Clustered Local Projections for Time-Varying Models*

Ana María Herrera†, Elena Pesavento‡, Alessia Scudiero§

April 30, 2026

Abstract

We propose a *clustered* local projection (*clustered LP*) method to estimate impulse response functions in a class of time-varying models where parameter variation is linked to a low-dimensional matrix of observables. We show that the clustered LP recovers the conditional average response when the driving variables are exogenous and a weighted average of the conditional marginal effects when they are endogenous. We propose an iterative estimation method that first classifies the data using k-means, estimates impulse response functions via GMM, and evaluates differences across clustered LP estimates. Our Monte Carlo simulations illustrate the ability of *clustered LP* to approximate the conditional average response function. We employ our technique to examine how uncertainty influences the transmission of a contractionary monetary policy shock to the 5- and 10-year U.S. nominal Treasury yields. Our estimation results suggest macroeconomic and monetary policy uncertainty operate through complementary but distinct channels: the former primarily amplifies the risk compensation embedded in the term premium, while the latter governs the speed and persistence with which markets revise their expectations about the future rate path following a monetary policy shock.

JEL Classification: C32, E17, E42, E52, E60, E63.
Keywords: Time-varying parameter VAR, time-varying local projections, clustered local projections, monetary policy transmission, uncertainty.

*We thank seminar participants at the 2024 SNDE conference, the 2024 Midwest Econometrics Group Conference and the Fall 2025 Midwest Macroeconomics Meeting for helpful comments. This paper is based on research supported by the NSF under Grants No. SES-2417534 and SES-2417535. The authors have no conflict of interest to declare.

†University of Kentucky, Email: amherrera@uky.edu
‡Emory University, Email: epesave@emory.edu. - Corresponding author.
§Emory University, Email: alessia.scudiero@emory.edu

1 Introduction

Models with time-varying parameters are commonly used by applied macroeconomists and policymakers to study how shocks affect economic aggregates. Their appeal stems from the realization that many macroeconomic time series employed in policy analysis exhibit some form of nonlinearity. For instance, some researchers posit that inflation disagreement tends to increase during periods of high inflation, which, in turn, renders monetary policy less effective (see, e.g., Dong et al. (2024)). Others suggest that changes in the size or direction of shocks hitting the economy may lead to variation in the transmission of such shocks to aggregate economic activity (see, e.g., Barnichon, Debortoli, and Matthes (2022)).

A key aspect that differentiates alternative time-varying models is the law of motion that governs how the parameters evolve over time. In state-dependent models, some regime-switching models, and threshold autoregressive models, the link between parameters and external economic variables is stated explicitly, but the parameters are restricted to take values in a finite set of regimes. Other approaches, such as traditional time-varying vector autoregression models (TVP-VARs) and time-varying local projections (TVP-LP), allow the parameters to vary freely over time according to a parametric law of motion. However, they leave the connection to economic conditions unspecified. In practice, many empirical studies assume a specific law of motion for the parameters in the estimation step and, later, interpret the estimated parameter path in connection with the evolution of specific economic conditions. For example, Inoue, Rossi, and Wang (2024) suggests that their estimated government spending multipliers are correlated with the level of public debt, while alternative regimes in Markov-Switching (MS) models are often interpreted in terms of economic expansions and recessions.

In this paper, we propose a new estimator that can be used to study impulse response functions in a broad class of time-varying parameter models, *clustered local projections (clustered LP)*. In our framework, time variation is linked to a low-dimensional matrix of observables that the researcher deems relevant in explaining the evolution of the parameters via a (possibly)

nonlinear function. A novelty of our framework is the use of an iterative estimator, where the data are first classified into clusters (periods) using *kmeans*, the impulse responses for each cluster are estimated via local projections –thus allowing for heterogeneity over time–, and an evaluation step assesses whether the responses are statistically different between clusters. A crucial ingredient of the clustered LP estimator is the use of a machine learning method to group together time periods when observables (e.g., economic policy uncertainty, unemployment) are similar. This contrasts with the typical literature on state-dependent models, where it is common to determine a-priori the number of states and the cutoff by grouping the data into samples below and above the mean of one variable of interest (e.g., average unemployment in Ramey and Zubairy (2018)).

Our paper contributes to the literature on local projections in several aspects. From a theoretical perspective, we build on the work of Gonçalves et al. (2024a) and extend the idea of the state-dependent local projections to a version with many low-dimensional states. We show that the clustered LP recovers the conditional average response (CAR) when the variables that drive the time-variation are exogenous and a weighted average of the conditional marginal effects (CMR) when they are endogenous. In the latter case, impulse response estimands from the clustered LP have a causal interpretation as in Kolesár and Plagborg-Møller (2025). That is, the clustered LP estimands provide a scalar causal summary of the nonlinear effect in each cluster.

From a methodological point of view, the proposed clustered LP combines a popular machine learning algorithm (k-means) with the local projection method proposed by Jordà (2005); hence, the number and partition of the states are not defined prior to the estimation. Instead, the data-driven iterative procedure consists of three steps. First, a classification step where the k-means algorithm is used to classify the data into groups (clusters) according to the similarity in the driving variables. Second, an estimation step where the local projections for all clusters and horizons are estimated jointly via a system of Generalized Method of Moments (GMM). A third step evaluates whether the impulse response functions are different across all pairs of

clusters using a Wald test. We iterate on these three steps until we reject the null of equality for all clusters for a given horizon of interest.

Using *clustered LP*, we inquire how effective monetary policy shocks are in altering the medium- and long-term U.S. nominal yields during uncertain times. We build on the work of De Pooter et al. (2021), Tillman (2020) and Bauer, Lakdawala, and Mueller (2021) who find that monetary policy is less effective during periods of high monetary policy uncertainty (MPU) and (Aastveit, Natvik, and Sola, 2017) who focus on periods of high/low macroeconomic uncertainty. We deviate from their work by allowing time variation to depend on the evolution of both uncertainty measures and by letting our data-driven method determine the grouping. The iterative procedure classifies the data into four groups (low macro uncertainty-low MPU, low macro uncertainty-moderate MPU, moderate macro uncertainty-high MPU and high macro uncertainty-low MPU), which reflects the fact that the two uncertainty measures do not move in synchrony. The results suggest that on impact and in the very short-run, macroeconomic uncertainty is the main driver of heterogeneity in the response of the 5- and 10-year yields to monetary policy shocks. Nevertheless, on longer horizons, the MPU drives yield responsiveness. In fact, twelve months after the monetary policy shock, the response of the yields is an order of magnitude lower during times of high MPU. Focusing on uncertain times, we find evidence suggesting that the expectations component plays a key role in accounting for heterogeneity during times of high MPU, whereas the response of the term premium is a key determinant during periods of high macro uncertainty. All in all, our results illustrate how clustered LP responses provide a more multifaceted characterization of the interaction between monetary policy and uncertainty than state-dependent models.

Related Literature. *Clustered LP* are most closely related to two strands of literature that estimate impulse responses in unstable environments via local projections. The first strand links the existence of parameter instability to an observable variable such as the state of the economy (Ramey and Zubairy, 2018), the direction of fiscal intervention Barnichon, Debortoli, and Matthes (2022), or the degree of uncertainty or disagreement about inflation expectations

(Klepacz (2021) and Falck, Hoffmann, and Hürtgen (2021)). An advantage of using these state-dependent models is that the researcher can explicitly investigate whether an observable macroeconomic variable of interest drives the evolution of the parameters. However, a disadvantage is that they restrict the form of time variation.

The second strand assumes that the evolution of the parameters is independent of the observables and occurs in a deterministic fashion (Inoue, Rossi, and Wang (2024) and Maung and Yoshida (2025)) and the estimation is performed through local projections. Assuming a particular law of motion has several advantages: it provides a very flexible way to capture different forms of nonlinearity, it simplifies the estimation process by limiting the number of parameters to be estimated in a highly nonlinear model, and, by design, it ensures that the parameters vary gradually over time. Nevertheless, such an assumption has a limitation: the cost of remaining agnostic about the underlying source of the time variation is that the researcher may not directly learn about the economic reasons behind the changes in parameters.

Our framework is also related to a broad and expanding TVP–VAR literature that builds on Primiceri (2005) and Cogley and Sargent (2005). However, our framework differs in three important respects. First, instead of modeling parameter evolution as independent random walks, we allow the parameters to vary as nonparametric functions of observables (Z_t), thus explicitly linking parameter dynamics to economic conditions. Second, while TVP–VAR models assume a specific stochastic law of motion, we do not make assumptions about the functional form that links Z_t with the evolution of model parameters. Third, our frequentist approach relies on LP for estimation, whereas TVP-VAR models are estimated using Bayesian methods. Similarly, our framework shares some features with the work of Fischer et al. (2023), Chan, Eisenstat, and Strachan (2020) and Hauzenberger et al. (2023) where the parameters' law of motion is linked to an observable predictor or a latent shifter. Yet, these papers also use Bayesian estimation methods.

Finally, our empirical framework bears some resemblance to Markov Switching (MS) models (e.g., Hamilton (1989) and Diebold, Lee, and Weinbach (1994)) where the parameters are

assumed to switch over time according to some (possibly unobserved) variable, Z_t. However, in contrast to MS models where the number of states is determined a priori, the classification step of our algorithm uses k-means to select the number of states and partition the data accordingly. Our model is similar, although conceptually different, to a functional-coefficient autoregressive model (see, e.g. Chen and Tsay (1993)). Unlike the standard functional-coefficient we allow the parameters to be driven by an external variable rather than lagged dependent variable. More importantly, we focus on local projections regressions as we are interested in estimating impulse responses.

Outline. The remainder of the paper proceeds as follows. Section 2 outlines the *clustered LP* framework and describes the estimator. Section 3 illustrates the performance of the estimator via Monte Carlo simulations for alternative DGPs; Section 4 uses *clustered LP* to inquire about the effect of monetary policy shocks on the 5- and 10-year yields during uncertain times. Section 5 concludes with a brief summary of the paper and possible future steps.

2 Clustered local projections

In this section, we propose a fast and easy-to-implement method, *clustered LP*, that leverages the ease of implementation of LP and k-means to estimate time-varying impulse response functions when time variation is driven by an observable variable of interest. Our framework relies on a first step where heterogeneity over time is revealed, allowing us to classify the data into groups, and then the responses to impulse response functions are estimated via LP.

2.1 Theoretical Framework

Our baseline specification is a time-varying local projection where the evolution of the parameters is driven by a variable (or low-dimensional set of variables), Z_{t-1}, given by:

$$y_{t+h} = \beta_t^h(Z_{t-1})\varepsilon_t + \gamma_t^h(Z_{t-1})'\mathbf{w_t} + v_{t+h} \tag{1}$$

where $t = 1, \ldots, T$; y_{t+h} is a scalar endogenous variable of interest h periods ahead, ε_t is the structural shock of interest, β_t^h denotes the time-varying impulse response at horizon h, $\mathbf{w_t}$ is a vector of control variables including deterministic terms and lags of endogenous variables, and v_{t+h} is the local projection residual. The notation $\beta_t^h(Z_{t-1})$, and $\gamma_t^h(Z_{t-1})$ makes explicit that the model parameters vary as unknown functions of the external variables Z_t at time $t-1$. The timing of the variables Z_{t-1} follows the literature on state dependent models by assuming that the value of Z in the previous time period affects the current value of the parameters.

We make the following two assumptions:[1]

Assumption 1 (Identification)**.** *ε_t is an observed structural shock such that for each t:*

(a) ε_t is independent of $\mathcal{F}_{t-1} = \sigma(y_{t-j}, \mathbf{w_t}, Z_{t-j}, \varepsilon_{t-j} : j \geq 1)$ and of $(\varepsilon_{t+1}, \ldots, \varepsilon_{t+H})$;

(b) ε_t is continuously distributed on an interval $I \subseteq \mathbb{R}$ with mean zero and positive, finite variance.

Assumption 2 (Estimation)**.** *y_t and Z_t are strictly stationary and ergodic.*

Assumption 1 states the conditions required for the identification result in Proposition 1. It requires full independence of ε_t from $\mathcal{F}_{t-1}$, not merely mean independence; this rules out GARCH-type conditional heteroskedasticity but does not restrict the unconditional variance to be constant. Assumption 2 is not required for identification but ensures that the sample OLS estimator converges to the population quantity β_k^h. Under Assumptions 1 and 2, the distribution of ε_t is time-invariant (by stationarity of the system), so ε_t is in fact i.i.d. and constant variance holds as a consequence rather than as a separate restriction.

[1]The use of more primitive conditions is outside the scope of this paper; for example, Chen and Tsay (1993) gives conditions for geometric ergodicity for a univariate version of our model when $Z_t = Y_{t-p}$.

For ease of exposition, in the remainder of the paper, we assume that the shock ε_t has been appropriately identified and we focus on the response of the variable y_t at time $t+h$ to a shock of size δ in ε_t. Nevertheless, the *clustered LP* approach we propose is flexible and can easily accommodate alternative identification strategies, such as instrumental variables or externally identified shocks, as in our empirical application in Section 4.

In our theoretical framework, we consider two DGPs. In the first, we focus on the case where Z_{t-1} is a low-dimensional set of exogenous driving variables. In the second, we allow for endogenous Z_{t-1}. In both cases, we first classify the observations into K 'groups/clusters' where the driving variable, Z_{t-1}, takes on similar values. Note that the timing of Z_{t-1}, which is crucial for our model, implies that it is predetermined with respect to the shock at time t.

Suppose that the data has been classified into K clusters using k-means clustering. Then, the *clustered LP* for the variable i at horizon h is given by

$$y_{t+h} = \sum_{k=1}^{K} D_k \, \beta_k^h \, \varepsilon_t + \sum_{k=1}^{K} D_k \, \gamma_k^h \, \mathbf{w_t} + v_{t+h}, \tag{2}$$

where $D_k = \mathbf{1}\{Z_{t-1} \in \mathcal{C}_k\}$ is the indicator for the cluster k (with $\mathcal{C}_k$ the k-th k-means cluster on the support of Z_{t-1}), $\mathbf{w_t}$ collects all controls, and the i subindex for variable i is omitted for simplicity.

Following Gonçalves et al. (2024a), we start by defining the causal objects of interest in our time-varying setting. Let $y_{t+h}(e)$ denote the potential outcome obtained by setting $\varepsilon_t = e$. More precisely, the outcome admits the nonparametric structural representation

$$y_{t+h} \;=\; \psi_h(\varepsilon_t,\, U_{h,t+h}), \tag{3}$$

where $\psi_h(\cdot,\cdot)$ is an unknown measurable function and $U_{h,t+h}$ collects all variables other than ε_t that causally affect y_{t+h}: initial conditions (y_{t-1} and lags), the path of the driving variable $(Z_{t-1}, Z_t, \ldots, Z_{t+h-1})$, and all other disturbances entering the outcome between periods t and

$t+h$. The *Conditional Average Structural Function* for cluster k is given by

$$\Psi_k^h(e) \equiv \mathbb{E}[y_{t+h} \mid \varepsilon_t = e,\ Z_{t-1} \in \mathcal{C}_k], \tag{4}$$

We define two causal objects. The *Conditional Average Response* for cluster k to a shock of fixed size δ is given by

$$\mathrm{CAR}^h(\delta, k) \equiv \mathbb{E}\Big[y_{t+h}(\varepsilon_t + \delta) - y_{t+h}(\varepsilon_t) \,\Big|\, Z_{t-1} \in \mathcal{C}_k\Big], \tag{5}$$

and the *Conditional Marginal Response* is given by $CMR^h(k) = \frac{CAR^h(\delta,k)}{\delta} = CAR^h(1,k)$.

In the context of a state dependent model, Gonçalves et al. (2024a) shows that the state-dependent LP estimates the CAR when the state is exogenous. Thus, when Z_{t-1} is exogenous, the *clustered LPs* recover the CAR. More precisely, for $h = 0$ the *clustered LP* estimates the impact effect of the shock ε_{1t} conditional on the value of Z_{t-1} for that particular cluster. For longer horizons, $\hat{\beta}_k^h$ estimates the average response over all possible future paths of the state between time t and $t+h$. When Z_{t-1} is endogenous, the estimates still have a casual interpretation. In particular, in the state-dependent case, Gonçalves et al. (2024a) shows that under normal errors, linear LPs recover the CMR. Similarly, the *clustered LP* recovers the CMR. When normality fails, Kolesár and Plagborg-Møller (2025) demonstrate that linear LPs admit a causal interpretation even when the underlying model is nonlinear. This is also the case here.

The following proposition formally extends these results to our *clustered LP*.

Proposition 1. *Consider the clustered local projection* (2). *Fix any cluster* $k \in \{1, \ldots, K\}$ *formed by k-means clustering of* Z_{t-1}.

Part (i) - General DGP. *Suppose Assumption 1 holds and the following regularity conditions are satisfied for each cluster k:*

(a) $g_{h,k}(e) \equiv \mathbb{E}[y_{t+h} \mid \varepsilon_t = e, D_k = 1]$ *is locally absolutely continuous in* e *on* I*;*

(b) $\mathbb{E}[|g_{h,k}(\varepsilon_t)|(1 + |\varepsilon_t|) \mid D_k = 1] < \infty$ *and* $\int_I \omega_k(e)|g'_{h,k}(e)|\,de < \infty$*, where* ω_k *is defined below.*

Then for each cluster k *and all horizons* $h \geq 0$*:*

$$\beta_k^h \;=\; \int_I \omega_k(e)\, \Psi_k^{h\prime}(e)\, de, \tag{6}$$

where $\omega_k(e) \;\equiv\; \frac{\mathrm{Cov}\left(\mathbf{1}\{\varepsilon_t \geq e\}, \varepsilon_t \,\middle|\, D_k=1\right)}{\mathrm{Var}(\varepsilon_t | D_k=1)}$*. The weight function* ω_k *is: (i) nonnegative and integrates to one; (ii) hump-shaped, peaking near* $\mathbb{E}[\varepsilon_t \mid D_k = 1]$*; and (iii) determined solely by the conditional distribution of* ε_t *given cluster* k*, not by the outcome variable or horizon* h*.*

Part (ii) - Exogenous Z*.* *Suppose, in addition to Assumption 1, that* $\{Z_s : s \geq t\} \perp \varepsilon_t$*, then for each cluster* k *and all horizons* $h \geq 0$*:*

$$\beta_k^h \;=\; \frac{\mathrm{CAR}^h(\delta, k)}{\delta} \;=\; \mathrm{CAR}^h(1, k), \tag{7}$$

which equals $\mathrm{CMR}^h(k)$*, the conditional marginal response for cluster* k*.*

Proposition 1 is a population-level result that holds for any partition of Z_{t-1} into K clusters. The data-generating process need not have discrete regimes: when $\beta^h(Z_{t-1})$ varies continuously, the clustered LP provides a piecewise-constant approximation and Proposition 1 ensures that each piece retains a causal interpretation. The proof of Proposition 1 can be found in Appendix A; the formal approximation properties of this estimator as K grows with T are left for future work.

Some remarks are in order here. First, we follow the usual timing convention of state dependent models: $D_k = \mathbf{1}\{Z_{t-1} \in \mathcal{C}_k\}$ is a function of Z_{t-1}, which is predetermined with respect to ε_t. Under Assumption 1 (a), ε_t is independent of $\mathcal{F}_{t-1}$ and hence independent of the

past, so $\varepsilon_t \perp\!\!\!\perp D_k$ for all k. This makes the within-clustered LP algebraically equivalent to a standard LP estimated on the subsample (i.e., cluster) defined by Z_{t-1}. Second, we do not make any assumptions regarding the functional form of the true LP within the cluster. However, we note that when Z_{t-1} is exogenous, the evolution of Z is independent of the structural shock ε_t, so that future state variables do not respond to ε_t. In this case the potential outcome is linear in e and a linear regression within each cluster recovers CAR. When Z_{t-1} is endogenous, Part (i) applies Proposition 1 of Kolesár and Plagborg-Møller (2025) to the within-cluster subsample so the estimator admits a causal interpretation. Given the timing of the model, the clustered LP will recover the CAR only at impact. For $h > 0$, the OLS estimand can be expressed as a weighted average of marginal effects $\Psi_k^{h\prime}(e)$ with weights $\omega_k(e)$ that can be estimated from the data. When the weights are positive, the estimand inherits the sign of the CMR within the cluster whenever $\Psi_k^{h\prime}(e)$ has a constant sign.[2] Lastly, a researcher interested in recovering CAR could employ the nonparametric approach of Gonçalves et al. (2024b).

2.2 Estimator

The *clustered LP* estimator consists of three iterative steps: classification, estimation, and evaluation. Note that the estimation of β_k^h requires, in addition to the identification conditions in Assumption 1, that the data are strictly stationary and ergodic (Assumption 2) so that the sample OLS/GMM estimator converges to the population quantity characterized in Proposition 1.

Classification step: We first partition the whole sample into groups using k-means. More specifically, we start with a maximum number of clusters, K, and use the k-means algorithm to classify Z_{t-1} into groups (clusters) based on their similarity.[3]

Estimation step: Then, building on Inoue, Jordà, and Kuersteiner (2024), we estimate

[2]See the discussion of Kolesár and Plagborg-Møller (2025) and the replies to the discussants

[3]For more details on Lloyd's algorithm, see Lloyd, Stuart P. "Least Squares Quantization in PCM." IEEE Transactions on Information Theory, vol. 28, 1982, pp. 129–137.

the local projection for all clusters and horizons jointly via a system Generalized Method of Moments (GMM). Specifically, we rewrite equation (2) in matrix form as a system of $H+1$ equations:

$$\boldsymbol{y}_t(H) = \mathbf{C}(\mathbf{H})\mathbf{x}_t + \boldsymbol{v}_t(H) \tag{8}$$

where $\mathbf{D}_t' := [D_{1t} \dots D_{Kt}]$, $\mathbf{w_t}$ is an $m \times 1$ vector that contains control variables,

$$\mathbf{C}(H) = \begin{bmatrix} \mathbf{B}(H) & \mathbf{\Gamma}(H) \end{bmatrix}, \quad \mathbf{x}_t = \begin{bmatrix} \mathbf{D}_t \otimes \varepsilon_t \\ \mathbf{D}_t \otimes \mathbf{w_t} \end{bmatrix}, \tag{9}$$

$\boldsymbol{y}_t(H)$, and $\boldsymbol{v}_t(H)$ are vectors with variables stacked over the horizons $h = 0, \dots, H$, such that $\boldsymbol{y}_t(H) = (y_t, y_{t+1}, \dots, y_{t+H})^\top \in \mathbb{R}^{(H+1)\times 1}$ and $\boldsymbol{v}_t(H) = (v_t, v_{t+1}, \dots, v_{t+H})^\top \in \mathbb{R}^{(H+1)\times 1}$. $\mathbf{B}(\mathbf{H})$ and $\mathbf{\Gamma}(\mathbf{H})$ are matrices that collect the parameters stacked accordingly, and the first $K \times (H+1)$ parameters, β_k^h, correspond to the impulse response estimates for each of the K clusters. The population moment conditions of the system of clustered LP can be written as

$$\mathbb{E}\left[\mathbf{x}_t \cdot (\boldsymbol{y}_t(H) - \mathbf{C}(\mathbf{H})\mathbf{x_t})'\right] = 0.$$

Defining the moment function as $m_t(\theta) = \mathbf{x}_t \otimes (\boldsymbol{y}_t(H) - \mathbf{C}(\mathbf{H})\mathbf{x_t})$, allows us to express the GMM estimator as

$$\hat{\theta} = \arg\min_{\theta} \left[\frac{1}{N}\sum_{t=t_0}^{T^*} m_t(\theta)\right]^\top \hat{V}^{-1} \left[\frac{1}{N}\sum_{t=t_0}^{T^*} m_t(\theta)\right] \tag{10}$$

where $N = T^* - t_0$, t_0 denotes the first observation available after accounting for lags in the control set, $T^* = T - H - 1$, and $\hat{V}$ denotes the optimal weighting matrix, correcting for heteroskedasticity and autocorrelation. Joint estimation of the IRFs using this GMM setup is useful, as it provides an estimate of the covariance matrix, $\hat{\Omega}_\beta$, a key ingredient in the construction of a joint test on the IRFs to determine the optimal number of clusters in the next step.

Evaluation step: After estimating the IRFs for the K clusters, we evaluate whether they differ across clusters via a series of pairwise Wald tests.[4] Specifically, given a horizon of interest to the analyst, $(\tilde{H})$, which may differ from H, let the null hypothesis of equality between the responses at horizon $h = 0, 1, \ldots, \tilde{H}$ for clusters k and k' be given by

$$H_0 : \beta_k^h - \beta_{k'}^h = 0 \text{ for } h = 0, 1, 2, ..., \tilde{H}.$$

Let $\beta = [\ \beta_k^0 \ \ \ldots \ \ \beta_k^H \ \ \beta_{k'}^0 \ \ \ldots \ \ \beta_{k'}^{\tilde{H}}]'$, then the Wald statistic can be expressed as

$$W \ = \ (R\hat{\beta})'\big(R\,\hat{\Omega}_\beta\,R'\big)^{-1}(R\hat{\beta}) \ \sim \ \chi^2_{\tilde{H}+1} \quad \text{under } H_0.$$

To control for multiple testing across all $\frac{K(K-1)}{2}$ possible cluster pairs, we apply a Bonferroni correction by adjusting the significance level to $\alpha_{\text{adj}} = \alpha/[K(K-1)/2]$. Each pairwise comparison is evaluated against the corresponding critical value $\chi^2_{r,1-\alpha_{\text{adj}}}$. If we fail to reject the null for a pair, then we repeat the procedure with $K-1$ clusters. We iterate until we reject the null for all pairs. At each iteration of the procedure, the Bonferroni correction ensures that the probability of any false rejection among the pairwise Wald tests is at most α. Since the procedure performs at most $K-1$ iterations, the probability that any false rejection occurs is at most $(K-1)\alpha$. In practice, this bound is conservative as the procedure typically stops before exhausting all $K-1$ stages.

It is important to note that the iterative procedure does not aim to recover a "true" number of clusters. In general, the data-generating process need not have discrete regimes as the IRF $\beta_t^h(Z_{t-1})$ may vary continuously with the driving variable. In this case, the clustered LP provides a piecewise-constant approximation, and the selected $\hat{K}$ reflects the number of groups whose IRFs are statistically distinguishable at the given sample size and significance level. Two clusters with similar, but not identical, impulse responses will be merged if the difference is

[4] See Kilian and Vigfusson (2011) for a similar pairwise test in a linear setup.

not detectable. Conversely, as the sample size grows and estimation precision improves, finer differences become detectable and $\hat{K}$ may increase.

3 Illustrative Simulations

This section presents simulation results that illustrate the performance of the *clustered LP*. We focus on the case where Z_t is exogenous, thus our objects of interest are the number of states selected by the Wald test and the $CAR^h(\delta, k)$.

We consider a simplified bi-variate model given by:

$$\begin{cases} x_t = \varepsilon_t^x, \\ y_t = \beta_t\left(Z_{t-1}\right)x_t + \gamma_{1,t}(Z_{t-1})\, y_{t-1} + \gamma_{2,t}(Z_{t-1})\, y_{t-2} + \varepsilon_t^y, \end{cases} \tag{11}$$

where we assume the shock of interest, ε_t^x, is observed; the intercepts have been normalized to zero; the innovations, $(\varepsilon_t^x, \varepsilon_t^y)^\top \sim \mathcal{N}(\mathbf{0}, V)$ with $V = I_2$, and parameter evolution is driven by a low-dimensional set of exogenous variables Z_{t-1}.

We present simulations where time-variation is driven by three different DGPs. We employ $M = 10000$ Monte Carlo replications and set $T = 2000$. As is common in time-varying models, we are interested in the IRFs conditional on the time when the shock hits and the initial conditions given by Z_{t-1}. Therefore, to simulate the object of interest, we proceed as follows.

Step 1: Baseline time series for $y_t^{(m)}$. For each Monte Carlo replication ($m = 1, 2, \ldots, M$), we generate a time series $\{y_t^{(m)}\}_{t=1}^T$ from the initial conditions, $(\beta_1^{(m)}, \gamma_{1,1}^{(m)}, \gamma_{2,1}^{(m)})$ using the DGP for Z_t and the model (11). We discard 10,000 observations to ensure a stationary distribution.

Step 2: Computation of the CAR. The object of interest is the response of y_{t+h}, for $h = 0, 1, .., H$, to a shock of size $\delta = 1$ hitting ε_t^x at time t where, to mimic the object of interest in time-varying models, t is any point in the sample. Thus, for every t, we simulate two paths for the outcome variable: a counterfactual baseline path where $x_t = \varepsilon_t^x$ and a perturbed path

where $x_t = \delta + \varepsilon_t^x$. Then, for each group k we compute

$$CAR^h(\delta, k) = \frac{1}{M} \sum_{m=1}^{M} \left[y_{t+h}^{(m)}(\varepsilon_t^x + \delta) - y_{t+h}^{(m)}(\varepsilon_t^x) \,\middle|\, z_{t-1}^{(m)} \in \mathcal{C}_k \right].$$

Step 3: Clustered LP estimation. The clustered LP estimates are obtained using the iterative algorithm described in section 2.2. The initial number of groups is set to $K = 10$, the number of horizons for the Wald test is set to $\tilde{H} = 5$, and the significance level is $\alpha = 0.05$.

3.1 Smooth Transition Threshold Models

We consider two smooth transition threshold models.

Univariate Case: Let the DGP for the time-varying parameters be given by

$$\boldsymbol{\psi}_t(z_{t-1}) = \sum_{k=1}^{K} \xi_k(z_{t-1})\, \boldsymbol{\psi}_k + \boldsymbol{\eta}_t$$

where $\boldsymbol{\psi}_t = (\beta_t,\, \gamma_{1,t},\, \gamma_{2,t})'$ and $\boldsymbol{\eta}_t = (\eta_{1,t},\, \eta_{2,t},\, \eta_{3,t})'$, the number of regimes is $K = 4$, $\xi_k(z_{t-1})$ are weights associated with each regime, and $\eta_{i,t}$, $i = 1, 2, 3$ are uncorrelated i.i.d. normal errors with variance set to 0.0009.[5] To mimic time dependence – often found in macro and financial data – the threshold variable z_t is drawn from a stationary ARMA(p_z, m_z) process:

$$z_t = c_z + \sum_{i=1}^{p_z} \phi_i^z\, z_{t-i} + \varepsilon_t^z + \sum_{j=1}^{m_z} \theta_j^z\, \varepsilon_{t-j}^z, \qquad \varepsilon_t^z \sim \mathcal{N}(0, \sigma_e^2).$$

The calibration uses $p_z = 2$, $m_z = 3$, $c_z = 0$, AR parameters $(\phi_1^z, \phi_2^z) = (0.6, 0.3)$, MA parameters $(\theta_1^z, \theta_2^z, \theta_3^z) = (0.8, 0.7, 0.4)$, and $\sigma_e = 1$.

Four latent regimes are defined by the empirical quartiles of z_t, giving three thresholds (τ_1, τ_2, τ_3) at the 25th, 50th, and 75th percentiles. To ensure stationarity, the regime-specific

[5] A small variance is required for the model to be stationary.

parameters are set to

$$\begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \\ \gamma_{1,1} & \gamma_{1,2} & \gamma_{1,3} & \gamma_{1,4} \\ \gamma_{2,1} & \gamma_{2,2} & \gamma_{2,3} & \gamma_{2,4} \end{bmatrix} = \begin{bmatrix} -1.9 & -0.5 & 0.2 & 0.8 \\ 0.7 & 0.4 & 0.9 & 1.2 \\ 0.1 & 0.2 & -0.1 & -0.3 \end{bmatrix}$$

and the change between regimes is governed by the transition logistic function

$$G_k(z_{t-1};\lambda) \;=\; \frac{1}{1+\exp\big(-\lambda\,[\,z_{t-1}-c_k\,]\big)}\,, \qquad \lambda>0, \tag{12}$$

where setting $\lambda = 5$ ensures a smooth transition between non absorbing regimes, $(c_1, c_2, c_3) = (-4.3359, -0.5981, 3.5717)$. The regime weights are constructed sequentially as $\xi_1(z_{t-1}) = 1 - G_1(z_{t-1};\lambda)$, $\xi_k(z_{t-1}) = G_{k-1}(z_{t-1};\lambda) - G_k(z_{t-1};\lambda)$, for $k = 2, 3$,and $\xi_4(z_{t-1}) = G_3(z_{t-1};\lambda)$, so that $\xi_k \in [0,1]$ and $\sum_{k=1}^{4}\xi_k(z_{t-1}) = 1$ for all z_{t-1}.

Bivariate Case: The second DGP we consider assumes that Z_{t-1} is bivariate and generated by a stationary VARMA(2,3). As in the univariate case, we assume that there are four regimes and that the transition across regimes for each of the variables in Z_t is a logistic transition function. We use of a single threshold for each driving variable to generate four states (see the Online Appendix B for details).

Table 1: Frequency of Clusters from Iterative Procedure

$\hat{K}$	Univariate Threshold	Bivariate Threshold	Absolute Value
2	0%	0%	13.51%
3	0%	12.9%	73.14%
4	91.4%	86.2%	12.8%
5	7.3%	0.8%	0.55%
6	1.1%	0.07%	0%
7	0.2%	0.04%	0%

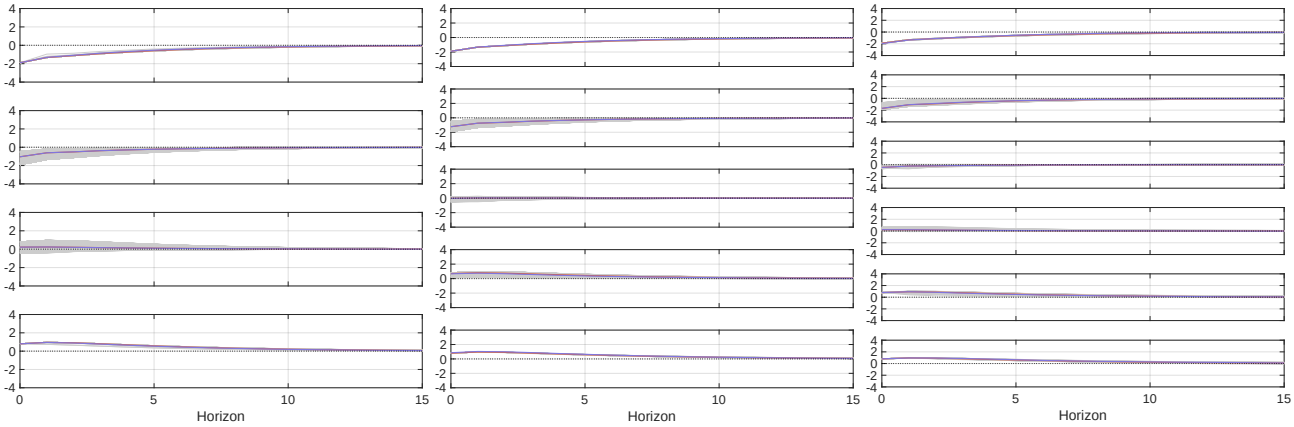
Notes: This table reports the frequency of the number of clusters, $\hat{K}$, estimated with the iterative procedure across 10,000 Monte Carlo replications for the univariate and bivariate threshold DGPs, as well as for the absolute value DGP.

Table 1 reports the frequency of $\hat{K}$ selected. The procedure selects four clusters 91.4% (86.2%) of the time in the univariate (bivariate) model. When the a number of clusters selected differs from the true ($K = 4$), the procedure selects a larger number of clusters ($\hat{K} = 5$ for 7.3% of the simulations) for the univariate DGP, but favors a parsimonious model for the bivariate DGP ($\hat{K} = 3$ for 12.9% of the simulations). Recall that the iterative procedure involves a classification step that is based on Z_{t-1} and an evaluation step that tests differences in the IRFs up to an horizon of interest for the researcher. Therefore, the number of clusters selected can differ from the number of clusters in Z_{t-1}. We will return to this point in section 3.3.

The purpose of the clustered LP is to reduce the dimensionality of the estimation problem so as to facilitate computation of the conditional average response in general time-varying setups while providing a causal summary of the time-varying effects. Figure 1 illustrates the IRF (gray) for each value of z_{t-1} and the partition of the IRFs when the number of groups is set to the true $K = 4$ or the next most likely number of groups ($K = 5, 6$) selected by the iterative procedure. The figure also plots the $CAR^h(1, k) \in C_k$ (red), and the average response estimated by the clustered LP (purple) for each group. Two insights are derived from this figure. First, as expected, the iterative procedure divides the data into different subsamples based on Z_{t-1} and the differences in the IRFs across groups up to $\tilde{H}$. Due to the stationary nature of the DGP, the largest differences are observed on impact. Second, the clustered LP accurately recover the average conditional response for each group. Note how the clustered LP estimates are indistinguishable from the $CAR^h(1, k) \in C_k$.

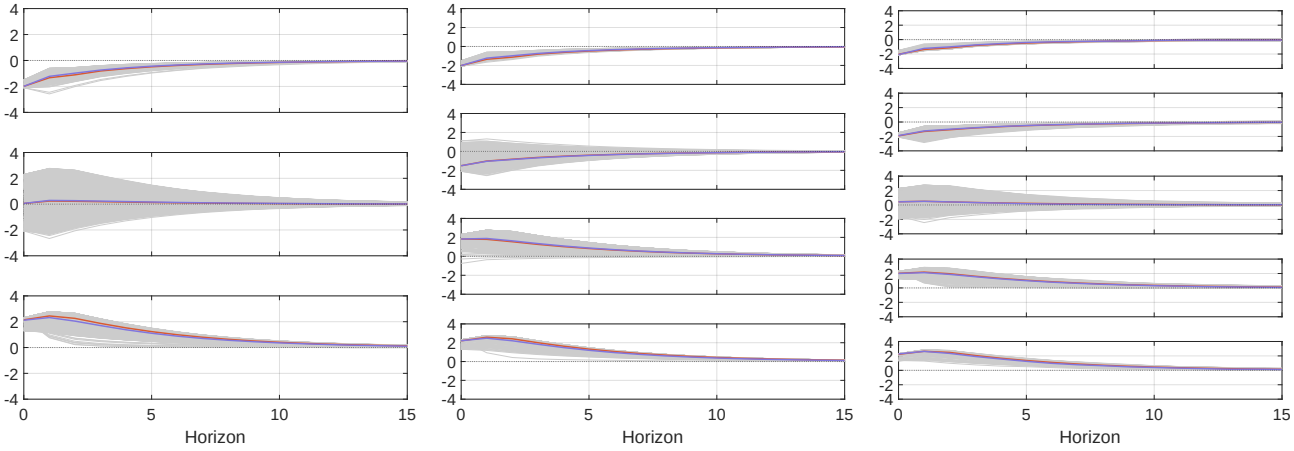
To illustrate a key advantage of our approach, the ability to have a low-dimensional multivariate driving force Z_{t-1}, Figure 2 plots the simulation results for the bivariate smooth threshold model. The figure shows the IRF (gray) for each value of z_{t-1} and the partition of the IRFs when the number of groups is set to the true $K = 4$, the $CAR^h(1, k) \in C_k$ (red), and the average response estimated by the clustered LP (purple) for each group. The figure also reports simulation results for the next most likely number of groups ($K = 3, 5$) selected by the iterative procedure.

Figure 1: Univariate Threshold Model



Notes: This figure illustrates the true impulse response functions for each Z_{t-1} in gray, the clustered LP estimate in (purple), and the $CAR^h(\delta = 1, k)$ (in red) for each partition ($K = 4$ in the left panel, $K = 5$ in the middle panel, and $K = 6$ in the right panel).

Figure 2: Bivariate Threshold Model



Notes: This figure illustrates the true impulse response functions for each Z_{t-1} in gray, the clustered LP estimate in (purple), and the $CAR^h(\delta = 1, k)$ (in red) for each partition ($K = 3$ in the left panel, $K = 4$ in the middle panel, and $K = 5$ in the right panel).

Figure 2 illustrates how the iterative algorithm partitions the data into different subsamples, correspondingly, into different groups of IRFs for the bivariate DGP. The grouping of the IRFs (gray) appears to be closely linked to the impact responses. In addition, the overlap of the clustered LP and the $CAR^h(1,k)$ for each $k \in C_k$ evidences how the proposed estimator recovers the CAR. Lastly, in simulations with a smaller $\hat{K}$ (see left panel where $\hat{K}=3$), the estimator groups together very different dynamic IRFs, which results in clustered LP estimates that are close to zero in the short-run. In contrast, in cases where the number of groups is greater (see the right panel where $\hat{K}=5$), the estimation produces groups similar CARs.

3.2 Alternative Nonlinear transformation of Z_{t-1}

Consider now the case where the autoregressive parameters evolve in a smooth and continuous manner, but in an asymmetric fashion. The DGP is given by:

$$\beta_t(z_{t-1}) = -0.4 + 0.7|z_{t-1}| + \eta_{\beta,t}, \tag{13}$$

$$\gamma_{1,t}(z_{t-1}) = -0.2 + 0.5|z_{t-1}| + \eta_{g1,t}, \tag{14}$$

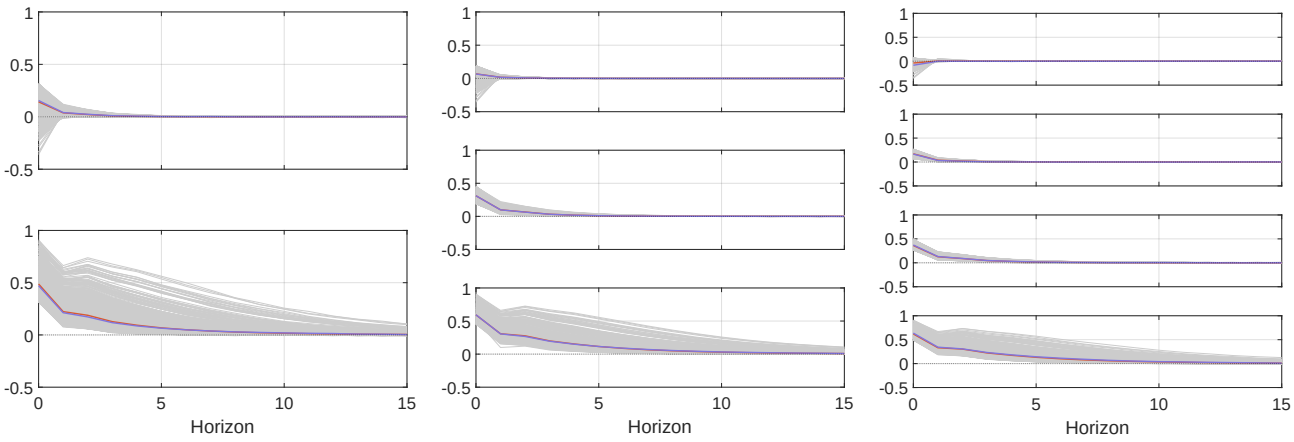
$$\gamma_{2,t}(z_{t-1}) = -0.1 + 0.2|z_{t-1}| + \eta_{g2,t}, \tag{15}$$

where $(\eta_{\beta,t},\ \eta_{g1,t},\ \eta_{g2,t}) \sim \mathcal{N}(\mu\mathbf{1}, \sigma_{\mathrm{p}}^2\mathbf{I})$, are serially uncorrelated with $\mu = 0$, and $\sigma_{\mathrm{p}}^2 = 0.0009$. As in the univariate smooth transition model, $\{z_t\}$ is assumed to follow an ARMA(p_z, m_z) with AR parameters $(\phi_1^z, \phi_2^z) = (0.6, 0.3)$, MA parameters $(\theta_1^z, \theta_2^z, \theta_3^z) = (0.8, 0.7, 0.2)$, $\mu_z = 1$ and $\sigma_z = 0.003$. Although this DGP does not feature discrete regime switching, it serves to illustrate how alternative forms of time-variation can still be meaningfully grouped into a few clusters.

The third column of Table 1 reports the selection frequencies for this DGP where the nonlinear transformation of the driving variable is $|z_{t-1}|$. Note that in this case, where time-variation evolves smoothly and there is no true number of clusters, the iterative procedure favors a small number of clusters: the classification frequency is 13.51% for $K = 2$, 73.14% for

$K = 3$, and 12.8% for $K = 4$. The selection frequency for $K > 4$ is negligible.

Figure 3: Absolute Value Model



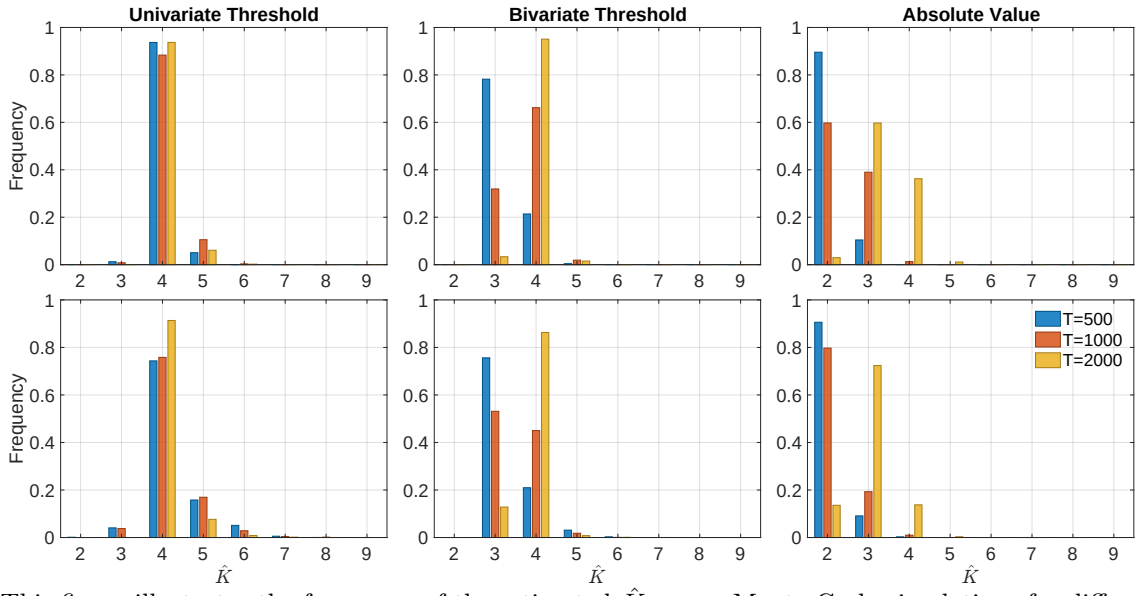
Notes: This figure illustrates the true impulse response functions for each Z_{t-1} in gray, the clustered LP estimate in (purple), and the $CAR^h(\delta = 1, k)$ (in red) for each partition ($K = 2$ in the left panel, $K = 3$ in the middle panel, and $K = 4$ in the right panel).

Figure 3 displays simulation results for this alternative DGP . The response corresponding to each value of z_{t-1} is shown in gray, the $CAR^h(1, k)$ for each $k \in C_k$ is colored red, and the clustered LP estimates are colored purple. The key takeaways from these simulations echo the insights derived from the smooth threshold models. The clustered LP estimates provide a good approximation of the conditional average responses. Due to the stationarity of the DGP, the effect of the shock to ε_t^x dies quickly; therefore, differences between responses across groups are concentrated on impact and on the short term. In summary, the three DGPs illustrate how the clustered LP provides a causal summary of the time-varying effect of interest.

3.3 Identifying Heterogeneity in Small Samples

Our iterative algorithm partition has two components, the initial grouping of the data via k-means on Z_{t-1} and the evaluation step that compares IRFs between groups, leading to the selection of responses that are statistically different given a horizon of interest $\tilde{H}$. Figure 4 reports the frequency of $\hat{K}$ for each of the DGPs, setting $\tilde{H} = 0, 5$ in the evaluation step, and samples of size $T = 500, 1000, 200$.

Figure 4: Frequency of $\hat{K}$ across DGPs for different sample sizes



Notes: This figure illustrates the frequency of the estimated $\hat{K}$ across Monte Carlo simulations for different sample sizes and two values of H in the evaluation step.

For the threshold DGPs, when $\tilde{H} = 0$ the frequency of $\hat{K} = 4$ is higher than for $\tilde{H} = 5$, especially for smaller sample sizes. In other words, if the evaluation step (i.e., the Wald test) is based only on the impact response, then the differences in the IRFs are mainly due to the distinct initial Z_{t-1}. In contrast, when the horizon used in the evaluation step increases ($\tilde{H} = 5$), the system dynamics play a larger role in accounting for differences in the IRFs. For the multivariate threshold and the absolute value DGPs, a key takeaway is that the smaller the number of observations, the more likely the procedure is to select a small number of clusters. This suggests that the iterative algorithm is conservative when the data is limited and avoids splitting the sample into clusters with few observations.

4 The Uncertainty Channel of Monetary Policy Transmission to Treasury Yields

A central question in macroeconomics is whether *monetary policy uncertainty* (i.e., uncertainty around the central bank's course of action) affects the transmission of monetary policy

to financial assets. Several empirical studies have found evidence of an uncertainty channel whereby heightened uncertainty dampens the response of U.S. Treasury yields to monetary policy shocks (see Tillman (2020), Bauer, Lakdawala, and Mueller (2021), and De Pooter et al. (2021)). Specifically, using an event study around FOMC dates, De Pooter et al. (2021) show that the pass-through of monetary policy to medium- and long-term yields is stronger when monetary policy uncertainty is low, with the effect operating mainly through the term premium. They attribute this effect to investors taking larger positions when uncertainty is low; therefore, when a shock occurs, investors are forced to adjust more abruptly, pushing the term premium and yields up by a greater amount. Similar results are found by Bauer, Lakdawala, and Mueller (2021) who link the muted response to a signal extraction problem. That is, under high uncertainty, investors attribute less weight to signals from the Fed and, consequently, monetary policy is less effective. Using state-dependent local projections, Tillman (2020) finds that monetary pollicy transmission to medium- and long-term yields is weaker under high uncertainty. Flight to safety, in the form of heightened investor appetite for the safety of locking in long-term bonds over repeatedly rolling over short-term debt, drives down the compensation required holding longer maturities.

Less is known about how *macroeconomic uncertainty* (i.e., uncertainty about the macroeconomic environment at a given point in time) affects the effectiveness of monetary policy. While increases in *macroeconomic uncertainty* may be linked to increased *monetary policy uncertainty*, other factors such as geopolitical tensions, fiscal policy uncertainty, and real economic shocks can lead to greater macroeconomic uncertainty. Hence, measures of monetary policy and macroeconomic uncertainty do not necessarily move in a synchronous manner. For instance, following the onset of the Global Financial Crisis, the Fed signaled its commitment to expansionary policy, and monetary policy uncertainty declined even as uncertainty about the macroeconomic outlook remained elevated. In contrast, monetary policy uncertainty remained high during the post-COVID inflation surge, while uncertainty about the macroeconomic environment dropped.

We employ our clustered LP method to investigate whether the existing literature on the uncertainty channel of monetary policy transmission to U.S. treasury yields overlooked potential interactions between these different types of uncertainty, which could give rise to time-varying responses. Our approach has several advantages relative to the event studies and state-dependent methods: it does not require the researcher to impose a-priori restrictions about the classification and number of the states; it groups the data taking into account the behavior of both macroeconomic and monetary policy uncertainty, and it does not take a stance on the functional form that drives the time variation but explicitly links this time variation to changes in uncertainty.

4.1 Data Description and Preliminaries

We use monthly data for the U.S. that span February 1988 to December 2023. The daily off-the-run, nominal Treasury zero-coupon bond yields (Gürkaynak, Sack, and Wright (2007)) are obtained from the Federal Reserve.[6] The decomposition of daily yields into a term premium, a future short-term interest rate expectations component, and a residual follow Christensen, Diebold, and Rudebusch (2011); the data are obtained from the Federal Reserve Bank of San Francisco.[7] To construct monthly series, we aggregate the yield, term premium, and expectations component by averaging the monthly observations.

To measure macroeconomic uncertainty, we use the monthly 3-period ahead macroeconomic uncertainty index by Jurado, Ludvigson, and Ng (2015). This index captures the common variation in *macroeconomic uncertainty* (defined as the conditional volatility of the purely unforecastable component of a series) across many macroeconomic variables, such as real output and income, employment and hours, real retail, manufacturing, and trade sales, and consumer spending. We employ Husted, Rogers, and Sun (2017) *monetary policy uncertainty* index, a monthly textual-based index based on articles published in the New York Times, Wall Street

[6]The data are available at https://www.federalreserve.gov/pubs/feds/2006/200628/200628abs.html.

[7]The data are available at: https://www.frbsf.org/research-and-insights/data-and-indicators/treasury-yield-premiums/

Journal, and Washington Post. We standardize both uncertainty indices to facilitate comparison. Data for the CPI inflation (constructed from monthly differences in the CPI series, in percent), and the unemployment rate are obtained from FRED.

The monetary policy shock is identified using a high-frequency approach based on Bauer and Swanson (2023), who use changes in federal funds futures prices in a narrow window around FOMC announcements. Even though we focus on the response of the 5- and 10-year Treasury yields, the Online Appendix C reports results for the 1-year and 2-year yields. We also present results for alternative identification approaches; namely the daily change of 2-year Treasury yields on FOMC days as in Tillman (2020) and the Bauer and Swanson (2023) FOMC frequency series, both aggregated to a monthly series following the weighting scheme in Kilian (2024).

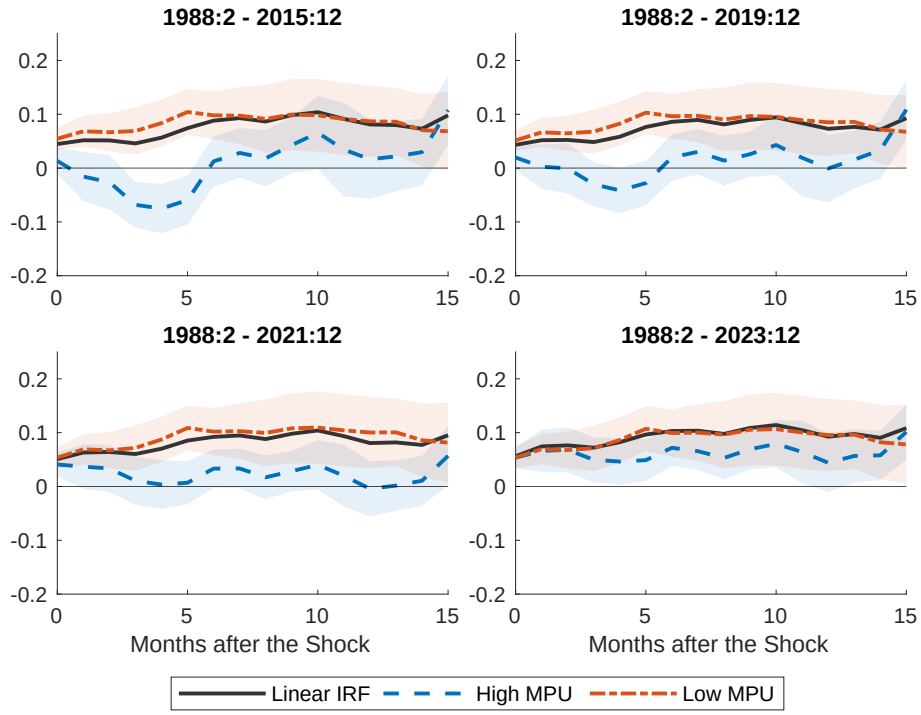
To gather some preliminary insights regarding the connection between *macroeconomic uncertainty*, *monetary policy uncertainty* –hereafter *MPU* and *MacroUncer*, respectively– and the effect of monetary policy shocks on the 5-year yield, we estimate the state-dependent LP:

$$y_{t+h} = I_{t-1}\left[\alpha^h_{(1)} + \beta^h_{(1)}\varepsilon_t + \boldsymbol{\pi}^h_{(1)}\mathbf{w}_t\right] + (1 - I_{t-1})\left[\alpha^h_{(0)} + \beta^h_{(0)}\varepsilon_t + \boldsymbol{\pi}^h_{(0)}\mathbf{w}_t\right] + u_{t+h},$$

where y_{t+h} denotes the 5-year Treasury yield at horizon h, ε_t is the monetary policy shock, $\mathbf{w}_t$ is a vector of control variables, I_{t-1} is a binary state indicator that takes on the value of 1 when $z_{t-1} > 0$ (z_{t-1} equal MPU_{t-1} or $MacroUncer_{t-1}$) and zero otherwise. $\mathbf{w}_t$ includes the first lags of CPI inflation, unemployment, and the 5-year yield. To investigate possible time-variation in the responses, we estimate the model using four expanding subsamples (1988:2-2015:12, 1988:2-2019:12, 1988:2-2021:12, and 1988:2-2023:12) with the last period corresponding to the full sample. For comparison, we report the estimate IRF for a linear LP specification.

Figure 5 shows the response to a one standard deviation monetary policy shock (5.5 basis points) during the high (dashed blue line) and low (dot-dashed red line) *MPU* states, as well as the response estimated with the linear LP (black solid line). Shaded areas represent 68%

Figure 5: State-Dependent LP: Monetary Policy Uncertainty



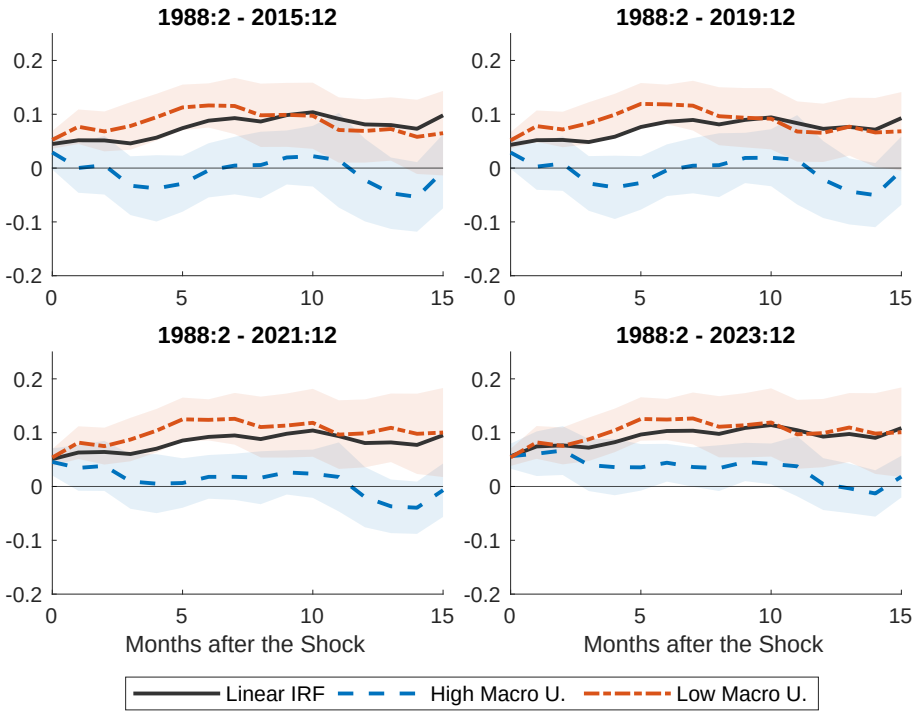
Notes: State-dependent Impulse Response Functions of the 5-year yield to a one-standard deviation monetary policy shock under different monetary policy uncertainty regimes, above (blue line) and below (red line) the mean, with 68% confidence intervals. The impulse response function of the linear model (in black) is plotted for comparison. Each panel represents a different sample.

confidence intervals.[8] The difference between the response in high- and low-uncertainty states appears to be time-varying. Whereas estimates for the earlier samples reveal a clear difference between the IRFs in both states (negative during high MPU and positive during low MPU), this difference is muted in the full sample. This suggests a change in the interaction between MPU and the transmission of monetary policy since (or during) the Covid-19 pandemic. Moreover, the linear LP overestimates the effectiveness of monetary policy during high MPU times.

The estimates reported in Figure 6 are indicative of time-variation in the interaction between *MacroUncer* and the transmission of monetary policy shocks to the 5-year yield. Note how the gap between the IRFs for low- and high-uncertainty states decreases as we extend the sample. For the majority of the subsamples, the response in times of high *MacroUncer* is sta-

[8]Confidence intervals are computed using HAC standard errors.

Figure 6: State-Dependent LP: Macroeconomic Uncertainty



Notes: State-dependent Impulse Response Functions of the 5-year yield to a one-standard deviation monetary policy shock under different macroeconomic uncertainty regimes, above (blue line) and below (red line) the mean, with 68% confidence intervals. The impulse response function of the linear model (in black) is plotted for comparison. Each panel represents a different sample.

tistically insignificant or small at most horizons, while positive when *MacroUncer* is low. Once the Covid-19 period is included (1988:2-2021:12 and 1988:2-2023:12), the response on impact and at short horizons turns positive.

To summarize, state-dependent LP estimates suggest that the uncertainty channel of monetary policy transmission is time-varying and multifaceted, with high monetary policy and macroeconomic uncertainty attenuating the effectiveness of monetary policy.

4.2 How effective is monetary policy in uncertain times?

We investigate the role of both forms of uncertainty of the yields response to monetary policy via the clustered local projection model described in (2), where Z_{t-1} includes the *MPU* and the *MacroUncer* indices, and the controls comprise lagged CPI inflation, lagged unemployment

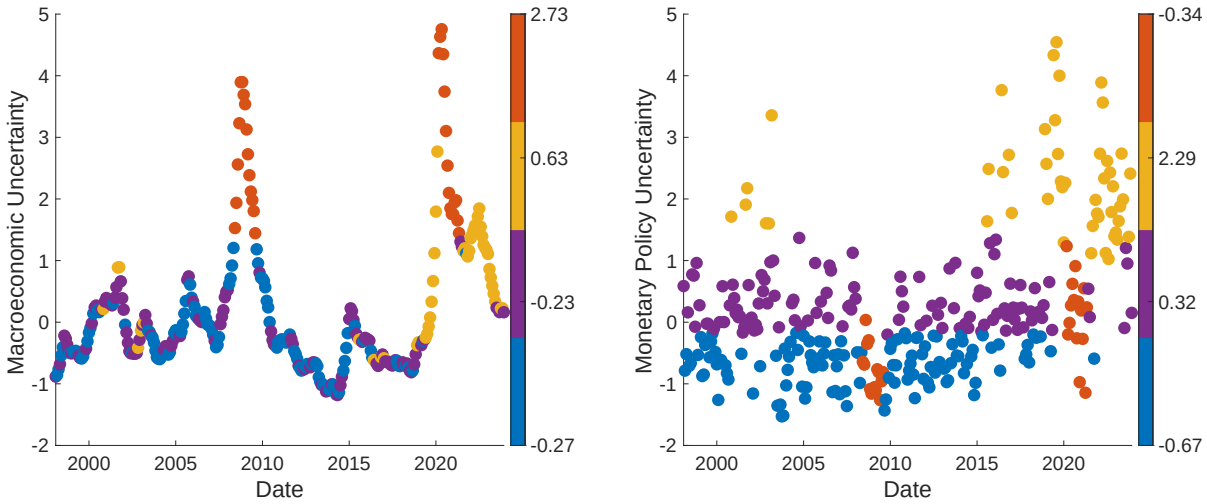
and the first lag of the dependent variable. The iterative procedure starts with 10 clusters and, testing the null of pairwise equality for $h = 0, 1, ..., 15$ in the third step of the algorithm, yields a total of $\hat{K} = 4$ clusters.

Table 2: Estimated cluster means and their labeling

Cluster	**Mean Macro U.**	**Mean MPU**
1: Low macro uncertainty, low MPU	-0.27	-0.67
2: Low macro uncertainty, moderate MPU	-0.23	0.32
3: Moderate macro uncertainty, high MPU	0.63	2.29
4: High macro uncertainty, low MPU	2.73	-0.34

Table 2 summarizes the means for each of the four clusters and their labeling. Recall that we standardize both indices, thus positive values represent uncertainty above the historical mean (hereafter high uncertainty) and negative values represent uncertainty below the mean (hereafter low uncertainty). Note how the different groups have distinctive characteristics. For instance, cluster one corresponds to macro and monetary policy uncertainty below the mean whereas cluster three corresponds to moderate macro and high monetary policy uncertainty.

Figure 7: Classification of Macroeconomic and Monetary Policy Uncertainty Indices



Notes: This figure reports the classification of the macroeconomic (left panel) and monetary policy uncertainty (right panel) indices into the $K = 4$ clusters.

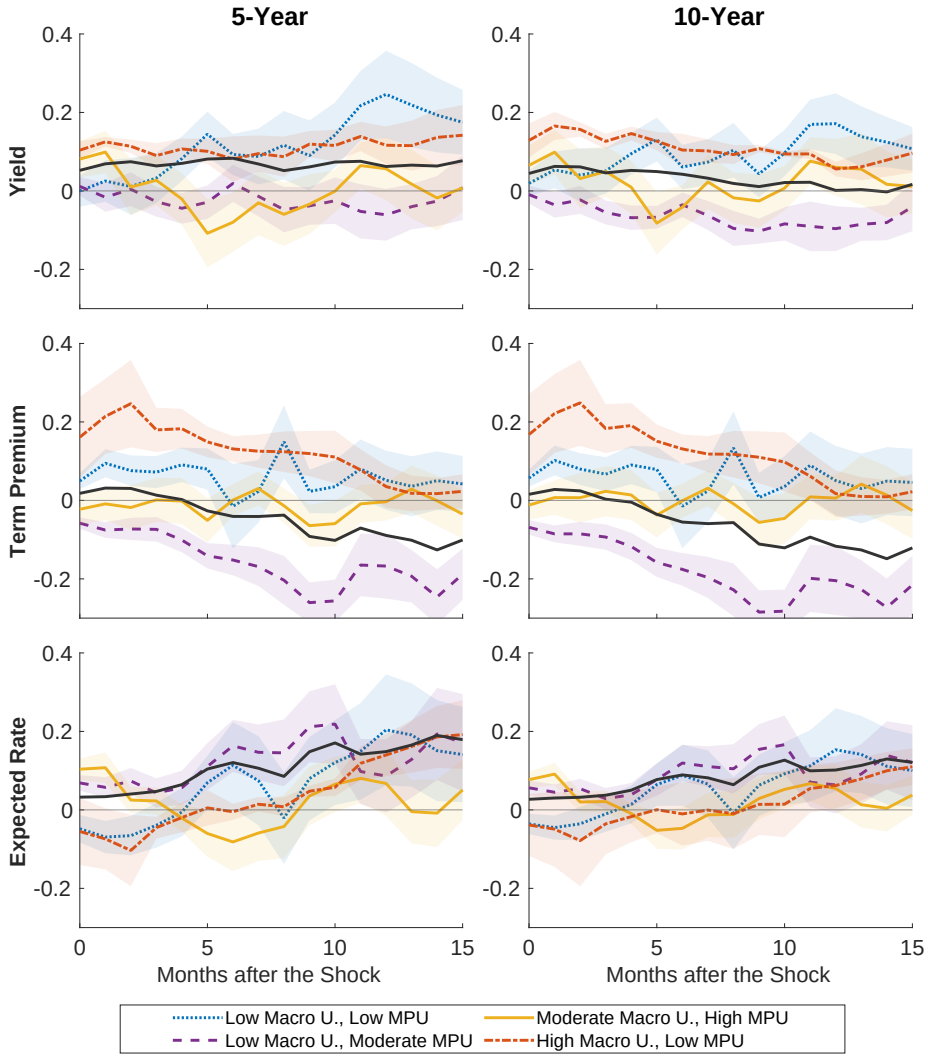
To visualize how the historical data are partitioned by the iterative procedure, the scatterplots in Figure 7 map the uncertainty data series to the four clusters. As the figure illustrates,

some periods of high *MacroUncer* (e.g., the Great Recession and the Covid-19 pandemic) were also periods of low *MPU* (cluster 4 in red). In contrast, the months preceding and following the pandemic correspond to high *MPU* and moderate *MacroUncer* (cluster 3 in yellow). This cluster includes the period when banks took initial write-downs (October 2007) and the Fed's announcement of QE2 (November 2010). Furthermore, note that clustering based on only one of the two uncertainty indices–as effectively done in the state-dependent models–would overlook periods where monetary policy and macroeconomic uncertainty do not move in a synchronous manner.

We now turn to the cluster LP estimates. Figure 8 depicts the IRFs of the 5- and 10-year U.S. Treasury nominal yields and their components to a one–standard deviation contractionary monetary policy shock for each of the four clusters. For comparison, we also plot the response estimated by a linear LP (solid black line). Our results reveal an interesting pattern regarding the relative importance of each type of uncertainty for the transmission of monetary policy shocks. On impact and in the very short-run, *MacroUncer* is the main driver of heterogeneity in the response of the yields. In fact, the impact responses during high *MacroUncer* (clusters 3 and 4 in solid yellow line, and dash-dotted red line, respectively) are positive and statistically significant, whereas the responses in times of low *MacroUncer* (clusters 1 and 2 in dotted blue line, and dashed purple line, respectively) are not. However, at longer horizons, *MPU* drives the responsiveness of yields. Note how the response of the yield is an order of magnitude higher twelve months after the shock when *MPU* is low (clusters 1 and 4) than when it is high (clusters 2 and 3). Of particular interest is the linear LP estimate, which corresponds to the average IRF over the sample and underestimates (overestimates) the response of the yields during low *MPU* and overestimates the response during low *MacroUncer* and moderate *MPU*.

To dig deeper into the mechanisms that drive this heterogeneity, we estimate the responses of the term premium and the expected rate component. On impact, the effect of a contractionary monetary policy shock on the term premiums is heterogeneous across clusters, but the average responses of the expectations component are similar for clusters with high (low)

Figure 8: The Effect of Contractionary Monetary Policy Shocks Across Uncertainty Regimes



Notes: The figure reports clustered Local Projections Impulse Response Functions of the 5- and 10-year yield and its components to a one-standard deviation monetary policy shock under different monetary policy and macroeconomic uncertainty regimes. Shaded areas represent 68% confidence intervals. The estimated linear LP response is plotted in black.

MPU. To gather some insight regarding the role of heightened uncertainty in the transmission of monetary policy, let us focus on the states where either of the indices is high. Comparing the average responses of the yields and the expectations components during times of high *MPU* and moderate *MacroUncer* (solid yellow line) reveals that short-term dynamics are driven by the expectations component. In contrast, the positive response of the term premium explains the short-run increase experienced by the yields during times of high *MacroUncer* (dash-dotted red

line); indeed, the increase in the term premium more than offsets the decline in the expected rate on impact and on the first three months following the shock. Moreover, the linear LP underestimates (overestimates) the response of the term premium (expectations during times of low *MPU*.

Our estimates suggest that the mechanisms through which macroeconomic and monetary policy uncertainty affects each component of the yield differ. The behavior of the term premium is consistent with the view that macroeconomic uncertainty amplifies the risk compensation required by investors following a contractionary shock. In times of high macroeconomic uncertainty, the distribution of possible future states is more widespread: growth, inflation, and economic policy could swing in multiple directions, making the future path of short-term rates inherently harder to predict. Even a surprise rate hike may increase market disagreement—some investors may fear an overshoot into recession, while others may interpret it as a signal of stronger inflation risks—further increasing perceived risk around the future rate path and pushing the term premium upward. This explains why clusters 3 (4), characterized by moderate (high) *MacroUncer*, display the largest term premium responses. Nevertheless, in times of similar macroeconomic uncertainty, *MPU* is associated with a dampening force on the term premium, consistent with the flight-to-safety channel highlighted by Tillman (2020): when the future course of policy is unclear, long-term securities become relatively more attractive as safe havens, compressing the additional compensation demanded by investors. This is reflected in the smaller term premium response of cluster 3 relative to cluster 4, and of cluster 2 relative to cluster 1.

Taken together, these results suggest that macroeconomic and monetary policy uncertainty operate through complementary but distinct channels: the former primarily amplifies the risk compensation embedded in the term premium, while the latter governs the speed and persistence with which markets revise their expectations about the future rate path following a monetary policy shock. Overall, this analysis demonstrates the potential of our IRF estimation approach to capture the nonlinearity in the interaction between macroeconomic and monetary policy

uncertainty, offering insights that would be difficult to obtain from standard two-state models conditioning on a single uncertainty measure or a linear LP.

5 Conclusion

This paper proposed a new method to estimate impulse response functions in time-varying parameter models, which we term clustered local projections (LP). We showed that the clustered LP recover the conditional average response CAR if the set of variables that drive the time-variation are exogenous and the conditional marginal response CMR when they are endogenous. Thus, the impulse response estimands from the clustered LP provide a causal summary of the nonlinear effect in each cluster.

From a methodological perspective, we proposed an iterative estimation approach that comprises three steps: classification using the k-means algorithm, joint estimation of impulse responses for all clusters and horizons via local projections, and evaluation. Monte Carlo simulations demonstrated that the clustered LP estimator recovers accurately CAR in smooth threshold models, as well as in specifications in which the parameters depend on the driving variable in an asymmetric fashion. They also illustrated the performance of the iterative algorithm in smaller samples and when alternative horizons are employed in the evaluation step.

We used clustered LP to study how effective contractionary monetary policy is in shifting the 5- and 10-year US treasury nominal yields during uncertain times. Our estimates provided evidence that the mechanisms whereby macroeconomic and monetary policy uncertainty affect the U.S. Treasury yields differ. On the one hand, during times of high macroeconomic uncertainty, contractionary monetary policy is associated with a significant increase in the term premium, which, in turn, accounts for the increase in the yields. On the other hand, during times of high monetary policy uncertainty, contractionary policy leads to an increase in expected rates on impact and shortly after the shock, but to a decline at longer horizons. This response, in con-

junction with a moderate increase in the term premium, accounts for the short-lived response of the yields.

References

Aastveit, Knut Are, Gisle James Natvik, and Sergio Sola. 2017. "Economic uncertainty and the influence of monetary policy." *Journal of International Money and Finance* 76:50–67.

Barnichon, Regis, Davide Debortoli, and Christian Matthes. 2022. "Understanding the size of the government spending multiplier: It's in the sign." *The Review of Economic Studies* 89 (1):87–117.

Bauer, Michael D, Aeimit Lakdawala, and Philippe Mueller. 2021. "Market-Based Monetary Policy Uncertainty." *The Economic Journal* 132 (644):1290–1308.

Bauer, Michael D. and Eric T. Swanson. 2023. "A Reassessment of Monetary Policy Surprises and High-Frequency Identification." *NBER Macroeconomics Annual* 37:87–155.

Chan, Joshua C.C., Eric Eisenstat, and Rodney W. Strachan. 2020. "Reducing the state space dimension in a large TVP-VAR." *Journal of Econometrics* 218 (1):105–118.

Chen, Rong and Ruey S. Tsay. 1993. "Functional-Coefficient Autoregressive Models." *Journal of the American Statistical Association* 88 (421):298–308.

Christensen, Jens H.E., Francis X. Diebold, and Glenn D. Rudebusch. 2011. "The Affine Arbitrage-Free Class of Nelson-Siegel Term Structure Models." *Journal of Econometrics* 164(1):4–20.

Cogley, Timothy and Thomas J Sargent. 2005. "Drifts and volatilities: monetary policies and outcomes in the post WWII US." *Review of Economic dynamics* 8 (2):262–302.

De Pooter, Michiel, Giovanni Favara, Michele Modugno, and Jason Wu. 2021. "Monetary policy uncertainty and monetary policy surprises." *Journal of International Money and Finance* 112:102323.

Diebold, Francis X., Jae-Ha Lee, and Gretchen Weinbach. 1994. "Regime Switching with Time-Varying Transition Probabilities." In *Nonstationary Time Series Analysis and Cointegration*, edited by Colin Hargreaves, Advanced Texts in Econometrics. Oxford: Oxford University Press, 283–302.

Dong, Ding, Zheng Liu, Pengfei Wang, and Min Wei. 2024. "Inflation disagreement weakens the power of monetary policy." Tech. rep., Federal Reserve Bank of San Francisco.

Falck, E., M. Hoffmann, and P. Hürtgen. 2021. "Disagreement about inflation expectations and monetary policy transmission." *Journal of Monetary Economics* 118:15–31.

Fischer, Manfred M., Niko Hauzenberger, Florian Huber, and Michael Pfarrhofer. 2023. "General Bayesian time-varying parameter vector autoregressions for modeling government bond yields." *Journal of Applied Econometrics* 38 (1):69–87.

Gonçalves, Sílvia, Ana María Herrera, Lutz Kilian, and Elena Pesavento. 2024a. "State-dependent local projections." *Journal of Econometrics* :105702.

———. 2024b. "Nonparametric Local Projections¨" *Working Paper* .

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright. 2007. "The U.S. Treasury yield curve: 1961 to the present." *Journal of Monetary Economics* 54 (8):2291–2304. URL `https://www.sciencedirect.com/science/article/pii/S0304393207000840`.

Hamilton, James D. 1989. "A new approach to the economic analysis of nonstationary time series and the business cycle." *Econometrica: Journal of the econometric society* :357–384.

Hauzenberger, N., F. Huber, G. Koop, and J. Mitchell. 2023. "Bayesian Modeling of TVP-VAR Using Regression Trees." *Working Paper* .

Husted, Lucas, John Rogers, and Bo Sun. 2017. "Monetary Policy Uncertainty." *International Finance Discussion Papers 1215* 1215.

Inoue, Atsushi, Òscar Jordà, and Guido M. Kuersteiner. 2024. "Inference for Local Projections." *Federal Reserve Bank of San Francisco Working Paper 2024-29* .

Inoue, Atsushi, Barbara Rossi, and Yiru Wang. 2024. "Local projections in unstable environments." *Journal of Econometrics* :105726.

Jordà, Ò. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95 (1):161–182.

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. 2015. "Measuring Uncertainty." *American Economic Review* 105 (3):1177–1216. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20131193`.

Kilian, Lutz. 2024. "How to construct monthly VAR proxies based on daily surprises in futures markets." *Journal of Economic Dynamics and Control* 168:104966. URL `https://www.sciencedirect.com/science/article/pii/S0165188924001581`.

Kilian, Lutz and Robert J. Vigfusson. 2011. "Are the responses of the U.S. economy asymmetric in energy price increases and decreases?" *Quantitative Economics* 2 (3):419–453.

Klepacz, Matthew. 2021. "Price Setting and Volatility: Evidence from Oil Price Volatility Shocks." *Board of Governors of the Federal Reserve System International Finance Discussion Papers* .

Kolesár, Michal and Mikkel Plagborg-Møller. 2025. "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly." *Journal of Business & Economic Statistics* 43 (4):737–754.

Maung, Kenwin and Nathaniel Yoshida. 2025. "Nonparametric Time-Varying Local Projections." Working paper, Rutgers University.

Primiceri, G. 2005. "Time Varying Structural Vector Autoregressions and Monetary Policy." *Review of Economic Studies, 72* :821–852.

Ramey, V. A. and S. Zubairy. 2018. "Government spending multipliers in good times and in bad: Evidence from U.S. historical data." *Journal of Political Economy 126* :850–901.

Tillman, Peter. 2020. "Monetary Policy Uncertainty and the Response of the Yield Curve to Policy Shocks." *Journal of Money, Credit and Banking* 52 (4):803–833.

A Appendix

Proof of Proposition 1

Part (i): Because the clustered LP (2) is fully interacted and by the FWL theorem together with Assumption 1, the OLS estimand β_k^h equals the within-cluster regression coefficient:

$$\beta_k^h \;=\; \frac{\mathbb{E}[y_{t+h}\,\varepsilon_t \mid D_k = 1]}{\mathbb{E}[\varepsilon_t^2 \mid D_k = 1]}. \tag{16}$$

This is the population OLS coefficient from regressing y_{t+h} on ε_t using only observations in the subsample $\{t : D_k = 1\}$. In addition, since $D_k = \mathbf{1}\{Z_{t-1} \in \mathcal{C}_k\}$ is a deterministic function of Z_{t-1} and ε_t is independent of $\mathcal{F}_{t-1}$ (Assumption 1(a)), we have $\varepsilon_t \perp\!\!\!\perp Z_{t-1}$ and therefore $\varepsilon_t \perp\!\!\!\perp D_k$ for all k. The assumption 1 and the timing of the clustering also ensure that a conditional independence assumption is satisfied so that $\varepsilon_t \perp\!\!\!\perp U_{h,t+h} \mid D_k = 1$.

Using the structural representation (3) and the conditional independence:

$$\begin{aligned} g_{h,k}(e) \;&\equiv\; \mathbb{E}[y_{t+h} \mid \varepsilon_t = e,\, D_k = 1] \\ &=\; \mathbb{E}[\psi_h(e, U_{h,t+h}) \mid \varepsilon_t = e,\, D_k = 1] \\ &=\; \mathbb{E}[\psi_h(e, U_{h,t+h}) \mid D_k = 1] \;=\; \Psi_k^h(e). \end{aligned} \tag{17}$$

The second equality substitutes (3). The third equality uses $\varepsilon_t \perp\!\!\!\perp U_{h,t+h} \mid D_k = 1$: since $U_{h,t+h}$ is independent of ε_t given $D_k = 1$, conditioning on $\varepsilon_t = e$ does not alter the distribution of $U_{h,t+h}$. Thus, the within-cluster regression function $g_{h,k}$ identifies the conditional average structural function Ψ_k^h. The proof of Proposition 1 follows directly from Proposition 1 of Kolesár and Plagborg-Møller (2025) given our Assumption 1. In particular,

$$\beta_k^h \;=\; \frac{\mathbb{E}[y_{t+h}\,\varepsilon_t \mid D_k = 1]}{\mathbb{E}[\varepsilon_t^2 \mid D_k = 1]} \;=\; \int_I \omega_k(e)\, g'_{h,k}(e)\, de, \tag{18}$$

where

$$\omega_k(e) \;=\; \frac{\mathrm{Cov}\Big(\mathbf{1}\{\varepsilon_t \geq e\},\, \varepsilon_t \,\Big|\, D_k = 1\Big)}{\mathrm{Var}(\varepsilon_t \mid D_k = 1)}. \tag{19}$$

By Kolesár and Plagborg-Møller (2025), Proposition 1, $\omega_k(e) \geq 0$ for all e, $\int_I \omega_k(e)\, de = 1$, and ω_k is hump-shaped, peaking near $\mathbb{E}[\varepsilon_t \mid D_k = 1]$. Furthermore, ω_k depends only on the distribution of $\varepsilon_t \mid D_k = 1$, not on y_{t+h} or h.

Substituting the identification result (17) into (18):

$$\beta_k^h \;=\; \int_I \omega_k(e)\, \Psi_k^{h\prime}(e)\, de, \tag{20}$$

establishing (6).

Part (ii): Part (ii) follows from Part (i) by showing that the conditional average structural function $\Psi_k^h(e)$ is linear in e when Z_t is exogenous. When $Z_t \perp\!\!\!\perp \varepsilon_t$ for all t, the future path $\{Z_t, Z_{t+1}, \ldots, Z_{t+h-1}\}$ does not depend on the realization $\varepsilon_t = e$. Hence all time-varying coefficients $\{\beta_t^h(Z_{t-1}), \gamma_t^h(Z_{t-1})\}_{h=0}^H$ are independent of e. This is the direct analogue of Gonçalves et al. (2024a), where linearity of the potential outcome in e follows from the absence of e-dependence in the future states. It follows that $\psi_h(e, U_{h,t+h})$ is linear in e, e.g. $\psi_h(e, U_{h,t+h}) = A_h(U_{h,t+h}) + \Lambda_h(U_{h,t+h})\, e.$, where $A_h(U_{h,t+h})$ and $\Lambda_h(U_{h,t+h})$ are a random variables that depend on $U_{h,t+h}$ but not on e or δ. Specifically, $\Lambda_h(U_{h,t+h})$ aggregates the cumulative effect of the shock through the chain of time-varying coefficients along the path from t to $t+h$. Then

$$y_{t+h}(e+\delta) - y_{t+h}(e) \;=\; \Lambda_h(U_{h,t+h})\;\delta, \tag{21}$$

In addition, by linearity, the conditional average structural function is

$$\Psi_k^h(e) \;=\; \mathbb{E}[\psi_h(e, U_{h,t+h}) \mid D_k = 1] \;=\; \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid D_k = 1] \cdot e, \tag{22}$$

so its derivative $\Psi_k^{h\prime}(e) = \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid D_k = 1]$ is *constant* in e. Substituting into (6) from Part (i):

$$\beta_k^h \;=\; \int_I \omega_k(e)\, \Psi_k^{h\prime}(e)\, de \;=\; \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid D_k = 1] \cdot \underbrace{\int_I \omega_k(e)\, de}_{=1} \;=\; \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid Z_{t-1} \in \mathcal{C}_k]. \tag{23}$$

Taking expectations in (21) conditional on $Z_{t-1} \in \mathcal{C}_k$:

$$\mathrm{CAR}^h(\delta, k) \;=\; \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid Z_{t-1} \in \mathcal{C}_k] \cdot \delta. \tag{24}$$

Therefore

$$\beta_k^h \;=\; \mathbb{E}[\Lambda_h(U_{h,t+h}) \mid Z_{t-1} \in \mathcal{C}_k] \;=\; \frac{\mathrm{CAR}^h(\delta, k)}{\delta} \;=\; \mathrm{CAR}^h(1, k) \;=\; \mathrm{CMR}^h(k) \tag{25}$$

establishing (7).