicity, we assume that the target variable is the variable at time $t + h$, $Y_h = Y_{t+h}$. Additionally, we assume here that the model parameters are constant over time. However, the examples could be adapted to incorporate time-varying parameters, provided the variation is compatible with the assumption of local stationarity for the revisions, as discussed in Section 2.4.3. For instance, one could allow for parameters that evolve slowly, ensuring that there is effectively no temporal variation between periods $t-1$ and t. It is easy to see that this type of time variation in parameters translates into time varying loadings in the factor representation for the revisions.

2.3.1 Rational expectations and SVMA model

Suppose a representative agent uses a Structural Vector Moving Average (SVMA) model to forecast the target variable Y_{t+h} (See, e.g., Plagborg-Møller (2019), for how a large class of economic models can be represented in this form). The model for the target variable is thus one of the equations of the SVMA:

$$Y_t = \Theta(L)\varepsilon_t,$$

where ε_t is a vector of structural shocks and $\Theta(L)$ is a lag polynomial (for simplicity here assumed to be of order greater than the number of horizons H) whose coefficients represent the structural impulse responses of the target variable to each structural shock at the corresponding horizon.

It is easy to see that in this case the expectation revision between times $t-1$ and t for the target variable at the h-th horizon is given by

$$X_{ht} = \widehat{Y}_{t+h|t} - \widehat{Y}_{t+h|t-1} = \theta_h' \varepsilon_t,$$

where $\widehat{Y}_{t+h|t}$ is the conditional mean implied by the SVMA model. This implies a factor structure for the revisions with no idiosyncratic errors, $X_{ht} = \lambda_{ht}' F_t$, where the factors are the structural shocks, $F_t = \varepsilon_t$, and the loadings are the associated structural impulse responses at horizon h, $\lambda_{ht} = \theta_h$.

We thus see that, if the agent uses a model that can be expressed as a SVMA, our procedure can recover the latent vector of structural shocks and the structural impulse responses. Note that we can in principle recover multiple structural shocks from only observing the expectation revisions for one of the variables in the system (provided the number of horizons is larger than the number of shocks).

2.3.2 Rational expectations and factor model

Factor models are frequently employed to model and forecast macroeconomic and financial variables (see, e.g., Stock and Watson, 2002). For instance, interest rates are typically represented using a factor model that captures the joint dynamics of rates across various maturities (Diebold and Li, 2006). A factor model for the target variable suggests a corresponding factor structure for the expectation revisions. For example, consider the dynamic factor model

$$\begin{aligned} Y_t &= \gamma' \beta_t + v_t \\ \beta_t &= \Phi \beta_{t-1} + \varepsilon_t. \end{aligned} \tag{2}$$

The dynamic Nelson and Siegel model (see, e.g., Diebold and Li, 2006) for a specific interest rate maturity Y_t could be written in this form, with three latent factors in the vector β_t and a specific parameterization for γ that depends on the maturity. Stock and Watson (2007)'s model of inflation is also a special case of (2), with β_t

scalar and $\gamma = \Phi = 1$. The expectation for the target variable at horizon h is

$$Y_{t+h|t} = \gamma' \Phi^h \beta_t,$$

which means that the revision of the expectation for Y_{t+h} made between times $t-1$ and t is given by

$$X_{ht} = \gamma' \Phi^h (\beta_t - \Phi \beta_{t-1}) = \gamma' \Phi^h \varepsilon_t.$$

We thus once again obtain a factor structure for the revisions with no idiosyncratic errors, $X_{ht} = \lambda_{ht}' F_t$, where the factors $F_t = \varepsilon_t$ coincide with the shocks in the state equation specifying the law of motion for β_t and the loadings $\lambda_{ht}' = \gamma' \Phi^h$ are the impulse responses implied by the model. In this setting, therefore, our method can recover the true impulse responses and can also back out the number of dynamic factors in agents' models by only observing how agents revise expectations for one variable (e.g., one interest rate maturity) across different horizons.

2.3.3 Information rigidities

To see how existing theories of expectation formation with information rigidities could give rise to a factor structure for expectation revisions, consider, e.g., the noisy-information model in Coibion and Gorodnichenko (2015), where the target variable follows an AR(1) process

$$Y_t = \rho Y_{t-1} + \varepsilon_t,$$

with ε_t Gaussian white noise. Here the (single) true shock is ε_t and the true impulse response at horizon h is given by ρ^h. The theory assumes that a representative agent observes a noisy signal of Y_t,

$$Z_t = Y_t + v_t,$$

with v_t a Gaussian white noise process independent of ε_t. The agent forecasts using the Kalman filter, so the expectation for the h-th horizon made at time t is

$$\widehat{Y}_{t+h|t} = \rho^h \widehat{Y}_{t|t},$$

where

$$\widehat{Y}_{t|t} = GZ_t + (1-G)\widehat{Y}_{t|t-1}$$

and G is the Kalman gain, which captures the degree of information rigidity. The revision is then given by

$$X_{ht} = \widehat{Y}_{t+h|t} - \widehat{Y}_{t+h|t-1} = \rho^h(\widehat{Y}_{t|t} - \widehat{Y}_{t|t-1}).$$

We thus again obtain a factor structure for the expectation revisions with no idiosyncratic errors, $X_{ht} = \lambda_{ht} F_t$, where the loadings $\lambda_{ht} = \rho^h$ correspond to the true impulse responses and the factor is the "filtered shock" $F_t = \widehat{Y}_{t|t} - \widehat{Y}_{t|t-1} = G(Z_t - \widehat{Y}_{t|t-1})$, that is, the surprise from the Kalman filter updating equation that the agent uses to extract the signal from the noise.

2.4 Assumptions

The model for the expectation revisions in (1) can be written in vector notation as:

$$\underset{H\times 1}{X_t} = \underset{H\times r}{\Lambda_t}\ \underset{r\times 1}{F_t} + \underset{H\times 1}{e_t}\ , \tag{3}$$

where $\Lambda_t = (\lambda_{1t}, \ldots, \lambda_{Ht})'$ is the matrix of loadings (i.e., the perceived impulse responses), F_t is the vector of common factors (i.e., the perceived shocks) and e_t is the vector of idiosyncratic errors. Our method is based on *local* (in time) PCA, and thus the objects of interest are *local* covariance matrices, defined as follows:

$$\Sigma_t = E\left[X_t X_t'\right], \qquad \Sigma_{F,t} = E\left[F_t F_t'\right], \qquad \Sigma_{e,t} = E\left[e_t e_t'\right]. \tag{4}$$

Our method relies on the following assumptions, which we group into different

types to facilitate the discussion of their practical implications.

2.4.1 Shocks and idiosyncratic errors

Assumption 1 (Shocks and idiosyncratic errors):

(a) The shocks F_t have unconditional mean zero.

(b) The shocks F_t are cross-sectionally independent but can be serially correlated (stationary). Moreover, $cov(e_{ht}, F_{jt-k}) = 0$ for any h, j, t and k.

(c) The errors e_t can be serially correlated (stationary), and $e_t \sim (0, \Sigma_{e,t})$, where $\Sigma_{e,t}$ is a diagonal matrix with uniformly bounded eigenvalues for all $t = 1, \ldots, T$.

Comments. Assumption 1(a) requires shocks to have unconditional mean zero. This seems plausible, given that we are dealing with expectation revisions and a non-zero mean would indicate forecast bias, which is unlikely to be present in the expectations data that economists typically analyze.[7] Assumption 1(b) is the standard PCA assumption that assumes the factors to be independent of each other and of the errors. Note that we don't require serial independence of shocks and errors. While the examples in Section 2.3 imply serially uncorrelated shocks, it is possible that in practice extracted shocks present some serial correlation. We show in simulations that serial correlation in either the shocks or the idiosyncratic errors has no effect on the performance of the tvHPCA method (in terms of bias of estimated impulse responses, correlation between true and estimated shocks and residual mean squared error). The key requirement of Assumption 1 is assumption 1(c) that the errors are uncorrelated across horizons. We note that it is possible in principle to allow for a small number of nonzero off-diagonal entries of $\Sigma_{e,t}$, but we do not further investigate this possibility in the paper. The plausibility of this assumption should be assessed in any given application. While the examples in Section 2.3 imply a factor structure

[7] A non-zero, time-invariant mean can be accommodated by extracting shocks from demeaned X_{ht} and then recovering the nonzero-mean shocks by adding back the impulse-response-weighted average of the X_{ht} means.

for the revisions without idiosyncratic errors, errors can arise for different reasons: changing composition of survey respondents across time and horizons (when considering consensus expectations); rounding of expectations; considering different types of expectations or merging different datasets. All these examples of idiosyncratic errors are plausibly uncorrelated across horizons. Weak correlation in the errors across horizons is less of a concern if the number of available horizons H is large, since it is well-known that in this case PCA is robust, see e.g., Bai and Wang (2016).

Practical Implications. In practical terms, Assumption 1 implies that, although serial correlation in shocks and errors does not impede the methodology's ability to recover shocks and impulse responses, the assumption of uncorrelated idiosyncratic errors across horizons is important when the number of horizons is small. In scenarios where there are concerns regarding weak cross-sectional correlation among idiosyncratic errors, one potential remedy is to increase the number of horizons. This can be achieved, for instance, by carefully merging different expectations datasets, as demonstrated in our empirical application. Any strong correlation among idiosyncratic errors will ultimately be absorbed by the extracted shocks.

2.4.2 Normalizations

As in standard factor models, there is a rotational indeterminacy in the identification of shocks and impulse responses.[8] This can be resolved by imposing a normalization either on the shocks or the impulse responses, depending on the objectives of the analysis. We consider the following examples of normalizations.

Assumption 2 (Normalizations):

(a) *Unit-local shock variance normalization:* Set $\Sigma_{F,t} = I_r$, where I_r is the identity matrix.

(b) *Unit-effect normalization:* Set $\Lambda_t'\Lambda_t = I_r$, such that the shocks have unit effect on all horizons.

[8]For any $r \times r$ nonsingular matrix A_t it follows that $\lambda_{ht}'F_t = (A_t^{-1}\lambda_{ht})'(A_t'F_t)$ and therefore shocks and impulse responses are not separately identified.

(c) *Unit-impact normalization:* Set $\Lambda_t[1, h] = 1$ for $h = 1, \cdots, H$, such that the shocks have unit effect only on the first horizon (i.e., on impact).

Practical Implications. The practical implication of Assumption 2 is that one must choose whether to normalize the shocks or the impulse responses based on the primary focus of the analysis. If the emphasis is on the impulse responses, the shocks can be normalized to have a unit local variance (Assumption 2(a)), allowing the impulse responses to reflect the dynamic effects of a one-standard-deviation shock. Conversely, if the goal is to recover the shocks, the impulse responses can be normalized so that the shock magnitudes are interpretable (Assumption 2(b)). Alternatively, one can impose a unit-impact normalization on the impulse responses (Assumption 2(c)), which rescales the shocks to produce a unit effect on the first horizon at each point in time. This normalization facilitates comparisons of impulse response functions over time.

2.4.3 Time-varying impulse responses and nonparametric estimation

The time-varying impulse responses are assumed to be deterministic functions of time:

$$\lambda_t = \lambda(t/T) \qquad \text{and} \qquad \Lambda_t = \Lambda(t/T). \tag{5}$$

Such rescaling is common in nonparametric estimation (see e.g. Robinson, 1989, Cai, 2007). The idea is that, as the number of observations increases in the rescaled time framework, we "observe" the process on an increasingly dense grid on the unit interval, and letting $T \to \infty$ allows for the impulse responses to be consistently estimated under infill asymptotics, see e.g. Motta et al. (2011), Su and Wang (2017). Note that, since the λ_{ht} in (1) is time-varying, X_{ht} is no longer stationary, but locally stationary in the sense of Dahlhaus et al. (2019), i.e., behaving in an approximately stationary manner within short time periods. With the new notation, the locally stationary model reads:

$$X_{t,T} = \Lambda(t/T)\, F_t + e_t, \tag{6}$$

where time variation in impulse responses gives rise to a triangular array $X_{t,T}$. This framework allows us to analyze the dynamics of $X_{t,T}$ locally by using a stationary approximation[9] given by:

$$X_t(u) = \Lambda(u)\, F_t + e_t, \tag{7}$$

where $X_t(u)$ is a locally stationary equivalent of $X_{t,T}$ and where $\Lambda(t/T) \approx \Lambda(u)$. Note that $X_t(u)$ is not observed in practice but is used as a theoretical construct. For ease of exposition we henceforth drop the double subscript and simply write X_t.

Our procedure is based on the following nonparametric estimator of the local covariance matrix of X_t:

$$\widehat{\Sigma}(u) = T^{-1} \sum_{t=1}^{T} K_b(t/T - u) X_t X_t', \tag{8}$$

where $K_b(\cdot) = K(\cdot/b)/b$ is a kernel function, with bandwidth b acting as a choice parameter that determines the amount of data to be used in estimation.

We make the following assumptions.

Assumption 3 (Time-varying impulse responses and kernel estimation):

(a) $\lambda_{ht} = \lambda_h(t/T)$, $t = 1, \ldots, T$, where for $h = 1, \ldots, H$, $\lambda_h(\cdot) : [0,1] \to \mathbb{R}$ is an unknown piece-wise continuous function of the rescaled time t/T. $\lambda_h(\cdot) : [0,1] \to \mathbb{R}$ is twice continuously differentiable for any $h = 1, \ldots, H$.

(b) For all $u \in (0,1)$ it holds that $rank(\Lambda(u)) = r$, where $\Lambda(u)$ is defined in eq.(7).

(c) The kernel function $K : \mathbb{R} \to \mathbb{R}^+$ is a symmetric continuously differentiable probability density function with compact support $[-1,1]$ normalized such that $\int K(z)dz = 1$.

(d) As $T \to \infty$ $b \to 0$, such that $Tb \to \infty$.

[9] The idea is to approximate $X_{t,T}$ by $X_{t,T} = X_t(t/T) \approx X_t(u)$, where $X_t(u)$ is a stationary equivalent of $X_{t,T}$.

Comments. Assumption 3 is similar to the assumptions in Su and Wang (2017). Assumption 3(a) requires the time-varying impulse responses to be smooth (i.e., continuously differentiable) functions of time. Assumption 3(b) states that the number of shocks is fixed across time. The assumption could be relaxed, but this would come at the cost of losing interpretability and the ability to label the shocks across time. Assumption 3(c) is the standard assumption in the nonparametric literature, applied here to estimation of the local covariance matrix, requiring the chosen kernel to be a symmetric probability density function. In the paper we use the Epanechnikov kernel $K(x) = 0.75(1 - x^2)\mathbb{1}\{|x| \leq 1\}$, rescaled when necessary to ensure consistency even on the boundary points, where only the data on one side are available.[10] Assumption 3(d) states typical conditions on the bandwidth b that ensure that the local covariance matrix can be consistently estimated locally in time, see e.g. Motta et al. (2011) and Su and Wang (2017).

Practical implications. Assumption 3 has several practical implications. First, to evaluate the validity of the assumption of a constant number of shocks over time, we recommend applying the approach proposed by Onatski (2010) locally in time (detailed in section 2.5). Second, since shocks and impulse responses are identified only up to a sign,[11] ensuring continuity in time of the impulse responses necessitates additional methodological steps. The algorithm presented in the following section addresses this by employing two sequential sub-algorithms: Algorithm A extracts shocks and time-varying impulse responses, while Algorithm B further ensures continuity of the impulse responses. Third, the time dimension T must be sufficiently large to recover time-varying impulse responses. In cases where T is small, the method can still be applied under the assumption of time-invariant impulse responses, reducing to the algorithm in Zhang et al. (2022). Finally, the continuity of impulse responses and the assumption of a stable number of shocks are crucial for enabling structural interpretations of the extracted shocks through the lens of economic theory. For

[10] This is equivalent to applying the boundary kernel for the boundary regions, see e.g. Li and Racine (2023) and is the same as the boundary kernel applied in Su and Wang (2017). The rescaling is necessary to achieve consistency of the corresponding estimates on the boundary points by ensuring that the first moment of the kernel is always normalized to 1.

[11] The normalization to address scale is managed by Assumption 2.

instance, as discussed in the introduction, in a context with expectation revisions regarding price and quantity, a shock consistently yielding positive impulse responses for both price and quantity may be interpreted as a demand shock, whereas a shock that shows positive responses for quantity and negative responses for price could be viewed as a supply shock. If shocks were permitted to disappear and re-emerge or exhibited abrupt changes in impulse responses, it would be challenging to discern whether these represent the same shock or a new one.

2.4.4 Local incoherence

The final assumption is known as the incoherence condition, a standard concept in the matrix completion literature (see e.g., Candès and Recht, 2009, Zhang et al., 2022). This condition ensures that the information contained in the row and column spaces of the covariance matrix is not concentrated in too few rows or columns. This is important because it enables the original high-dimensional matrix to be approximated by a lower-dimensional one.

Given the time-varying nature of our approach, we need to apply this incoherence condition locally, i.e., a "local incoherence" assumption.

Assumption 4 (Local Incoherence): There exists a constant c such that the following holds for $t = 1, \ldots, T$:

$$\max_{1 \leq h \leq H} \|\nu_h \Lambda_t (\Lambda_t' \Lambda_t)^{-1/2}\|_2^2 \leq c, \tag{9}$$

where ν_h denote the h-th standard basis of appropriate dimension with h-th coordinate equal to 1 and other coordinates being 0s and, for a vector v, $\|v\|_2$ denotes its l_2-norm.

Comments. Condition in (9) is a slightly modified version of the incoherence condition found in e.e.g Candès and Recht (2009), Zhang et al. (2022) to ensure it holds under all normalizations stated in our Assumption 2. Specifically, we re-normalize the loadings such that they are orthonormal regardless of the chosen normalization. While Assumption 4 is high-level, it relates to the more familiar concept

of "strong factors" (see, e.g., the discussion in Agarwal et al., 2023) and to the "pervasiveness" assumption found in the economics literature, see e.g. Onatski (2012), and Fan et al. (2013). These assumptions ensure that the common component can be discerned from the idiosyncratic component of the sample variance. For example, under the unit-local shock variance normalization in Assumption 2(a), the condition corresponds to the pervasiveness assumption $\psi_{\min}(\Lambda_t'\Lambda_t/H) > c > 0$, where $\psi_{\min}(M)$ denotes the smallest eigenvalue of matrix M.

Practical implications. In practice, Assumption 4 means we need to be wary of "weak factors" - scenarios where many impulse responses are zero or all responses are close to zero. This may occur if expectations are infrequently revised, or if the bandwidth for the nonparametric estimation of the local covariance matrix is too small to capture sufficient variability in the data. We examine and provide practical guidance on how to diagnose potential violations of this assumption and their possible effects in one of our simulations in Section 3. There we identify the worst-case scenario - tailored to our empirical application - as one marked by a low signal-to-noise ratio (a measure that can be computed in applications), along with a high proportion of zero impulse responses and quickly decaying remaining responses.

2.5 Choosing the number of shocks

Before implementing the algorithm described in section 2.6, it is necessary to determine the number of shocks, which we assume to be constant over time (see Assumption 3(b)). We recommend applying a local version of the procedure proposed by Onatski (2010), both as a guide for selecting the number of shocks and as a verification of Assumption 3(b). Although we do not formally establish the theoretical validity of this approach, we present simulation evidence in section 3 suggesting that Onatski (2010)'s test is capable of recovering multiple shocks, even when the number of horizons is very small.

This approach is grounded in the observation that the largest "idiosyncratic" eigenvalues of the sample covariance matrix tend to cluster around a single point, while the "systematic" eigenvalues - corresponding to the number of shocks - diverge

to infinity. An estimator for the number of shocks is derived from the differences between consecutive eigenvalues. By applying this approach locally, we can ascertain the number of shocks at a specific point in time, with the overall number of shocks determined by the maximum number of local shocks observed across time periods.

The procedure is outlined as follows:

1. Calculate the local eigenvalues from the eigendecomposition of the local covariance matrix, $\psi_{1,t}, \cdots, \psi_{H,t}$,

2. Select r_{max} as a preliminary maximum number of shocks we are interested in testing for (2 in our empirical application),

3. Setting $j = r_{max}+1$, regress $\psi_{j,t}, \cdots, \psi_{j+4,t}$ on a constant and $(j-1)^{2/3}, \cdots, (j+3)^{2/3}$,

4. Set $\delta_t = 2|\hat{\beta}_t|$, where $\hat{\beta}_t$ is the slope coefficient from the above regression,

5. Compute $r(\delta_t) = \max\{i \leq r_{max} : \psi_{i,t} - \psi_{i+1,t} \geq \delta_t\}$, and $r(\delta_t) = 0$ if $\psi_{i,t} - \psi_{i+1,t} < \delta_t$ for all $i \leq r_{max}$

6. If $r_{max} \neq r(\delta_t)$, set $j = r(\delta_t) + 1$ and repeat from step 2 onward, otherwise select $r_t = r(\delta_t)$ as the number of local shocks,

7. Set $r = \max_t\{r_t\}$ to determine the number of shocks in our data across time.

2.6 Estimation of shocks and impulse responses via tvHPCA

This section outlines the estimation of shocks, time-varying impulse responses, and idiosyncratic variances for a specified number of shocks, r. To understand the small-sample challenges we face, consider the tvPCA method discussed by Motta et al. (2011) and Su and Wang (2017). This approach applies PCA to the estimated local covariance matrix rather than the global one. While heteroskedastic errors are not a problem when both the time dimension T and the horizon H are large, inconsistency can arise under heteroskedasticity when H is fixed and small, as highlighted for PCA by Paul (2007), Johnstone and Lu (2009), and Onatski (2012).

To get an intuition for why this happens, note that, under the normalization $\Sigma_{F,t} = I_r$, we have:

$$\Sigma_t = E(X_t X_t') = \Lambda_t \Lambda_t' + \Sigma_{e,t}, \tag{10}$$

where $\Sigma_{e,t}$ exhibits heteroskedasticity. When $H \to \infty$ and the factor is pervasive (i.e. $\Lambda_t \Lambda_t' \to \infty$), then $\Lambda_t \Lambda_t'$ becomes the dominant term in (10). This ensures that the principal eigenvector and eigenvalue of Σ_t are close to those of $\Lambda_t \Lambda_t'$. If H is fixed, however, the first principal component eigenvector will put relatively large weight on the component of X_t with the largest idiosyncratic variance. Such an eigenvector, however, might be unrelated to any of the columns of Λ_t, which is necessary for identification of factors and loadings.

Assuming that $\Sigma_{e,t}$ is diagonal, heteroskedasticity introduces bias into the diagonal elements of the sample covariance matrix. A common approach to address this issue is to use a diagonal-deletion Singular Value Decomposition (SVD), which involves setting the diagonal elements of the sample covariance matrix to zero before applying the SVD (see e.g. Florescu and Perkins (2016)). However, Zhang et al. (2022) note that this approach can fundamentally alter the singular subspace which determines the factors and their loadings, distancing it from the singular subspace corresponding to the true factors and loadings. Zhang et al. (2022) propose an algorithm, HeteroPCA, for estimating the factor model in the presence of heteroskedasticity when H and T are fixed. The idea is to iteratively impute the diagonal entries of the sample covariance matrix by the diagonals of its low-rank approximation.

The following algorithm extends the procedure in Zhang et al. (2022) to allow for time-varying loadings (i.e., impulse responses).

The tvHPCA algorithm:

The following algorithm delivers estimates of the shocks $\widehat{F}_t$, the time-varying impulse responses $\widehat{\Lambda}_t$, the errors $\widehat{e}_t = X_t - \widehat{\Lambda}_t \widehat{F}_t$, and the diagonal matrix of time-varying idiosyncratic variances $\widehat{\Sigma}_{e,t}$. Define the operator $\Delta(\Sigma(u)) = \Sigma(u) - D(\Sigma(u))$, where $D(\cdot)$ denotes the diagonal operator that returns the matrix containing only the diagonal entries. For a chosen number of shocks r and for each $t = 1, \ldots, T$, perform the following steps:

A. Preliminary estimation of shocks and impulse responses

Step A1: Calculate the local covariance matrix $\widehat{\Sigma}(t/T)$ according to (8). Set maximum number of iterations M.

Step A2: Initialize the local tvHPCA algorithm at $m = 0$ by setting the diagonal entries of $\widehat{\Sigma}(t/T)$ to zero,

$$N_t^{(0)} := \Delta(\hat{\Sigma}(t/T)), \qquad m = 0.$$

Step A3: Perform Singular Value Decomposition (SVD) on $N_t^{(m)}$ and denote its rank-r approximation by $\widetilde{N}_t^{(m)}$. Specifically, for $H \geq r$:

$$N_t^{(m)} = \sum_{i=1}^{H} \psi_{it}^{(m)} u_{it}^{(m)} (v_{it}^m)', \qquad \psi_1^{(m)} \geq \cdots \geq \psi_H^{(m)} \geq 0,$$

$$\widetilde{N}_t^{(m)} = \sum_{i=1}^{r} \psi_{it}^{(m)} u_{it}^{(m)} (v_{it}^m)', \tag{11}$$

where $\psi_{it}^{(m)}$ is the i-th largest singular value (i.e., the square root of the eigenvalue) of $N_t^{(m)\prime} N_t^{(m)}$, and $u_{it}^{(m)}$ and $v_{it}^{(m)}$ are the eigenvectors of $N_t^{(m)} N_t^{(m)\prime}$ and $N_t^{(m)\prime} N_t^{(m)}$, respectively.

Step A4: Update $N_t^{(m+1)} = D(\widetilde{N}_t^{(m)}) + \Delta(N_t^{(m)})$, that is, replace the diagonal entries of $N_t^{(m)}$ by those of $\widetilde{N}_t^{(m)}$:

$$N_{hj,t}^{(m+1)} = \begin{cases} N_{hj,t}^{(m)} = \widetilde{N}_{hj,t}^{(m)}, & h = j; \\ \widehat{\Sigma}_{hj}(t/T), & h \neq j. \end{cases}$$

Step A5: Calculate the convergence distance as the maximum change in eigenvalues across horizons and time. Stop if the convergence distance is less than a predefined threshold (we select 10^{-3}),[12] or if $m = M$, otherwise set

[12] A smaller threshold choice will typically lead to more accurate convergence at the cost of slower

$m = m + 1$ and return to Step A3 and continue to iterate.

Step A6: Upon convergence, set $\widetilde{\lambda}_{it} = u_{it}^{(m)}$ for $i = 1, \ldots, r$ as the estimated normalized impulse responses, which for a given t are identified up to the scale and sign. The estimated common covariance matrix is given by $N_t^{(m+1)}$. The diagonal error covariance matrix estimator is:

$$\widehat{\Sigma}_{e,t} = \widehat{\Sigma}(t/T) - N_t^{(m+1)}.$$

Algorithm A above can be seen as an extension of the original HeteroPCA of Zhang et al. (2022) to account for time-variation in the factor loadings, and it can further be interpreted as the projection gradient descent (PGD) for the following rank-constrained (nonconvex) optimization problem:

$$\min_{rank(\widetilde{N}_t) \leq r} \|\Delta(\widehat{\Sigma}_t - \widetilde{N}_t)\|_F^2, \tag{12}$$

where for a matrix M we write $\|M\|_F = (\sum_{h,j} M_{hj}^2)^{1/2}$ to denote its Frobenius norm and $\widehat{\Sigma}_t$ and $\widetilde{N}_t$ are defined in (8) and (11) respectively.[13] The algorithm requires choosing a bandwidth b for the estimation of the local sample covariance matrix. Given that the estimation is done via local singular value decomposition followed by several steps of refinements, the bias-variance trade-off necessary to discuss a theoretically optimal bandwidth is non-standard. We therefore follow Su and Wang (2017) and use a data-driven way of selecting an optimal bandwidth by cross-validation, which we describe in the appendix.

Next, to unify the identification of the sign of the impulse responses *across time*, we leverage the fact that the impulse responses are continuous in time. Without loss of generality, we therefore make an additional assumption that the impulse responses are on average positive across time[14], which identifies the entire path of the impulse

compute times. Our choice of 10^{-3} can be adjusted depending on the application. In our experience a smaller threshold did not change our estimates but resulted in significantly longer compute times.

[13] All existing convergence results for PGD do not apply to nonconvex optimization problems such as the one in (12), while Zhang et al. (2022) provide theoretical guarantees for their algorithm.

[14] It is also possible to fix the sign of the shocks's correlation with some external variable instead.

responses for each shock up to sign. Specifically, for $t = 2, \ldots, T$, we perform the following steps:

B. Ensuring continuity of impulse responses.

Step B1: For a given shock $k = 1 \ldots, r$, compare the estimated impulse responses at time $t-1$ from step A and assign the sign according to the condition below that ensures continuity across time:

$$\text{sign}(\widetilde{\lambda}_{k,t}) = \begin{cases} \text{sign}(\widetilde{\lambda}_{k,t-1}) & \text{if } ||\widetilde{\lambda}_{k,t} - \widetilde{\lambda}_{k,t-1}|| \leq ||\widetilde{\lambda}_{k,t} + \widetilde{\lambda}_{k,t-1}|| \\ -\text{sign}(\widetilde{\lambda}_{k,t-1}) & \text{otherwise} \end{cases} \tag{13}$$

where $||\cdot||$ is the L^2 norm.

Step B2: Estimate the latent shocks associated with $\widetilde{\Lambda}_t = (\widetilde{\lambda}_{1t}, \ldots, \widetilde{\lambda}_{Ht})'$ by least squares as

$$\widetilde{F}_t = \widetilde{\Lambda}_t \widetilde{\Lambda}_t' (\widetilde{\Lambda}_t X_t)^{-1}, \tag{14}$$

which will not generally have unit variance per shock across time.

Step B3: For $k = 1, \ldots, r$, estimate the time-varying variance of the k-th shock using a local constant kernel regression[15]:

$$\widetilde{F}_{k,t}^2 = \widehat{\mu}_k(t/T) + w_{k,t}, \tag{15}$$

to obtain the estimated standard deviation of shock k at time t as

$$\widehat{\sigma}_{k,t} = \max(\sqrt{\widehat{\mu}_k(t/T)}, 10^{-6}). \tag{16}$$

Step B4(a): The normalization in Assumption 2(a) is imposed by scaling the shocks and the associated impulse responses as:

$$\widehat{F}_t = D(\widehat{\sigma}_t)^{-1} \widetilde{F}_t, \qquad \widehat{\Lambda}_t = D(\widehat{\sigma}_t) \widetilde{\Lambda}_t, \tag{17}$$

We therefore suggest that the choice should be driven by the application at hand.

[15] For simplicity, we use the same bandwidth we used for local estimation of the sample covariance matrix in (8).

where $D(\widehat{\sigma}_t)$ is the diagonal matrix of the estimated shock standard deviations, obtained from Step B3.

Step B4(b): The normalization in Assumption 2(b) is imposed by setting:

$$\widehat{F}_t = \widetilde{F}_t, \qquad \widehat{\Lambda}_t = \widetilde{\Lambda}_t \tag{18}$$

as the impulse responses are eigenvectors that already have unit L^2- norm.

Step B4(c): The normalization in Assumption 2(c) is imposed by setting:

$$\widehat{F}_t = D(\widetilde{\Lambda}_t[1,:])\widetilde{F}_t, \qquad \widehat{\Lambda}_t = D(\widetilde{\Lambda}_t[1,:])^{-1}\widetilde{\Lambda}_t, \tag{19}$$

where $D(\widetilde{\Lambda}_t[1,:])$ is the first row of Λ_t as a diagonal matrix.

2.7 Confidence intervals for impulse responses

Confidence intervals for impulse responses can be obtained using the bootstrap. The following is a schematic description of a residual-based wild bootstrap procedure that takes into account the uncertainty in both shocks and impulse responses. The algorithm assumes that the shocks and the idiosyncratic errors of the factor model are serially uncorrelated.

1. Apply tvHPCA on the original revisions data to obtain the shocks $\widehat{F}_t$, the time-varying impulse responses $\widehat{\Lambda}_t$, the idiosyncratic errors $\widehat{e}_t = X_t - \widehat{\Lambda}_t\widehat{F}_t$ and their diagonal covariance matrix $\widehat{\Sigma}_{e,t}$.

2. Calculate the standardized errors $\widehat{\varepsilon}_t = \widehat{\Sigma}_{e,t}^{-1/2}\widehat{e}_t$.

3. Generate B samples of time points randomly with replacement for each $t = 1, ..., T$ to obtain: $\{t_b\}_{b=1}^{B}$.

4. Use the resampled time points to generate B samples of bootstrapped standardized errors $\widehat{\varepsilon}_t$ and shocks $\widehat{F}_t$ to obtain: $\{\varepsilon_t^b\}_{b=1}^{B}$, and $\{F_t^b\}_{b=1}^{B}$.

5. Combine the bootstrapped standardized errors with the estimated impulse responses, the bootstrapped shocks, and the estimated volatilities, to create B bootstrap samples of data: $\{X_t^b = \widehat{\Lambda}_t F_t^b + \widehat{\Sigma}_{e,t}^{1/2} \varepsilon_t^b\}_{b=1}^B$.

6. Apply tvHPCA on each of the B bootstrap samples and retain the impulse responses estimates: $\{\widehat{\Lambda}_t^b\}_{b=1}^B$.

7. Calculate boostrap confidence intervals for the impulse responses using their empirical distribution across bootstrap samples.

3 Simulations

This section analyses the performance of our method in a similar setting as our empirical application. We first consider time-invariant impulse responses and show that: 1) HPCA corrects the poor performance of PCA under heteroskedasticity; 2) HPCA is robust to serial correlation in shocks or idiosyncractic errors; 3) weak factors have a lesser impact on the ability to recover shocks and on the average bias of the estimated impulse responses, but result in increased bias for some impulse response estimates; 4) the method can in principle recover more than one shock even though the number of horizons is small. Second, we show that bootstrap confidence intervals for impulse responses tend to undercover in small samples, the more so the higher the serial correlation in the shocks. Third, we focus on time-varying impulse-responses and show that our method can recover the time variation, with some deterioration in performance at the boundaries of the sample. All simulations below are based on 1000 Monte Carlo replications.

3.1 HPCA vs. PCA

We generate data from the factor model (1), with $r = 1, H = 7, T = 100$, equal impulse responses across horizons $\lambda_h = 1$ for all h, shock $F_t \sim i.i.d.N(0,1)$ and different variance for the first idiosyncratic error $\sigma_1 = 1$, $\sigma_h = 0.5$ for $h = 2, ..., 7$. Since in this design heteroskedasticity mainly affects estimation of the impulse response for

the first horizon, we only present results for λ_1. Figures 1 and 2 show that HPCA improves on the performance on PCA. Figure 1 reports the estimates of λ_1 across the simulations, which show that HPCA corrects the bias of PCA in estimation of the impulse response.

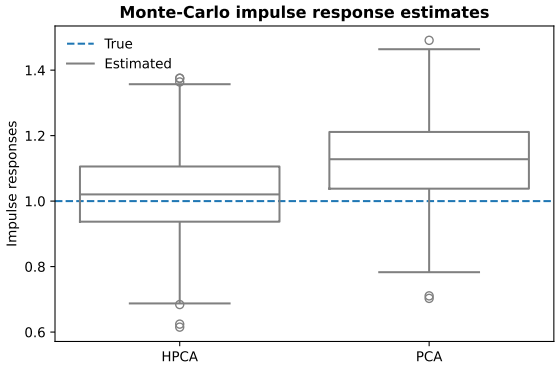

Figure 1: First impulse response estimates: HPCA vs. PCA.

The left panel of Figure 2 shows for both HPCA and PCA the distribution of normalized residual MSEs[16] across simulations, while the right panel shows the distribution of correlations between the estimated shock and the true shock across simulations. The figure shows that HPCA dominates PCA, by yielding lower average MSE and higher average correlation between estimated and true shock.

[16]The normalized residual Mean Squared Error is obtained by summing the squares of the in-sample residuals of the HPCA/PCA model divided by the true idiosyncratic volatility.

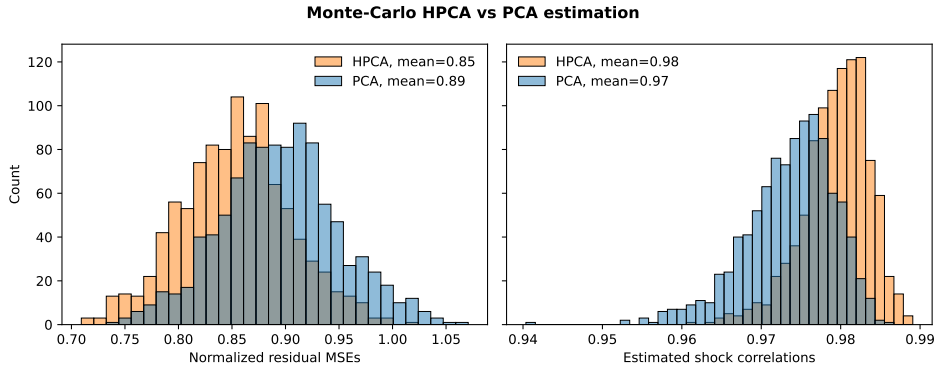

Figure 2: Residual MSEs and estimated/true shock correlation: HPCA vs. PCA.

3.2 Robustness to serial correlation

Here we modify the simulation design in section 3.1 to allow for either a serially correlated shock or serially correlated idiosyncratic errors. Regarding serial correlation in the shock, we simulate the shock as an AR(1) process with autoregressive coefficient 0.7 (chosen to match the extracted shock in our empirical application).

Figures 3 and 4 show estimates of the first impulse response and the distributions of MSEs and correlations between estimated and true shocks across simulations for HPCA, either with a serially independent or a serially correlated shock. The figures show that serial correlation in the shock does not affect the performance of HPCA (as the mean MSE and correlation across simulations are unchanged).

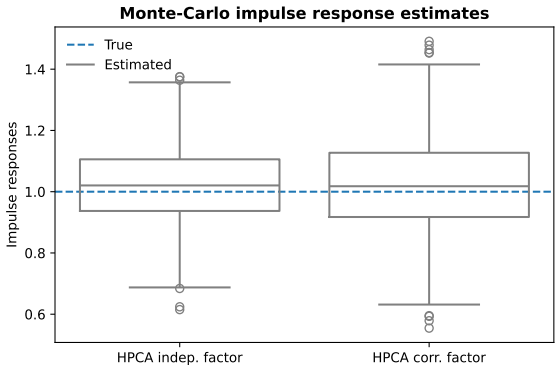

Figure 3: First impulse response estimates: independent vs. serially correlated shock.

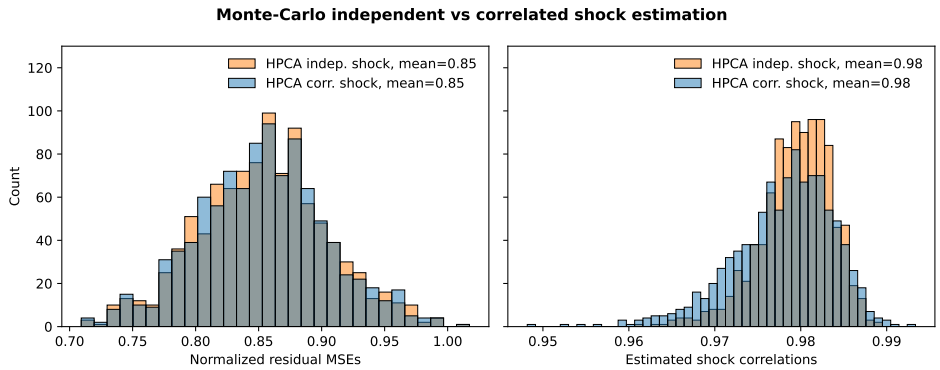

Figure 4: Residual MSEs and estimated/true shock correlation: independent vs. serially correlated shock.

Regarding serial correlation in the idiosyncratic errors, we modify the simulation design in section 3.1 so that the idiosyncratic error is an AR(1) with autoregressive coefficient $\rho = 0.5$ (rescaling the standard deviation of the error by $(1 - \rho^2)^{0.5}$ to maintain the same unconditional variance as in the independent case).

Figures 5 and 6 report the estimates of the first impulse response and the distributions of MSEs and correlations between estimated and true shock across simulations for HPCA either with independent or serially correlated errors. The figures show that serial correlation in the errors does not affect the performance of HPCA (as the mean MSE and correlation across simulations are unchanged).

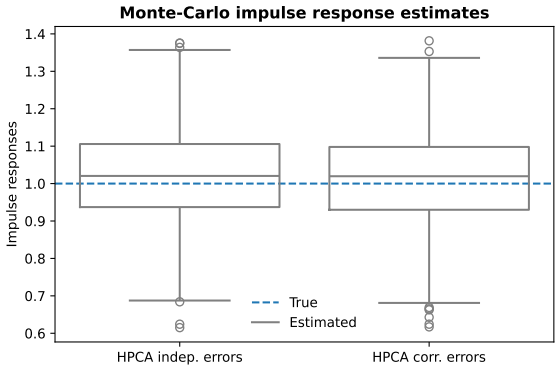

Figure 5: First impulse response estimates: independent vs. serially correlated errors.

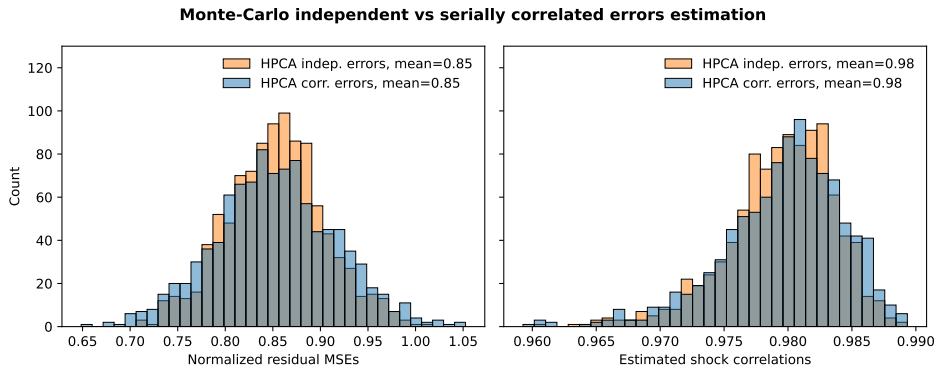

Figure 6: Residual MSEs and estimated/true shock correlation: independent vs. serially correlated errors.

3.3 Impact of weak factors

In this section, we aim to evaluate the implications of violating Assumption 4, by examining the effects of weak factors on our HPCA method. We build on the literature concerning the finite sample performance of PCA by first relating factor strength to the signal-to-noise ratio (SNR), a quantity that can be computed in applications. We then go beyond the existing literature by exploring whether, keeping the SNR constant, the shape of the impulse response functions matters - specifically, their rate of decay to zero and the number of zero responses. We consider the following definition of SNR, which extends the measure considered by Maldonado and Ruiz (2021) to account for heteroskedastic noise:

$$SNR_t = \frac{\sum_{h=1}^{H} \lambda_{h,t}^2}{\max_{h=1}^{H} \sigma_{h,t}^2}, \tag{20}$$

which can be computed in applications by plugging in estimates of the impulse responses $\lambda_{h,t}^2$ and idiosyncratic variances $\sigma_{h,t}^2$. The SNR can be viewed as a way to quantify the amount of information available for estimating the latent shocks and impulse responses. A higher SNR indicates that we can expect more accurate estimates from our HPCA algorithm.

We select Monte Carlo parameters once again to match our application, with

$r = 1, H = 7, T = 100$. For simplicity, we set all idiosyncratic variances to 1, so that the SNR is the sum of squared impulse responses. In our empirial application, the SNR varies significantly over time, with the lowest SNR, representing the worst-case scenario, being just above 5. We thus set SNR=5 and vary the shapes of impulse-response functions based on two characteristics: 1) the slope of the non-zero impulse responses (ranging from 0 to 1), and 2) the number of impulse responses that are exactly equal to zero. Figure 7 illustrates two examples of impulse response functions with the same SNR: the left panel depicts a flat response with a slope of 0, while the right panel shows a response that linearly decreases to zero with a slope of 1.

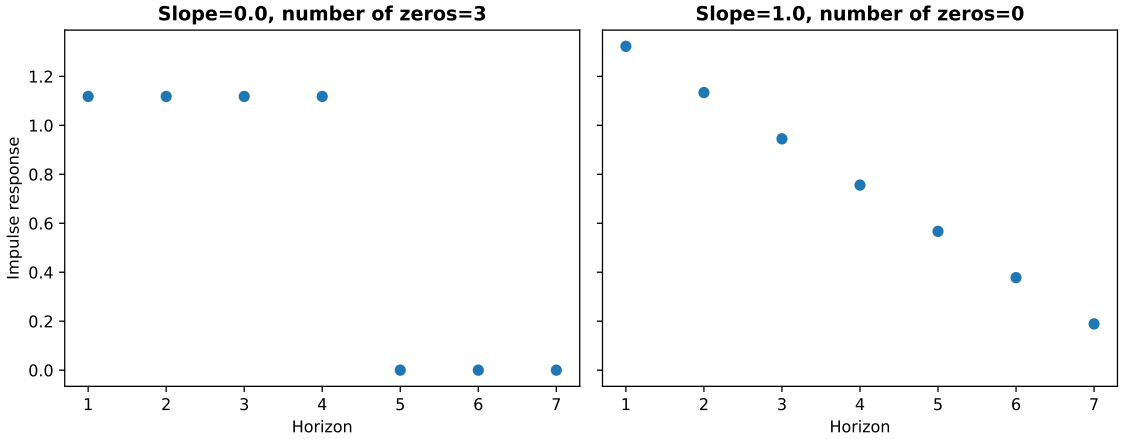

Figure 7: Two examples of impulse response functions with $SNR = 5$.

We assess the performance of our HPCA estimation in terms of the residual MSE, the correlation between estimated and true shock, the average (across horizons) impulse response bias, and the maximum (across horizons) impulse response bias. Figure 8 shows the summary of the performance across a matrix of different shapes for the impulse responses, while keeping the SNR fixed at the worst-case scenario of SNR=5.

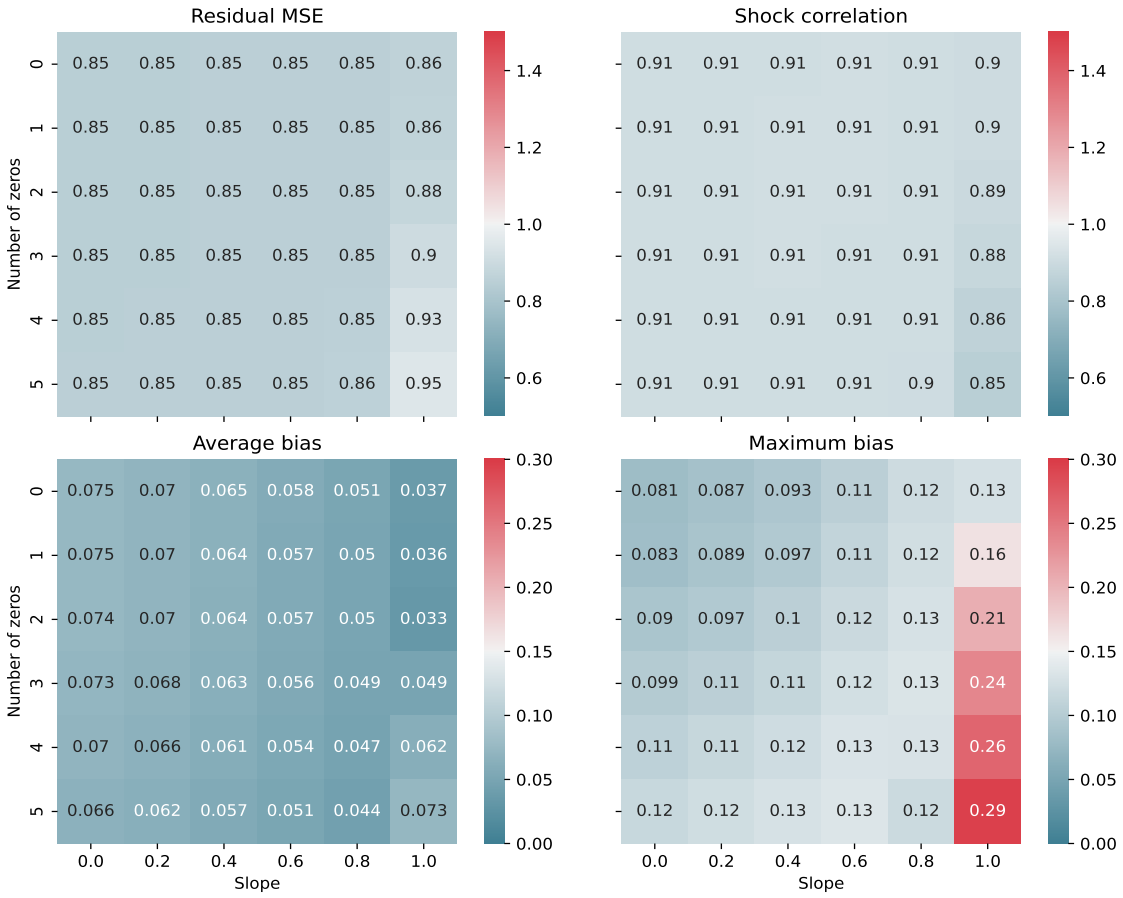

Figure 8: HPCA summary statistics across a matrix of IRF parametrizations.

Overall, we observe a good performance of HPCA, which is generally not affected by the shape of the IRF in terms of MSE, correlation between true and estimated shock and average bias for impulse responses across horizons. However some impulse response estimates become more biased when the IRF has a steep slope and many zeros, consistent with the incoherence assumption that covariance matrices should not concentrate information in just a few rows and columns.

3.4 Ability to recover multiple shocks

We generate data from model (1) with $r = 2, H = 7, T = 100$, with the two shocks in $F_t \sim i.i.d.N(0,1)$, independent of each other, impulse responses given by $\lambda_h = (1, -(h/H))$, and heteroskedasticity given by $\sigma_h = 0.6 - 0.4(h/H)$, $h = 1, ..., H$. Across 1000 Monte Carlo replications, we apply Onatski's (2010) procedure for determining the number of shocks. The procedure correctly identifies the presence of two shocks 82% of the time, and incorrectly identifies the presence of one shock 18% of the time. This indicates that the procedure is in principle able to uncover the presence of more than one shock, even in small samples where the cross-sectional dimension H is very small and there is heteroskedasticity.

3.5 Bootstrap coverage

We investigate the coverage rates of the bootstrap procedure described in section 2.7 for obtaining confidence intervals for impulse responses. The simulation design is the same as in section 3.1, assuming serially uncorrelated factors and idiosyncratic errors. For each Monte Carlo replication, we generate bootstrap confidence intervals using $B = 1000$ bootstrap iterations. The empirical coverage rates for the confidence interval for the fist impulse response across Monte Carlo replications (averaged across horizons) are 93% for a nominal 95%, 87.7% for 90% and 83% for 85%. We thus see a slight tendency to undercover.

We then introduce serial correlation and generate the shock as an AR(1) with autoregressive coefficient 0.7. In this case, the undercoverage of the confidence interval is clearer, with empirical coverages 85.1% for a nominal 95%, 77.9% for 90% and 72.5% for 85%. This is unsuprising, since the bootstrap assumes no serial correlation.

3.6 Estimation of time-varying impulse responses

Here we introduce time-varying impulse responses in a model with two shocks and analyze the ability of tvHPCA to recover time-varying impulse responses.

We generate data as in (1), with $H = 7, T = 500, r = 2$. The time-varying impulse

responses are given by $\lambda_{h,t} = [\sin(2\pi(h/H + t/T)), e^{-1}\sin(2\pi((h+2)/H + t/T))]$, such that the second shock explains $1/e$ the variation of the first shock. The time-varying idiosyncratic volatilities are given by $\sigma_{h,t} = [3 + \sin(2\pi(h/H + t/T))]/5$. The two shocks in F_t are i.i.d. N(0, 1), independent of each other. We note that these parameterizations imply a signal-to-noise ratio SNR (defined in equation (3.3)) for the first factor approximately stable around 5.5, and approximately 0.8 for the second factor. We thus expect the second factor and its impulse responses to be less precisely estimated.

Indeed, we find that both shocks are generally well recovered by the tvHPCA estimation, although the first shock is more precisely estimated (0.94 correlation with the true first shock) than the second shock (0.72 correlation).

Figure 9 shows the nonparametric estimates of the time-varying impulse responses together with pointwise 95% confidence intervals. As expected, the impulse responses to the first shock are more precisely estimated than those to the second shock. The empirical confidence interval coverage rates (averaged across all horizons) are 97.2% and 81.8% for the impulse responses to the first and second shocks, respectively, for a nominal 95% coverage. Figure 9 reveals that our method is generally able to recover the patterns of time variation in impulse responses, with some deterioration in performance at the beginning and end of the sample.

A possible explanation for the deterioration in performance at the sample boundaries is that we use a locally-constant regression in estimating the local covariance matrix, which introduces some bias. This is a known issue with locally-constant estimation, and a typical solution is to use a locally-linear estimation instead. This is however challenging in our context of estimating covariance matrices, as the positive definiteness constraint cannot be easily enforced in locally-linear estimators.[17]

[17] E.g., Chen and Leng (2015) discuss the bias of locally-constant estimation of covariance matrices. They propose a locally-linear estimator, however this method leverages the Cholesky decomposition which is dependent on the order of the variables and is not generally applicable.

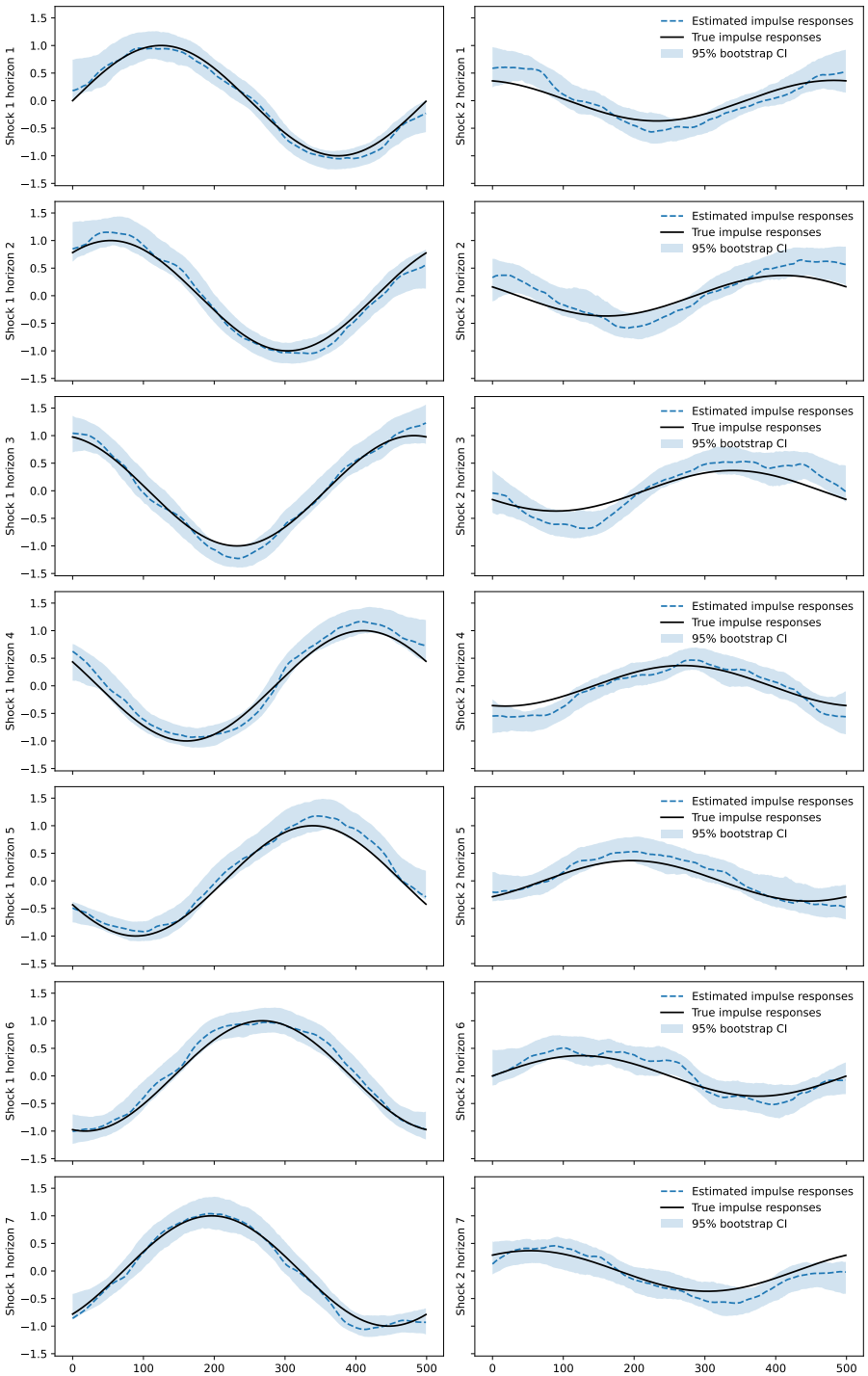

Figure 9: Simulation bootstrap 95% confidence intervals for all impulse responses.

4 Empirical application: perceived shocks and impulse responses of inflation

In this section we use our tvHPCA method to extract historical perceived shocks and impulse response functions from time series of expectations data on inflation.

4.1 Data

Our primary data source is the consensus expectations from the Blue Chip Economic Indicators (BCEI) survey of professional forecasters. The consensus expectations are the average of the individual expectations across all survey participants at any given point in time. We focus on the expectations of quarterly CPI inflation (annualized rate, percentage). In each calendar month, the survey reports expectations for the quarters of the current and next calendar years. The number of available horizons thus decreases throughout the year and the largest number of horizons available every month is five.[18] From the BCEI we thus consider a balanced panel of monthly expectations for five horizons, for which we use indices $h = 0, ..., 4$ to highlight the fact the first point in the term structure of horizons is a nowcast ($h = 0$) whereas the remaining points are forecasts of the future four quarters ($h = 1, ..., 4$).[19]

In addition to the short- and medium-horizons expectations described above, we would like to include measures of long-term inflation expectations. However, a limitation of the BCEI is that long-term inflation expectations are surveyed less

[18]For instance, a survey conducted in January has eight horizons (2 full years), while a survey conducted in December of the same year has only the current quarter and the four quarters of the next calendar year.

[19]Some attention should be paid to issues of timing and information sets in the BCEI. The survey is usually published in the middle of each calendar month, shortly after the CPI data release. The information set of the forecasters typically contains the previous month's CPI release, although this is not necessarily guaranteed (for instance the forecaster may submit their response prior to the CPI release). One thus needs to establish if the expectation provided in a given month constitutes observed data, a nowcast or a forecast. For example, for the September survey the latest CPI release is that of August and therefore Q3 CPI inflation is not yet observed. This means that the Q3 expectation in September is a nowcast (corresponding to $h = 0$) and the revision is computed as the change relative to the August survey expectation for Q3.

frequently than short- and medium-term expectations.[20]

For long horizons, we therefore opt to use the inflation expectations series published monthly by the Federal Reserve Bank of Cleveland (CF thereafter). This series is obtained as a model-based composite of BCEI surveys and data (the latest CPI release, treasury yields, and inflation swap prices). We find this to be the most appealing option as one can interpret these expectations as the result of updating the infrequent professional survey expectations for long horizons using high-frequency data. The CF expectations are updated each month immediately after the CPI release; therefore, their timing and information set are aligned with those of the BCEI survey.[21] From the CF inflation expectations series, we consider the expectation revisions of 2-year 3-year and 5-year 5-year CPI inflation. These correspond approximately to medium- and long-term inflation expectations, and they extend the time span captured by the expectations from one year (for the BCEI) to 10 years.

To summarize, combining expectations data from BCEI and CF gives a balanced panel of CPI inflation expectations monthly revisions from February 1982 to July 2023 for a term structure of seven horizons: the 0th-4th quarterly horizons and the 2-year 3-year and 5-year 5-year horizons.[22]

A possible concern of augmenting the BCEI expectations with the CF long-term expectations is that the results are driven by the CF expectations. Figure 10 provides reassurance that this is not the case, as we see that the (local) sample covariance matrix from the merged dataset has similar eigenvalues as when considering only BCEI data.

[20]Long-term expectations from the Survey of Professional Forecasters (SPF) are also similarly not available every month. There are some market-based measures of inflation expectations available at higher frequencies, including for long horizons. The breakeven inflation rate (the difference between yields on nominal and real U.S. debt of similar maturities), for example, is often quoted as a measure of inflation expectations. However, this measure is not ideal for our analysis because it is confounded with the inflation risk premium.

[21]Note that the data prior to 2009 is constructed ex-post by CF using real-time data.

[22]Borağan Aruoba (2020) similarly combines different data sources to construct a term structure of expectations (not revisions), but for the different purpose of linking asset prices to inflation expectations. We note that we obtain a smaller set of points in the term structure than Borağan Aruoba (2020), because for our purpose it is paramount that the expectations from different data sources are based on aligned information sets.

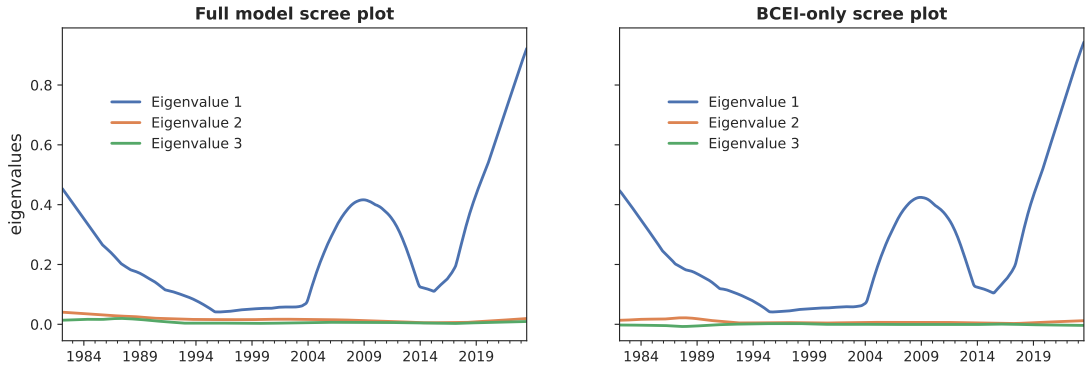

Figure 10: The three largest eigenvalues of the local sample covariance matrix from merged BCEI and CF data (left panel) and BCEI data only (right panel).

4.2 Results

We now present the key findings from our empirical analysis. The bandwidth for the tvHPCA was determined through cross-validation, resulting in a value of $b = 0.121$ (additional details can be found in the appendix). It is important to note that confidence intervals for impulse responses should be interpreted with caution due to the serial correlation of the extracted shocks in our application, which our simulations indicate may lead to some undercoverage.

4.2.1 One perceived shock, highly correlated with inflation surprises

To determine the number of shocks, we apply the procedure described in Section 2.5, based on a local (in time) version of the test proposed by Onatski (2010).

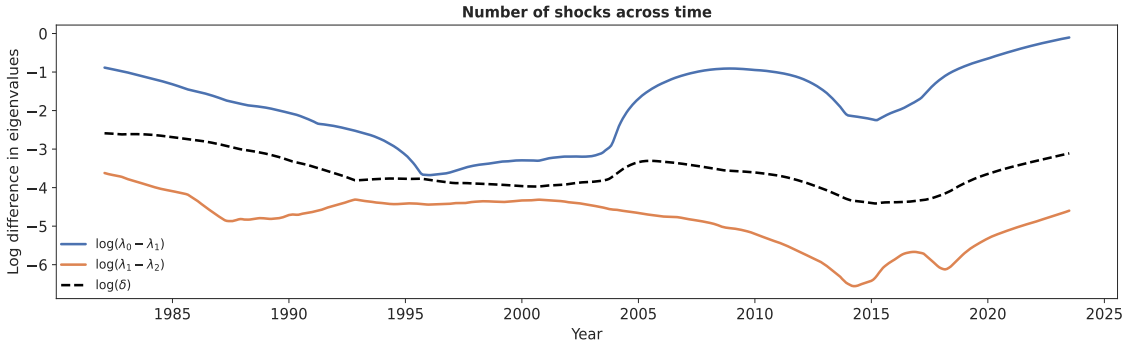

Figure 11: Results of the test for the number of factors developed in Onatski (2010) applied across time. The solid blue line depicts the log difference between the 1st largest eigenvalue and the "0" factors case across time; the orange solid line depicts the log difference between the 1st and 2nd largest eigenvalues. The dashed black line represents the critical values at each point in time.

From Figure 11 we can see clear evidence of 1 shock consistently across time. This is one of our main stylized facts. We then extract the shock by applying tvHPCA using the normalization in Assumption 2(a). To gain insight into the nature of the extracted shock, we try to relate it to actual data for the period under consideration. First, we find that the shock is highly correlated with the 3-month 3-month annualized rate of inflation (correlation 0.73) over the time period February 1982 to July 2023.[23] This could be interpreted as showing a fairly high correlation between the shock and the surprises from a model that forecasts inflation as being constant (e.g., at a given target). We then estimate and compute surprises from Stock and Watson (2007)'s Unobserved Component Stochastic Volatility (UCSV) model over the same period[24] for quarter-on-quarter annualized CPI inflation. We find an even higher correlation between our shock and the surprises from this model (the correlation equals 0.81 over the whole sample, and 0.85 when only considering data up to

[23] 3-month 3-month inflation is calculated using headline CPI data from the U.S. Bureau of Labor Statistics.

[24] The solution to the UCSV model is computed by Markov Chain Monte Carlo (MCMC) using a diffuse prior for the initial condition and $\gamma = 0.2$ as the sole model parameter. The Matlab code from Chan (2018) can be accessed at `https://joshuachan.org/code/code_spectest.html`. The surprises are the estimated model's residuals.

2020).

The high correlation between the shock and the surprises from the UCSV model can be seen in Figure 12. The figure also shows a change in pattern post-pandemic, with shocks of persistently larger magnitude than the surprises from the UCSV model. This could be interpreted as suggesting that agents in our expectation data kept underestimating the persistence of inflation in the post-pandemic period, and thus perceived a sequence of positive shocks, whereas the UCSV model more accurately characterized the persistence of inflation in the data.

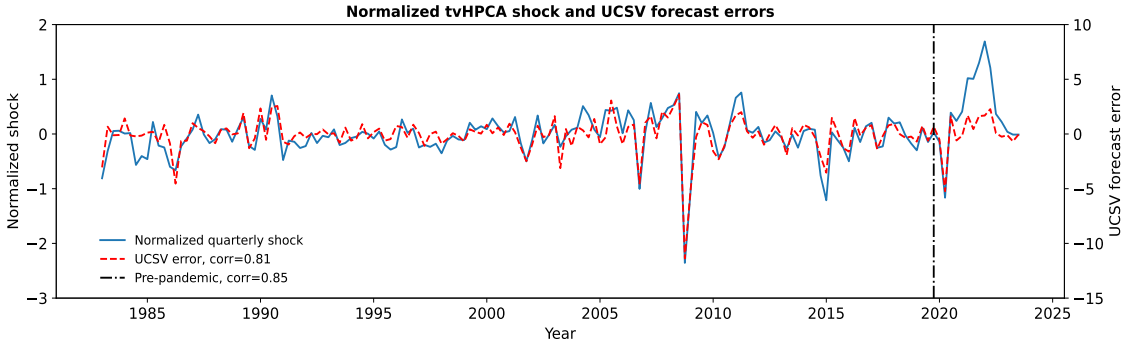

Figure 12: Perceived shock vs. surprises from Stock and Watson's (2007) model

4.2.2 Secular decrease in the perceived persistence of the shock

Below we show our estimates of the loadings across horizons and over time, obtained under the normalization 2(a) of a unit local standard deviation of the shock.

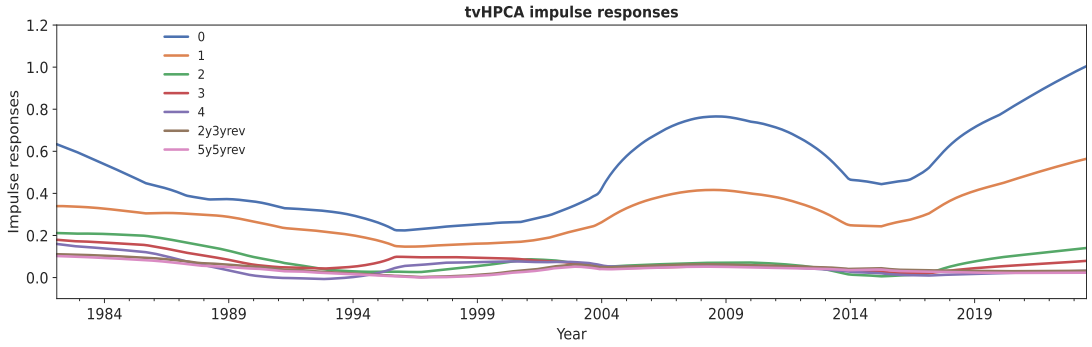

Figure 13: The graph presents the estimated impulse responses across the term structure of horizons and over time.

Figure 13 shows clear evidence of time variation in impulse responses across time. The impulse response function (capturing the dynamic effects of the shocks across different horizons) at a given point in time can be obtained as a vertical slice from the figure. These slices suggest time-varying shapes of the impulse response function, with the effects of the shock generally decreasing with the horizon. The apparent non-monotonicity at longer horizons in the 1990s is not statistically significant (see Figure 14, which plots 95% confidence intervals for selected horizons between 1993-1999).[25]

[25]Monotonicity requires that the 4th horizon impulse response lie between the 2nd and 5-year 5-year horizon impulse response. The point estimate for the 4th horizon impulse response is below that of the 5-year 5-year horizon in the early 1990s and exceeds that of the 2nd horizon in the late 1990s. However, the overlap in confidence intervals shows that these deviations are not statistically significant.

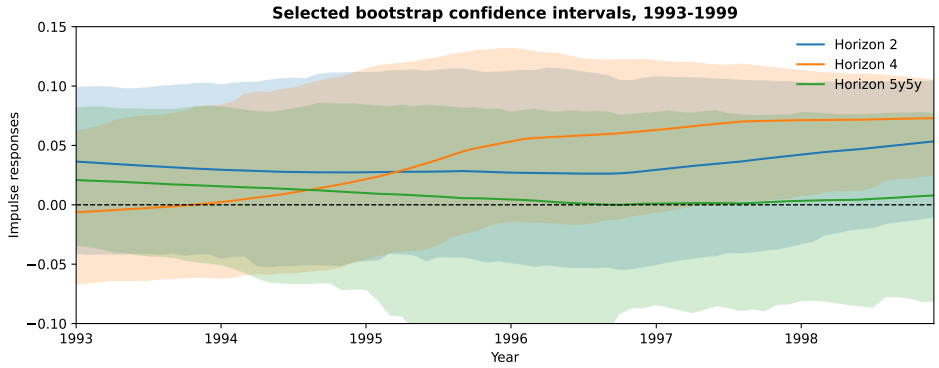

Figure 14: 95% bootstrap confidence intervals for impulse responses at selected horizons, 1993-1999.

Because of the unit-standard-deviation normalization for the shock used in the figure, the magnitude of the impulse responses reflects the volatility of expectation revisions at that horizon. We see that this volatility increased around the financial crisis of 2009 and at the end of our sample, which includes the Covid pandemic and the post-pandemic recovery. The difference between the impulse responses at a point in time contains information about the perceived persistence of shocks, indicating for example quickly decaying responses to shocks around 2009.

Because the standard deviation of the shocks changes over time, the unit-standard deviation normalization has too many moving parts to enable a clear comparison of impulse responses over time. We thus recompute the impulse responses using the unit-impact normalization in Assumption 2(c) and plot in Figure 15 the impulse response functions for three selected times: during the Volker disinflation period of 1986, during the financial crisis of 2009 and during the high-inflation period of 2022. The figure also reports bootstrap confidence intervals at each point in time, computed as described in Section 2.7.

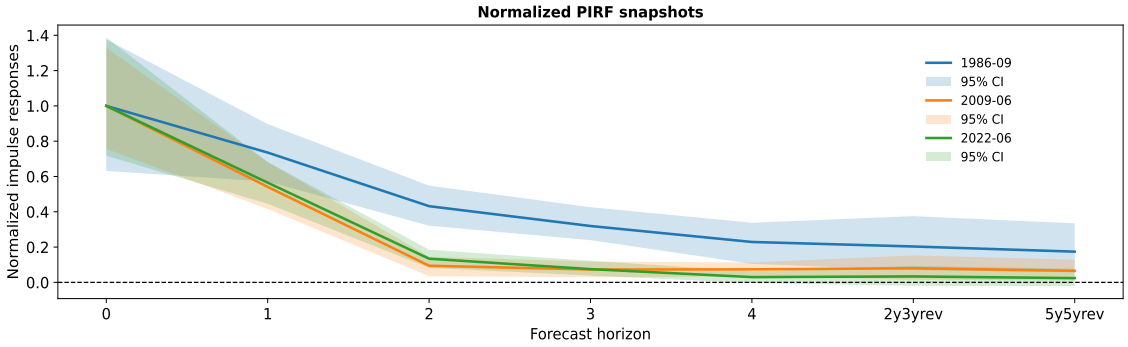

Figure 15: Perceived impulse response functions during the Volker disinflation, the financial crisis of 2009 and the 2022 high-inflation period

We see that during the Volker disinflation years, long-term inflation expectations were deanchored, in the sense that the perceived persistence of the shock remained high at long horizons. While in 2009 and 2022 the short-term inflation volatility was higher (as reflected by the large impulse responses for the 0th horizon in Figure 13), long-term inflation expectations remained anchored. In fact, the long-term impulse response has decreased over time for these snapshots.[26] It is worth noting that, although long-term inflation expectations were at risk of becoming deanchored in 2022, our data imply a historically low persistence of the shock at long horizons (subject to the discussed caveats of nonparametric estimation at the sample boundaries).

The finding of time-varying (and rapidly decaying) impulse responses also suggests that agents do not use Stock and Watson (2007)'s UCSV model to produce forecasts, in spite of our finding in the previous section that our extracted shock is highly correlated with the surprises from this model. This is because the UCSV model implies a flat impulse response function (as can be deduced from equation (2) and the discussion thereafter, the UCSV's model-implied impulse response equals 1 at all horizons), which we do not observe in the data.

[26]This is consistent with the finding of Stock and Watson (2007) - and extends it to the current period - that transitory shocks to inflation have become more volatile/relevant while permanent shocks to inflation have become less volatile/relevant.

4.2.3 A possible narrative about the 2022 high-inflation episode

Putting together the findings in the previous two sections, we can provide a possible narrative for what happened during the post-pandemic high-inflation episode (again subject to the caveat of possible increased bias at the sample boundaries).

In general, our method can provide an answer to the policy-relevant question: if we see a large change in long-term inflation expectations, is it because agents perceived a large shock or because they expected the shock to persist (i.e., a notion of deanchoring)? The method can answer this question by disentangling the two latent sources. The two subpanels of Figure 16 plot the perceived impulse response functions at two points in time during 1986 and 2022 (left panel) and the full time series of the extracted shock (right panel, smoothed using a 12-month moving average).

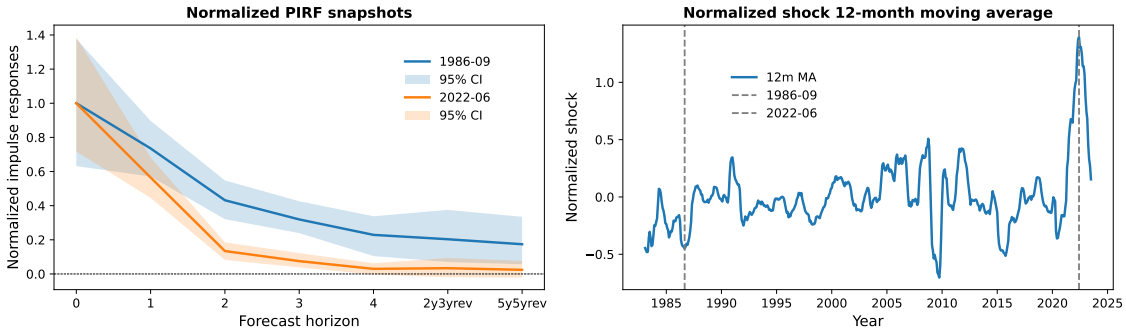

Figure 16: Perceived impulse response functions and corresponding shocks: 1986 vs. 2022

The figure reveals a clear contrast between the two dates: during 1986 agents perceived a shock of normal magnitude by historical standards, but believed the shock to be highly persistent (deanchoring), whereas in 2022 the perceived shock was unprecedentedly large, but agents believed that its effects would essentially disappear within a year (anchoring).

5 Conclusion

This paper illustrated how utilizing the horizon dimension found in various expectations datasets, along with a focus on expectation revisions, enables the extraction of novel empirical measures of time-varying beliefs about shocks and impulse responses. The core concept involves fitting a time-varying factor model to the panel of revisions across different horizons and time periods, which produces shocks (the factors) and time-varying impulse responses (the loadings). Our nonparametric approach is based on weak assumptions and is specifically designed to address the small-sample characteristics of these datasets.

The versatility of our method allows it to be adapted to answer different economic questions and to accommodate the unique opportunities and limitations associated with different types of expectations data. Here we focused on a balanced panel with a sufficiently long time dimension, enabling the analysis of time-varying impulse responses, which is typically achievable when examining aggregate expectations. However, given the small-sample nature of our method, it could be further adapted to expectations at the individual level, to investigate heterogeneity in beliefs about shocks and their dynamic effects.

Moreover, we note the potential to explore an additional dimension of the data that we did not exploit in this paper: expectations related to multiple variables. We briefly addressed how the extraction of shocks and impulse responses for various variables could help in providing a structural interpretation of these shocks. Considering this additional dimension of the data could further enhance our understanding of the expectation formation process, including whether agents respond to the same shocks when forecasting different variables and whether their beliefs align with economic theory. We leave the exploration of these important questions for future research.

Appendix: Bandwidth selection

This appendix presents the details of the cross-validation used to determine the optimal bandwidth b for nonparametric estimation of the sample covariance matrix. Given that the estimation is done via local singular value decomposition followed by several steps of refinements, the bias-variance trade-off necessary to discuss a theoretically optimal bandwidth is non-standard. We therefore follow Su and Wang (2017) and use a data-driven way of selecting an empirically optimal bandwidth by a version of cross-validation (CV). Specifically, we select the empirically optimal bandwidth $\widehat{b}^*$ by solving the following minimization problem:

$$\min_b CV(b) = \frac{1}{Tp} \sum_{h=1}^{p} \sum_{s=1}^{T} \left[X_{hs} - \widehat{\lambda}_{hs}^{(-s)} \widehat{F}_t^{(-s)} \right]^2, \tag{21}$$

where $\widehat{\lambda}_{hs}^{(-s)}$ and $\widehat{F}_t^{(-s)}$ are the versions of $\widehat{\lambda}_{hs}$ and $\widehat{F}_t$ respectively with the s-th time series removed when performing the local PCA analysis. In the context of our application, the graph of the CV distance, defined in (21), is presented below with the resulting optimal bandwidth $\widehat{b}^* = 0.121$.

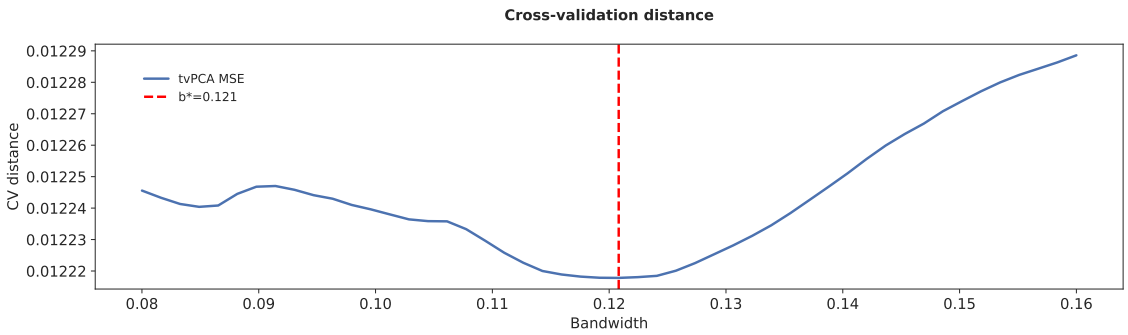

Figure 17: Plot of CV distance, defined in eq. (21). The vertical red dashed line represents the empirically optimal bandwidth corresponding to the minimum along the CV curve.

References

Agarwal, A., Agarwal, A., and Vijaykumar, S. (2023). Synthetic combinations: A causal inference framework for combinatorial interventions. *Advances in Neural Information Processing Systems*, 36:19195–19216.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

Bai, J. and Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, 212(1):78–96.

Bai, J., Ng, S., et al. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, 3(2):89–163.

Bai, J. and Wang, P. (2016). Econometric analysis of large factor models. *Annual Review of Economics*, 8(1):53–80.

Bianchi, F., Ludvigson, S. C., and Ma, S. (2022). Belief distortions and macroeconomic fluctuations. *American Economic Review*, 112(7):2269–2315.

Blanchard, O. J., L'Huillier, J.-P., and Lorenzoni, G. (2013). News, noise, and fluctuations: An empirical exploration. *American Economic Review*, 103(7):3045–3070.

Borağan Aruoba, S. (2020). Term structures of inflation expectations and real interest rates. *Journal of Business & Economic Statistics*, 38(3):542–553.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1):163–188.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found Comput Math*, 9:717–772.

Carvalho, C., Eusepi, S., Moench, E., and Preston, B. (2023). Anchored inflation expectations. *American Economic Journal: Macroeconomics*, 15(1):1–47.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica: Journal of the Econometric Society*, pages 1305–1323.

Chan, J. C. (2018). Specification tests for time-varying parameter models with stochastic volatility. *Econometric Reviews*, 37(8):807–823.

Chen, Z. and Leng, C. (2015). Local linear estimation of covariance matrices via cholesky decomposition. *Statistica Sinica*, pages 1249–1263.

Choi, I. (2012). Efficient estimation of factor models. *Econometric theory*, 28(2):274–308.

Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678.

Dahlhaus, R., Richter, S., and Wu, W. B. (2019). Towards a general theory for nonlinear locally stationary processes. *Bernoulli*, 25(2).

Del Negro, M. and Otrok, C. (2008). Dynamic factor models with time-varying parameters: measuring changes in international business cycles. *FRB of New York Staff Report*, (326).

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2):337–364.

Diebold, F. X., Piazzesi, M., and Rudebusch, G. D. (2005). Modeling bond yields in finance and macroeconomics. *American Economic Review*, 95(2):415–420.

Enders, Z., Kleemann, M., and Müller, G. J. (2021). Growth expectations, undue optimism, and short-run fluctuations. *Review of Economics and Statistics*, 103(5):905–921.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(4):603–680.

Florescu, L. and Perkins, W. (2016). Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959. PMLR.

Gürkaynak, R. S., Sack, B., and Swanson, E. (2005). The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models. *American economic review*, 95(1):425–436.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Li, Q. and Racine, J. S. (2023). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Maldonado, J. and Ruiz, E. (2021). Accurate confidence regions for principal components factors. *Oxford Bulletin of Economics and Statistics*, 83(6):1432–1453.

Miranda-Agrippino, S. and Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, 13(3):74–107.

Motta, G., Hafner, C. M., and von Sachs, R. (2011). Locally stationary factor models: Identification and nonparametric estimation. *Econometric Theory*, 27(6):1279–1319.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642.

Plagborg-Møller, M. (2019). Bayesian inference on structural impulse response functions. *Quantitative Economics*, 10(1):145–184.

Robinson, P. M. (1989). *Nonparametric estimation of time-varying parameters*. Springer.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2006). Forecasting with many predictors. *Handbook of economic forecasting*, 1:515–554.

Stock, J. H. and Watson, M. W. (2007). Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics*, volume 2, pages 415–525. Elsevier.

Su, L. and Wang, X. (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics*, 198(1):84–101.

Zhang, A. R., Cai, T. T., and Wu, Y. (2022). Heteroskedastic pca: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80.