TI 2024-016/III
Tinbergen Institute Discussion Paper

A Score-Driven Filter for Causal Regression Models with Time-Varying Parameters and Endogenous Regressors

Francisco Blasques[1]
Noah Stegehuis[2]

1 Vrije Universiteit Amsterdam and Tinbergen Institute

2 Vrije Universiteit Amsterdam and Tinbergen Institute

A Score-Driven Filter for Causal Regression Models with Time-Varying Parameters and Endogenous Regressors*

Francisco Blasques[1, 2] and *Noah Stegehuis*[1, 2, †]

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]Tinbergen Institute, The Netherlands

Abstract

This paper proposes a score-driven model for filtering time-varying causal parameters through the use of instrumental variables. In the presence of suitable instruments, we show that we can uncover dynamic causal relations between variables, even in the presence of regressor endogeneity which may arise as a result of simultaneity, omitted variables, or measurement errors. Due to the observation-driven nature of score models, the filtering method is simple and practical to implement. We establish the asymptotic properties of the maximum likelihood estimator and show that the instrumental-variable score-driven filter converges to the unique unknown causal path of the true parameter. We further analyze the finite sample properties of the filtered causal parameter in a comprehensive Monte Carlo exercise. Finally, we reveal the empirical relevance of this method in an application to aggregate consumption in macroeconomic data.

Keywords: observation-driven models, time-varying parameters, causal inference, endogeneity, instrumental variables

---

*This paper has been presented at the 2023 International Association for Applied Econometrics Annual Conference in Oslo (June 2023) and the International Conference on Econometrics and Statistics in Tokyo (August 2023). The authors thank participants at those conferences, as well as Paolo Gorgi, Siem Jan Koopman, Andre Lucas and Janneke van Brummelen for helpful comments and suggestions. Francisco Blasques and Noah Stegehuis acknowledge the financial support of the Dutch Science Foundation (NWO) under grant Vidi.195.099.

†Corresponding author: Noah Stegehuis (n.stegehuis@vu.nl).

1 Introduction

Establishing causal relationships between relevant variables is fundamental in economics and other fields of science. For example, causal inference is key for understanding of the effects of Fiscal and Monetary policy in macroeconomics. It is however well known that standard predictive methods used in the econometric, statistical and machine learning literature will typically fail to uncover causal relationships when dealing with observational data, due to regressor endogeneity. As a result, considerable effort has been made to develop new and effective causal inference techniques. This effort, carried over the last few decades, as been recently recognized by the 2021 Nobel Memorial Prize in Economic Sciences awarded to Joshua Angrist and Guido Imbens. In particular, a range of solutions have been proposed in the literature for causal inference of observational data that typically suffers from endogeneity issues. Important examples include the use of instrumental variables (IV) (Angrist, Imbens, & Rubin, 1996), difference in differences (Ashenfelter & Card, 1985; Bertrand, Duflo, & Mullainathan, 2004) and regression discontinuity design (Thistlethwaite & Campbell, 1960) among others (e.g. synthetic control (Abadie & Gardeazabal, 2003), propensity score matching (Rosenbaum & Rubin, 1983)).

When dealing with time-series or panel data, it is important to note however that these causal inference methods generally attempt to uncover causal relationships that are assumed to be *time-invariant*. They are not designed to keep track of time-varying causal relationships, and do not focus on modeling that time-variation or producing dynamic forecasts of future cause and effect interactions. This can be, of course, a shortcoming in a number of applications. Much like in other scientific domains, in economics causal relationships can change over time both qualitatively and quantitatively. For example, the effectiveness of different fiscal and monetary policies may change substantially over time as they depend on the historical political, social, economic, technological and institutional context. In many practical applications, parameters can thus be time-varying and the need may arise to filter such parameters in order to track the evolution of the true causal effect and potentially forecast it.

The literature on regression models featuring time-varying parameters is a rich one, and so is the closely related literature on filtering techniques. Both are however fundamentally focused on predictive models, designed to improve predictive and/or forecast accuracy. An interesting exception includes the identification of the time-varying effects of monetary policy on macroeconomic variables by Kang, Ratti, and Yoon (2015); Pereira and Lopes (2014), which is motivated by the findings of several studies that the effect of monetary policy actions changed over the years (Boivin, Kiley, & Mishkin, 2010).

In this paper we propose a novel score-driven filtering method (Creal, Koopman, & Lucas, 2013; Harvey, 2013) featuring instrumental variables to estimate the time-varying parameter in a regression model in which regressors are endogenous. We model the endogeneity with a control function approach (e.g. Heckman and Robb (1985); Wooldridge (2015)) if suitable instruments are available. We show consistency of the two-step Maximum Likelihood Estimator (MLE) and show that the filter with estimated parameters converges to the unique path of the true unknown causal parameter when endogeneity is present. In a Monte Carlo simulation study we analyse the finite sample properties for various types and levels of endogeneity. The procedure is straightforward to implement, and works considerably well when valid instruments are available. In simulations we observe that the filter manages to uncover the true path even in non-stationary settings like a structural break at an unknown point in time.

Similar to IV methods in general, our proposed method can potentially handle endogeneity originating from multiple causes. Important ones include: (i) simultaneity, which emerges when two variables are contemporaneously causally linked to each other (Haavelmo, 1943; Kennan, 1989); (ii) omitted variables, which occurs when regressors are correlated with other relevant regressors that are omitted from the regression model (Wooldridge, 2009); and (iii) measurement errors in relevant regressors (Bound, Brown, & Mathiowetz, 2001). Each of these conditions, and several others, result in endogeneity which renders predictive models unsuitable as tools for distilling causal effects (Haavelmo, 1943; Wooldridge, 2002).

Alternative regression models with time-varying parameters and endogenous regressors have been proposed in the literature. C.-J. Kim (2006) and Y. Kim and Kim (2011) propose joint and two-step estimation procedures of the Kalman Filter with a similar control function approach to handle endogeneity. Another approach that has recently been proposed by Giraitis, Kapetanios, and Marcellino (2021) uses a kernel based technique to estimate a time-varying IV estimator. Inoue, Rossi, and Wang (2022) propose a time-varying IV framework for Local Projections, based on the work of Müller and Petalas (2010), who show that for nonlinear non-Gaussian parameter-driven models with moderate time-variation, the sample information can be approximated by a linear Gaussian model that contains the scores of the likelihood as observations. This fact further motivates our use of score functions in the filter that estimates the time-varying parameter. The difference in performance between the score-driven filters and the method proposed by Müller and Petalas (2010) has been analysed in Calvori, Creal, Koopman, and Lucas (2017) in the context of a time-varying parameter test.

Our score-driven filter stands out in its simplicity of implementation and ability to produce robust and reliable path estimates in nonlinear non-Gaussian settings. Compared to the parameter-driven models, such as the Kalman filter, observation-driven models like score-driven models are easier to implement and computationally less intensive, especially in nonlinear non-Gaussian parameter settings (Koopman, Lucas, & Scharth, 2016). Compared to non-parametric kernel methods, our parametric score-driven approach will naturally stand out in small-sample problems and forecasting exercises. Naturally, non-parametric methods can have advantages in terms of flexibility, but they will also they require a choice of hyperparameters such as kernel related bandwidths.

Our filter fits more generally in the class of (quasi) score driven filters, as introduced by Creal et al. (2013), Harvey (2013) and Blasques, Francq, and Laurent (2023). The score driven method has been shown to have optimality properties over other methods such as the Kalman Filter. In particular, the updating scheme for a time-varying parameter is optimal in information theoretic sense if and only if it contains the score of the likelihood (Blasques, Koopman, & Lucas, 2015, 2018; Blasques, Lucas, & van Vlodrop, 2021; Beutner, Lin, & Lucas, 2023). A similar optimality result holds for an analogous time-varying parameter approach in the context of the general method of moments (GMM) framework (Creal, Koopman, Lucas, & Zamojski, 2018). Although the score-driven approach is a very general structure and encompasses many different models with time-varying parameters, not much attention has yet been paid to adopt this procedure to time-varying regression models, with the exception of Blasques, Koopman, and Lucas (2020), Gorgi, Koopman, and Schaumburg (2017), in which autoregressive frameworks are considered. The time-varying parameter regression model with a score driven update has however been investigated by Thiele and Harvey (2013), from an angle of time-varying correlation and in a finance setting by Blasques, Francq, and Laurent (2022), who introduce a time-varying beta model with GARCH dynamics. Nevertheless, thus far no score-driven filter has been developed that can be used in the case of regressor endogeneity.

The rest of the paper is structured as follows. In Section 2 we describe the model and introduce the Instrumental Variables Score filter (IV-score). In Section 3 we analyse the stochastic properties of the filter. In Section 4 we derive the asymptotic properties of the maximum likelihood estimator. Section 5 we devote to an extensive simulation study to show filter behaviour. In Section 6 we demonstrate the emperical relvance by appling the filter to estimate the excess sensitivity of consumption to income.

2 Causal score-driven filtering model

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series generated according to

$$y_t = \beta_t x_t + \varepsilon_t, \tag{1}$$

where $\{x_t\}_{t\in\mathbb{Z}}$ is a stochastic regressor, $\{\beta_t\}_{t\in\mathbb{Z}}$ is a time-varying parameter, and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a mean-zero identically distributed error term with density p_ε, indexed by static parameter vector λ. We assume the regressors are endogenous, hence the usual exogeneity condition fails $\mathbb{E}(\varepsilon_t|x_t) \neq 0$. The specific cause of endogeneity can be left unspecified, but it could emerge from the usual culprits, ranging from simultaneity between y_t and x_t, omitted variables h_t which are correlated with the regressor x_t, a measurement error in the regressor x_t, functional form misspecification, etc.

In the quasi score-driven approach, the filter for the time-varying parameter β_t is given by

$$\beta_{t+1} = \omega + \alpha s_t + \gamma \beta_t, \tag{2}$$

for fixed unknown parameters ω, α, γ, and where s_t is the scaled score,

$$s_t = S_t \cdot \nabla_t \qquad \nabla_t = \frac{\partial \ln p_y(y_t|x_t, \beta_t; \boldsymbol{\lambda})}{\partial \beta_t}, \tag{3}$$

with S_t being a scaling matrix and $p_y(y_t|\beta_t, x_t; \lambda) = p_\varepsilon(\tilde{\varepsilon}_t; \lambda)$, where $\tilde{\varepsilon}_t = y_t - \beta_t x_t$ is the prediction error. For example, for normally distributed homoskedastic disturbances $p_\varepsilon(\varepsilon_t; \lambda) = N(0, \sigma_\varepsilon^2)$, we get $\nabla_t = x_t(y_t - \beta_t x_t)\sigma_\varepsilon^{-2}$, giving rise to the filter

$$\beta_{t+1} = \omega + \alpha \sigma_\varepsilon^{-2} x_t (y_t - \beta_t x_t) + \gamma \beta_t. \tag{4}$$

Intuitively, this specification of the score ensures that the parameter will be updated to reduce the latest prediction error. In this way the parameter stays up-to-date and provides the closest model fit based on the most recent observations. Clearly other distributions for the error term can be chosen and the parameter update will naturally be adjusted.

When regressors are exogenous, the static parameters of the filtering equation in (2), collected in the vector $\boldsymbol{\theta} = (\omega, \alpha, \gamma, \sigma_\varepsilon^2)'$, can be consistently estimated by maximum likelihood, and ultimately, a causal time-varying β_t can be adequately filtered (see e.g. Blasques, Gorgi, Koopman, and Wintenberger (2018)). This means essentially that the filtered parameter $\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T)$ initialized at some value $\hat{\beta}_1$ in a correctly specified model, will converge to its true unobserved value β_t as both t and T diverge to infinity,

$$|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\beta}_1) - \beta_t| \xrightarrow{p} 0 \quad \text{as } (t, T) \to \infty. \tag{5}$$

Taking t to infinity is required so that the effect of the (almost surely) incorrect initialization of the filter at $\hat{\beta}_1$ vanishes. This is ensured by establishing the so-called *invertibility* of the filter, which requires the filter to have fading memory. Taking the sample size T to infinity ensures that the MLE of the static parameters $\hat{\boldsymbol{\theta}}_T$ converges to $\boldsymbol{\theta}_0$.

Unfortunately, this same score filtering technique will fail to uncover the causal β_t when the regressor x_t is endogenous. Indeed, when the exogeneity condition $\mathbb{E}[\varepsilon_t|x_t] \neq 0$ fails due to simultaneity, omitted variables, or other factors, the filter convergence in (5) will no longer hold. In order to handle the problem of regressor endogeneity, we take a two step instrumental variable approach. This method requires the existence of valid instruments $\{z_t\}$, that are used to estimate the first stage regression

$$x_t = \pi z_t + u_t \qquad\qquad u_t \overset{i.i.d}{\sim} N(0, \sigma_u^2), \tag{6}$$

where $\mathbb{E}[u_t|z_t] = 0$ and the true parameter $\pi_0 \neq 0$. In a second step, we take a control function (CF) approach to correct for endogeneity. Like in a static model (see Wooldridge

(2015)), correlation between the regressor and error term in (1) can be modelled as $\varepsilon_t = \tau u_t + \eta_t$, where we let η_t is an identically and independently distributed (i.i.d.) innovation regardless of the type of endogeneity. As such, the relation between y_t and x_t is given by $y_t = \beta_t x_t + \tau u_t + \eta_t$. Analogous to the static control function approach, we add the first stage fitted residuals to the structural equation to replace the unknown first stage errors. As a result, after obtaining the residuals $\hat{u}_t = x_t - \hat{\pi}_T z_t$ from the first-stage regression, where $\hat{\pi}_T$ denotes the estimate of π, the second stage regression model is given by

$$y_t = \beta_t x_t + \tau \hat{u}_t + \tilde{\eta}_t \tag{7}$$

where $\tilde{\eta}_t$ denotes the residual counterpart of the error term η_t. Further, the Gaussian causal IV-score filtering equation that assumes $\eta_t \overset{i.i.d}{\sim} N(0, \sigma^2)$, takes the form

$$\beta_{t+1} = \omega + \alpha\sigma^{-2} x_t (y_t - \beta_t x_t - \tau(x_t - \hat{\pi}_T z_t)) + \gamma\beta_t \tag{8}$$

where $\hat{\pi}_T$ has been estimated ex ante through equation (6), and we collect the static parameters in $\boldsymbol{\theta} = (\omega, \alpha, \gamma, \tau, \sigma^2)'$.

This control function approach bears similarity to Terza, Basu, and Rathouz (2008), who highlight the importance of using a CF in nonlinear linear-index models to avoid inconsistent estimation. Note that simply replacing the regressors with the fitted values of the first stage in the filter in (4), a more common 2SLS approach, will not work. In time-invariant models, the knowledge that the prediction error is zero on average and that the loss function should be minimised is enough. But in this setting, next to its direct role in loss function minimisation, the prediction error also drives the filter, hence a good estimate for it is crucial. It is necessary therefore to control for the movements in the regressors as much as possible, by adding the first stage residuals to the structural equation. Not doing that, will result in a prediction error and hence a filtered path being governed by unobserved movements in x_t. In the case of a highly relevant but omitted variable for example, the prediction error and filter will mimic the omitted variable rather than distill the causal effect.

We note that all of the subsequent results extend easily to the case in which models defined in (1), (6) and (7) have an intercept $a \in \mathbb{R}$ so that $y_t = a + \beta_t x_t + \eta_t$. For simplicity, we let $a = 0$ and assume that the data is demeaned. We note that the current model can also be easily extended to allow for a time-varying parameter π_t, by constructing a multivariate filter similar to Blasques, Francq, and Laurent (2022). We leave this extension for future research. For theoretical simplicity we focus on the causal time-varying parameter of interest β_t and continue using a static π.

3 Stochastic Properties of the Filter

We collect all data at time t in the vector $Y_t := (y_t, x_t, z_t)'$. We denote the sample history of this vector by $Y^{1:t} := \{Y_1, \ldots, Y_{t-1}, Y_t\}$, and entire history of this vector stretching to the infinite past by $Y^t := \{\ldots, Y_{t-1}, Y_t\}$. We start from the general filter definition (2) to allow for a more flexible framework that includes more general models with different error distributions, model specifications, among others. We denote the filtered sequence, that depends on the sample data, by $\{\hat{\beta}_t(Y^{1:t-1}, \boldsymbol{\theta}, \pi, \hat{\beta}_1)\}_{t\in\mathbb{N}}$, with short hand $\hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)$ also sometimes denoted by $\hat{\beta}_t$. The filtered sequence initialized in the infinite past is denoted by $\{\tilde{\beta}_t(Y^{t-1}, \boldsymbol{\theta}, \pi)\}_{t\in\mathbb{Z}}$ or simply $\tilde{\beta}_t := \tilde{\beta}_t(\boldsymbol{\theta}, \pi)$. If the model is correctly specified and both $\boldsymbol{\theta}_0$ and π_0 are the true parameters, then $\beta_t^o := \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)$ is the true stochastic time-varying parameter. Finally, we let $\boldsymbol{\lambda}$ contain all model and density specific parameters (in equation (8) $\boldsymbol{\lambda} = (\tau, \sigma^2)'$) and let $\boldsymbol{\theta} = (\omega, \alpha, \gamma, \boldsymbol{\lambda})$ and $k = \dim(\boldsymbol{\theta})$. For a random variable $x(\boldsymbol{\theta})$ possibly depending on $\boldsymbol{\theta} \in \Theta$, we further let $\|x(\cdot)\|_n^{\Theta} := (\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} |x(\boldsymbol{\theta})|^n)^{1/n}$ and define the stochastic function

$$\Lambda_t^*(\boldsymbol{\theta}, \boldsymbol{\theta}^*, \pi) := \sup_{\beta^* \in \mathcal{F}_{\boldsymbol{\theta}^*}} |\gamma + \alpha \partial s(\beta, Y_t, \pi; \boldsymbol{\theta})/\partial\beta|_{\beta=\beta^*}| \tag{9}$$

for some $\pi \in \mathbb{R}$. Then $\Lambda_t(\boldsymbol{\theta}, \pi) := \Lambda_t^*(\boldsymbol{\theta}, \boldsymbol{\theta}, \pi)$ is the special case where $\boldsymbol{\theta}^* = \boldsymbol{\theta}$.

Invertibility

Lemma 1 gives general conditions for invertibility and bounded moments of the filter. Invertibility ensures that the effect of the initialisation of the filter vanishes in the limit, meaning that the filtered sequence converges to its unique limit for any given initialisation $\hat{\beta}_1$. Lemma 1 further establishes stationarity and ergodicity (SE) of the limiting sequence, which will be used when deriving the asymptotic properties of the maximum likelihood estimator. Note that this result holds irrespective of whether the model is correctly specified. If these conditions are satisfied, the resulting filter will converge to the same limit sequence, regardless of its initial value, uniformly on Θ.

Lemma 1 (Invertibility)**.** *Let $\Theta \subset \mathbb{R}^k$ be compact, and let the elements in $\{Y_t\}_{t\in\mathbb{Z}}$ be SE sequences. Let $\hat{\pi}_T \in \mathbb{R}$ and assume there exists some $\hat{\beta}_1 \in \mathcal{F}$ such that*

(i) $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} |s(\hat{\beta}_1, Y_t, \hat{\pi}_T; \boldsymbol{\theta})| < \infty$

(ii) $\mathbb{E}\log \sup_{\boldsymbol{\theta}\in\Theta} \Lambda_t(\boldsymbol{\theta}, \hat{\pi}_T) < 0$

Then the sequence $\{\hat{\beta}_t\}_{t\in\mathbb{N}}$ converges e.a.s. to a unique limit SE sequence $\{\tilde{\beta}_t\}_{t\in\mathbb{Z}}$ uniformly on Θ , i.e. $\|\hat{\beta}_t(\boldsymbol{\theta}, \pi^*, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)\|_{\Theta} \xrightarrow{e.a.s.} 0$ as $t \to \infty$.

In Corollary 1 we formulate the invertibility conditions for the specific case of the Gaussian IV-score filter in equation (8) (the conditions for the score filter with exogenous regressors in equation (4) follow with $\tau = 0$). The invertibility conditions (i) and (ii) are trivially satisfied

Corollary 1 (Invertibility for Gaussian IV-score). *Let $\{y_t\}_{t\in\mathbb{Z}}, \{x_t\}_{t\in\mathbb{Z}}, \{z_t\}_{t\in\mathbb{Z}}$ be SE sequences and let $\hat{\pi}_T$ be the first stage estimate. Let $\boldsymbol{\theta} = (\omega, \alpha, \gamma, \tau, \sigma^2)' \in \Theta$ where $\Theta \subset \mathbb{R}^5$ is compact and assume there exists some $\hat{\beta}_1 \in \mathcal{F}$ such that*

- *(i)* $\mathbb{E}\log^+ \sup_{\boldsymbol{\theta}\in\Theta} |\sigma^{-2}x_t(y_t - \hat{\beta}_1 x_t - \tau(x_t - \hat{\pi}_T z_t))| < \infty$
- *(ii)* $\mathbb{E}\log \sup_{\boldsymbol{\theta}\in\Theta} |\gamma - \alpha\sigma^{-2}x_t^2| < 0$

Then the sequence $\{\hat{\beta}_t\}_{t\in\mathbb{N}}$ produced by the filtering equation (8) converges e.a.s. to a unique limit SE sequence $\{\tilde{\beta}_t\}_{t\in\mathbb{Z}}$ uniformly on Θ , i.e. $\|\hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)\|_\Theta \xrightarrow{e.a.s.} 0$ as $t \to \infty$.

Bounded moments

As we shall see in Section 4, beyond filter invertibility, the MLE consistency proof that we establish will also require that the limit filter has bounded moments when evaluated at the true parameter $(\pi_0, \boldsymbol{\theta}_0) \in \mathbb{R} \times \Theta$. According to Lemma 1, the limit filter evaluated at $(\pi_0, \boldsymbol{\theta}_0)$, satisfying the recurrence $\beta_{t+1} = \omega_0 + \alpha_0\sigma_0^{-2}x_t(y_t - \beta_t x_t - \tau_0(x_t - \pi_0 z_t)) + \gamma_0\beta_t$ converges to the same unique solution as

$$\beta_{t+1} = \omega_0 + \alpha_0\sigma_0^{-2}x_t\eta_t + \gamma_0\beta_t. \tag{10}$$

This is a stochastic recurrence equation of the type $\beta_{t+1} = \phi_0(x_t, \eta_t, \beta_t)$ with derivative given by $\partial\phi_0(x_t, \eta_t, \beta_t)/\partial\beta = \gamma_0$ where ϕ_0 is a function defined by the parameter vector $(\pi_0, \boldsymbol{\theta}_0)$. Lemma 2 and Corollary 2 establish bounded moments for this limit process.

Lemma 2 (Limit filter moments). *Let $\{y_t\}_{t\in\mathbb{Z}}, \{x_t\}_{t\in\mathbb{Z}}, \{z_t\}_{t\in\mathbb{Z}}$ be SE sequences. Suppose $\exists\, n_\beta > 0$ such that $\|\phi_0(x_t, \eta_t, \beta_t)\|_{n_\beta} < \infty$, and $\sup_{(\beta^*, Y)\in\mathcal{F}\times\mathcal{Y}} |\partial\phi_0(x_t, \eta_t, \beta_t)/\partial\beta| < 1$. Then $\|\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)\|_{n_\beta} < \infty$.*

Corollary 2 (Limit filter moments for Gaussian IV-score). *Let $\{y_t\}_{t\in\mathbb{Z}}, \{x_t\}_{t\in\mathbb{Z}}, \{z_t\}_{t\in\mathbb{Z}}$ be SE sequences. Suppose $\exists\, n_\beta > 0$ such that $(\mathbb{E}|x_t\eta_t|^{n_\beta})^{1/n_\beta} < \infty$ and $|\gamma_0| < 1$. Then the limit sequence $\|\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)\|_{n_\beta} < \infty$.*

4 Asymptotic Properties of the Maximum Likelihood Estimator

An advantage of the score-driven time-varying parameter models is that the static parameter $\boldsymbol{\theta}$ can be estimated by a straight forward maximum likelihood (ML) procedure. We define the ML estimator as

$$\hat{\boldsymbol{\theta}}_T(\pi) \in \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_T(\boldsymbol{\theta}, \pi, \hat{\beta}_1) \tag{11}$$

where

$$\ell_T(\boldsymbol{\theta}, \pi, \hat{\beta}_1) = \frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}, \pi, \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)) \tag{12}$$

and $\ell_t(\boldsymbol{\theta}, \pi, \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1) := \log p_y(y_t|\hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1), x_t, z_t, \boldsymbol{\theta}, \pi)$. Define $\ell_0(\boldsymbol{\theta}, \pi_0) := \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))$ and $\ell_\infty(\boldsymbol{\theta}, \pi_0) = \mathbb{E}[\ell_0(\boldsymbol{\theta}, \pi_0)]$. To establish consistency in similar spirit to Blasques, Gorgi, et al. (2018), assume that the following conditions hold:

(C1) The DGP which satisfies equations (6) to (8) with $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \Theta$ admits a stationary solution and $\{x_t\}_{t\in\mathbb{Z}}$, $\{z_t\}_{t\in\mathbb{Z}}$ are SE sequences.

(C2) $\mathbb{E}|\ell_0(\boldsymbol{\theta}_0, \pi_0)| < \infty$

(C3) For any $\boldsymbol{\theta} \in \Theta$, $\ell_0(\boldsymbol{\theta}_0, \pi_0) = \ell_0(\boldsymbol{\theta}, \pi_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

(C4) The invertibility conditions (i) and (ii) of Lemma 1 are satisfied for the compact set $\Theta \subset \mathbb{R}^k$

(C5) $E\|\ell_0(\boldsymbol{\theta}, \pi_0) \vee c\|_\Theta < \infty$ for some $c \in \mathbb{R}$ such that $c < \ell_\infty(\boldsymbol{\theta}_0, \pi_0)$

(C6) $\hat{\pi}_T \xrightarrow{a.s.} \pi_0$ as $T \to \infty$

As noted in Blasques, Gorgi, et al. (2018), we obtain strong consistency following Wald (1949). Theorem 1 states that the maximum likelihood estimator evaluated at some initialisation $\hat{\beta}_1$ and at the first stage estimate $\hat{\pi}_T$ is consistent for the true unknown parameter. Corollary 3 provides the same consistency result but applied to the special case of our Gaussian IV-score model.

Theorem 1 (Consistency)**.** *Let the conditions* $(C1) - (C6)$ *hold. Then the maximum likelihood estimator is strongly consistent*

$$\hat{\boldsymbol{\theta}}_T(\hat{\pi}_T, \hat{\beta}_1) \xrightarrow{a.s.} \boldsymbol{\theta}_0, \qquad T \to \infty \tag{13}$$

for any initialisation $\hat{\beta}_1 \in \mathcal{F}$.

Corollary 3 (Consistency MLE of Gaussian IV-Score)**.** *Let the process* $\{y_t\}_{t\in\mathbb{Z}}$ *be generated by the model in equations equations* (6) *to* (8) *with* $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \Theta$ *such that* $|\gamma_0| < 1$ *and* $\sigma_0^2 > 0$ *and let the sequences* $\{x_t\}_{t\in\mathbb{Z}}$, $\{z_t\}_{t\in\mathbb{Z}}$ *be SE. Furthermore let* Θ *be compact such that* $\mathbb{E}\log|\gamma - \alpha\sigma^{-2}x_t^2| < 0$ *and* $\sigma > 0$ $\forall\boldsymbol{\theta} \in \Theta$. *Then the maximum likelihood estimator* $\hat{\boldsymbol{\theta}}_T(\hat{\pi}_T, \hat{\beta}_1)$ *with any initialisation* $\hat{\beta}_1 \in \mathcal{F}$ *is strongly consistent.*

Building on the consistency of the MLE, we can also provide a convergence result of the causal time-varying parameter of interest. Proposition 1 assures that the filter converges to the true time-varying parameter almost surely, and is a direct result of the previously established invertibility of the filter and consistency of the MLE.

Proposition 1 (Path Convergence at $\hat{\pi}_T$)**.** *If the model is correctly specified and if the invertibility conditions* $(i), (ii)$ *of Lemma 1 hold, if* $\mathbb{E}[\log^+ \|\tilde{\beta}_1\|_\Theta] < \infty$ *and* $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *then the estimated filter converges to the true unobserved causal time-varying parameter* $\{\beta_t^o\}_{t\in\mathbb{Z}}$, *when evaluated at* $\hat{\pi}_T$

$$|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T, \hat{\beta}_1) - \beta_t^o| \xrightarrow{a.s.} 0 \text{ as } T \geq t \to \infty \tag{14}$$

for any initialisation $\hat{\beta}_1 \in \mathcal{F}$.

5 Simulation Study

In the following simulation study we investigate the performance of the new IV-score filter compared to a regular score filter, as specified in (8) and (1) respectively, for a time-varying parameter regression model in which the regressor is endogenous. In terms of endogeneity, we consider the case of a stationary omitted variable that is correlated with the regressor. The omitted variable is generated with various specifications, such as a sine function that could be representative of an unobserved business cycle or other cyclical omitted elements or simply by an AR(1) process or a process with a break.

We consider the following DGP in which h_t is the omitted variable and the true time-varying parameter β_t as well as the instrument z_t, are generated by stationary AR(1) processes for given parameters $\delta, \sigma_\eta, \psi, \pi, \sigma_u, \sigma_h$.

$$\begin{aligned}
y_t &= \beta_t x_t + \underbrace{\delta h_t + \sigma_\eta N(0,1)}_{\varepsilon_t} \\
\beta_t &= 0.1 + 0.95\beta_{t-1} + \sqrt{0.1}N(0,1) \\
x_t &= \psi h_t + \pi z_t + \sigma_u N(0,1) \\
z_t &= 0.2 z_{t-1} + \sqrt{(1-0.2^2)}N(0,1)
\end{aligned}$$

$$\begin{array}{l|l}
\text{DGP 1} & h_t = 0.3h_{t-1} + N\left(0, (1-0.3^2)\sigma_h^2\right) \\
\text{DGP 2} & h_t = \sqrt{2}\sigma_h \sin(t/25) \\
\text{DGP 3} & h_t = \begin{cases} -1 & t \leq \lfloor T/2 \rfloor \\ 1 & t > \lfloor T/2 \rfloor \end{cases}
\end{array}$$

We generate the path for β_t once, and subsequently draw the data M=1000 times, estimate both filters and evaluate them by taking the mean MSE over the whole path. With the parameters $\delta, \psi, \pi, \sigma_h$ we can increase and decrease the level of endogeneity, while we fix $\sigma_\eta = 1, \sigma_u = 0.1$. The true path for β_t is chosen with a high persistency parameter, which is generally challenging to filter. The omitted variable is simulated in various ways: with the sine function to represent an AR(1) (DGP 1), a cyclical unobserved variable (DGP 2) and a midway break (DGP 3).

5.1 Uncovering the True Causal Parameter

In Figure 1 and Table 1 we present the results for DGP 1 for the case where $\psi = 1, \delta = -4, \pi = 1$ and $\sigma_h = 5$. This is a configuration of the parameters for which the bias is substantial and the illustration is clear. The plots of the path is constructed by taking medians at each point t from all simulations. In the plots we include the true path (blue), the static OLS (red) and IV (green) estimators as estimated on the whole sample, and the median IV-score and score (orange) paths over all simulations.

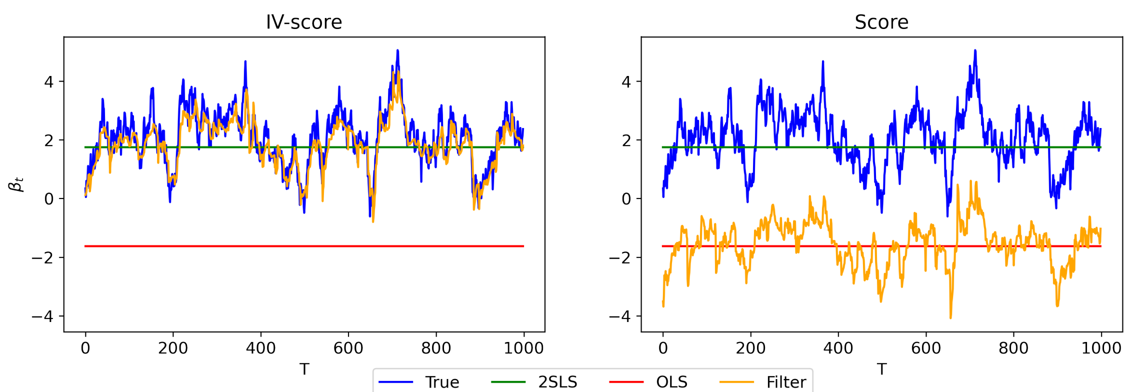


Figure 1: Estimated paths of the causal parameter with DGP 1 where h_t follows an AR(1).

Comparing the mean-squared errors (MSEs) of the filtered paths relative to the true causal paths (Table 1) it becomes clear that the IV-score filter outperforms the score filter without difficulty, highlighting that taking endogeneity into account is crucial. The endogeneity bias is visible in Figure 1 through the large gap between the static OLS and IV estimators. The OLS estimator suggests a consistent negative relation, while much of the causal parameter is actually positive, something that is correctly picked up on by the IV estimator. This bias in the OLS estimator is carried over to the score-driven model in the time-varying case. In contrast, the IV-score filter captures well the true causal relation.

At a first glance it seems that the aforementioned static OLS-IV bias shift is the only

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}_0$	$\hat{\sigma}_\eta^2$	$\hat{\tau}$	b_0	b_1	MSE
OLS							0.129	-1.635	14.477
IV						-3.740	0.022	1.762	0.979
score	-0.069	0.258	0.615	-3.483	20.254				13.590
IV-score	0.081	0.123	0.954	0.1618	5.664	-3.761			**0.329**

Table 1: DGP 1: Maximum Likelihood Estimates and MSE

difference between the score and IV-score filtered paths. But not only is it shifted, the score filter does not follow the same dynamics as the true parameter does. This is clearly shown in the scatter plots of the true parameter versus the filtered parameters displayed in Figure 2. Clearly, the IV-score filter moves along with the true parameter in the right directions (high when high).

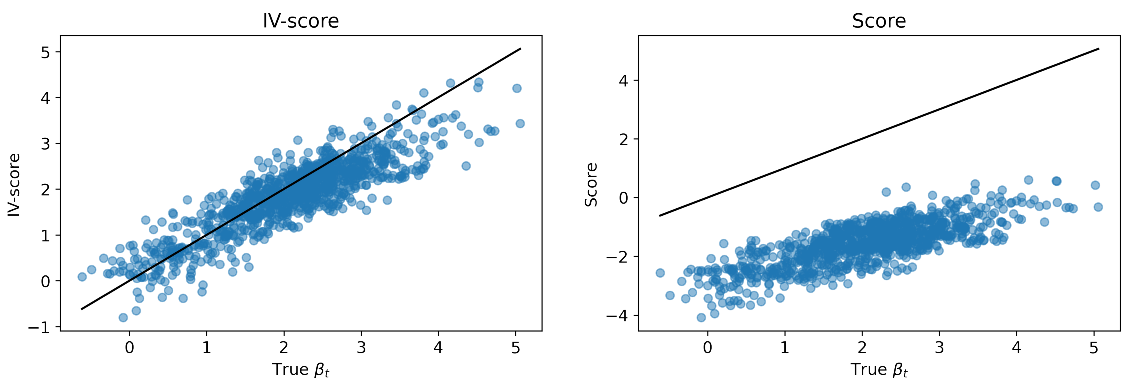


Figure 2: Scatter plot of IV-score and score filtered values vs. true β_t for DGP 1 where h_t follows a sine function.

In Figure 4 in the supplementary appendix we observe that modelling the omitted variable differently, for example following sine dynamics (DGP 2) or with a structural break (DGP 3), does not really affect the performance of either filter. Even in the case of a midway break in which we have a non-stationary omitted variable, we still observe similar behaviour.

Another crucial observation is that this DGP is parameter-driven, while our (IV)score method is observation-driven. This shows that misspecification in this sense does not affect the performance of our observation-driven approach. For more discussion on this comparison, see Koopman et al. (2016).

5.2 Robustness Check

In this section we investigate the robustness of the method. In particular, we explore whether the method breaks down with (i) a non-stationary causal parameter (DGP 4 & 5), (ii) when the instrument is not fully exogeneous (DGP 6), and (iii) when the observations have higher variance (DGP 7). All resulting figures can be found in the supplementary appendix. In each of these three scenario's we consider an omitted variable that is generated by an AR(1). All DGP's are based on DGP 1 with the following adaptations.

DGP 4	$\beta_t = \begin{cases} -1 & t \leq \lfloor T/2 \rfloor \\ 1 & t > \lfloor T/2 \rfloor \end{cases}$
	$\sigma_h = 1$
DGP 5	$\beta_t = 1 + \beta_{t-1} + \sqrt{0.1}N(0,1)$
DGP 6	$h_t = 0.3h_{t-1} + 3z_t + N\left(0, (1-0.3^2)\sigma_h^2\right)$
DGP 7	$\sigma_\eta^2 = 25$

5.2.1 Non-stationary Causal Parameter

Figure 5 contains the resulting path for a midway break in β_t, jumping from a negative value to positive in the middle of the sample. Even in this non-stationary setting, the IV-score filter manages to capture the change in the parameter quite closely, with some variation around the true line. The score filter on the other hand does show some indication of a break in levels, but remains negative throughout the whole sample due to the bias. When we generate a more severe case of non-stationarity by letting β_t follow a random walk (Figure 6), we still find that the filter performs considerably well in most cases. We do however observe an increased number of computational issues such as convergence problems and non-invertible results, up to 10% of all simulations.

5.2.2 Contaminated Instrument

When the instrument is not exogeneous but also correlated with the omitted variable, the IV-score filter breaks down as expected. In Figure 7 the paths are both visibly biased, although the IV-score filter still benefits from some exogeneous movements in z_t to get closer the the true path. This figure clearly illustrates how the performance depends on how the static IV estimator relates to the true path. The lower the bias in IV, the better the performance of the related IV-score filter.

5.2.3 Large Error Variance

When the variance of the error term is increased to $\sigma_\eta^2 = 25$, the result is a more flattened filter as visible in Figure 8. This is to be expected, as the prediction error in the score will have larger outlying values and the corresponding parameter α will adjust accordingly making the overall filter less responsive to any changes in the score, also those caused by the true parameter.

6 Application

We now illustrate our score driven filter for endogenous regressors by estimating time-varying excess sensitivity of consumption to income. Originally, Hall (1978) hypothesised that consumption (C_t) was a random walk, while later Campbell and Mankiw (1989) proposed an extended model by assuming that a fraction of the population (λ) consumes out of their current income (Y_t). As a result, a part of the changes in income drive changes in consumption, expressed as

$$\Delta C_t = \alpha + \lambda \Delta Y_t + \varepsilon_t.$$

Due to the potential endogeneity of income, Campbell and Mankiw (1989) suggested using lagged variables as instruments. It was only recently, that Bhatt, Kundan Kishor, and Marfatia (2020) highlighted that although these instruments have been argued to be exogeneous, they are in fact weak instruments. Instead, these authors proposed to use the lagged 1-step-ahead Greenbook forecasts of changes in real disposable income, as these are simultaneously highly correlated with the real disposable income at time t (as opposed to the previously used instruments), and exogeneous due to the forecasting nature of the variable. Thus eliminating any undesirable contemporaneous effects.

The parameter of interest is λ, which measures the causal impact that changes in income have over changes in aggregate consumption. By sub-sample estimation (1978-1999 and 2000-2010), Bhatt et al. (2020) find that λ is much smaller in the most recent sub-sample than in the sample based on the years before 2000, which motivates their choice for taking a time-varying approach. Similarly, we estimate time varying parameter λ_t using our score-driven causal filter, while adding a lag of ΔC_t with a fixed parameter[1] to capture the effect of changes in income on consumption growth. The structural model equation we estimate is

$$\Delta C_t = \alpha + \lambda_t \Delta Y_t + \phi \Delta C_{t-1} + \varepsilon_t, \tag{15}$$

[1] Bhatt et al. (2020) also add ΔC_{t-1} to their model, however with a time-varying parameter ϕ_t. Such a multivariate filter is also possible in our framework, but for simplicity and brevity we take a constant parameter as we wish to merely illustrate the use of our filter for the parameter of interest λ_t.

where use the Greenbook forecasts (Y_t^{GB}) as instruments.

Figure 3 plots the growth rates of aggregate consumption, national income, as well as the GreenBook forecasts. Similar to Bhatt et al. (2020), we take quarterly observations of real consumption expenditures per capita on nondurable goods and services and real per capita disposable income from 1978-2010 in the U.S., as well as the the one quarter forecast of real disposable income from the Greenbook reports[2].

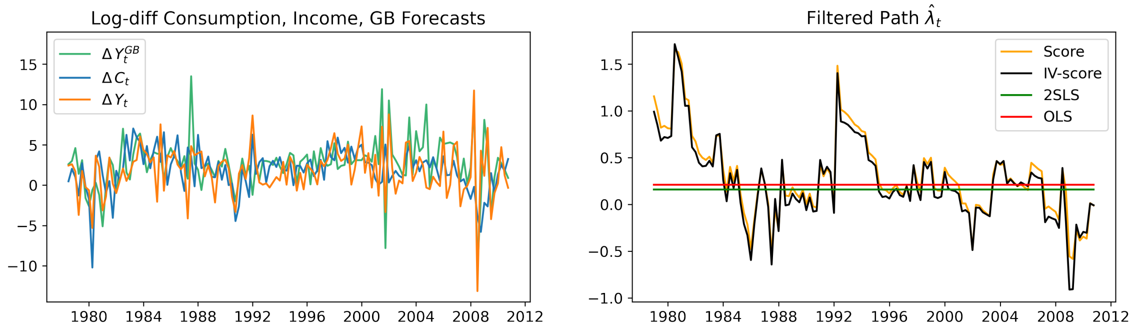


Figure 3: Data and filtered path of excess sensitivity parameter λ_t

In the right panel of Figure 3 we plot the static OLS and IV estimators and the resulting score driven filtered paths of λ_t, accounting for endogeneity and not (IVGAS and GAS respectively). From these estimated paths we observe that sensitivity of consumption to income seems to have steadily decreased during the 80s. Over the next decade the sensitivity increased somewhat until the sudden drop in the wake of the 2001 crisis. There seems to have been speedy recovery however, until the plunge due to the 2008 financial crisis.

An important observation is that endogeneity does not seem to be a problem as the static OLS and IV are not vastly different. This is indeed confirmed by a Hausman test with a p-value of 0.55. However, it should be noted that these static estimators and the Hausman statistic measure an average effect and that any local or temporary endogeneity is averaged out over the entire sample. Comparing the IVGAS and GAS paths, we also find that they overlap for a large part of the quarters in the sample, but do differ substantially in a few instances. Most notably, during the period 2008Q4-2009Q2 right after the start of the global financial with the bankruptcy of Lehman Brothers, the estimated effect by IVGAS is twice as large as that by GAS. In the end, the level of endogeneity can also vary over time. Although without error bounds or standard errors of estimated parameters no conclusions can be made about the significance of this result, Bhatt et al. (2020) find however that their estimated parameter belonging to the first stage residuals (our τ) is significant, hence that in a model with time variation endogeneity in fact is a concern.

[2]For a detailed description of the source and type of data, see Bhatt et al. (2020).

7 Conclusion

In this paper we have introduced a score-driven filter for time-varying regression parameters that can be applied when regressors are endogenous. We have established invertibility of the filter, consistency of the MLE and proven filter convergence to the true unobserved path. In a simulation study we have shown that the behaviour of the regular score filter gives undesirable results while the IV-score, in the presence of suitable instruments, uncovers the true underlying path of the time varying parameter. We have also shown in simulations that for non-stationary parameter processes such as structural breaks and random walks the filter shows appropriate behaviour. Nevertheless, caution should be taken with extrapolating this result, as non-stationary paths are not included in the theoretical framework.

Further research will focus on establishing confidence bounds for the filtered path. With current assumptions, the derivation of asymptotic normality of the MLE is complicated by the fact that the likelihood function does not have guaranteed bounded moments. With asymptotic normality, a test for time variation and a direct Hausman test for endogeneity could also be developed. The score driven IV method enriched with a confidence bound procedure and these hypothesis tests could significantly add by presenting new results in many fields of applied research.

Appendix: Proofs

Proof of Lemma 1

This lemma is an application of Proposition 3.2 of Blasques, van Brummelen, Koopman, and Lucas (2022), to which we refer for a proof.

Proof of Lemma 2

This lemma is an application of Proposition 3.3 of Blasques, van Brummelen, et al. (2022), to which we refer for a proof.

Proof of Theorem 1

Proof. For notational simplicity, write $\hat{\ell}_T(\boldsymbol{\theta}, \pi, \hat{\beta}_1) := \frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}, \pi, \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1))$ and $\tilde{\ell}_T(\boldsymbol{\theta}, \pi_0) := \frac{1}{T}\sum_{t=1}^{T} \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))$. We will prove this Theorem in the following parts:

(P1) The model is identifiable: $\ell_\infty(\boldsymbol{\theta}_0, \pi_0) > \ell_\infty(\boldsymbol{\theta}, \pi_0)$ for any $\boldsymbol{\theta} \in \Theta, \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

(P2) The function $\hat{\ell}_T(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)$ with first stage estimator $\hat{\pi}_T$ converges a.s. to $\tilde{\ell}_T(\boldsymbol{\theta}, \pi_0)$ uniformly over $\boldsymbol{\Theta}$. That is,
$\|\hat{\ell}_T(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\ell}_T(\boldsymbol{\theta}, \pi_0)\|_\Theta \stackrel{a.s.}{\to} 0$ as $T \to \infty$.

(P3) For any $\varepsilon > 0$, the following inequality holds a.s.

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\theta}\in B^c(\boldsymbol{\theta}_0,\varepsilon)} \hat{\ell}_T(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) < \ell_\infty(\boldsymbol{\theta}_0, \pi_0)$$

where $B^c(\boldsymbol{\theta}_0, \varepsilon) = \Theta \setminus B(\boldsymbol{\theta}_0, \varepsilon)$ with $B(\boldsymbol{\theta}_0, \varepsilon) = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\| < \varepsilon\}$

(P4) The result in **(P3)** implies strong consistency

(P1): Existence of $\ell_\infty(\boldsymbol{\theta}_0, \pi_0)$ is guaranteed by **C2** and by **C5** we have that either $\ell_\infty(\boldsymbol{\theta}, \pi_0) \in \mathbb{R}$ or $\ell_\infty(\boldsymbol{\theta}, \pi_0) = -\infty$. Then

$$\begin{aligned} \ell_0(\theta, \pi_0) - \ell_0(\theta_0, \pi_0) = \log p_\eta \Big(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)x_t - \tau(x_t - \pi_0 z_t); \lambda\Big) \\ - \log p_\eta \left(y_t - \beta_t^o x_t - \tau_0(x_t - \pi_0 z_t); \lambda_0\right) \end{aligned} \tag{16}$$

where $\tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)$ is the limit sequence, and $\beta_t^o = \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)$ is the true TVP (due to correct specification of the filter).

$$\begin{aligned} \ell_0(\theta, \pi_0) - \ell_0(\theta_0, \pi_0) &= \frac{\log \Big(p_\eta(y_t - \tilde{\beta}_t x_t - \tau(x_t - \pi_0 z_t); \lambda\Big)}{p_\eta\left(y_t - \beta_t^o x_t - \tau_0(x_t - \pi_0 z_t); \lambda_0)\right)} \\ &\leq \frac{p_\eta\Big(y_t - \tilde{\beta}_t x_t - \tau(x_t - \pi_0 z_t); \lambda\Big)}{p_\eta(y_t - \beta_t^o x_t - \tau_0(x_t - \pi_0 z_t); \lambda_0)} - 1 \end{aligned} \tag{17}$$

The case of an equal sign here is ruled out, since the densities are not the same for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. And, since $p_\eta(y_t - \beta_t^o x_t - \tau_0(x_t - \pi_0 z_t); \lambda_0)$ is the true conditional density we have

$$\mathbb{E}[\mathbb{E}\left[\ell_0(\boldsymbol{\theta}, \pi_0) - \ell_0(\boldsymbol{\theta}_0, \pi_0)|Y^t\right]] < \mathbb{E}\left[\mathbb{E}\left[\left.\frac{p_\eta\Big(y_t - \tilde{\beta}_t x_t - \tau(x_t - \pi_0 z_t); \lambda\Big)}{p_\eta(y_t - \beta_t^o x_t - \tau_0(x_t - \pi_0 z_t); \lambda_0)}\right| Y^t\right]\right] - 1 = 0$$

As a result,

$$\ell_\infty(\boldsymbol{\theta}, \pi_0) - \ell_\infty(\boldsymbol{\theta}_0, \pi_0) = \mathbb{E}[\mathbb{E}\left[\ell_0(\boldsymbol{\theta}, \pi_0) - \ell_0(\boldsymbol{\theta}_0, \pi_0)|Y^t\right]] < 0 \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

(P2): In this part of the proof we allow the likelihood function and the filter both to depend on another value of π. Then the log likelihood contributions under the assumption of Gaussian errors is given by

$$\ell_t(\boldsymbol{\theta}, \pi^*, \beta_t(\boldsymbol{\theta}, \pi^{**})) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y_t - \beta_t(\boldsymbol{\theta}, \pi^{**})x_t - \tau(x_t - \pi^* z_t))^2}{\sigma^2}$$

We have

$$
\begin{aligned}
&\|\hat{\ell}_T(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\ell}_T(\boldsymbol{\theta}, \pi_0)\|_\Theta \qquad (18)\\
&\qquad = \sup_{\boldsymbol{\theta}\in\Theta} \left| \frac{1}{T}\sum_{t=1} \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)) - \frac{1}{T}\sum_{t=1} \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)) \right| \\
&\qquad \leq \sup_{\boldsymbol{\theta}\in\Theta} \frac{1}{T}\sum_{t=1} \left| \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)) - \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)) \right| \\
&\qquad \leq \frac{1}{T}\sum_{t=1} \left\| \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)) - \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)) \right\|_\Theta ,
\end{aligned}
$$

We will show that each summand converges a.s. to zero, so that by the Stolz-Cesaro theorem the sample average does too. We can write

$$
\begin{aligned}
&\|\ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)) - \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))\|_\Theta \qquad (19)\\
&\qquad \leq \|\ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)) - \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T))\|_\Theta \\
&\qquad + \|\ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)) - \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))\|_\Theta \\
&\qquad + \|\ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) - \ell_t(\boldsymbol{\theta}, \pi_0, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))\|_\Theta
\end{aligned}
$$

Applying a Mean Value Theorem (MVT) to each of these terms we get

$$
\begin{aligned}
&= \left\| \frac{\partial \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \beta_t^*(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1))}{\partial f} \right\|_\Theta \|\hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)\|_\Theta \\
&+ \left\| \frac{\partial \ell_t(\boldsymbol{\theta}, \hat{\pi}_T, \tilde{\beta}_t^*(\boldsymbol{\theta}, \pi))}{\partial f} \right\|_\Theta \|\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)\|_\Theta \\
&+ \left\| \frac{\partial \ell_t(\boldsymbol{\theta}, \tilde{\pi}, \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))}{\partial \pi} \right\|_\Theta \|\hat{\pi}_T - \pi_0\|_\Theta
\end{aligned}
$$

where

- $\beta_t^*(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)$ is a point between $\hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)$ and $\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)$
- $\tilde{\beta}_t^*(\boldsymbol{\theta}, \pi)$ is a point between $\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)$ and $\tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)$
- $\tilde{\pi}$ is a point between $\hat{\pi}_T$ and π_0.

Writing out each derivative we obtain

$$= \left\| \frac{x_t(y_t - \beta_t^*(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)x_t - \tau(x_t - \hat{\pi}_T z_t))}{\sigma^2} \right\|_\Theta \|\hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)\|_\Theta$$

$$+ \left\| \frac{x_t(y_t - \tilde{\beta}_t^*(\boldsymbol{\theta}, \pi)x_t - \tau(x_t - \hat{\pi}_T z_t))}{\sigma^2} \right\|_\Theta \|\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)\|_\Theta$$

$$+ \left\| -\frac{\tau z_t(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)x_t - \tau(x_t - \tilde{\pi} z_t))}{\sigma^2} \right\|_\Theta \|\hat{\pi}_T - \pi_0\|_\Theta$$

Adding and subtracting $\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)$ to the first term, and expanding the second term further,

$$\leq \left[\underbrace{\left\| \frac{x_t(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)x_t - \tau(x_t - \hat{\pi}_T z_t))}{\sigma^2} \right\|_\Theta}_{SE} + \underbrace{\left\| \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T) - \beta_t^*(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) \right\|_\Theta}_{\xrightarrow{e.a.s.} 0} \right.$$

$$\left. \times \underbrace{\left\| \frac{x_t^2}{\sigma^2} \right\|_\Theta}_{SE} \right] \underbrace{\|\hat{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T)\|_\Theta}_{\xrightarrow{e.a.s.} 0}$$

$$+ \underbrace{\left\| \frac{x_t(y_t - \tilde{\beta}_t^*(\boldsymbol{\theta}, \pi)x_t - \tau(x_t - \hat{\pi}_T z_t))}{\sigma^2} \right\|_\Theta}_{SE} \underbrace{\left\| \frac{\partial \tilde{\beta}_t(\boldsymbol{\theta}, \tilde{\pi})}{\partial \pi} \right\|_\Theta}_{SE} \underbrace{|\hat{\pi}_T - \pi_0|}_{\xrightarrow{a.s.} 0}$$

$$+ \underbrace{\left\| -\frac{\tau z_t(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)x_t - \tau(x_t - \tilde{\pi} z_t))}{\sigma^2} \right\|_\Theta}_{SE} \underbrace{|\hat{\pi}_T - \pi_0|}_{\xrightarrow{a.s.} 0}$$

where $\tilde{\pi}$ is a point between $\hat{\pi}_T$ and π_0. The derivative process of stationary $\tilde{\beta}_t$ is SE as shown in the technical appendix of Blasques, Koopman, and Lucas (2014). We now have products of SE sequences with e.a.s. vanishing terms, which vanish e.a.s. by Lemma 2.1. in Straumann and Mikosch (2006). Note that the analogous result $\xi_n^* v_n \xrightarrow{a.s.} 0$ as $n \to \infty$ for stationary sequence v_n with a bounded log moment and $\xi_t^* \xrightarrow{a.s.} 0$ also holds.

(P3): We will show that (**P3**) holds for $\tilde{\ell}_T(\boldsymbol{\theta}, \pi_0)$ since by (**P2**) $\hat{\ell}_T(\boldsymbol{\theta}, \hat{\pi}_T, \hat{\beta}_1)$ is asymptotically equivalent to $\tilde{\ell}_T(\boldsymbol{\theta}, \pi_0)$.

Fix a $\boldsymbol{\theta}^*$. Then for a decreasing sequence $\{\varepsilon_i\}_{i\in\mathbb{N}}$ s.t. $\lim_{i\to\infty} \varepsilon_i = 0$, the sequence $\{\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*, \varepsilon_i)} \ell_0(\boldsymbol{\theta}, \pi_0)\}_{i\in\mathbb{N}}$ is non-increasing and greater than $\ell_0(\boldsymbol{\theta}^*, \pi_0)$ for every i. Considering this, and the fact that $\lim_{i\to\infty} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*, \varepsilon_i)} \ell_0(\boldsymbol{\theta}, \pi_0) = \ell_0(\boldsymbol{\theta}^*, \pi_0)$ by continuity, we conclude that $\sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*,\varepsilon_i)} \ell_0(\boldsymbol{\theta},\pi_0) \downarrow \ell_0(\boldsymbol{\theta}^*,\pi_0)$. This together with $\mathbb{E}\sup_{\boldsymbol{\theta}\in\Theta} \ell_0(\boldsymbol{\theta},\pi_0) < \infty$ which is implied by (**C5**), we can apply the Monotone Convergence Theorem to conclude that

$$\lim_{i\to\infty} \mathbb{E} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*,\varepsilon_i)} \ell_0(\boldsymbol{\theta},\pi_0) = \ell_\infty(\boldsymbol{\theta}^*,\pi_0)$$

By (**P1**) we have that $\ell_\infty(\boldsymbol{\theta}_0,\pi_0) > \ell_\infty(\boldsymbol{\theta}^*,\pi_0)$ so that for all $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}_0$ there exists a $\varepsilon_{\boldsymbol{\theta}^*} > 0$ such that

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}^*,\varepsilon_{\boldsymbol{\theta}^*})} \ell_0(\boldsymbol{\theta},\pi_0) < \ell_\infty(\boldsymbol{\theta}_0,\pi_0)$$

The set $B^c(\boldsymbol{\theta}_0,\varepsilon)$ is compact and is covered by the balls $\{B(\boldsymbol{\theta},\varepsilon_{\boldsymbol{\theta}}) : \boldsymbol{\theta} \in B^c(\boldsymbol{\theta}_0,\varepsilon)\}$. Let $B(\boldsymbol{\theta}_1,\varepsilon_1),\ldots,B(\boldsymbol{\theta}_p,\varepsilon_p)$ be a finite subcover with $\sup_{k=1,\ldots,p}\varepsilon_k < \varepsilon$. Then, for any $n \in \mathbb{N}$, we have

$$\sup_{\boldsymbol{\theta}\in B^c(\boldsymbol{\theta}_0,\varepsilon)} \tilde{\ell}_T(\boldsymbol{\theta},\pi_0) \leq \bigvee_{k=1}^{p} \frac{1}{T}\sum_{t=1}^{T} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_k,\varepsilon_k)} \tilde{\ell}_t(\boldsymbol{\theta},\pi_0)$$

Taking limits on both sides of the equation gives

$$\limsup_{T\to\infty} \sup_{\boldsymbol{\theta}\in B^c(\boldsymbol{\theta}_0,\varepsilon)} \tilde{\ell}_T(\boldsymbol{\theta},\pi_0) \leq \bigvee_{k=1}^{p} \mathbb{E} \sup_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_k,\varepsilon_k)} \ell_0(\boldsymbol{\theta},\pi_0) < \ell_\infty(\boldsymbol{\theta}_0,\pi_0)$$

(P4): Suppose there exists an $\varepsilon > 0$ such that $\limsup_{T\to\infty} ||\hat{\boldsymbol{\theta}}_T(\hat{\pi},\hat{\beta}_1) - \boldsymbol{\theta}_0|| \geq \varepsilon$. Then

$$\sup_{\boldsymbol{\theta}\in B^c_{\varepsilon_{\boldsymbol{\theta}_0}}} \hat{\ell}_T(\boldsymbol{\theta},\hat{\pi}_T,\hat{\beta}_1) \geq \hat{\ell}_T\left(\hat{\boldsymbol{\theta}}_T(\hat{\pi}_T,\hat{\beta}_1),\hat{\pi}_T,\hat{\beta}_1\right) \geq \hat{\ell}_T(\boldsymbol{\theta}_0,\hat{\pi}_T,\hat{\beta}_1)$$

by definition of the two-step MLE, for all $T \in \mathbb{N}$. By (**P2**) we have that $\hat{\ell}_T(\boldsymbol{\theta},\hat{\pi},\hat{\beta}_1)$ is asymptotically equivalent to $\tilde{\ell}_T(\boldsymbol{\theta},\pi_0)$, which leads to

$$\begin{aligned}
\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in B^c_{\varepsilon_{\boldsymbol{\theta}_0}}} \hat{\ell}_T(\boldsymbol{\theta},\hat{\pi}_T,\hat{\beta}_1) &\geq \lim_{T\to\infty} \hat{\ell}_T(\boldsymbol{\theta}_0,\hat{\pi}_T,\hat{\beta}_1)\\
&\simeq \lim_{T\to\infty} \tilde{\ell}_T(\boldsymbol{\theta}_0,\pi_0)\\
&= \ell_\infty(\boldsymbol{\theta}_0,\pi_0).
\end{aligned}$$

However, we established in (**P3**) that this event has probability zero. Therefore, $\limsup_{T\to\infty} ||\hat{\boldsymbol{\theta}}_T(\hat{\pi}_T,\hat{\beta}_1) - \boldsymbol{\theta}_0|| < \varepsilon$ with probability 1 for any $\varepsilon > 0$. This concludes the proof. □

Proof of Corollary 3

Proof. In order to show consistency of the MLE for the Gaussian IV-score model, we show that the assumptions imply each of the conditions *C1-C6* from Theorem 1. Let $K = \frac{1}{2}\log(2\Pi)$ where Π is used here to denote the number pi (3.14..) that appears in the Gaussian density. Then we have the following expressions for the likelihood contributions and the IV-score filter

$$\ell_t(\boldsymbol{\theta}, \pi, \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)) = -K - \frac{1}{2}\log(\sigma) - \frac{1}{2}\sigma^{-2}(y_t - \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)x_t - \tau(x_t - \pi z_t))^2 \tag{20}$$

$$\hat{\beta}_{t+1}(\boldsymbol{\theta}, \pi, \hat{\beta}_1) = \omega + \alpha\sigma^{-2}x_t(y_t - \hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)x_t - \tau(x_t - \pi z_t)) + \gamma\hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1) \tag{21}$$

where $\hat{\beta}_1(\boldsymbol{\theta}, \pi, \hat{\beta}_1) = \hat{\beta}_1$. Furthermore, we define

$$\ell_0(\boldsymbol{\theta}, \pi) = -K - \frac{1}{2}\log(\sigma) - \frac{1}{2}\sigma^{-2}(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi)x_t - \tau(x_t - \pi z_t))^2 \tag{22}$$

where $\tilde{\beta}_t(\boldsymbol{\theta}, \pi)$ is the stationary limit of $\hat{\beta}_t(\boldsymbol{\theta}, \pi, \hat{\beta}_1)$.

(C1) The DGP admits a stationary solution if $\beta_t^o, x_t, z_t, \eta_t$ are stationary sequences by Proposition 4.3 in Krengel (1985). Therefore we need to show that $\beta_t^0 = \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)$ admits a stationary solution, since the rest is assumed to be stationary. The true parameter follows the process $\beta_{t+1}^0 = \omega_0 + \alpha_0\sigma_0^2 x_t\eta_t + \gamma\beta_t^0$ which is SE whenever $E|\gamma_0| < 1$.

(C2) $\mathbb{E}|\ell_0(\boldsymbol{\theta}_0, \pi_0)| < \infty$ holds for the Gaussian density.

$$\mathbb{E}|\ell_0(\boldsymbol{\theta}_0, \pi_0)| = \mathbb{E}|-K - \frac{1}{2}\log(\sigma_0) - \frac{1}{2}\sigma_0^{-2}(y_t - \beta_t^0 x_t - \tau_0 u_t)^2| \tag{23}$$

$$= \mathbb{E}|-K - \frac{1}{2}\log(\sigma_0) - \frac{1}{2}\sigma_0^{-2}\eta_t^2| < \infty \tag{24}$$

(C3) We have to show that $\ell_0(\boldsymbol{\theta}_0, \pi_0) = \ell_0(\boldsymbol{\theta}, \pi_0)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\pi = \pi_0$.

Denote the Gaussian density function by $f(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{y-\mu}{\sigma})^2)$. Note that $f(y|\mu_0, \sigma_0) = f(y|\mu, \sigma)$ for any y if and only if $\mu = \mu_0$ and $\sigma = \sigma_0$. Therefore, for our Gaussian likelihood in equation (22) with $\sigma = \sigma_0$, we have $\mu = \mu_0$ if and only if

$$\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)x_t + \tau_0(x_t - \pi_0 z_t) = \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)x_t + \tau(x_t - \pi_0 z_t) \tag{25}$$

using the first stage equation $x_t = \pi_0 z_t + u_t$ we can rearrange this to

$$0 = \pi_0(\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0))z_t + (\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) + (\tau_0 - \tau))u_t$$

As $u_t \perp z_t$ this can only hold for any z_t, u_t if the terms premultiplying z_t and u_t are both zero. Solving the system of these two equations we get

$$\pi_0(\tau_0 - \tau) = 0,$$

with solutions $i)\,\tau = \tau_0$ and $ii)\,\pi_0 = 0$ and possibly $\tau \neq \tau_0$ (which means also $\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) = \tau - \tau_0 \neq 0$). It remains to verify what these solutions imply for the rest of the elements in $\boldsymbol{\theta}$. To investigate both solutions simultaneously, define $\kappa := \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)$, so that for solution $i)\,\kappa = 0$ and $ii)\,\kappa = \tau - \tau_0 \neq 0$ for any t. Then the filter at time $t+1$,

$$
\begin{aligned}
\kappa &= \tilde{\beta}_{t+1}(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_{t+1}(\boldsymbol{\theta}, \pi_0) \\
\kappa &= (\omega_0 + \alpha_0 \sigma_0^{-2} x_t \eta_t + \gamma_0 \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) \\
&\quad -(\omega + \alpha \sigma_0^{-2} x_t \underbrace{(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) x_t - \tau(x_t - \pi_0 z_t))}_{\tilde{\eta}_t} + \gamma \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) \\
\kappa &= (\omega_0 - \omega) + \frac{(\alpha_0 \eta_t - \alpha \tilde{\eta}_t)}{\sigma_0^2} x_t + (\gamma_0 \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \gamma \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)) \\
0 &= (\omega_0 - \omega) + \frac{(\alpha_0 \eta_t - \alpha \tilde{\eta}_t)}{\sigma_0^2} x_t + (\gamma_0 - \gamma) \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) + (\gamma - 1)\kappa
\end{aligned}
$$

Since the filter depends on the past, x_t and $\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)$ are independent, meaning that all premultiplying terms should be zero for this to hold. This gives rise to the following set of equations:

$$
\begin{aligned}
(\omega_0 - \omega) + (\gamma - 1)\kappa &= 0 \\
\alpha_0 \eta_t - \alpha \tilde{\eta}_t &= 0 \\
\gamma_0 - \gamma &= 0.
\end{aligned}
$$

The second equation gives us

$$
\begin{aligned}
0 = \alpha_0 \eta_t - \alpha \tilde{\eta}_t &= \alpha_0 \eta_t - \alpha(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) x_t - \tau(x_t - \pi_0 z_t)) \\
&= (\alpha_0 - \alpha)\eta_t - \alpha \left[\left(\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) + \tau_0 - \tau \right) x_t - (\tau_0 \pi_0 - \tau \pi_0) z_t \right] \\
&= (\alpha_0 - \alpha)\eta_t - \alpha \left[\left(\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) + \tau_0 - \tau \right) \pi_0 - (\tau_0 \pi_0 - \tau \pi_0) \right] z_t \\
&- \alpha \left(\tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0) + \tau_0 - \tau \right) u_t \\
&= (\alpha_0 - \alpha)\eta_t \implies \alpha_0 - \alpha = 0.
\end{aligned}
$$

Then we have (including the previously established values)

$$
\begin{aligned}
\sigma &= \sigma_0 \\
\pi &= \pi_0 \\
\gamma &= \gamma_0 \\
\alpha &= \alpha_0 \\
(\omega_0 - \omega) + (\gamma_0 - 1)(\tau - \tau_0) &= 0 \\
\pi_0(\tau - \tau_0) &= 0.
\end{aligned}
$$

Since $\pi_0 \neq 0$, we finally obtain that $\tau = \tau_0, \omega = \omega_0$ meaning that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

(C4) This is a direct result of Corollary 1.

(C5) Implied by Guassian density and boundedness of the parameter space

$$
\begin{aligned}
\ell_0(\boldsymbol{\theta}, \pi_0) &= -K - \frac{1}{2}\log(\sigma) - \frac{1}{2}\sigma^{-2}(y_t - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)x_t - \tau(x_t - \pi_0 z_t))^2 \\
&\leq -K - \frac{1}{2}\log(\sigma) \qquad \text{a.s.}
\end{aligned}
$$

Then

$$
\sup_{\boldsymbol{\theta}\in\Theta} \ell_0(\boldsymbol{\theta}, \pi_0) \leq -K - \frac{1}{2}\sup_{\boldsymbol{\theta}\in\Theta} \log\sigma < \infty,
$$

with probability 1 by the compactness of Θ. This also implies $E\|\ell_0(\boldsymbol{\theta}, \pi_0) \vee c\|_\Theta < \infty$.

(C6) Given the DGP with Gaussian i.i.d. errors, the first stage OLS estimator is strongly consistent (Anderson & Taylor, 1976). Note however that strong consistency also holds in more general cases, such as in dynamic models (Anderson & Taylor, 1979) or model with other less restrictive distributional assumptions (Lai et al., 1979; Christopeit & Helmes, 1980). □

Proof of Proposition 1

Proof. We will provide the proof by splitting it up in two parts.

(P1) $|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \pi_0, \hat{\beta}_1) - \beta_t^o| \xrightarrow{a.s.} 0$ as $T \geq t \to \infty$

We assume correct specification, which implies that $\beta_t^o = \tilde{\beta}(\boldsymbol{\theta}_0, \pi_0)$. To make dependencies explicit, denote the whole updating equation by $\beta_{t+1}(\boldsymbol{\theta}, \pi) = \phi(\beta_t(\boldsymbol{\theta}, \pi), Y_t, \pi; \boldsymbol{\theta}))$. Let $B_\varepsilon(\boldsymbol{\theta}) = \{\tilde{\boldsymbol{\theta}} \in \Theta : \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\| \leq \varepsilon\}$ be a compact neighbourhood of $\boldsymbol{\theta} \in \Theta$ with radius ε for some $\varepsilon > 0$. Since $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$, there exists some integer $N \in \mathbb{N}$ such that $\hat{\boldsymbol{\theta}}_T \in B_{\varepsilon_{\boldsymbol{\theta}_0}} := B_\varepsilon(\boldsymbol{\theta}_0)$ for any $T \geq t \geq N$.

We define the stationary sequence $\rho_t(\pi) := \sup_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in B_{\varepsilon_{\boldsymbol{\theta}_0}}\times B_{\varepsilon_{\boldsymbol{\theta}_0}}} \Lambda_t^*(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\pi)$ for some $\pi \in \Pi$. For any non-increasing sequence of constants $\{\varepsilon_i\}_{i\in\mathbb{N}}$ such that $\lim_{i\to\infty}\varepsilon_i = 0$, the sequence $\{\sup_{(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in B_{\varepsilon_i(\boldsymbol{\theta})}\times B_{\varepsilon_i(\boldsymbol{\theta})}} \log\Lambda_1^*(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\pi)\}$ is a non-increasing sequence (as $i \to \infty$). By condition (ii) of Lemma 1 we have that $E\sup_{(\theta_1,\theta_2)\in\Theta\times\Theta}\log\Lambda_1^*(\theta_1,\theta_2,\pi) \in \mathbb{R}^- \cup \{-\infty\}$. We can apply the monotone convergence theorem to conclude that

$$E\lim_{i\to\infty}\sup_{(\theta_1,\theta_2)\in B_{\varepsilon_i}(\boldsymbol{\theta})\times B_{\varepsilon_i}(\boldsymbol{\theta})}\log\Lambda_1^*(\theta_1,\theta_2,\pi) = E\log\Lambda_1(\boldsymbol{\theta},\pi)$$

Therefore, for every $\theta\in\Theta$ for which $E\log\Lambda_1(\boldsymbol{\theta}) < 0$, there exists an ε_θ such that

$$E\sup_{(\theta_1,\theta_2)\in B_{\varepsilon_\theta}(\boldsymbol{\theta})\times B_{\varepsilon_\theta}(\boldsymbol{\theta})}\log\Lambda_1^*(\theta_1,\theta_2,\pi) < 0$$

meaning that $E\log\rho_1(\pi) < 0$. With this result we can apply Lemma 2.5.5. from Straumann (2005) to conclude that $\prod_{k=1}^t \rho_k(\pi) \xrightarrow{e.a.s.} 0$, a result that will be used in the final part of the proof.

Then for $t > N$

$$\begin{aligned}
|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0,\hat{\beta}_1) - \beta_t^o| &= |\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0,\hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}_0,\pi_0)| \\
&\le |\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0,\hat{\beta}_1) - \tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0)| + |\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0,\pi_0)| \\
&\le \|\hat{\beta}_t(\boldsymbol{\theta},\pi_0,\hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta},\pi_0)\|_\Theta + |\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0,\pi_0)|
\end{aligned}$$

The first term converges e.a.s. to zero, by Lemma 1. The second term can be expanded and bounded by

$$\begin{aligned}
|\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0,\pi_0)| &= |\phi(\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0),Y_t,\pi_0;\hat{\boldsymbol{\theta}}_T)) - \phi(\tilde{\beta}_{t-1}(\boldsymbol{\theta}_0,\pi_0),Y_t,\pi_0;\boldsymbol{\theta}_0))| \\
&\le \underbrace{|\phi(\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0),Y_t,\pi_0;\hat{\boldsymbol{\theta}}_T)) - \phi(\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0),Y_t,\pi_0;\boldsymbol{\theta}_0))|}_{v_t(\hat{\boldsymbol{\theta}}_T)} \\
&\quad + |\phi(\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0),Y_t,\pi_0;\boldsymbol{\theta}_0)) - \phi(\tilde{\beta}_{t-1}(\boldsymbol{\theta}_0,\pi_0),Y_t,\pi_0;\boldsymbol{\theta}_0))|
\end{aligned}$$

By the Mean Value Theorem, the last term is bounded by

$$\sup_{\beta^*\in\mathcal{F}_{B_{\varepsilon_{\boldsymbol{\theta}_0}}}} |\dot{\phi}(\beta^*,Y_t,\pi_0;\boldsymbol{\theta}_0)|\,|\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_{t-1}(\boldsymbol{\theta}_0,\pi_0)|$$

This leads to the fact that

$$\begin{aligned}
|\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0,\pi_0)| &\le \rho_t(\pi_0)|\tilde{\beta}_{t-1}(\hat{\boldsymbol{\theta}}_T,\pi_0) - \tilde{\beta}_{t-1}(\boldsymbol{\theta}_0,\pi_0)| + v_t(\hat{\boldsymbol{\theta}}_T) \\
&\le \sum_{i\le t}\prod_{k=i+1}^t \rho_k(\pi_0)v_i(\hat{\boldsymbol{\theta}}_T)
\end{aligned}$$

by iterating backwards. With $\mathbb{E}[\log^+ \|\tilde{\beta}_1\|_\Theta] < \infty$ we can apply Lemma 2.5.2. of Straumann (2005), and conclude that the whole term converges almost surely. Moreover, by continuity of ϕ, $v_i(\hat{\boldsymbol{\theta}}_T) \xrightarrow{a.s.} 0$ for any $i \leq t \leq T$ as $t \to \infty$, meaning that this upper bound converges to zero a.s., as each summand vanishes. This concludes the proof. $\square$

(P2) $|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}, \hat{\beta}_1) - \beta_t^o| \xrightarrow{a.s.} 0$ as $T \geq t \to \infty$
Again we consider $N \in \mathbb{N}$ such that $\hat{\boldsymbol{\theta}}_T \in B_{\varepsilon_{\boldsymbol{\theta}_0}} := B_\varepsilon(\boldsymbol{\theta}_0)$ for any $T \geq t \geq N$. Then for $t > N$

$$
\begin{aligned}
&|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T, \hat{\beta}_1) - \beta_t^o| = |\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)| \\
&\leq |\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T)| + |\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T) - \tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \pi_0)| + |\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)| \\
&\leq \|\hat{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T, \hat{\beta}_1) - \tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T)\|_\Theta + \|\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \hat{\pi}_T) - \tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \pi_0)\|_\Theta + |\tilde{\beta}_t(\hat{\boldsymbol{\theta}}_T, \pi_0) - \tilde{\beta}_t(\boldsymbol{\theta}_0, \pi_0)|
\end{aligned}
$$

The first term vanishes e.a.s. by an application of Lemma 1. The last term is shown to converge to zero a.s. in **(P1)**. The remaining term can be shown to vanish a.s. by an application of the MVT, similar to the proof of **(P2)** in Theorem 1.

$$
\|\tilde{\beta}_t(\boldsymbol{\theta}, \hat{\pi}_T) - \tilde{\beta}_t(\boldsymbol{\theta}, \pi_0)\|_\Theta \leq \underbrace{\left\| \frac{\partial \tilde{\beta}_t(\boldsymbol{\theta}, \pi^*)}{\partial \pi} \right\|_\Theta}_{SE} \underbrace{|\hat{\pi}_T - \pi_0|}_{\xrightarrow{a.s.} 0}
$$

where π^* is a point between $\hat{\pi}_T$ and π_0. Since the derivative term is SE, we can apply our extension of Lemma 2.1 of Straumann and Mikosch (2006) to almost sure convergence (as mentioned in the proof of **(P2)** in Theorem 1) to conclude this term vanishes almost surely. This concludes the proof of the Theorem.

References

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, *93*(1), 113–132.

Anderson, T., & Taylor, J. B. (1976). Strong consistency of least squares estimates in normal linear regression. *The Annals of Statistics*, 788–790.

Anderson, T., & Taylor, J. B. (1979). Strong consistency of least squares estimates in dynamic models. *The annals of Statistics*, *7*(3), 484–489.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455.

Ashenfelter, O., & Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics*, *67*(4), 648–660.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, *119*(1), 249–275.

Beutner, E. A., Lin, Y., & Lucas, A. (2023). *Consistency, distributional convergence, and optimality of score-driven filters* (Tech. Rep.). Tinbergen Institute.

Bhatt, V., Kundan Kishor, N., & Marfatia, H. (2020). Estimating excess sensitivity and habit persistence in consumption using greenbook forecasts. *Oxford Bulletin of Economics and Statistics*, *82*(2), 257–284.

Blasques, F., Francq, C., & Laurent, S. (2022). *Autoregressive conditional betas* [Working paper].

Blasques, F., Francq, C., & Laurent, S. (2023). Quasi score-driven models. *Journal of Econometrics*, *234*(1), 251–275.

Blasques, F., Gorgi, P., Koopman, S. J., & Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, *12*(1), 1019–1052.

Blasques, F., Koopman, S. J., & Lucas, A. (2014). *Maximum likelihood estimation for generalized autoregressive score models* (Tech. Rep.). Tinbergen Institute Discussion Paper.

Blasques, F., Koopman, S. J., & Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, *102*(2), 325–343.

Blasques, F., Koopman, S. J., & Lucas, A. (2018). Amendments and corrections: 'information-theoretic optimality of observation-driven time series models for continuous responses'. *Biometrika*, *105*(3), 753–753.

Blasques, F., Koopman, S. J., & Lucas, A. (2020). Nonlinear autoregressive models with optimality properties. *Econometric Reviews*, *39*(6), 559–578.

Blasques, F., Lucas, A., & van Vlodrop, A. C. (2021). Finite sample optimality of score-driven volatility models: Some monte carlo evidence. *Econometrics and Statistics*, *19*, 47–57.

Blasques, F., van Brummelen, J., Koopman, S. J., & Lucas, A. (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics*, *227*(2), 325–346.

Boivin, J., Kiley, M. T., & Mishkin, F. S. (2010). How has the monetary transmission mechanism evolved over time? In *Handbook of monetary economics* (Vol. 3, pp. 369–422). Elsevier.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Elsevier.

Calvori, F., Creal, D., Koopman, S. J., & Lucas, A. (2017). Testing for parameter instability across different modeling frameworks. *Journal of Financial Econometrics*, *15*(2), 223–246.

Campbell, J. Y., & Mankiw, N. G. (1989). Consumption, income, and interest rates: Reinterpreting the time series evidence. *NBER macroeconomics annual*, *4*, 185–216.

Christopeit, N., & Helmes, K. (1980). Strong consistency of least squares estimators in linear regression models. *The Annals of Statistics*, *8*(4), 778–788.

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777–795.

Creal, D., Koopman, S. J., Lucas, A., & Zamojski, M. (2018). Generalized autoregressive method of moments.

Giraitis, L., Kapetanios, G., & Marcellino, M. (2021). Time-varying instrumental variable estimation. *Journal of Econometrics*, *224*(2), 394–415.

Gorgi, P., Koopman, S. J., & Schaumburg, J. (2017). Time-varying vector autoregressive models with structural dynamic factors. *Tinbergen Institute, The Netherlands and Aarhus University, Denmark*, 17.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 1–12.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. *Journal of political economy*, *86*(6), 971–987.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series* (Vol. 52). Cambridge University Press.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, *30*(1), 239-267. Retrieved from https://www.sciencedirect.com/science/article/pii/0304407685901393 doi: https://doi.org/10.1016/0304-4076(85)90139-3

Inoue, A., Rossi, B., & Wang, Y. (2022). Local projections in unstable environments: How effective is fiscal policy?

Kang, W., Ratti, R. A., & Yoon, K. H. (2015). Time-varying effect of oil market shocks on the stock market. *Journal of Banking & Finance*, *61*, S150–S163.

Kennan, J. (1989). Simultaneous equations bias in disaggregated econometric models. *The Review of Economic Studies*, *56*(1), 151–156.

Kim, C.-J. (2006). Time-varying parameter models with endogenous regressors. *Economics letters*, *91*(1), 21–26.

Kim, Y., & Kim, C.-J. (2011). Dealing with endogeneity in a time-varying parameter model: joint estimation and two-step estimation procedures. *The Econometrics Journal*, *14*(3), 487–497.

Koopman, S. J., Lucas, A., & Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics*, *98*(1), 97–110.

Krengel, U. (1985). *Ergodic theorems.* Berlin: De Gruyter.

Lai, T. L., Robbins, H., & Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression ii. *Journal of multivariate analysis*, *9*(3), 343–361.

Müller, U. K., & Petalas, P.-E. (2010). Efficient estimation of the parameter path in unstable time series models. *The Review of Economic Studies*, *77*(4), 1508–1539.

Pereira, M. C., & Lopes, A. S. (2014). Time-varying fiscal policy in the us. *Studies in Nonlinear Dynamics & Econometrics*, *18*(2), 157–184.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Straumann, D. (2005). *Estimation in conditionally heteroscedastic time series models.* Springer-Verlag, Berlin.

Straumann, D., & Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, *34*(5), 2449–2495.

Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics*, *27*(3), 531–543.

Thiele, S., & Harvey, A. (2013). *Time varying regression and changing correlation.* Retrieved from https://personal.vu.nl/s.j.koopman/projects/ScoreWorkShop2013/Steve%20Thiele%20talk.pdf (Score Workshop Vrije Universiteit Amsterdam)

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, *51*(6), 309.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, *20*(4), 595–601.

Wooldridge, J. M. (2002). M-estimation. In *Econometric analysis of cross section and panel data.* Cambridge: MIT Press.

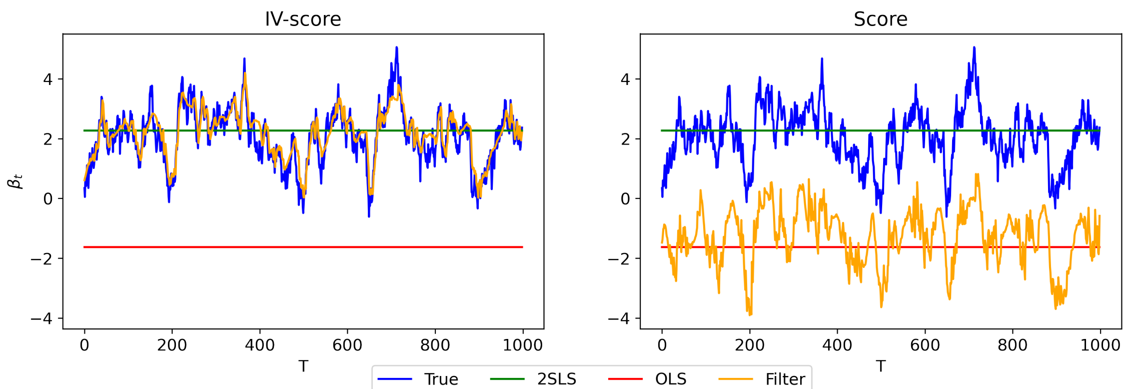
Wooldridge, J. M. (2009). Omitted variable bias: the simple case. *Introductory Econometrics: A Modern Approach. Mason, OH: Cengage Learning*, 89–93.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, *50*(2), 420–445.

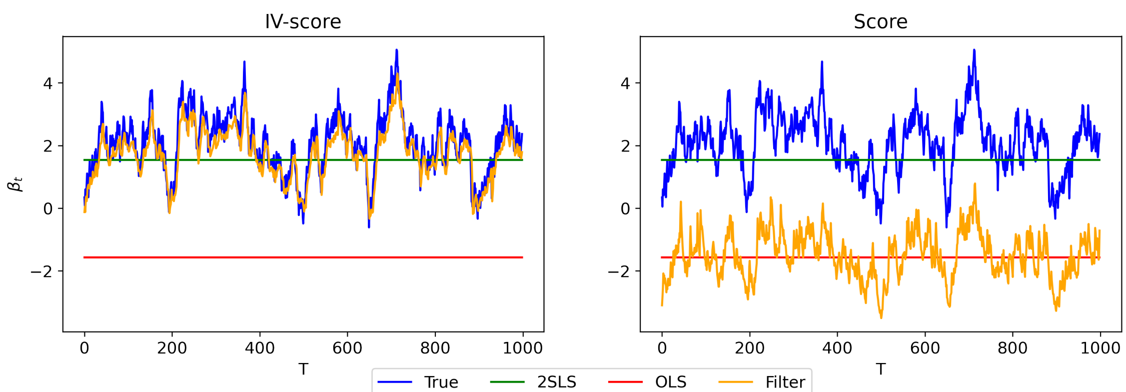
Supplementary Appendix

A Score-Driven Filter for Causal Regression Models with Time-Varying Parameters and Endogenous Regressors

Francisco Blasques and *Noah Stegehuis*



(a) h_t following sine dynamics (DGP 2).



(b) h_t with a midway break (DGP 3).

Figure 4: Estimated paths of the causal parameter for DGPs 2 and 3.

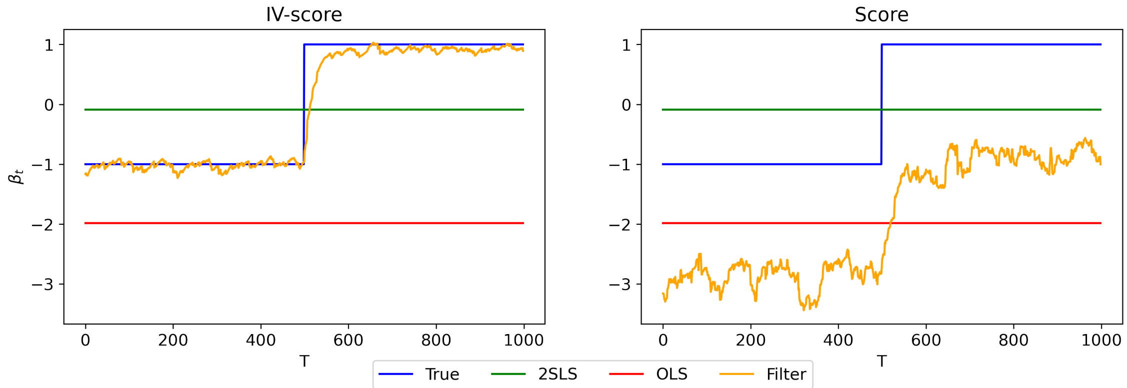


Figure 5: β_t with a midway break (DGP 4).

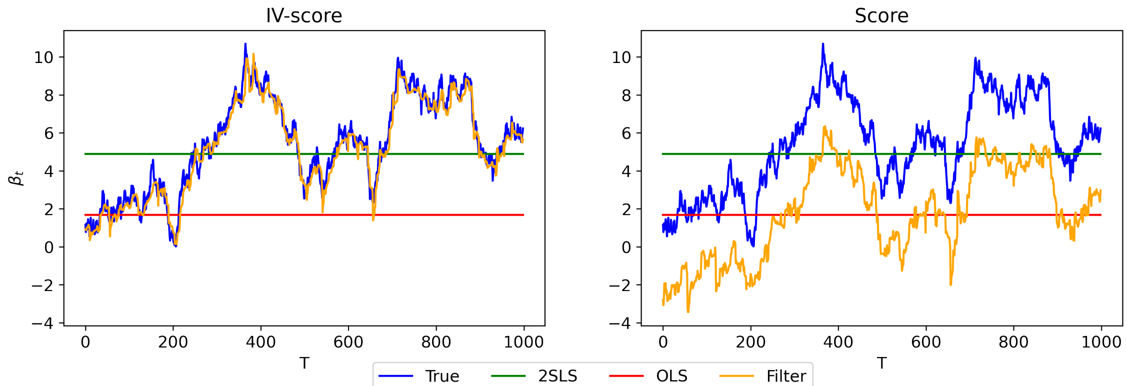


Figure 6: β_t following a random walk (DGP5).

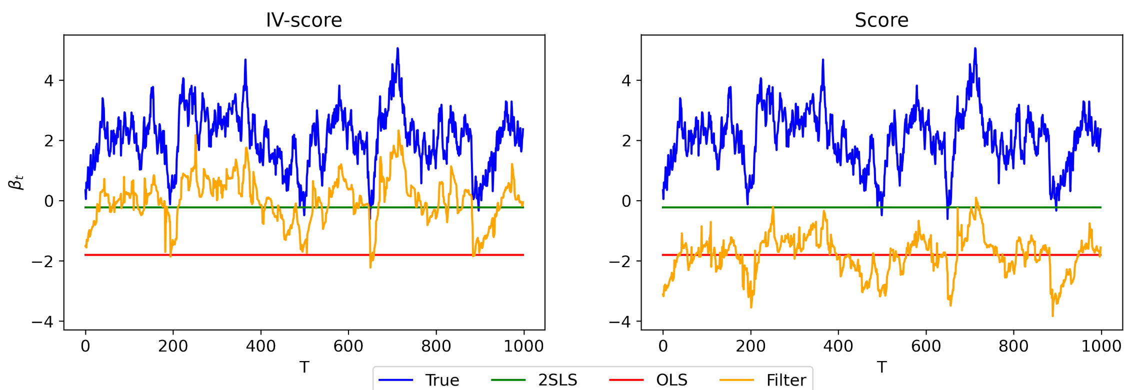


Figure 7: Filtering with an endogenous instrument (DGP 6).

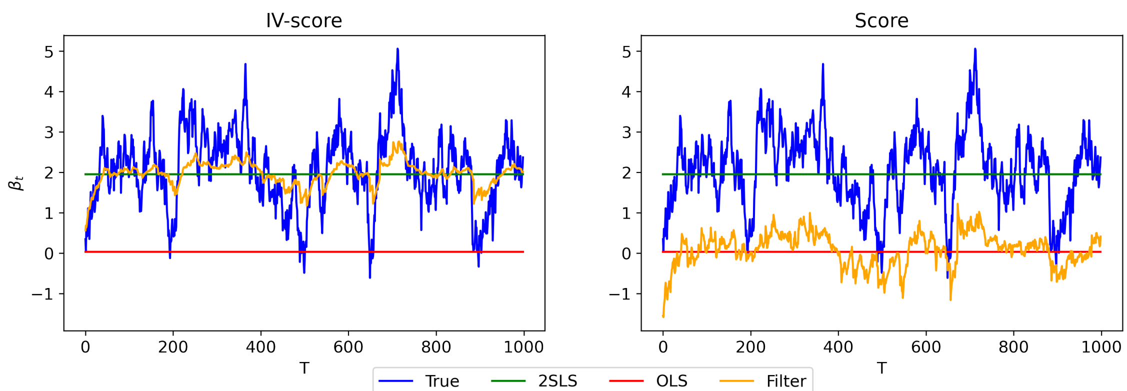


Figure 8: Filtering with large error variance (DGP7).