

Creal, Drew; Koopman, Siem Jan; Lucas, André; Zamojski, Marcin

Working Paper

Generalized Autoregressive Method of Moments

Tinbergen Institute Discussion Paper, No. 15-138/III

Provided in Cooperation with:

Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Creal, Drew; Koopman, Siem Jan; Lucas, André; Zamojski, Marcin (2015) : Generalized Autoregressive Method of Moments, Tinbergen Institute Discussion Paper, No. 15-138/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:

<https://hdl.handle.net/10419/130481>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Generalized Autoregressive Method of Moments

Drew Creal¹

Siem Jan Koopman^{2,3}

André Lucas^{2,3}

Marcin Zamojski^{2,3}

¹ *The University of Chicago Booth School of Business, United States;*

² *Faculty of Economics and Business Administration, VU University Amsterdam;*

³ *Tinbergen Institute, the Netherlands.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at <http://www.tinbergen.nl>

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Generalized Autoregressive Method of Moments^{*}

DREW CREAL,¹ SIEM JAN KOOPMAN,² ANDRÉ LUCAS,² AND MARCIN ZAMOJSKI²

¹*University of Chicago, Booth School of Business*

²*VU University Amsterdam and Tinbergen Institute*

This version: October 7, 2015

ABSTRACT

We introduce a new estimation framework which extends the Generalized Method of Moments (GMM) to settings where a subset of the parameters vary over time with unknown dynamics. To filter out the dynamic path of the time-varying parameter, we approximate the dynamics by an autoregressive process driven by the score of the local GMM criterion function. Our approach is completely observation driven, rendering estimation and inference straightforward. It provides a unified framework for modeling parameter instability in a context where the model and its parameters are only specified through (conditional) moment conditions, thus generalizing approaches built on fully specified parametric models. We provide examples of increasing complexity to highlight the advantages of our method.

Keywords: dynamic models, time-varying parameters, generalized method of moments, non-linearity.

JEL classification codes: C10, C22, C32, C51.

^{*}We thank Eric Ghysels, Lars Hansen, Mark Watson, and participants of the Dynamic Models driven by the Score of Predictive Likelihoods Workshops in Amsterdam (Jan 2013) and La Laguna (Jan 2014), the SoFiE Workshop in Cambridge (Apr 2014), the NBER-NSF Time Series Conference at the FED St. Louis (Sep 2014), the Econometric Society European Winter Meeting (Dec 2014), the 8th Annual SoFiE Conference in Aarhus (Jun 2015), the Montréal 2015 World Congress of the Econometric Society (Aug 2015), the 2015 European Finance Association Annual Meeting in Vienna (Aug 2015), and the European Economic Society 2015 meeting in Mannheim (Aug 2015) for helpful comments. Creal gratefully acknowledges financial support from the William Ladany Faculty Scholar Fund at the University of Chicago Booth School of Business. Lucas and Zamojski are grateful to the Dutch National Science Foundation (NWO, grant VICI453-09-005) for financial support. Koopman acknowledges support from CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78) at Aarhus University, Denmark, and funded by the Danish National Research Foundation; as well as support from the European Union Seventh Framework Programme (FP7-SSH/2007-2013, grant agreement 320270 - SYRTO). All numerical results presented in this paper were obtained with OxMetrics 7/Python/Julia at the Dutch National Cluster (LISA).

1 Introduction

We introduce a new estimation framework which extends the Generalized Method of Moments (GMM) of Hansen (1982) to allow for time-variation in a subset of the parameters. Our approach only requires the researcher to specify a set of conditional moment conditions and a set of parameters that are believed to vary over time. Given these moment conditions, we approximate the unknown dynamics of the time-varying parameter by an autoregressive process whose shocks are linear transformations of the scaled gradient of the conditional GMM objective function. This adjusts the parameters in a (local) steepest descent direction using the model’s objective function at time t . The resulting dynamics for the time-varying parameter are observation-driven, making estimation of the model straightforward. We label our approach as the Generalized autoregressive Method of Moments (GaMM) and provide several empirical applications that illustrate its usefulness.

GaMM directly builds on GMM; see Hansen (1982). GMM is appealing because it provides a unified framework for estimation and testing using only a vector of moment conditions. Moment conditions are often derived from economic theory, and express an economic agent’s conditional expectations over future outcomes given an appropriate information set. GMM does not require the researcher to specify the entire data generating process, which economic theory often does not provide. However, many economic models do require additional flexibility to match key features of the data, such as time-varying conditional means, conditional heteroscedasticity, or regime shifts. Research over the past decade has emphasized the economic importance of capturing these features by introducing latent variables or time-varying parameters into the model; for surveys in macro and financial economics, see, e.g. Shephard (2005), Hamilton (2010), and Fernández-Villaverde and Rubio-Ramírez (2013).

Despite the widespread appeal of GMM estimators, their extension to handle time-varying parameter models typically requires simulation based estimators as discussed further below. Our approach using observation-driven GaMM dynamics for the time-varying parameters offers a complementary and easy-to-implement alternative to these procedures.

We provide three empirical applications that illustrate the usefulness of GaMM estimation. Our examples highlight settings where traditional techniques are either difficult to implement or no alternative technique is readily available. These applications include: (i) estimation of sta-

ble distributions with time-varying scale parameters where no closed-form density exists, thus making likelihood based estimation a challenge; (ii) consumption-based asset pricing models with myopic expectations formation and unstable structural risk aversion parameters. A web appendix contains further material for a setting with time-varying parameters in a linear regression model with an endogenous regressor.

GaMM factor dynamics use the gradient of the local GMM objective function to determine next period's value of the time-varying parameter. We show that this factor recursion based on the gradient satisfies local optimality properties, even if the moment conditions are misspecified. In particular, the GaMM dynamics result in parameter changes that improve the local quadratic GMM objective function formulated in the (possibly misspecified) moment conditions. Similar optimality results were established for the more specialized maximum likelihood framework with the generalized autoregressive score dynamics of Creal et al. (2011, 2013) and Harvey (2013); see Blasques et al. (2015). GaMM factor dynamics have an additional advantage because they are observation-driven in the sense of Cox (1981): parameters vary over time as a function of lagged dependent variables and exogenous variables. Next period's parameter values are perfectly predictable given the current information set. The recursive nature of the estimation problem is similar to generalized autoregressive conditional heteroscedasticity (GARCH) models. Estimation and inference is relatively straightforward in the GaMM framework and does not require simulation.

Our paper is related to two strands of literature. The first deals with the estimation of parameter-driven (state space) models with partially specified conditional observation densities. For traditional parameter-driven state space models with fully specified observation densities, see for instance Kim and Nelson (1999) and Durbin and Koopman (2012). Our interest lies in the setting where the observation density is only partially specified. Method of moments (or more generally minimum distance) estimators become more attractive than fully specified parametric likelihood methods in this setting. Procedures for estimating the unknown, static parameters of parameter-driven models by method of moment estimators include the simulated method of moments of McFadden (1989), the efficient method of moments by Gallant and Tauchen (1996), indirect inference as in Gouriéroux et al. (1993), and the recent extension of GMM to latent variables by Gallant et al. (2014). Except for a few special cases, estimation of the latent variables in parameter-driven models can become quite involved. Estimation of the latent, time-varying parameters in models

whose conditional observation densities are only partially specified is even more challenging. Two contributions in this setting are the re-projection method proposed by Gallant and Tauchen (1998) in conjunction with their efficient method of moments (EMM) and the approach by Gallant et al. (2014) that uses sequential Monte Carlo methods. In contrast to the above approaches, the unknown path of the time-varying parameters in our GaMM framework follows directly as a by-product of the (fairly straightforward) estimation of the model's static parameters.

Second, our paper extends the literature on method of moments estimation of fully-specified observation-driven models. The most prominent example of this is Gaussian quasi-maximum likelihood (QML) estimation of GARCH models in the absence of conditional normality; see for example the overview of Francq and Zakoïan (2010). QML estimation of GARCH models can be viewed as a special case of GaMM estimation, with the latter offering more flexibility to include additional conditional moment conditions. The GaMM estimation framework also generalizes other observation-driven approaches proposed in the recent literature. In particular, when the conditional moment conditions are the scores of a fully parametric likelihood function, GaMM encompasses the generalized autoregressive score approach of Creal et al. (2011, 2013) and Harvey (2013). Consequently, GaMM nests many popular econometric models including the GARCH model of Engle (1982) and Bollerslev (1986), the ACD model of Engle and Russell (1998), as well as many new models for time-varying parameters under fat-tails and mixed observation densities; see the references in Creal et al. (2011, 2013) as well Harvey and Luati (2014), Lucas et al. (2014), and Creal et al. (2014). In addition, our framework gives rise to new time-varying parameter models that have not been studied before.

The remainder of this paper is organized as follows. In Section 2, we describe the basic methodology. Section 3 contains examples to illustrate the relevance of GaMM for applied work. In Section 4, we discuss penalized extensions of the methodology that allow for improved finite sample properties of the estimator. Section 5 concludes.

2 Methodology

2.1 A motivating example

We start with a motivating example. Consider the problem of estimating the mean μ of a random variable y_t using the moment condition $E[y_t - \mu] = 0$. The standard GMM objective

function for this problem is $(\sum_{t=1}^T (y_t - \mu))^2$, with solution $\hat{\mu} = T^{-1} \sum_{t=1}^T y_t$. Assume the true mean of y_t changes at time $\tau + 1$, such that $E[y_t] = \mu_0$ for $t = 1, \dots, \tau$, and $E[y_t] = \mu_1$ for $t = \tau + 1, \dots, T$, where $\mu_0 < \mu_1$. Using the same (unconditional) moment condition to estimate μ and abstracting for a moment from potential finite sample issues, the full-sample GMM estimate is too high for the first part of the sample, and too low for the second compared to the true conditional mean. Furthermore, if μ_0 is substantially below μ_1 , then $(y_t - \hat{\mu})$ is negative on average for $t = 1, \dots, \tau$, and positive for $t = \tau + 1, \dots, T$. Put differently, the moment condition evaluated for the observation at the time t provides a signal about the direction in which to adjust $\hat{\mu}$ to obtain a better fit to the data.

If consecutive observations y_t give a persistent signal that the current estimate $\hat{\mu}$ of μ is too high, it may be advisable to temporarily lower the value of $\hat{\mu}$, as decreasing $\hat{\mu}$ at time t is likely to reduce the predictive variance of $y_{t+1} - \hat{\mu}$. The converse holds if the data signals that the current estimate $\hat{\mu}$ of μ is too low. Thus, if the true parameter varies slowly over time or only changes incidentally, an adjustment based on time t 's moment condition helps reduce the criterion function at time $t + 1$. It is precisely this persistence in 'misfit' that we exploit in the Generalized autoregressive Method of Moments (GaMM) dynamics.

To introduce the GaMM dynamics for a time-varying parameter f_t , consider a GMM criterion function for the observation at time t only, i.e., $E_{t-1}[y_t - f_t]^2$, where the conditional mean f_t replaces the unconditional mean μ , and the conditional expectation $E_{t-1}[\cdot]$ replaces its unconditional counterpart $E[\cdot]$. Taking the derivative of this objective function with respect to f_t and evaluating it at the t -th observation rather than taking the expectation, we obtain

$$s_t = -2(y_t - f_t). \quad (1)$$

We use this gradient s_t of the time t objective function to formulate autoregressive dynamics for the time-varying parameter f_t . For example, with autoregressive dynamics of order one, we set

$$f_{t+1} = \omega \cdot (I - B_1) + B_1 f_t + A_1 s_t, \quad (2)$$

where ω , A_1 , and B_1 are static parameters that need to be estimated and that describe the dynamic behavior of f_t . It is easy to generalize this specification to include more lags of f_t and s_t (see Creal et al., 2013), non-linearity, structural time series dynamics (see Harvey and Luati,

2014), or fractional integration (see Janus et al., 2014).

It is evident that \mathbf{f}_t in (2) has observation-driven dynamics. Given information up to time t , the parameter \mathbf{f}_{t+1} is known as it only depends on y_t, y_{t-1}, \dots which is similar to a GARCH model. This makes the proposed GaMM methodology computationally fast and renders parameter estimation and inference straightforward. We emphasize that the GaMM factor dynamics are not arbitrary. In Section 2.3, we describe the optimality properties of these dynamics.

From the example above, define the vector of static parameters as $\boldsymbol{\theta} = (\omega, A_1, B_1)^\top$. To estimate $\boldsymbol{\theta}$, we need unconditional moment conditions which we obtain these by instrumenting the conditional moment condition $E_{t-1}[y_t - \mathbf{f}_t]$. We propose the unconditional moment conditions

$$E[(y_t - \mathbf{f}_t) \otimes (1, \mathbf{f}_{t-1}, s_{t-1})^\top] = 0, \quad (3)$$

with \otimes denoting the Kronecker product. Due to the presence of the constant term in the vector of instruments, the original (conditional) moment condition $E[y_t - \mathbf{f}_t]$ also needs to hold unconditionally. GaMM thus provides a natural extension of the static GMM moment conditions that we started out with. GaMM dynamics exploit any persistence in the misfit of the original moment condition $E_{t-1}[y_t - \mathbf{f}_t] = 0$ by including in (3) the autocorrelation of the misfit of the cross-product between $s_t = -2(y_t - \mathbf{f}_t)$ and s_{t-1} . The fact that the unconditional expectation of this cross-product needs to be zero in (3) forces the dynamic scheme in (2) to remove as much autocorrelation in $(y_t - \mathbf{f}_t)$ as possible.

The distributional properties of the GMM estimator for $\boldsymbol{\theta}$ in a framework with GaMM dynamics turn out to be straightforward. Equation (3) fits into the GMM framework of Hansen (1982) under standard assumptions. Therefore, consistency and asymptotic normality of the estimator for $\boldsymbol{\theta}$ follow easily, as do the optimal weighting matrices for two-stage feasible GMM estimation of $\boldsymbol{\theta}$. In the remaining sections, we provide the formal background of the intuitive results presented in this section and show the performance of GaMM estimation in a range of different settings.

2.2 GaMM dynamics

To introduce Generalized autoregressive Methods of Moments (GaMM) dynamics, consider the moment conditions

$$\mathbb{E} [\mathbf{g}_t(\mathbf{w}_t; f, \boldsymbol{\theta})] = 0, \quad (4)$$

where $\mathbf{g}_t : \mathbb{R}^m \times \mathcal{F} \times \Theta \rightarrow \mathbb{R}^K$, $\mathbf{w}_t \in \mathbb{R}^m$ is observed, and $f \in \mathcal{F}$ and $\theta \in \Theta$ denote parameter vectors that lie in the parameter spaces \mathcal{F} and Θ , respectively. In our case, we assume that f varies over time as f_t with unknown dynamics. We approximate the dynamics of f_t in an observation-driven way and denote the approximation as \mathbf{f}_t . As \mathbf{f}_t is observation-driven, it can be written as a function of past observations $\mathbf{w}_{t-1}, \mathbf{w}_{t-2}, \dots$

We now replace the unconditional moment condition (4) by its conditional counterpart

$$\mathbb{E}_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0. \quad (5)$$

The dynamic specification for \mathbf{f}_t starts by considering the GMM objective function at time t ,

$$\mathbb{E}_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})]^\top \boldsymbol{\Omega}_t \mathbb{E}_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})]. \quad (6)$$

where $\boldsymbol{\Omega}_t$ is a time $t - 1$ measurable weighting matrix that is positive semi-definite almost surely. The expectations in (6) are computed under the true time $t - 1$ conditional measure $F_{\mathbf{w}}$ of \mathbf{w}_t , for which (5) holds. To propagate \mathbf{f}_t forward to \mathbf{f}_{t+1} given the realization of \mathbf{w}_t , we take a scaled steepest descent step of (6) using an appropriate derivative of (6) evaluated at \mathbf{f}_t . The derivative concept used should account for the specific value of \mathbf{w}_t realized at time t . In our setting, the appropriate concept is given by the Fréchet derivative. It is directly related to the concept of the influence function of the estimator for \mathbf{f}_t given the objective function (6); see, e.g., Hampel et al. (2011).

To define the Fréchet derivative, consider a contaminated measure $F_{\mathbf{w}}^\epsilon = (1 - \epsilon)F_{\mathbf{w}} + \epsilon \delta_{\mathbf{w}_t}$, where $\delta_{\mathbf{w}_t}$ is the Dirac measure that puts unit mass on the realized value of \mathbf{w}_t . By considering $F_{\mathbf{w}}^\epsilon$ instead of $F_{\mathbf{w}}$, we account for the appropriate information in \mathbf{w}_t when updating \mathbf{f}_t to \mathbf{f}_{t+1} .

The first order condition corresponding to (6) evaluated at the measure F_w^ϵ is

$$(\mathbf{G}_t^\epsilon)^\top \Omega_t \mathbb{E}_{t-1}^\epsilon [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0, \quad \mathbf{G}_t^\epsilon = \mathbb{E}_{t-1}^\epsilon [\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}) / \partial \mathbf{f}_t^\top], \quad (7)$$

where $\mathbb{E}_{t-1}^\epsilon$ and $\mathbb{E}_{t-1} \equiv \mathbb{E}_{t-1}^0$ denote the expectations operators taken with respect to F_w^ϵ and $F_w^0 \equiv F_w$, respectively. The influence function of \mathbf{f}_t for the criterion function (6) is the Fréchet derivative of (7) in the direction $\delta_{\mathbf{w}_t}$, evaluated at $\epsilon = 0$. After some minor algebra, it is easy to show that the influence function equals $\mathbf{G}_t^\top \Omega_t \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})$, with $\mathbf{G}_t = \mathbf{G}_t^0$ for $\epsilon = 0$; see again Hampel et al. (2011) for more details. We scale this influence function by its inverse conditional covariance matrix to account for the curvature of the objective function and to obtain a Gauss-Newton type improvement when updating \mathbf{f}_t to \mathbf{f}_{t+1} . The resulting scaled step is

$$\nabla_t = -(\mathbf{G}_t^\top \Omega_t \mathbf{G}_t)^* \mathbf{G}_t^\top \Omega_t \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}), \quad (8)$$

where H^* denotes the Moore-Penrose pseudo-inverse of a general matrix H . The use of a pseudo-inverse rather than a regular inverse in equation (7) is important, because some of the moment and scaling matrices in a GMM context may be rank deficient in general. For example, this can arise when \mathbf{G}_t does not depend on either the data or the time-varying parameters; see the case of the stable distribution in Section 3.1, where most of the relevant matrices have rank one.

We could use the steps ∇_t directly in a random walk type updating scheme $\mathbf{f}_{t+1} = \mathbf{f}_t + \nabla_t$. Such dynamics are a special case of the more general autoregressive scheme

$$\mathbf{f}_{t+1} = \omega + \sum_{j=1}^p \mathbf{B}_j (\mathbf{f}_{t+1-j} - \omega) + \sum_{i=1}^q \mathbf{A}_i \mathbf{s}_{t+1-i}, \quad (9)$$

which we call GaMM(p, q) dynamics, where $\omega = \omega(\boldsymbol{\theta})$, $\mathbf{B}_j = \mathbf{B}_j(\boldsymbol{\theta})$, and $\mathbf{A}_i = \mathbf{A}_i(\boldsymbol{\theta})$ are appropriately sized vectors and matrices that depend on the static parameter vector $\boldsymbol{\theta}$, and

$$\mathbf{s}_t = -\mathbf{S}_t \nabla_t, \quad (10)$$

for some time $t - 1$ measurable (almost surely) positive semi-definite scaling matrix \mathbf{S}_t . The scaling matrix \mathbf{S}_t adjusts the direction of the step, for example, if one wants to annihilate the effect of $(\mathbf{G}_t^\top \Omega_t \mathbf{G}_t)^*$ in (8) to fully concentrate on the steepest descent direction. Additionally, we can use an efficient conditional weighting matrix $\Omega_t = \mathbb{E}_{t-1}[\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}) \mathbf{g}_t^\top(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})]^*$ in (6).

However, as Ω_t needs to be computed at each time t , this is computationally more demanding. In this paper, we set $\Omega_t = \mathbf{I}$ and $S_t = \mathbf{I}$, which provides good results in the empirical examples considered later on; see Section 3 and Supplemental Appendix A.

Under the assumption that equation (5) holds, $\{s_t\}_{t \in \mathbb{Z}}$ forms a martingale difference sequence $E_{t-1}[s_t] = 0$, where the expectation is taken with respect to the probability measure F_w . Due to properties of martingale differences, the unconditional mean of f_t in (9) is $E[f_t] = \omega$ if f_t is first-order stationary. To start up the recursion in (9), we can set $f_t = \omega$ for $t = 0, -1, \dots, -p+1$ and $s_t = 0$ for $t = 0, -1, \dots, -q+1$. Alternatively, the initial values can be estimated together with the static parameters in θ , though such estimates of initial values will not be consistent.

For the remainder of the discussion, we set $p = q = 1$ and consider the case of GaMM(1,1) dynamics with $A = A_1$ and $B = B_1$. If $A = B = 0$ and $f_1 = \omega$, GaMM dynamics reproduce a static parameter $f_t \equiv \omega$. The static parameter framework is thus a special case of GaMM. We can exploit this feature to write down a joint model for the static and dynamic parameters. Partition the vector of static parameters as $\theta = (\theta_f^\top, \theta_c^\top)^\top$ where θ_f contains the parameters governing the dynamics of f_t , i.e., ω , A , and B , and θ_c includes the remaining static parameters.

The joint GaMM(1,1) dynamics for the vector $\tilde{f}_{t+1} = (f_t^\top, \theta_c^\top)^\top$ are

$$\begin{aligned} \tilde{f}_{t+1} &= \begin{bmatrix} f_t \\ \theta_c \end{bmatrix} = \tilde{\omega} + \tilde{B} (\tilde{f}_{t-j+1} - \tilde{\omega}) + \tilde{A} \tilde{s}_{t-i+1}, \\ \tilde{B} &= \begin{bmatrix} B & C^B \\ 0 & 0 \end{bmatrix}, \quad \tilde{\omega} = \begin{bmatrix} \omega \\ \theta \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} A & C^A \\ 0 & 0 \end{bmatrix}, \end{aligned} \quad (11)$$

where \tilde{s}_t contains the derivatives with respect to the entire parameter vector \tilde{f}_t rather than f_t only. The matrices C^A and C^B allow the time-varying parameter f_t to also react to the score of the static parameters θ_c . This can be useful, for example, when modeling financial returns with constant mean but time-varying volatility. In that case the matrices C^A capture the leverage effect. To see this, consider a model with time-varying variance σ_t^2 and constant mean μ , i.e., $y_t = \mu + \varepsilon_t$, with $\varepsilon_t = \sigma_t z_t$, and $z_t \sim D(0, 1)$ for some distribution D with zero mean and variance one. We obtain that

$$g_t = \begin{bmatrix} y_t - \mu \\ (y_t - \mu)^2 - \sigma_t^2 \end{bmatrix} = \begin{bmatrix} \varepsilon_t \\ \varepsilon_t^2 - \sigma_t^2 \end{bmatrix}, \quad G_t = -\mathbf{I}. \quad (12)$$

Fixing $C^B = 0$, the GaMM(1,1) recursion for $\mathbf{f}_t = \sigma_t^2$ is

$$\begin{aligned}\sigma_{t+1}^2 &= (\sigma_t^2 - \omega)\mathbf{B} + \omega + \mathbf{C}^A \varepsilon_t + \mathbf{A} \varepsilon_t^2 - \mathbf{A} \sigma_t^2 \\ &= \underbrace{\omega(1 - \mathbf{B}) + \mathbf{A} \varepsilon_t^2 + (\mathbf{B} - \mathbf{A}) \sigma_t^2}_{\text{GARCH(1,1)}} + \underbrace{\mathbf{C}^A \varepsilon_t}_{\text{Leverage}}.\end{aligned}\quad (13)$$

Equation (13) coincides with the familiar GARCH(1,1) model of Engle (1982) and Bollerslev (1986) with an additional leverage effect $\mathbf{C}^A \varepsilon_t$. The leverage effect is similar in form and spirit to the optimal leverage effect of GARCH filters in a misspecified model setting as laid out in Nelson and Foster (1994). Allowing for C^B and/or C^A to be different from zero generalizes the score driven approach from Creal et al. (2013).

2.3 Local optimality of GaMM

In this section, we derive generic local optimality properties for the GaMM dynamics introduced in Section 2.2. Similar optimality properties were derived in Blasques et al. (2015) for the generalized autoregressive score model of Creal et al. (2011, 2013). In particular, Blasques et al. show that generalized autoregressive score updates improve the local Kullback–Leibler divergence between the true data density and the model density. They further show that any observation-driven update with similar optimality properties needs to be ‘score-equivalent’. All these results, however, are framed entirely in the setting of information theoretic optimality and Kullback–Leibler divergences. This follows directly from the use of the conditional log observation density as the criterion function in the generalized autoregressive score framework. In our current GMM context, optimality instead centers around the quadratic objective function of the moment conditions. As a result, the concepts and results in Blasques et al. need to be adapted accordingly. Interestingly, the optimality results in Blasques et al. (2015) hold whether or not the statistical model is correctly specified. Similarly, our results hold whether or not the moment conditions $E[\mathbf{g}(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$ are correctly specified.

We introduce the local GMM objective function

$$\mathcal{C}(t, \mathbf{f}_t, \mathcal{W}) = E_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}, \boldsymbol{\theta}) | \mathbf{w}_t \in \mathcal{W}]^\top \boldsymbol{\Omega}_{t-1} E_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}, \boldsymbol{\theta}) | \mathbf{w}_t \in \mathcal{W}]. \quad (14)$$

Equation (14) considers the behavior of the GMM objective function (6) for a restricted set $\mathcal{W} \subseteq \mathbb{R}^m$ of realizations of the random variable \mathbf{w}_t . If $\mathcal{W} = \mathbb{R}^m$, then (14) coincides with (6). For

concreteness, let w_t denote the realization of the random variable \mathbf{w}_t at time t . Our results are local in nature in that they focus on sets \mathcal{W} that contain the empirical realization w_t , i.e., $\{w_t\} \subset \mathcal{W}$. This approach aligns with the nature of observation-driven models, which intend to improve the ‘model fit’ close to the most recent observation w_t of \mathbf{w}_t . Though we hope that such an improvement also leads to a better fit of the model and moment conditions for other regions of the sample space, observation-driven update steps concentrate on making the model fit better near the most recent realizations of the data. Note that we do not require the moment conditions to be correctly specified, i.e., we do not require $\mathbb{E}_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$. In that sense we only focus on update steps that reduce the model’s misfit as measured in terms of the quadratic function of the moment condition. We introduce the following notions of optimality.

Definition 1. *Conditional on a realization $\mathbf{w}_t = w_t \in \mathcal{W}$ for a neighborhood \mathcal{W} of w_t , an update from \mathbf{f}_t to \mathbf{f}_{t+1} is called Realized Locally (RL) optimal if and only if*

$$\mathcal{C}(t, \mathbf{f}_t, \mathcal{W}) - \mathcal{C}(t, \mathbf{f}_{t+1}, \mathcal{W}) \geq 0, \quad (15)$$

for every (w_t, \mathbf{f}_t) . The update is called Conditionally Locally (CL) optimal if and only if

$$\int_{\mathcal{W}} [\mathcal{C}(t, \mathbf{f}_t, \mathcal{W}) - \mathcal{C}(t, \mathbf{f}_{t+1}, \mathcal{W})] dF_w(w_t) \geq 0, \quad (16)$$

for every \mathbf{f}_t , where F_w is the distribution of \mathbf{w}_t conditional on the information up to time $t - 1$ and conditional on $\mathbf{w}_t \in \mathcal{W}$.

Here, \mathbf{f}_{t+1} can be seen as a function of w_t . Note that \mathbf{f}_t does not depend on w_t . The concept of RL optimality considers updates from \mathbf{f}_t to \mathbf{f}_{t+1} that improve the local criterion $\mathcal{C}(t, \mathbf{f}_t, \mathcal{W})$ in a neighborhood \mathcal{W} of the realization w_t for given values of $\mathbf{w}_t = w_t$ and \mathbf{f}_t . It, thus, treats the realized data value w_t as given. CL optimality goes one step further and no longer conditions on the realization w_t , but rather takes the expectation over the entire neighborhood $\mathbf{w}_t \in \mathcal{W}$. This accounts for the fact that \mathbf{w}_t has an impact on \mathbf{f}_{t+1} through the update equation. Note that Definition 1 generalizes the definitions in Blasques et al. (2015) from a fully parametric setting to the semi-parametric setting of GMM. The proof of the following proposition can be found in Section Appendix A.

Proposition 1. *Let $\mathbf{g}_t(\mathbf{w}; \mathbf{f}, \boldsymbol{\theta})$ and $\mathbf{G}_t = \mathbf{G}_t(\mathbf{w}; \mathbf{f}, \boldsymbol{\theta})$ be continuous in all their arguments, and let $\mathbf{G}_t^\top \boldsymbol{\Omega}_{t-1} \mathbf{G}_t$ be positive definite for all $\mathbf{w}_t \in \mathcal{W}$ and given \mathbf{f}_t . Then there is a GaMM(1,1) update with*

$\omega = 0$, $\mathbf{A} = \mathbf{a} \cdot \mathbf{I}$ for $\mathbf{a} \in \mathbb{R}^+$, and $\mathbf{B} = \mathbf{I}$ that is both RL-optimal and CL-optimal.

Proposition 1 ensures that the score driven GaMM(1,1) factor dynamics improve the local GMM objective function (14) at each time step. These results hold without specifying the true conditional distribution of the data F_w and without assuming that the conditional moment conditions $E_{t-1} [\mathbf{g}(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$ are correctly specified: the GaMM dynamics still operate to minimize the local deviations from the moment conditions laid down by the econometrician by efficiently processing the new observations and updating \mathbf{f}_t to \mathbf{f}_{t+1} . Note that the positive semi-definiteness of the matrix $\mathbf{G}_t^\top \boldsymbol{\Omega}_{t-1} \mathbf{G}_t$ may be easily satisfied, depending on the specific setting. For example, if \mathbf{G}_t does not depend on \mathbf{w}_t such as in our example of stable distributions with time-varying scale, the positive semi-definiteness follows directly for small enough \mathcal{W} .

Along the same lines as the results in Blasques et al. (2015) for generalized autoregressive score models, the results for GaMM dynamics can be substantially extended to establish non-local optimality properties. Using the same arguments, we can derive optimality properties for the more general autoregressive scheme in (9) rather than the restrictive setting of $\omega = 0$, $\mathbf{A} = \mathbf{a} \cdot \mathbf{I}$, and $\mathbf{B} = \mathbf{I}$. Each of these properties continues to hold under model misspecification.

2.4 Choice of instruments

To estimate $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_c^\top, \boldsymbol{\theta}_f^\top)$, we augment the conditional moment conditions in (5) by the matrix of instruments $\mathbf{W}_t \in \mathbb{R}^{L \times K}$ to arrive at the unconditional moment conditions

$$E [\mathbf{W}_t \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0, \quad (17)$$

and the corresponding GMM objective function

$$\min_{\boldsymbol{\theta} \in \Theta} \bar{\mathbf{g}}_T^\top \bar{\boldsymbol{\Omega}}_T \bar{\mathbf{g}}_T, \quad \bar{\mathbf{g}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{W}_t \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}), \quad (18)$$

where $\bar{\boldsymbol{\Omega}}_T$ is a positive definite matrix. As usual, we can start by setting $\bar{\boldsymbol{\Omega}}_T = \mathbf{I}$ in a first stage estimation, and set $\bar{\boldsymbol{\Omega}}_T$ to be an estimate of $\text{Var} [T^{1/2} \bar{\mathbf{g}}_T]$ in a second stage; see Hansen (1982).

The matrix of instruments we propose equals $\mathbf{W}_t = (1, \mathbf{f}_{t-1}^\top, \mathbf{s}_{t-1}^\top)^\top \otimes \mathbf{I}$, such that

$$E [((1, \mathbf{f}_{t-1}^\top, \mathbf{s}_{t-1}^\top)^\top \otimes \mathbf{I}) \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = E [(1, \mathbf{f}_{t-1}^\top, \mathbf{s}_{t-1}^\top)^\top \otimes \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})]$$

$$\mathbb{E} \left[\begin{pmatrix} 1, \mathbf{f}_{t-1}^\top, \mathbf{s}_{t-1}^\top \end{pmatrix}^\top \otimes \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}) \right] = 0. \quad (19)$$

The instrumented moment conditions $\mathbb{E} [1 \otimes \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$, $\mathbb{E} [\mathbf{s}_{t-1} \otimes \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$, and $\mathbb{E} [\mathbf{f}_{t-1} \otimes \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0$ are intuitive in that they impose the unconditional moment condition and also exploit any autocorrelation in the sample values of the moment conditions. In particular, $(\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}) \otimes \mathbf{s}_{t-1})$ holds the cross products of the moment conditions \mathbf{g}_t and their lags \mathbf{g}_{t-1} via the lagged scores \mathbf{s}_{t-1} . If there is autocorrelation in \mathbf{g}_t , GMM adjusts \mathbf{A} and \mathbf{B} to push this autocorrelation closer to zero in line with the intuition underlying the GaMM dynamics.

A different way to motivate (19) is to consider what optimal instruments would be for estimating $\boldsymbol{\theta}$; see, e.g. Davidson and MacKinnon (1993) for a textbook treatment on optimal instruments. Deriving optimal instruments in a general setting is non-trivial. To facilitate the exposition, we make a simplifying assumption that the moment conditions are correctly specified in the sense that for all $t < s$ we have

$$\mathbb{E} [\mathbf{g}_s(\mathbf{w}_s; \mathbf{f}_s, \boldsymbol{\theta}) \mathbf{g}_t^\top(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})] = 0. \quad (20)$$

The optimal instruments are then given by

$$\mathbf{W}_t^\top = \boldsymbol{\Omega}_t \mathbb{E}_{t-1} \left[\frac{d \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top} \right], \quad (21)$$

with $\boldsymbol{\Omega}_t = \mathbb{E}_{t-1} [\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta}) \mathbf{g}_t^\top(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})]^\star$. Note that \mathbf{f}_t depends on $\boldsymbol{\theta}$ through the GaMM dynamics in (11). We make this explicit by writing $\mathbf{f}_t = \mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})$, to obtain

$$\frac{d \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top} = \frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} + \frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \mathbf{f}_t^\top} \frac{d \mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top},$$

such that

$$\mathbf{W}_t^\top = \boldsymbol{\Omega}_t \mathbb{E}_{t-1} \left[\frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right] + \boldsymbol{\Omega}_t \mathbf{G}_t \frac{d \mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top}, \quad (22)$$

with $\mathbf{G}_t = \mathbf{G}_t^0$ as defined in (7). Let $(\boldsymbol{\omega}^\top, \text{vec}[\mathbf{B}]^\top, \text{vec}[\mathbf{A}]^\top)^\top$ constitute the lower part of $\boldsymbol{\theta}$.

Then the last derivative in (22) follows the recursion

$$\begin{aligned}
\frac{d \mathbf{f}_{t+1}}{d \boldsymbol{\theta}^\top} &= \frac{d \mathbf{f}_{t+1}(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top} \\
&= \frac{d(\mathbf{I} - \mathbf{B})\boldsymbol{\omega}}{d \boldsymbol{\theta}^\top} + (\mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})^\top \otimes \mathbf{I}) \frac{d \text{vec}[\mathbf{B}]}{d \boldsymbol{\theta}^\top} + \\
&\quad (\mathbf{s}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})^\top \otimes \mathbf{I}) \frac{d \text{vec}[\mathbf{A}]}{d \boldsymbol{\theta}^\top} + \mathbf{B} \frac{d \mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top} + \\
&\quad \mathbf{A} \frac{\partial \mathbf{s}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} + \mathbf{A} \frac{\partial \mathbf{s}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \mathbf{f}_t^\top} \frac{d \mathbf{f}_t(\mathbf{w}_{t-1}; \mathbf{f}_{t-1}, \boldsymbol{\theta})}{d \boldsymbol{\theta}^\top} \\
&= \begin{pmatrix} 0 & \mathbf{I} - \mathbf{B} & \mathbf{f}_t^\top \otimes \mathbf{I} & \mathbf{s}_t^\top \otimes \mathbf{I} \end{pmatrix} + \mathbf{B} \frac{d \mathbf{f}_t}{d \boldsymbol{\theta}^\top} + \mathbf{A} \frac{\partial \mathbf{s}_t}{\partial \boldsymbol{\theta}^\top} + \mathbf{A} \frac{\partial \mathbf{s}_t}{\partial \mathbf{f}_t^\top} \frac{d \mathbf{f}_t}{d \boldsymbol{\theta}^\top}. \quad (23)
\end{aligned}$$

In empirical work, we expect GaMM dynamics to work well if time-variation in parameters is persistent and relatively slow. This implies \mathbf{A} will often be estimated close to zero and \mathbf{B} close to the identity matrix. Consequently, the last two terms in (23) are typically small relative to the first two terms. As long as $\mathbf{I} - \mathbf{B}$ is non-singular, the first term in (23) is equivalent to the matrix of instruments proposed earlier in (19). The instruments in (19) therefore account for the dominant sources of variation in the optimal instruments (22), while avoiding the use of a second recursion for the derivatives (23) during estimation. The main difference between (19) and (23) is the presence of $\mathbf{B} d \mathbf{f}_t / d \boldsymbol{\theta}^\top$, which induces additional smoothing of the instruments in \mathbf{W}_t as proposed in (19).

The instruments in (19) have computational advantages over optimal instruments. Components of (22) can be hard to compute for specific models, particularly given the need to compute conditional expectations. The instruments in (19) simplify the computational challenges because the time-varying parameter \mathbf{f}_t as well as its derivatives are always time $(t - 1)$ -measurable by construction. We demonstrate the usefulness of the instruments in (19) for different models in Section 3 using both simulated and empirical data.

2.5 Asymptotic distribution theory

GaMM dynamics fall entirely within the standard set-up of GMM estimation. The consistency and asymptotic normality results for the GMM estimator as in Hansen (1982) can therefore be applied directly, including the expression for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, under standard high-level regularity conditions. We make the following assumptions.

Assumption 1. Let $\boldsymbol{\Theta}$ denote the compact parameter space with interior $\text{int}(\boldsymbol{\Theta})$. The GMM objec-

tive function (18) is almost surely twice continuously differentiable and has a unique minimum at $\theta_0 \in \text{int}(\Theta)$.

Assumption 2. (i) A central limit theorem holds for $T^{\frac{1}{2}} \bar{\mathbf{g}}_T$, i.e., $T^{-\frac{1}{2}} \bar{\mathbf{g}}_T \xrightarrow{d} \mathbf{N}(0, \bar{\mathbf{V}})$, with $\bar{\mathbf{g}}_T$ as defined in (18), and $T \cdot \text{Var} [\bar{\mathbf{g}}_T] \rightarrow \bar{\mathbf{V}}$. (ii) Let $\bar{\mathbf{G}}_t = \text{d}(\mathbf{W}_t \mathbf{g}_t) / \text{d} \theta$, then $\mathbb{E}[\bar{\mathbf{G}}_t]$ exists and a uniform law of large numbers result holds for $T^{-1} \sum_{t=1}^T \bar{\mathbf{G}}_t \rightarrow \bar{\mathbf{G}}$. (iii) The matrix $\bar{\mathbf{\Omega}}_T$ converges in probability to a positive semi-definite matrix $\bar{\mathbf{\Omega}}$. (iv) The matrix $\bar{\mathbf{H}} = \bar{\mathbf{G}}^\top \bar{\mathbf{\Omega}} \bar{\mathbf{G}}$ is invertible.

These high-level conditions can be worked out for low-level conditions along the lines of for instance White (1996) for specific models. The following standard result now follows directly from Hansen (1982).

Proposition 2. Under Assumptions 1 and 2, the GMM estimator for the static parameters θ , including the parameters ω , \mathbf{A} and \mathbf{B} specifying the GaMM dynamics, is consistent and asymptotically normal, i.e.,

$$T^{1/2} (\hat{\theta} - \theta_0) \xrightarrow{d} \mathbf{N}(0, \mathbf{H}^{-1} \bar{\mathbf{D}} \mathbf{H}^{-1}), \quad \bar{\mathbf{D}} = \bar{\mathbf{G}}^\top \bar{\mathbf{\Omega}} \bar{\mathbf{V}} \bar{\mathbf{\Omega}} \bar{\mathbf{G}}. \quad (24)$$

The efficient weighting matrix is $\bar{\mathbf{\Omega}} = \bar{\mathbf{V}}^{-1}$, in which case the asymptotic covariance matrix collapses to $(\bar{\mathbf{G}}^\top \bar{\mathbf{\Omega}} \bar{\mathbf{G}})^{-1}$.

Though the differentiability conditions in Assumption 1 are typically straightforward to verify and the uniqueness of the optimum is typically imposed by assumption, Assumption 2 can be more cumbersome to verify. The applicability of a central limit theorem and a law of large numbers typically builds on stationarity and ergodicity requirements for the sequences $\{\mathbf{W}_t\}$ and $\{\mathbf{g}_t\}$, which in turn depend on the stationarity and ergodicity of the underlying data, and that of the time-varying parameter \mathbf{f}_t and of its derivatives with respect to θ .

A key difference here is the non-linearity of the GaMM transition equation describing the dynamics of \mathbf{f}_t as a function of the data, i.e., the filtering equations. Even if the data \mathbf{w}_t are stationary, ergodic, near epoch dependent, and have the appropriate moments, these properties are not necessarily inherited by the ‘filtered’ time-varying parameter \mathbf{f}_t and its derivatives. In order to obtain stationarity and ergodicity results for \mathbf{f}_t , the transition equation (2) needs to be studied separately. Formulating low-level conditions to establish such properties is hard at the current level of generality. We therefore stick with the high-level conditions in Assumption 2. For

particular models or model classes, lower level conditions are obtainable; see, e.g. Straumann and Mikosch (2006) for generalized autoregressive conditional heteroscedasticity models, or Blasques et al. (2014) for generalized autoregressive score models.

2.6 Penalized objective function

There are two settings where we could improve the GaMM approach further. First, given the autoregressive structure of GaMM dynamics, these dynamics may react too slowly to structural changes if such changes are sizable and abrupt. Second, we want to prevent GaMM from picking up noise rather than the signal when estimating the dynamics of the time-varying parameters \mathbf{f}_t .

The GaMM methodology can solve these problems by introducing a penalized objective function. Penalization has been considered previously, e.g., in the context of Maximum Likelihood Estimation (see Eggermont and LaRiccia, 2001) or duration models (see Rondeau et al., 2003). Penalties can be introduced as a means of incorporating *a priori* knowledge about qualitative features of the model. For instance, if we expect the estimated path of a time-varying parameter to be relatively smooth, we can introduce a penalty that takes higher values for rougher paths. In addition, statistical procedures like the Kalman filter can be regarded as penalized maximum likelihood procedures for linear regression under normality, placing a quadratic penalty on the magnitude of parameter changes from one period to the next.

Rather than augmenting our GMM objective function with a quadratic penalty, we propose a piecewise linear penalty function. The penalized procedure then weights a sequence of smaller departures as much as it does one single large departure. The penalized criterion function takes the form

$$\min_{\boldsymbol{\theta} \in \Theta} \bar{\mathbf{g}}_T^\top \bar{\boldsymbol{\Omega}}_T \bar{\mathbf{g}}_T + \frac{\lambda_T}{T} \sum_{t=1}^T \boldsymbol{\iota}^\top |\mathbf{s}_t|, \quad (25)$$

where $\boldsymbol{\iota}$ denotes a vector of ones, and $\lambda_T \geq 0$ is a smoothing parameter. The impact of the penalty term $T^{-1} \lambda_T \sum_{t=1}^T \boldsymbol{\iota}^\top |\mathbf{s}_t|$ is twofold. First, the penalty on $|\mathbf{s}_t|$ favors dynamics that have lower scores not only on average (case without penalty), but at every time t . Given the GaMM dynamic specification, this induces smoother paths and discourages the approach from capturing noise rather than the signal. Second, given the linear rather than quadratic tail shape of the penalty function in \mathbf{s}_t , the penalty favors a path with one large jump, as long as this jump can prevent a

large number of smaller steps with non-negligible s_t in other regions of the sample. This feature is particularly relevant in cases where there are incidental structural breaks in the paths of the parameter.

We assume that the scores s_t are stationary and ergodic and satisfy $E_{t-1} [\boldsymbol{\iota}^\top |s_t|] < \infty$, and furthermore that $T \cdot \lambda_T \rightarrow 0$. This ensures that the smoothing parameter vanishes at an appropriate rate in order not to impact the objective function asymptotically. To see this, note that $T \cdot \bar{\mathbf{g}}_T^\top \bar{\boldsymbol{\Omega}}_T \bar{\mathbf{g}}_T$ converges to a non-degenerate random variable under Assumptions 1 and 2. The above assumptions on $E_{t-1} [\boldsymbol{\iota}^\top |s_t|]$ and λ_T then ensure that the penalty term in (25) becomes negligible asymptotically compared to the first term. Further assumptions need to be developed to ensure consistency and asymptotic normality of penalized GaMM estimation. We leave this for future work.

3 Example applications of GaMM

In this section we apply GaMM to empirical examples of increasing complexity. Each example highlights a different feature of the methodology that is not easily dealt with in other generic observation-driven modeling frameworks, such as for example the generalized autoregressive score framework of Creal et al. (2013). An additional illustration of GaMM for time-varying linear regression models with endogeneity problems is provided in the Supplementary Appendix to this paper.

3.1 Stable distributions

Model

The use of α -stable distributions has a long history in finance. Mandelbrot (1963) and Fama (1965) show that it is possible to capture a number of stylized facts about financial returns using these distributions. Stable distributions continue to attract attention in the recent empirical and theoretical literature, see for example Garcia et al. (2011), and can be particularly convenient for modeling highly erratic data such as changes in electricity or energy prices.

A challenging aspect of α -stable distributions is that their density function is generally not known analytically. If the problem at hand requires that some of the distributional assumptions

are relaxed, for example, by allowing the scale of the stable distribution to vary over time, it is therefore hard to base parameter dynamics on the density of the stable distribution by, for example, using a generalized autoregressive score model as in Creal et al. (2013). Similarly, using standard volatility models such as generalized autoregressive conditional heteroscedasticity (GARCH) models does not seem to be appropriate either: as the stable distribution allows for the realization of much more extreme observations than the normal distribution, updating the scale of the stable distribution by the squared lagged observations (like in QML GARCH) results in significant instability in the filtered volatility paths and potentially a large bias in the estimated time-varying scale parameter.

The family of α -stable distributions is fully characterized by four parameters: the stability parameter $\alpha \in (0, 2]$ (which corresponds with a tail index of α^{-1}), the skewness parameter $\beta \in [-1, 1]$, the scale parameter $\sigma \in \mathbf{R}^+$, and the location parameter $\mu \in \mathbf{R}$. Stable distributions have $p < \alpha$ finite moments and thus infinite variance and heavy tails if $\alpha < 2$. Apart from a few specific cases—of which Cauchy ($\alpha = 1, \beta = 0$), Lévy ($\alpha = \frac{1}{2}, \beta = 1$), and Gaussian ($\alpha = 2, \beta = 0$) distributions are the most notable—closed form expressions for density functions of α -stable distributions are not known. Although the densities are not known in closed-form, the characteristic function of α -stable distributions is (i) readily available and has a simple and intuitive form. The α -stable model with time-varying scale provides a good example of how GaMM dynamics can be operationalized and lead to useful results since the characteristic function can be written as a set of moment conditions.

The conditional characteristic function $c(u; \alpha, \beta, \sigma_t, \mu)$ of a stable distribution $S(\cdot; \alpha, \beta, \sigma_t, \mu)$, has the following simple expression:

$$\log c(u; \alpha, \beta, \sigma_t, \mu) = \log E_{t-1} [e^{iuX_t}] \quad (26)$$

$$= \begin{cases} i\mu u - (\sigma_t |u|)^\alpha \{1 - i\beta \operatorname{sign}(u) \tan(\frac{1}{2}\alpha\pi)\}, & \alpha \neq 1, \\ i\mu u - \sigma_t |u| \{1 + \frac{2i}{\pi}\beta \operatorname{sign}(u) \log u\}, & \alpha = 1. \end{cases} \quad (27)$$

Given equation (27), we can therefore formulate the conditional moment condition

$$E_{t-1} [\mathbf{g}_t^c] = 0, \quad \mathbf{g}_t^c = \exp(iu_i X_t) - c(u_i; \alpha, \beta, \sigma_t, \mu), \quad i = 1, \dots, K, \quad (28)$$

where m denotes the number of grid points u_1, \dots, u_K . The dimensionality of the moment

condition vector depends on the choice of the u_i s. One extreme is to choose a continuum of u_i s. The number of moment conditions then becomes infinite and the parameters can be estimated with Continuous GMM (CGMM) as proposed by Carrasco and Florens (2000); see also Kotchoni (2012) for a good empirically oriented discussion of CGMM. In this section, we opt for a simpler approach and use a finite set of K grid points as in Yu (2004). Feuerverger and McDunnough (1981) show that by choosing a sufficiently dense and extended grid the asymptotic covariance matrix of the resulting GMM estimator can be made arbitrarily close to the Cramer-Rao bound. Feuerverger and McDunnough (1981) and Yu (2004) suggest that the points on the grid should be equidistant. We find that the choice of the grid (both in terms of its density and range) should depend on the size of the parameters α , β , σ_t , and μ ; for example, for large values of σ_t the grid should be much denser around zero than for smaller values of σ_t ; see also the discussion in Carrasco and Florens (2000).

For a given grid $\{u_1, \dots, u_K\}$, the moment conditions are constructed by stacking the real and imaginary parts of \mathbf{g}_t^c for different values of u_i ,

$$\mathbf{g}_t^\top = (\text{Re}(\mathbf{g}_t^c)^\top, \text{Im}(\mathbf{g}_t^c)^\top). \quad (29)$$

This produces $2K$ moment conditions and sets the maximum number of grid-points to $T/2$.

Given equations (27)–(28), the matrix \mathbf{G}_t from (7) is

$$\frac{\partial \mathbf{g}_t^c}{\partial \sigma_t} = \begin{cases} \mathbf{c}_t \times i\alpha |u\sigma_t|^\alpha \left\{ i + \beta \text{sign}(u) \tan\left(\frac{1}{2}\alpha\pi\right) \right\}, & \text{for } \alpha \neq 1, \\ \mathbf{c}_t \times (-|u\sigma_t|) \left\{ 1 + \frac{2}{\pi}i\beta \text{sign}(u) \log(u) \right\}, & \text{for } \alpha = 1, \end{cases} \quad (30)$$

where $c_t \equiv c(u; \alpha, \beta, \sigma_t, \mu)$. As the partial derivatives do not contain \mathbf{w}_t , we do not need to compute conditional expectations to obtain \mathbf{G}_t . Also the optimality results from Section 2.3 hold directly.

Simulation results

To study the performance of GaMM in this setting, we generate a time series of $T = 5000$ observations from a stable distribution with $\alpha_0 = 1.5$, $\beta_0 = -0.5$, $\mu_0 = -0.5$, and a time-varying scale $\sigma_{0,t}$ that varies between 1 and 12. We refer to the Section Appendix B for details about the sampling procedure. We estimate the static parameters $\boldsymbol{\theta}^\top = (\alpha, \beta, \mu, \omega, B, A)$ as

described in Section 2 using GaMM(1,1) dynamics. We guarantee positivity of the estimated scale σ_t by defining $\mathbf{f}_t = \log \sigma_t$.

Results based on 10,000 replications are shown in Panels A–D of Figure 1. We let the true scale parameter $\sigma_{0,t}$ follow four distinct paths: structural breaks (Panel A); slow gradual changes (Panel B); an autoregressive (AR) process of order 1, i.e., an AR(1) process (Panel C); and a dynamic pattern obtained by estimating the dynamic α -stable scale model for empirical S&P 500 returns over the last 30 years, as used in our empirical application later on (Panel D). At every point in time, we report the median estimate across all replications along with some quintiles to form Monte Carlo confidence bands. In all cases, the median scale estimate $\hat{\sigma}_t$ at time t is close to the true value $\sigma_{0,t}$. From the quantile bands around the median estimate, we see that the distribution of $\hat{\sigma}_t$ is more or less symmetric. The variability in the estimate of σ_t increases with the level of the scale. The latter phenomenon is intuitive: as the scale increases, the signal-to-noise ratio of the data decreases.

For each simulation, we summarize the bias and root mean squared error (RMSE) of the time-varying scale into a single summary statistic,

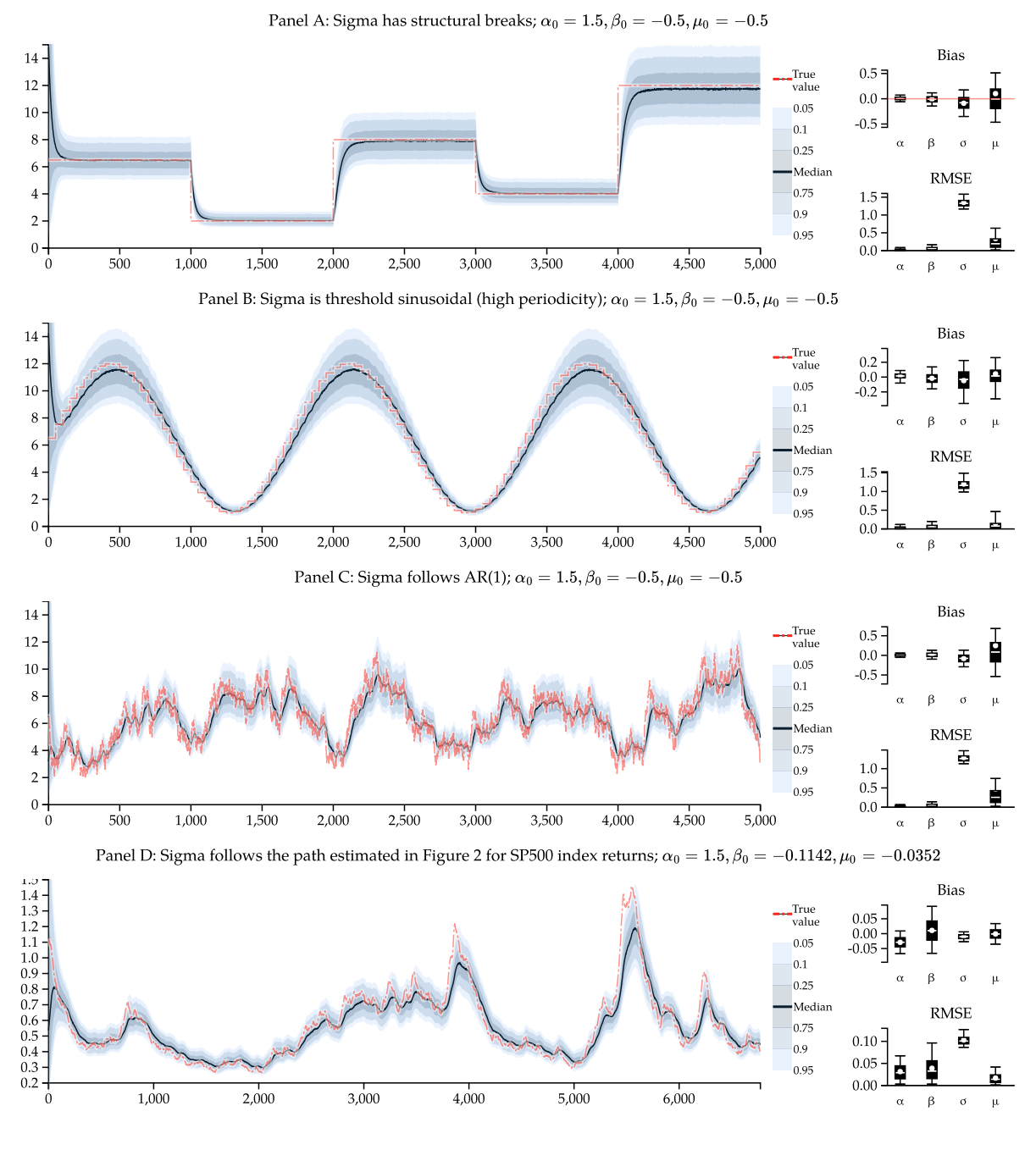
$$\text{Average Bias} = \frac{1}{T} \sum_1^T (\hat{\sigma}_t - \sigma_{t,0}), \quad \text{Average RMSE} = \left[\frac{1}{T} \sum_1^T (\hat{\sigma}_t - \sigma_{t,0})^2 \right]^{\frac{1}{2}}. \quad (31)$$

We also compute the bias and RMSE for the static parameters α , β , and μ . Box plots for the bias and RMSE across all simulations are presented to the right of each of the different panels of Figure 1. GaMM(1,1) overall produces unbiased estimates of the static parameters α , β , and μ . These parameters are typically also estimated with a low RMSE. The estimated path for the scale parameter also appears to be unbiased, though it tends to oscillate around the true value for any given replication. This results in a somewhat higher RMSE statistic. The oscillating behavior is typical for observation-driven models and filtering methods in general; see the discussion in Nelson and Foster (1994). It indicates that the GaMM(1,1) dynamics are able to capture the unknown true dynamics of the scale parameter and adapt to it based on the information in the data and the shape of the moment conditions.

Figure 1

Estimating Time-Varying Scale of a Stable Distribution with GaMM

This figure contains summary figures for a series of simulations in which returns are drawn from a stable distribution. We let the scale parameter, σ , vary in different ways. The remaining parameters are static and (apart from Panel D) are as follows: $\alpha_0 = 1.5, \beta_0 = -0.5, \mu_0 = -0.5$. In each panel, we show the true path of σ , a path of median estimates, and some quantile bands around the median path. The results are based on 10,000 replications. We report distributions of average biases and RMSEs for all parameters across replications. MSE (bias) for the time-varying scale is computed as the average of MSEs (biases) per replication.



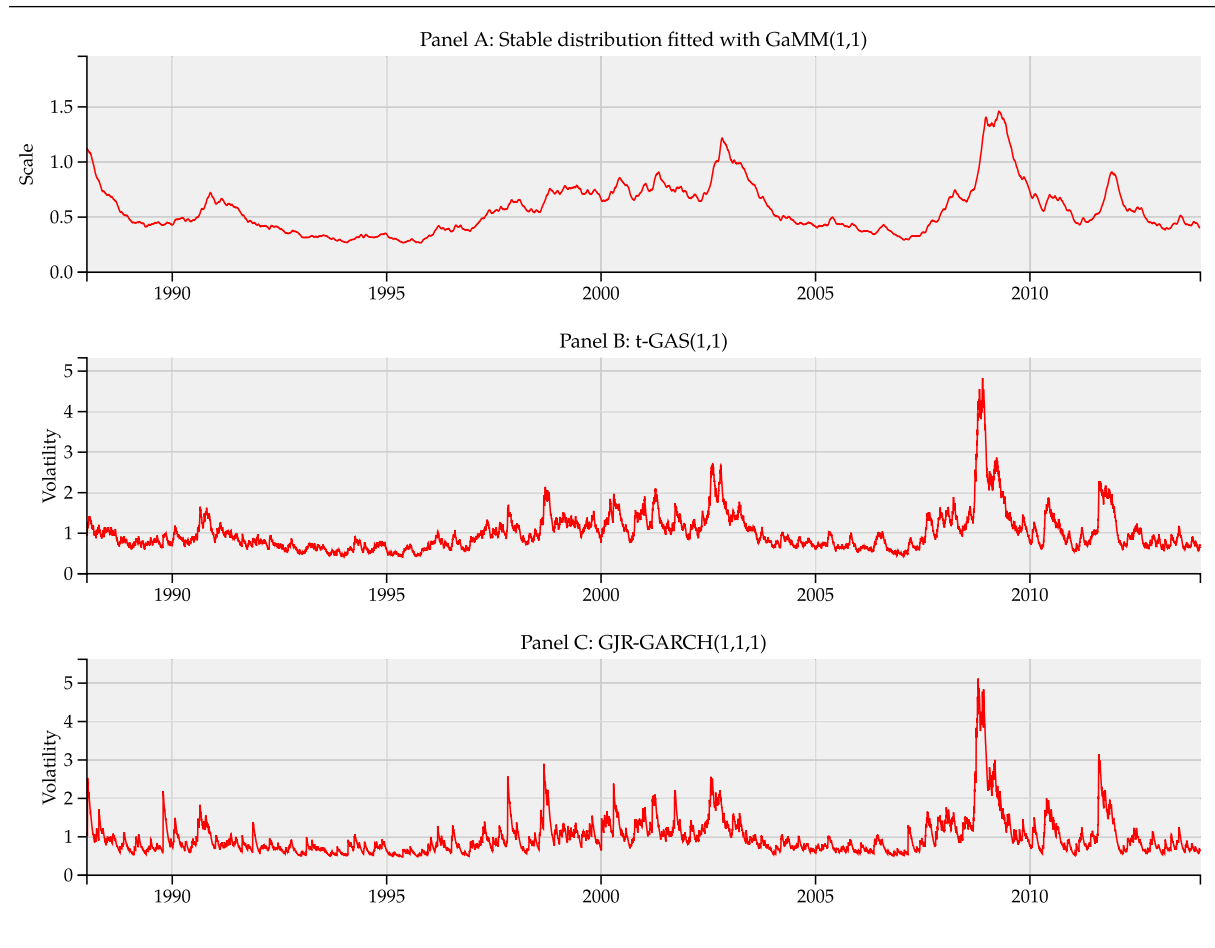
Empirical application

For our empirical illustration, we use S&P 500 daily returns over the period 1988–2013 and fit a time-varying scale stable distribution with GaMM(1,1) dynamics as above. The results are shown in Figure 2. We provide two benchmark estimates, namely a GJR-GARCH(1,1,1) model of Glosten et al. (1993), and the Student’s t -GAS(1,1) model of Creal et al. (2011) and Harvey (2013).

Figure 2

S&P 500 Returns as Draws From a Stable Distribution

This figure contains estimated time-varying volatility of daily returns on the S&P 500 index in 1988–2013. In Panel A, we use GaMM(1,1) to fit a time-varying scale stable distribution to the data. As benchmarks, panels B and C contain estimated time-varying volatility paths obtained with GJR-GARCH(1,1,1) and t -GAS(1,1).



Overall, the estimated paths of σ_t for the GJR-GARCH(1,1,1) and Student’s t -GAS(1,1) are similar. However, the dynamics for t -GAS(1,1) are less responsive to incidentally large observations, for example, at the end of 1989, or in 1997 and 1998. This stems from the fact the t -GAS(1,1)

results are based on the fat-tailed Student's t distribution with an estimated 5.76 degrees of freedom. The path produced with the stable distribution and GaMM(1,1) dynamics is smoother than that of the other two models. Given the assumption of a stable distribution, the GaMM(1,1) dynamics are much more cautious in ascribing the realization of a large positive or negative return to an increase in the time-varying scale parameter. Extreme absolute returns could be the result of the heavy-tailed (vs fat-tailed in Student's t) nature of the stable distribution with $\alpha < 2$. This effectively removes the smaller up and down movements in the volatility estimates compared to the two benchmark models. Also note that the magnitude of σ_t cannot be compared directly to that of the GJR-GARCH or GAS models, because the stable distribution does not have a finite second moment for $\alpha < 2$. The estimated values for $\hat{\alpha}$ and $\hat{\beta}$ are 1.58 and -0.04 , respectively. The former suggests heavy tails even after correcting for changes in scale, while the latter indicates that there is hardly any unconditional skewness.

All three paths of σ_t sketch a similar story for the volatility of returns and move in the same directions. This also holds for the much smoother time-varying scale of the stable distribution. It is worth noting, though, that the assumption of a stable distribution for the returns alters the relative magnitude of volatility over time between the different models. For instance, based on the GJR-GARCH and Student's t -GAS results, the maximum volatility over the 2008 financial crisis was twice as high compared to 2003 and to the subsequent European sovereign debt crisis. However, the GaMM estimates suggest there is only roughly a 50% difference in magnitude of the scale of the stable distribution between 2003 and 2008: much of the remainder increase is attributed to the heavy tails of the distribution.

3.2 Consumption CAPM with power utility

Model

In this application, we use GaMM to estimate a simple non-linear asset pricing model with time-varying risk-aversion coefficient. Building on the seminal work of Hansen and Singleton (1982), we consider the power utility function which produces the following Euler equations for pricing assets:

$$E_t [\beta (C_{t+1}/C_t)^{-\gamma} \mathbf{R}_{t+1}^x] = 1, \quad (32)$$

where \mathbf{R}_{t+1}^x denotes the vector of gross asset returns, β is a subjective discount factor, and γ represents curvature of the utility function as well as relative risk aversion. Both consumption and asset returns are assumed to be expressed in real terms. Euler equations in (32) imply a stochastic discount factor $M_t = \beta (C_{t+1}/C_t)^{-\gamma}$. Empirical estimates of the risk-aversion parameter γ are sensitive to the particular sample period, starting values, and instruments employed. Results of many studies suggest that time series estimates of γ are typically too high compared to risk-aversion estimates obtained from experimental data and that the simple model in (32) fails at explaining the cross-sectional variation in stock returns, e.g. see Savov (2011); Mehra and Prescott (1985); Chen and Ludvigson (2009); Lettau and Ludvigson (2009), or Ludvigson (2011) for a recent summary of the literature and developments in the field. Mehra and Prescott (1985) dubbed this phenomenon the equity premium puzzle. The reason for poor performance stems from the fact that consumption growth is too smooth relative to the variation in returns and thus the stochastic discount factor needs to be blown up through high γ and β .

Poor performance of the standard model in (32) can be addressed by allowing for habit formation which adds an additional source of variation to the stochastic discount factor, see for example Constantinides (1990); Campbell and Cochrane (1999), or Ludvigson (2011). The vastness of literature on habit formation shows it is generally accepted to think of relative risk-aversion as a time-varying quantity¹. Furthermore, stability of deep parameters in the simple structural model in (32) is already questioned by Ghysels and Hall (1990). Ghysels and Hall introduce a structural break test for γ , but do not find sufficient evidence to reject the hypothesis of a constant risk aversion parameter in their sample. These tests, however, may have low power against specific mean-reverting alternatives.

In this example, we consider the simple consumption CAPM model in (32). The relative risk-aversion, γ_t , is allowed vary in time and we filter it out by endowing the standard Euler equation with GaMM(1,1) dynamics. We assume that the shocks to risk-aversion are exogenous and that agents are myopic in the sense that they consider γ_t to remain fixed forever when making their decision at time t . This results in the Euler equation

$$E_t [\beta (C_{t+1}/C_t)^{-\gamma_t} R_{t+1}^x] = 1. \quad (33)$$

¹Allowing for habit formation implies that relative risk-aversion depends on γ and on the current values of consumption and consumption habit.

Using (33), we can directly employ the GaMM framework by taking $\mathbf{f}_t = \gamma_t$.

One complication in our current context concerns the estimation of \mathbf{G}_t , which in this case depends both on the consumption growth data, the return data, and on the time varying parameter \mathbf{f}_t . Computing this conditional expectation analytically is impossible in this case, as we do not know the distribution of the risky returns nor of consumption growth. Moreover, replacing the expectation by a sample average is also not appropriate: the conditional distribution of asset returns given consumption growth in equilibrium depends on the current (myopic) risk aversion parameter, or put differently, observations from previous time-periods strictly speaking are realizations from a distribution characterized by a different value of γ_t . Still, if \mathbf{f}_t varies sufficiently slowly, observations in the recent past can be informative about the curvature \mathbf{G}_t of the moment conditions now. We therefore estimate \mathbf{G}_t as an exponentially weighted moving average

$$\mathbf{G}_t = \lambda \mathbf{G}_{t-1} + (1 - \lambda) \frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \mathbf{f}_t}, \quad (34)$$

where we choose λ in the range $(0.98, 1.0)$. If \mathbf{f}_t varies sufficiently slowly, such an exponentially weighted moving average estimates the local curvature of the score accurately enough to provide an adequate form of scaling in the GaMM transition dynamics (9). This holds even though $\hat{\mathbf{G}}_t$ may not be a consistent estimate of \mathbf{G}_t . See Section Appendix C for further details.

Simulation results

For our simulations, we fix $\beta = 1$, which is close to its typical empirical estimate; see for instance Hansen et al. (2008) or Savov (2011). Furthermore, we endow the true risk-aversion parameter $\gamma_{0,t}$ either with a structural break or with exogenous AR(1) dynamics. Given a series of $\{\gamma_{0,t}\}$ we use the following DGP to simulate data:

$$\Delta c_t = \mu_c + \varepsilon_{ct} \quad \varepsilon_{ct} \sim \mathcal{N}(0, \sigma_\gamma^2) \quad (35)$$

$$R_t = e^{\gamma_{0,t} \Delta c_t} + \varepsilon_{Rt} \quad \varepsilon_{Rt} \sim \mathcal{N}(0, \sigma_R^2), \quad (36)$$

where we set $\mu_c = 0.041$, $\sigma_{ct} = 0.09$, and $\sigma_{Rt} = 0.1$. Estimation results using the GaMM(1,1) specification are presented in Figure 3. Note that the classical full sample GMM estimates of γ for the structural break and the AR(1) case are 11.10 and 12.25, respectively. For the case of the AR(1), this implies that 70% of the γ_t observations actually lie below the full sample GMM estimate. If

there is time-variation in γ_t , the full sample GMM estimates are thus severely biased towards the high-end realizations of the time-varying risk-aversion parameter which as we show later may be part of the explanation of why the equity premium puzzle arises.

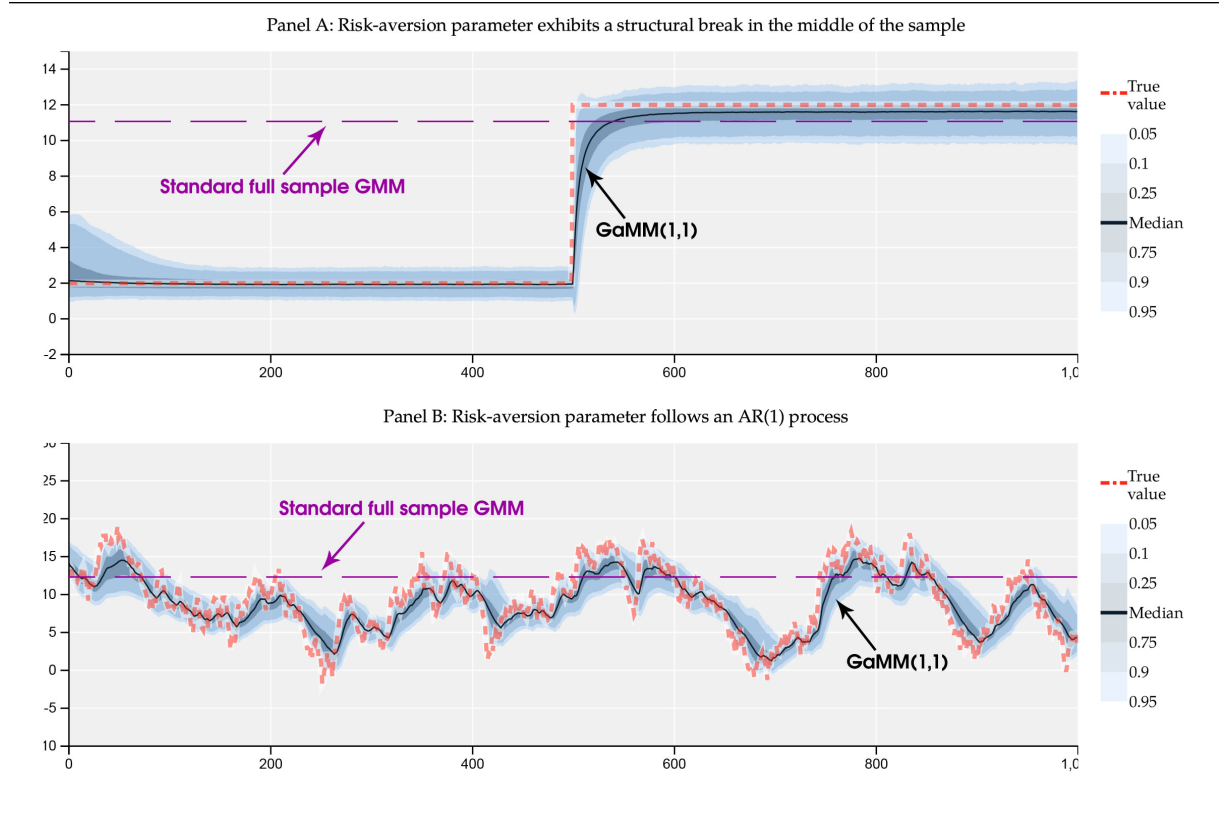
Figure 3

Time-Varying Risk Aversion in CCAPM With Power Utility

This figure illustrates performance of GaMM in estimating the risk-aversion parameter in CCAPM. We use the basic power utility specification:

$$1 = E_t \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma_t} R_{t+1} \right],$$

and assume that agents are myopic about the changes in risk-aversion parameter γ_t . Furthermore, in simulations we fix $\beta = 1$ and we do not estimate the discount factor. Results are based on 10,000 replications. Paths were estimated with GaMM(1,1). We juxtapose the true value of the parameter at time t with the median estimate across all replications. We also present 95, 90, and 50



If we consider the estimation results for the approach based on GaMM(1,1), we clearly see that the filtered path $\hat{\gamma}_t$ tends to follow the true path closely. In case of a large structural break, the estimator takes some time to adjust to the new setting. Overall, however, the path is able to capture both the episodes of high and low relative risk aversion. In case of the mean-reverting

AR(1) dynamics, the GaMM(1,1) approach also recovers the major up and down swings in γ_t .

Empirical equity premium results for U.S. data

We estimate the model for quarterly U.S. data between 1947 and 2015. The tests assets are comprised of the 3-Month Treasury Bill rate (from Board of Governors of the Federal Reserve System) and six equity portfolios double sorted on size and book-to-market as provided by Fama and French. Data on population size in the U.S. as well as expenditure on non-durable goods and consumption of services is provided by US. Bureau of Economic Analysis. Apart from the six Fama-French portfolios we obtain all remaining data from FRED². All series are deflated with an implicit price deflator (2009=100) which we calculate for the combined consumption (non-durables and services) series. We use two-step feasible GMM/GaMM and we report results from the second step only. Standard errors for parameter θ are estimated using the Newey and West (1987) HAC covariance matrix with truncation lag, p , chosen following the procedure proposed by Newey and West (1994). We denote them as $se_{\theta}^{(p)}$. In the results reported below, the model is estimated without additional conditioning information; see Ludvigson (2011) for a discussion of why this is appropriate. We note that adding standard instrumental variables to the model does not impact the results qualitatively.

We first estimated the model without time-varying relative risk-aversion and obtain the following estimates for γ and β . The discount rate is estimated at $\hat{\beta} = 1.42$ with $se_{\beta}^{(7)} = 0.14$ while the curvature parameter is $\hat{\gamma} = 133.06$ with $se_{\gamma}^{(7)} = 38.92$. The very high and imprecisely estimated value for the relative risk aversion parameter is in line with previous research (Savov, 2011; Lettau and Ludvigson, 2009). A similarly high value of β is reported by Lettau and Ludvigson (2009) for a shorter sample. Not only do we see the equity premium puzzle but the results suggest that agents value future utility more than the present one. A closer inspection of the data suggests that given the static model there are many ‘outliers’ (1949–1953, 1960, 1980, and 2008) which are both clustered in time and heavily bias the estimates upwards³.

²The specific series we use are: 3-Month Treasury Bill (TB3MS), population (B230RC0Q173SBEA), personal consumption expenditures on nondurable goods (PCND) and services (PCESV) with their corresponding implicit price deflators (DNDGRD3Q086SBEA and DSERRD3Q086SBEA respectively).

³It is worth noting what impact do β and γ have on moment conditions in a neighborhood of time t . *Ceteris paribus*, an increase in the discount rate increases both the mean and dispersion of moment conditions without affecting higher moments of their distribution. On the other hand, an increase in the curvature increases variance and skewness of moment conditions and thus lowers the mean. At the estimated values of parameters, the ‘outliers’ we identify bring an order of magnitude higher contributions to the criterion function and thus have huge leverage on the outcome.

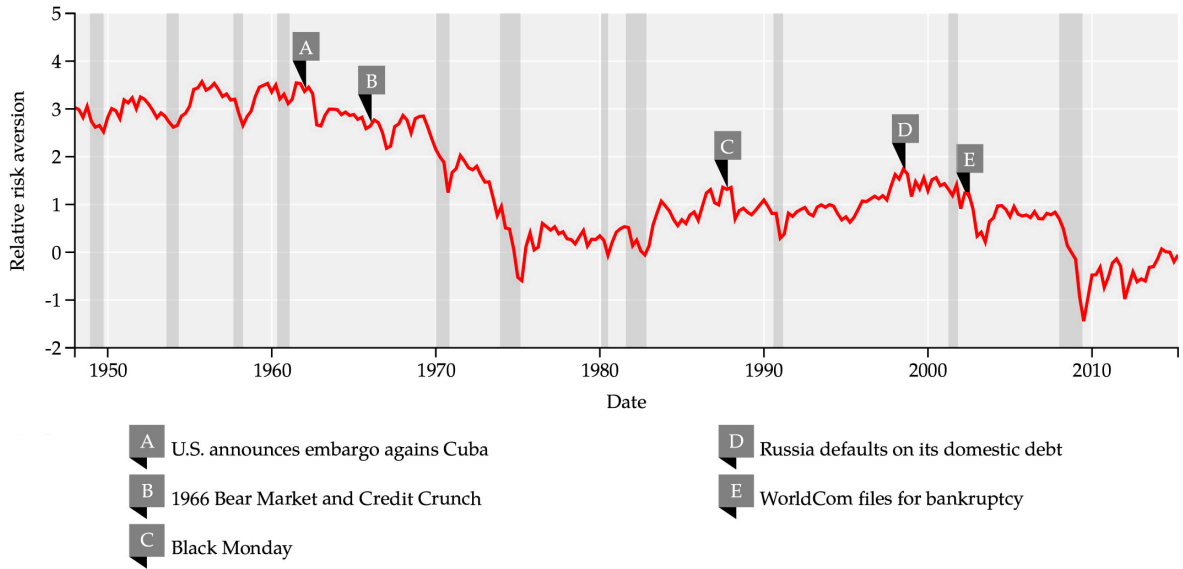
Figure 4

Equity Premium Puzzle

This figure illustrates performance of GaMM in estimating the time-varying relative risk-aversion parameter in CCAPM. We use the basic power utility specification:

$$1 = E_t \left[\beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma_t} R_{t+1} \right],$$

and assume that agents are myopic about the changes in risk-aversion parameter γ_t . Tests assets are comprised of the 3-Month Treasury Bill rate and six equity portfolios double sorted on size and book-to-market. We use consumption of non-durable goods and services. All series are deflated with an implicit price deflator (2009=100). Shaded regions correspond to NBER recessions while other relevant events are labeled separately.



In contrast, GaMM(1,1) produces reasonable values for both the subjective discount factor and the relative risk aversion. The discount factor is estimated at $\hat{\beta} = 0.98$ with $se_{\beta}^{(4)} = 0.001$. Figure 4 shows the estimated path of the relative risk-aversion parameter $\hat{\gamma}_t$ based on GaMM(1,1) dynamics together with NBER recession periods. Average value of the relative risk aversion is $\bar{\gamma}_t = 1.42$ (3.13 in 1950–1960, 1.06 in 1990–2000, and 0.55 in 2000–2010). Static parameters governing the GaMM dynamics are estimated as $\hat{\omega} = 3.03$ with $se_{\omega}^{(4)} = 0.75$, $\hat{B} = 1.00$ with $se_B^{(4)} = 0.006$, and finally $\hat{A} = 0.02$ with $se_A^{(4)} = 0.005$.

There are two components in the time-varying risk aversion parameter γ_t : a long-term and a short-term cycle. The short-term cycle appears to follow the business cycle. We find that during

recessions and sometimes even before the recession, risk-aversion is pushed downwards. Given the postulated utility framework and corresponding Euler equation, these pro-cyclical short-term fluctuations imply that agents adjust their consumption slowly and with a delay compared to reactions of financial markets. In other words, when a recession hits there is a period during which consumption is too high given the observable (negative) returns. In the current limited framework, this can only be explained by a lower risk-aversion parameter, which is why we see the drops in γ_t . After the recession ends, we observe risk aversion returning to its long-term path. This can happen either because agents adjusted their consumption expenditure or because markets recovered. These patterns are consistent with many of the phenomena and extensions to the basic model setting as reported in the literature, such as habit formation (Campbell and Cochrane, 2000), increasing leverage and shortening of investment time-horizon for households (Adrian and Shin, 2010), and loss-aversion (Kahneman and Tversky, 1984) where agents exhibit (in this case relative) risk-loving behavior in the loss domain and risk-aversion in the gain domain.

More interesting than the short-term cyclical behavior is the secular (long-term) pattern in risk-aversion. In particular, we notice a continuous decrease in risk aversion since the 1950s, when $\hat{\gamma}_t$ was higher possibly due to the post World War II recovery period. Comparing the first and the last decade in the sample shows the extent of the change. On the one hand, in both periods the average real consumption growth was between 0.4 and 0.5%. On the other hand, the real returns in the 1950s were more than three times as high as they were in the 2000s (7% versus 2%) based on data in Shiller (2005). Such a difference in realized returns given the relatively constant pattern of consumption growth can only be consistently explained in the proposed model if risk-aversion has decreased substantially: from the 1950s to the 2000s the GaMM(1,1) estimates of risk-aversion decrease from an average value of about 3.13 to a value of about 0.55, i.e. a full order of magnitude.

4 Example results for penalized GaMM

To illustrate the usefulness of the penalized version of GaMM from Section 2.6, we consider a simple linear regression model with an endogenous regressor x_t and an instrument z_t ,

$$y_t = \beta_{0,t} + \beta_{1,t}x_t + \varepsilon_{yt}, \quad \varepsilon_{yt} \sim N\left(0, \sigma_{\varepsilon_y}^2\right), \quad E[\varepsilon_{yt}\varepsilon_{xt}] = \rho, \quad (37)$$

$$x_t = \pi_{0,x} + \pi_{1,x}z_t + \varepsilon_{xt}, \quad \varepsilon_{xt} \sim N\left(0, \sigma_{\varepsilon_x}^2\right), \quad z_t \sim N\left(\mu_z, \sigma_z^2\right), \quad (38)$$

$$\beta_{i,t+1} = \pi_\beta \beta_{i,t} + \eta_{i,t}, \quad \eta_{i,t} \sim N(0, \sigma_{i,\eta}^2), \quad i = 0, 1. \quad (39)$$

If $\rho \neq 0$ and $\sigma_{\varepsilon_x}^2 > 0$, we have a standard endogenous regressor problem. Lower values of $\sigma_{\varepsilon_x}^2$ result in larger endogeneity biases when estimating $\beta_{1,t}$ by ordinary least-squares methods. Define $\mathbf{x}_t^\top = (1, x_t)$, $\mathbf{z}_t^\top = (1, z_t)$, and $\boldsymbol{\beta}_t^\top = (\beta_{0,t}, \beta_{1,t})$, and let $\boldsymbol{\Pi}_x$ be such that $\mathbf{x}_t^\top = \mathbf{z}_t^\top \boldsymbol{\Pi}_x + (0, \varepsilon_{xt})$. In the Supplementary Appendix to this paper, we provide an elaborate simulation experiment demonstrating the usefulness of GaMM in this setting compared to a standard Kalman Filtering approach: GaMM tracks the dynamic parameter well, while removing the endogeneity biases.

In this section, we focus on the effect of including the penalty function in the GMM objective function on the smoothness of the estimated path of $\boldsymbol{\beta}_t$. Given the availability of an instrument variable z_t , the obvious way to estimate $\boldsymbol{\beta}_t$ is via the conditional moment condition

$$E_{t-1} [\mathbf{z}_t (y_t - \mathbf{x}_t^\top \boldsymbol{\beta}_t)] = 0. \quad (40)$$

The complication here is that the parameters $\boldsymbol{\beta}_t$ are time-varying. The moment condition (40), however, lends itself directly to the GaMM framework by setting $\mathbf{f}_t = \boldsymbol{\beta}_t$.

We consider a setting where $\beta_{1,t}$ has a structural break at $t = t^*$. The result without the use of the penalty is presented in Panel A of Figure 5. A perfect fit to the structural break would require $s_{t^*} \neq 0$, while $s_t = 0$ for all $t \neq t^*$. The GaMM criterion function in (18) does not incorporate this prior knowledge about the properties of $\{s_t\}$. Panel A of Figure 5 shows the consequences of this. Because of the low rate of adjustment after the structural break, we accumulate a long sequence of observations for which the moment conditions are not minimized. In the end, we obtain a path that is negatively biased directly after the break, and positively biased before the break and long after the break. Given that $\bar{\mathbf{g}}_n$ in (18) considers the average (instrumented) moment condition across all times, it follows that the negative bias directly after the structural break offsets the other two (positive) biases.

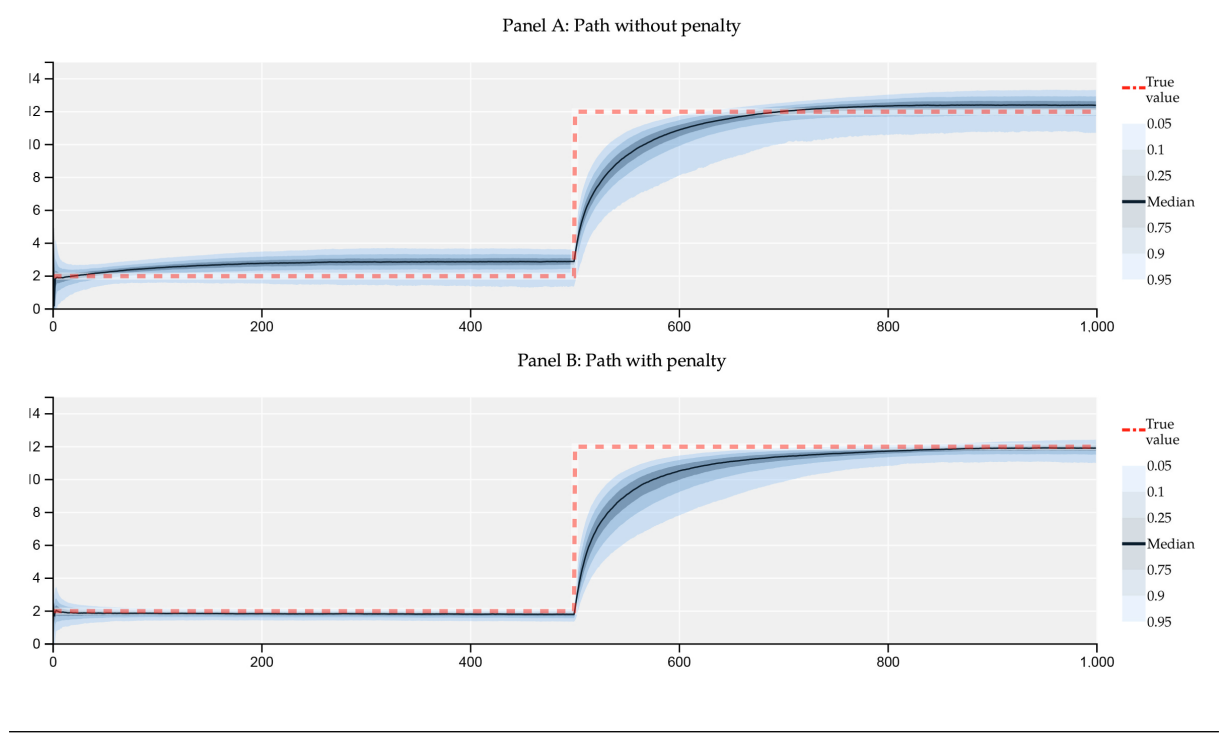
The effect of the penalty is presented in Panel B of Figure 5. We clearly see that the upward bias in the path before the time of the break has been almost entirely removed. The same holds with the upward bias after observation 700.

To obtain further insight, we run a Monte-Carlo simulation study for a variety of parameter settings using the penalized criterion function (25) with $\lambda_T = \{10, 5, 1, 0.6, 0.2, 0\}$, a variety

Figure 5

Penalizing the GaMM Criterion Function

In this figure we show how introduction of penalty impacts a filtered path. In Panel A, we see that due to the long adjustment time around the structural break, the path is biased upwards before and after the break. This happens because the upward bias outside of the break offsets the negative bias in the transition period. In Panel B, we introduce the proposed penalty which does not increase the rate of adjustment in the transition period. It does, however, remove the upward bias outside of this period.



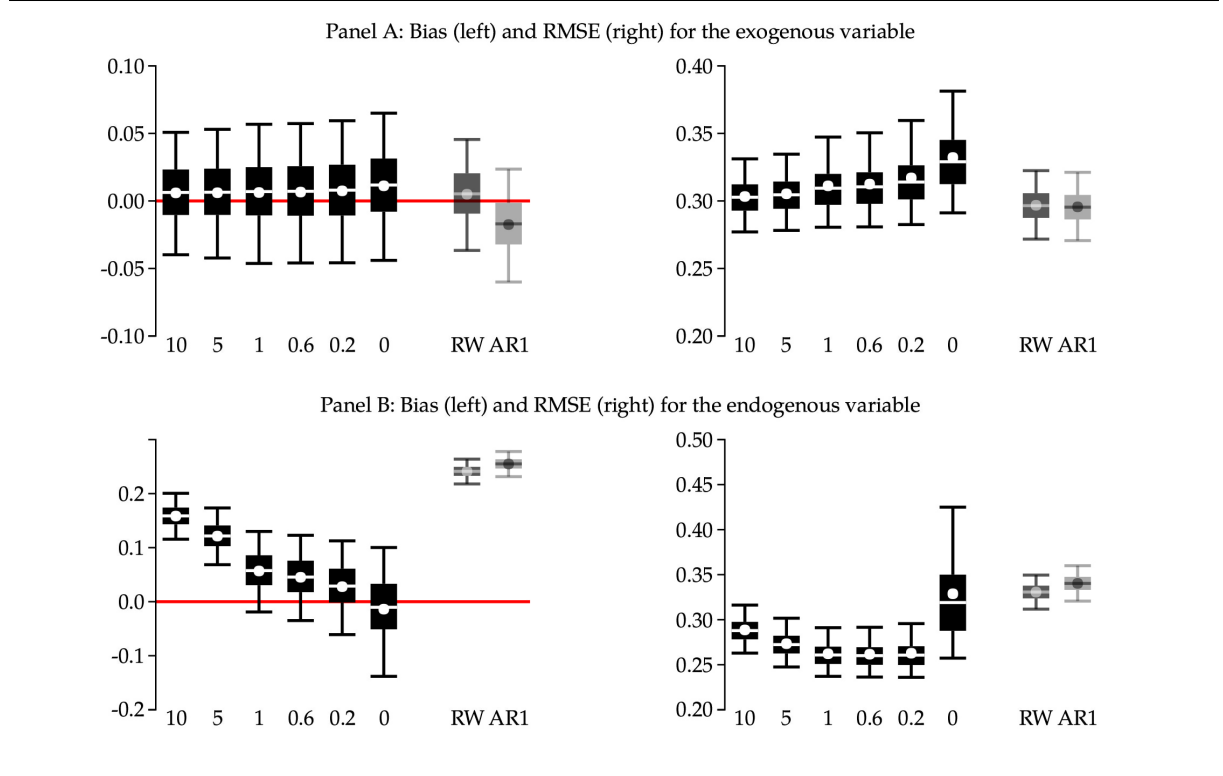
of sample sizes T , and a moderate endogeneity problem. We only report a subset of the results here. The full results can be found in the Supplementary Appendix. As a benchmark, we include maximum likelihood estimates of two linear state space models, one with AR(1) and one with random walk dynamics for the parameters. Both state space models are estimated using Kalman Filtering methods. We plot the impact of the penalty on the bias and RMSE of $\hat{\beta}_t$ in Figure 6. The figure contains the average bias and RMSE for the constant (top) and for the slope parameter (bottom). Within each panel, we see 6 boxplots for the different values of the smoothing parameter λ_T . The two grey boxplots correspond to the state space models.

Bias of the parameter path for the constant term, i.e. the exogenous regressor, obtained by GaMM is generally negligible and similar to both state-space models. However, the average bias in the path of the slope parameter (endogenous regressor) for both Kalman Filter methods is

Figure 6

Performance of GaMM for a Simple Endogeneity Problem With Penalized Criterion Function

The figure compares performance of GaMM (black) to Kalman Filter with AR1 (dark grey) or Random Walk (light grey) dynamics for coefficients. We simulate observations and true parameters from a linear regression model with an endogenous regressor. GaMM estimation was performed for different values of the smoothing parameter, $\lambda = \{10, 5, 1, 0.6, 0.2, 0\}$, and are presented in this order.



substantial. The GaMM methods perform much better in this respect. Biases are visible for very high values of $\lambda_T = 10$, but quickly vanish for smaller values of λ_T . Even for $\lambda_T = 10$, the average bias is smaller than for the state space models estimated with Kalman Filter, which does not account at all for the endogeneity problem. The penalized estimates also come with substantially lower RMSEs though this difference becomes smaller if the endogeneity problem is increased; see the additional results in the Supplementary Appendix.

We conclude that already small values of the penalty parameter λ_T such as $\lambda_T = 0.2$ result in considerable improvements in terms of RMSE. Even if bias is found to be a prime concern, we recommend that θ is first estimated using the penalized GaMM criterion, e.g., with $\lambda_T = 1$. The resulting estimates can then be used as starting values for θ when estimating under $\lambda_T = 0$. We

have found that this results in a numerically much more stable algorithm that is less susceptible to the potential issues of multiple local optima of the G(a)MM criterion function.

5 Conclusions

In this paper we proposed a new approach for modeling time-varying parameters in linear and non-linear econometric models identified through moment conditions. We call the approach the Generalized Autoregressive Method of Moments (GaMM) as it endows parameters that are identified via standard GMM (conditional) moment conditions with autoregressive dynamics based on local deviations of those same (conditional) moment conditions. The method goes substantially beyond previous observation-driven approaches and encompasses many of the previous observation-driven models found in the literature, including the generalized autoregressive score approach of Creal et al. (2013) and Harvey (2013). Being observation-driven, the method also falls directly within the generic GMM framework of Hansen (1982) in terms of the development of the appropriate asymptotic theory for the estimator.

Using a range of different examples, we illustrated the forcefulness of the new approach. To the best of our knowledge, we are the first to endow the class of stable distributions with an observation-driven time-varying scale in a way that is both computationally fast and intuitively appealing. The approach also turns out to work well in settings with endogeneity problems or in settings where we use Euler equations to identify our parameters of interest.

Interestingly, the approach can be further refined by using a penalized version of the GaMM criterion function. We showed how penalization can decrease the average root mean squared error of the estimated path of the time-varying parameter by allowing for quicker adjustments to large incidental parameter changes and structural breaks. Though we provided some first steps in this direction, the current paper also opens new research directions, for example to establish the optimal penalty parameter in the adjusted objective function. We look forward to further developments in this area.

References

Adrian, T. and H. S. Shin (2010). Liquidity and leverage. *Journal of Financial Intermediation* 19(3), 418–437.

- Blasques, F., S. J. Koopman, and A. Lucas (2014). Maximum likelihood estimation for generalized autoregressive score models. *Tinbergen Institute Discussion Paper* (TI 14-029/III).
- Blasques, F., S. J. Koopman, and A. Lucas (2015). Information theoretic optimality of observation driven time series models for continuous responses. *Biometrika*, forthcoming.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.
- Campbell, J. Y. and J. Cochrane (1999, April). Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *Journal of Political Economy* 107(2), 205–251.
- Campbell, J. Y. and J. H. Cochrane (2000). Explaining the poor performance of consumption-based asset pricing models. *The Journal of Finance* 55(6), 2863–2878.
- Carrasco, M. and J.-P. Florens (2000). Generalization of gmm to a continuum of moment conditions. *Econometric Theory* 16(06), 797–834.
- Chen, X. and S. C. Ludvigson (2009). Land of addicts? an empirical investigation of habit-based asset pricing models. *Journal of Applied Econometrics* 24(7), 1057–1093.
- Constantinides, G. M. (1990). Habit formation: A resolution of the equity premium puzzle. *Journal of political Economy*, 519–543.
- Cox, D. R. (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics* 8, 93–115.
- Creal, D., S. J. Koopman, and A. Lucas (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics* 29(4).
- Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* 28(5), 777–795.
- Creal, D., B. Schwaab, S. J. Koopman, and A. Lucas (2014). Observation driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics* 96(5), 898–915.
- Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*. Oxford University Press.

- Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods* (2 ed.). Oxford, UK: Oxford University Press.
- Eggermont, P. and V. LaRiccia (2001). *Maximum Penalized Likelihood Estimation*. Number v. 2 in Springer Series in Statistics. Springer-Verlag New York.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50(4), 987–1007.
- Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* 66(5), 1127–1162.
- Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business* 38(1), 34–105.
- Fernández-Villaverde, J. and J. Rubio-Ramírez (2013). Macroeconomics and Volatility: Data, Models, and Methods. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics: Tenth World Congress*, pp. 137–183. Cambridge, UK: Cambridge University Press.
- Feuerverger, A. and P. McDunnough (1981). On the efficiency of empirical characteristic function procedures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20–27.
- Francq, C. and J.-M. Zakoïan (2010). *GARCH Models: Structure, Statistical Inference and Financial Applications*. Wiley.
- Gallant, A. R., R. Giacomini, and G. Ragusa (2014). Generalized method of moments with latent variables. Working paper, Penn State University, Department of Economics.
- Gallant, A. R. and G. Tauchen (1996). Which moments to match? *Econometric Theory* 12, 65–81.
- Gallant, A. R. and G. Tauchen (1998). Reprojecting partially observed systems with application to interest rate diffusions. *Journal of the American Statistical Association* 93(441), 10–24.
- Garcia, R., E. Renault, and D. Veredas (2011). Estimation of stable distributions by indirect inference. *Journal of Econometrics* 161(2), 325–337.
- Ghysels, E. and A. Hall (1990). Are consumption-based intertemporal capital asset pricing models structural? *Journal of Econometrics* 45(1), 121–139.

- Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48(5), 1779–1801.
- Gouriéroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics* 8(S), 85–118.
- Hamilton, J. D. (2010). Macroeconomics and ARCH. In T. Bollerslev, J. R. Russell, and M. Watson (Eds.), *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, pp. 79–95. Cambridge, UK: Cambridge University Press.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (2011). *Robust statistics: the approach based on influence functions*, Volume 114. John Wiley & Sons.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.
- Hansen, L. P., J. C. Heaton, and N. Li (2008). Consumption strikes back? Measuring long-run risk. *Journal of Political Economy* 116(2), 260–302.
- Hansen, L. P. and K. J. Singleton (1982). Generalized instrumental variables estimation of non-linear rational expectations models. *Econometrica: Journal of the Econometric Society*, 1269–1286.
- Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press.
- Harvey, A. C. and A. Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, forthcoming.
- Janus, P., S. J. Koopman, and A. Lucas (2014). Long memory dynamics for multivariate dependence under heavy tails. *Journal of Empirical Finance* 29, 187–206.
- Kahneman, D. and A. Tversky (1984). Choices, values, and frames. *American Psychologist* 39(4), 341.
- Kim, C.-J. and C. R. Nelson (1999). *State Space Models with Regime Switching Classical and Gibbs Sampling Approaches with Applications*. Cambridge, MA: MIT Press.
- Kotchoni, R. (2012). Applications of the characteristic function-based continuum gmm in finance. *Computational Statistics & Data Analysis* 56(11), 3599–3622.

- Lettau, M. and S. C. Ludvigson (2009). Euler equation errors. *Review of Economic Dynamics* 12(2), 255–283.
- Lucas, A., B. Schwaab, and X. Zhang (2014). Conditional euro area sovereign default risk. *Journal of Business and Economic Statistics* 32(2), 271–284.
- Ludvigson, S. C. (2011). Advances in consumption-based asset pricing: Empirical tests. Technical report, National Bureau of Economic Research.
- Mandelbrot, B. B. (1963). The variation of certain speculative prices. *Journal of Business* 36(4), 394–419.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57(5), 995–1026.
- Mehra, R. and E. C. Prescott (1985). The equity premium: A puzzle. *Journal of Monetary Economics* 15(2), 145–161.
- Nelson, D. B. and D. P. Foster (1994). Asymptotic filtering theory for univariate arch models. *Econometrica* 62(1), 1–41.
- Newey, W. K. and K. D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.
- Newey, W. K. and K. D. West (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies* 61(4), 631–653.
- Rondeau, V., D. Commenges, and P. Joly (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis* 9(2), 139–153.
- Savov, A. (2011). Asset pricing with garbage. *The Journal of Finance* 66(1), 177–201.
- Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Oxford, UK: Oxford University Press.
- Shiller, R. (2005). *Irrational Exuberance*. Broadway Books. Currency/Doubleday.
- Straumann, D. and T. Mikosch (2006). Quasi-maximum-likelihood estimation in conditionally heteroskedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics* 34(5), 2449–2495.

- Weron, R. (1996). On the chambers-mallows-stuck method for simulating skewed stable random variables. *Statistics & probability letters* 28(2), 165–171.
- White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge university press.
- Yu, J. (2004). Empirical characteristic function estimation and its applications. *Econometric reviews* 23(2), 93–123.

Appendix A: Proofs

Proof of Proposition 1. Define

$$\tilde{\mathbf{g}}_t(\mathbf{f}) = \int_{\mathcal{W}} \mathbf{g}_t(w; \mathbf{f}, \boldsymbol{\theta}) \, dF_w(w), \quad (\text{A.1})$$

$$\tilde{\mathbf{G}}_t(\mathbf{f}) = \frac{\partial}{\partial \mathbf{f}^\top} \int_{\mathcal{W}} \mathbf{g}_t(w; \mathbf{f}, \boldsymbol{\theta}) \, dF_w(w), \quad (\text{A.2})$$

where $F_w(w)$ is the conditional distribution of \mathbf{w}_t given all information up to time $t-1$ and given $\mathbf{w}_t \in \mathcal{W}$. Also define $\mathbf{V} = \mathbf{G}_t^\top \boldsymbol{\Omega}_{t-1} \mathbf{G}_t$, which is positive definite for $w_t \in \mathcal{W}$ by assumption. If \mathcal{W} collapses to the singleton $\{w_t\}$, then $\tilde{\mathbf{g}}_t(\mathbf{f}_t) = \mathbf{g}_t(w_t; \mathbf{f}_t, \boldsymbol{\theta})$ and $\tilde{\mathbf{G}}_t(\mathbf{f}_t) = \mathbf{G}_t(w_t; \mathbf{f}_t, \boldsymbol{\theta})$. As $\mathbf{g}_t(\mathbf{w}; \mathbf{f}, \boldsymbol{\theta})$ is assumed to be continuously differentiable, also \mathbf{f}_{t+1} is a continuous function of w_t . We obtain

$$\begin{aligned} & \tilde{\mathbf{g}}_t(\mathbf{f}_t)^\top \boldsymbol{\Omega}_{t-1} \tilde{\mathbf{g}}_t(\mathbf{f}_t) - \tilde{\mathbf{g}}_t(\mathbf{f}_{t+1})^\top \boldsymbol{\Omega}_{t-1} \tilde{\mathbf{g}}_t(\mathbf{f}_{t+1}) \\ &= -2(\mathbf{f}_{t+1} - \mathbf{f}_t)^\top \tilde{\mathbf{G}}_t(\mathbf{f}_{t+1}^*)^\top \boldsymbol{\Omega}_{t-1} \tilde{\mathbf{g}}_t(\mathbf{f}_{t+1}^*) \\ &= 2\mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_{t+1}, \boldsymbol{\theta})^\top \boldsymbol{\Omega}_{t-1} \mathbf{G}_t \mathbf{V}^* \mathbf{A}^\top \tilde{\mathbf{G}}_t(\mathbf{f}_{t+1}^*)^\top \boldsymbol{\Omega}_{t-1} \tilde{\mathbf{g}}_t(\mathbf{f}_{t+1}^*), \\ &= 2a \int_{\mathcal{W}} \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_{t+1}, \boldsymbol{\theta})^\top \boldsymbol{\Omega}_{t-1} \cdot \mathbf{G}_t \mathbf{V}^* \tilde{\mathbf{G}}_t(\mathbf{f}_{t+1}^*)^\top \cdot \\ & \quad \boldsymbol{\Omega}_{t-1} \mathbf{g}_t(\mathbf{w}; \mathbf{f}_{t+1}^*, \boldsymbol{\theta}) \, dF_w(w), \end{aligned} \quad (\text{A.3})$$

where \mathbf{f}_{t+1}^* is a point between \mathbf{f}_t and \mathbf{f}_{t+1} . For $\mathcal{W} = \{w_t\}$, the integrand is strictly positive. Given the assumed continuity of $\mathbf{g}_t(\mathbf{w}; \mathbf{f}, \boldsymbol{\theta})$ and $\mathbf{G}_t(\mathbf{w}; \mathbf{f}, \boldsymbol{\theta})$ in \mathbf{w} and \mathbf{f} , and the subsequent continuity of \mathbf{f}_{t+1} as a function of w_t , the result follows immediately for a small enough ball \mathcal{W} around $\{w_t\}$. \square

Appendix B: Sampling from stable distributions

To simulate draws from a stable distribution, we follow the generalized Chambers-Mallows-Stuck procedure developed in Weron (1996, including erratum). Let v_u and w_u be two i.i.d. random variables drawn from a uniform distribution $U[0, 1]$. Define $v = \pi(v_u - \frac{1}{2})$ and $w = -\log w_u$. For given α, β, σ , and μ , we obtain a random vector $Z \sim S(\alpha, \beta, \sigma, \mu)$ from the trans-

formations

$$B = \frac{1}{\alpha} \operatorname{atan} \left[\beta \tan \left(\alpha \frac{\pi}{2} \right) \right], \quad (\text{B.1})$$

$$S = \left[1 + \beta^2 \tan^2 \left(\frac{1}{2} \alpha \pi \right) \right]^{\frac{1}{2\alpha}}, \quad (\text{B.2})$$

$$Z = \mu + \sigma \frac{S \sin(\alpha[v + B])}{\cos^{\frac{1}{\alpha}}(v)} \left(\frac{\cos(v - \alpha[v + B])}{w} \right)^{-1 + \frac{1}{\alpha}}, \quad (\text{B.3})$$

for $\alpha \neq 1$, and

$$B = \frac{\pi}{2} + \beta v, \quad (\text{B.4})$$

$$Z = \frac{2}{\pi} B \sigma \tan(v) - \frac{2}{\pi} \sigma \beta \log \left(\frac{\frac{\pi}{2} w \cos(v)}{B} \right) + \mu + \frac{2}{\pi} \beta \sigma \log(\sigma), \quad (\text{B.5})$$

for $\alpha \equiv 1$.

Appendix C: Estimating procedure for the derivative matrix

Two considerations need to be made when it is necessary to estimate \mathbf{G}_t and $\boldsymbol{\theta}$ jointly; see the example in Subsection 3.3.2. First of all, since in these models \mathbf{G}_t depends on the true values of $\boldsymbol{\theta}$ and \mathbf{f}_t at time t ; if $\mathbf{f}_{0,t=i} \neq \mathbf{f}_{0,t=j}$ then $\mathbf{G}_{t=i} \neq \mathbf{G}_{t=j}$ (parameter-value inconsistency). It follows that estimates of \mathbf{G}_t need to take into account in which neighbourhood of the parameter space \mathcal{F} the filtered parameter $\hat{\mathbf{f}}_{t-1}$ is located. More importantly, especially at the beginning of the estimation procedure for $\boldsymbol{\theta}$ it may very well be that while $\hat{\mathbf{f}}_{t=i} = \hat{\mathbf{f}}_{t=j}$, it is still true that $\mathbf{f}_{0,t=i} \neq \mathbf{f}_{0,t=j}$ (time-period inconsistency) and thus using the same value of \mathbf{G}_t for $t = i$ and $t = j$ should be avoided. If $\{\mathbf{f}_t\}$ is relatively slowly-varying, we can account for both issues by estimating \mathbf{G}_t as an exponentially weighted moving average. Let for some λ :

$$\mathbf{G}_t^{(i)} = \vec{\mathbf{G}}_t^{(i)} = \lambda \vec{\mathbf{G}}_{t-1}^{(i)} + (1 - \lambda) \frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \mathbf{f}_t}, \quad (\text{C.1})$$

where the superscript (i) denotes the i -th iteration of the optimizing algorithm and the arrow denotes direction in the time domain in which we apply the EWMA. In practice we find that λ can be chosen to be in the $(0.98, 1.0)$ interval. For $i = 0$, we initialize the recursion in (C.1) by setting $\vec{\mathbf{G}}_{t=0}^{(0)} = \mathbf{I}$. In order to update this initial guess for $\mathbf{G}_{t=0}$ we propose to apply the EWMA

in reverse, i.e. let:

$$\overleftarrow{\mathbf{G}}_t^{(i)} = \lambda \overleftarrow{\mathbf{G}}_{t+1}^{(i)} + (1 - \lambda) \frac{\partial \mathbf{g}_t(\mathbf{w}_t; \mathbf{f}_t, \boldsymbol{\theta})}{\partial \mathbf{f}_t}, \quad (\text{C.2})$$

$$\overrightarrow{\mathbf{G}}_{t=0}^{(i+1)} = \overleftarrow{\mathbf{G}}_{t=0}^{(i)}. \quad (\text{C.3})$$

Note that estimating \mathbf{G}_t in this manner does not increase computational burden significantly.

In second stage step of GaMM estimation, it is possible to replace this procedure and use non-parametric estimation method for \mathbf{G}_t . This is because given the estimate of $\hat{\boldsymbol{\theta}}$ from the first-step, the path $\{\hat{\mathbf{f}}_t\}$ should be relatively close to the true path and the time-period inconsistency is no longer an issue. We propose to use a Gaussian kernel to obtain $\mathbf{G}_t^{(i)}$ based on previously obtained $\{\mathbf{G}_t^{(i-1)}\}$ and $\{\mathbf{f}_t^{(i-1)}\}$. For $i = 0$, we initialize the kernel density smoother with $\{\mathbf{G}_t\}$ and $\{\mathbf{f}_t\}$ obtained at the first stage of the estimation.

Supplemental Appendix A: Time-varying linear regression models with endogenous covariates

Model

Linear regression models are standard tools in economic analysis. A typical concern is the endogeneity of one of the regressors with the error term due to omitted variable bias or measurement error.

$$y_t = \mathbf{x}_t^\top \boldsymbol{\beta}_t + \varepsilon_t$$

Given the availability of an instrument variable z_t , the obvious way to estimate $\boldsymbol{\beta}_t$ is via the conditional moment condition

$$\mathbb{E}_{t-1} [z_t (y_t - \mathbf{x}_t^\top \boldsymbol{\beta}_t)] = 0. \quad (\text{SA.1})$$

The complication here is that the parameters $\boldsymbol{\beta}_t$ are time-varying. The moment condition (SA.1), however, lends itself directly to the GaMM framework by setting $\mathbf{f}_t = \boldsymbol{\beta}_t$.

In contrast to the setting with stable distributions in Section 3.1 where \mathbf{G}_t did not depend on the data at time t , $\mathbf{G}_t = -\mathbb{E}_{t-1} [z_t \mathbf{x}_t^\top]$ now obviously does require us to compute the conditional expectation for each time t , even though the exogenous instruments z_t are assumed to lie in the conditioning information set used to compute this conditional expectation. We can obtain the matrix \mathbf{G}_t in a straightforward way by noting from (SA.3) that $\mathbb{E}_{t-1} [z_t \mathbf{x}_t^\top] = z_t z_t^\top \boldsymbol{\Pi}_x$. Define \mathbf{X} and \mathbf{Z} as the $T \times 2$ matrices with t th row equal to \mathbf{x}_t^\top and z_t^\top , respectively. Given the exogeneity of the instruments z_t , we can estimate \mathbf{G}_t in the standard way by $\hat{\mathbf{G}}_t^\top = z_t z_t^\top \hat{\boldsymbol{\Pi}}_x = z_t z_t^\top (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}$. The fact that we can use all observations to estimate $\hat{\boldsymbol{\Pi}}_x$ rather than only the observations up to time $t - 1$ follows from the exogeneity assumption for z_t and the assumption that $\boldsymbol{\Pi}_x$ is static.

Simulation results

Consider a simple linear model where all coefficients are time-varying:

$$y_t = \beta_{0,t} + \beta_{1,t}x_t + \varepsilon_{yt}, \quad \varepsilon_{yt} \sim N(0, \sigma_{\varepsilon_y}^2), \quad E[\varepsilon_{yt}\varepsilon_{xt}] = \rho, \quad (\text{SA.2})$$

$$x_t = \pi_{0,x} + \pi_{1,x}z_t + \varepsilon_{xt}, \quad \varepsilon_{xt} \sim N(0, \sigma_{\varepsilon_x}^2), \quad z_t \sim N(\mu_z, \sigma_z^2), \quad (\text{SA.3})$$

$$\beta_{i,t+1} = \pi_\beta \beta_{i,t} + \eta_{i,t}, \quad \eta_{i,t} \sim N(0, \sigma_{i,\eta}^2), \quad i = 0, 1. \quad (\text{SA.4})$$

If $\rho \neq 0$ and $\sigma_{\varepsilon_x}^2 > 0$, we have a standard endogenous regressor problem. For lower values of $\sigma_{\varepsilon_x}^2$, the size of the bias when estimating $\beta_{1,t}$ using standard least-squares based methods, is larger. Define $\mathbf{x}_t^\top = (1, x_t)$, $\mathbf{z}_t^\top = (1, z_t)$, and $\boldsymbol{\beta}_t^\top = (\beta_{0,t}, \beta_{1,t})$, and let $\boldsymbol{\Pi}_x$ be such that $\mathbf{x}_t^\top = \mathbf{z}_t^\top \boldsymbol{\Pi}_x + (0, \varepsilon_{xt})$. As a benchmark model we use an ordinary state space model consisting of equations (SA.2) and (SA.4) and estimated using standard Kalman Filter methods. This is close to comparing the performance of an instrumental variables (IV) estimator with an ordinary least squares (OLS) estimator in a static context, where in our dynamic context the Kalman Filter and the GaMM estimator take the roles of the OLS and IV estimators, respectively.

Our data generating process uses equations (SA.2)–(SA.4). The parameters we selected are $\rho = 0.5$, $\sigma_{\varepsilon_y}^2 = 1$, $\pi_x = 0.5$, $\mu_z = 0$, $\sigma_z^2 = 1$, and $\boldsymbol{\pi}_\beta^\top = [0.98, 0.98]^\top$. We vary σ_{ε_x} and σ_η to study the effect of different magnitudes of the endogeneity problem and of the time-variation of the coefficients. We use $\sigma_{\varepsilon_x}^2 = \{0.5, 4\}$ and $\sigma_\eta^2 = \{0.01, 0.25, 0.75\}$, where lower values of σ_{ε_x} result in larger biases of the least-squares based estimator. All simulations are repeated for $T = \{1000, 2500, 5000\}$ observations, with 1000 observations corresponding to approximately 4 years of daily data. We obtained similar results for simulations conducted with shorter time-series $T = \{250, 500\}$, as well as for other time-varying patterns for $\boldsymbol{\beta}_t$ than in equation (SA.4), including structural breaks and slowly varying sinusoid waves.

The results of 10,000 replications are presented in Figure SA.1 and Figure SA.2. The figures summarize a substantial amount of information. In the top panel (Panel A), we summarize the results for $\hat{\beta}_{0,t}$, while the bottom panel (Panel B) presents the results for the coefficient $\beta_{1,t}$ corresponding to the endogenous regressor. The results are presented as box-plots using the 10,000 simulated average bias and RMSE statistics as defined in equation (31). In the left-half of the plot, the endogeneity problem is moderate with a high value of $\sigma_{\varepsilon_x}^2 = 4.0$. In the right-hand half of the graph, the endogeneity problem is more severe with $\sigma_{\varepsilon_x}^2 = 0.5$. This becomes particularly

Figure SA.1

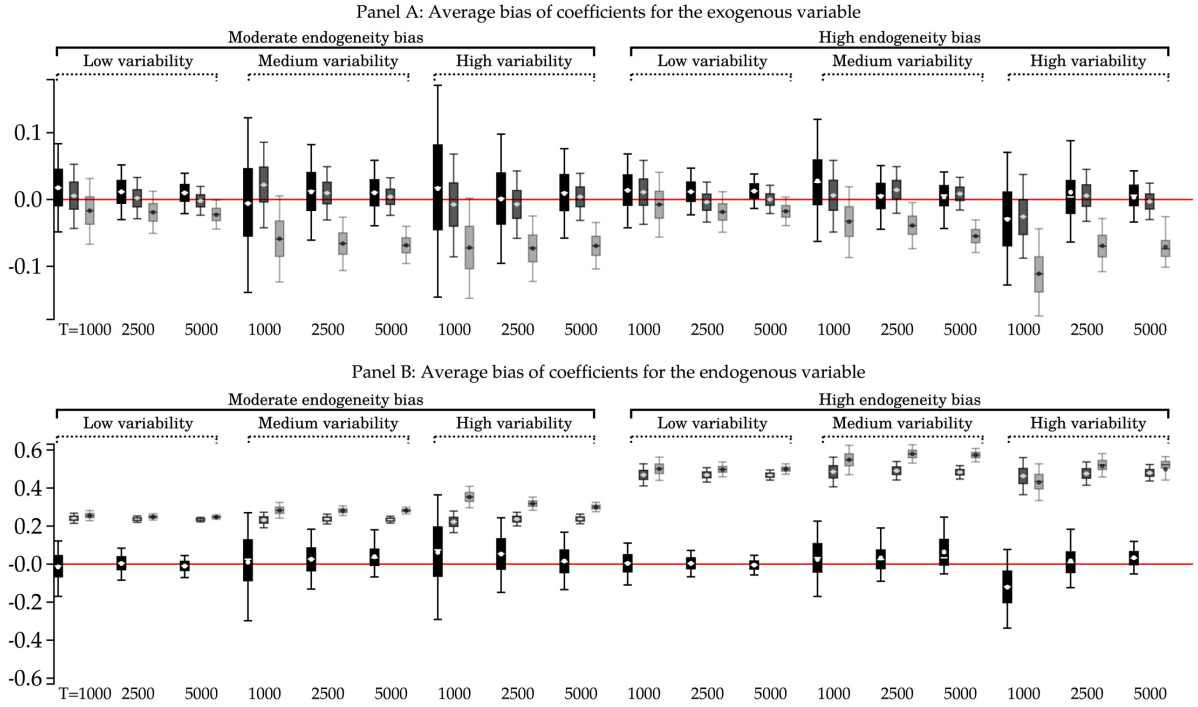
Bias of GaMM for a Simple Endogeneity Problem

The figure compares performance of GaMM (black) to Kalman Filter with AR1 (light grey) or Random Walk (dark grey) dynamics for coefficients. We simulate observations and true parameters from a simple linear model where all coefficients are time-varying:

$$\begin{aligned} y_t &= \beta_{0,t} + x_t \beta_{1,t} + \varepsilon_t \\ x_t &= 0.5z_t + \phi_t \\ \beta_{i,t+1} &= 0.98\beta_{i,t} + \eta_{i,t} \end{aligned}$$

$$E[\varepsilon_t \phi_t] = 0.5 \quad \varepsilon_t \sim N(0, 1), \quad \phi_t \sim N(0, \sigma_\phi) \quad z_t \sim N(0, 1) \quad \eta_{i,t} \sim N(0, \sigma_\eta)$$

We consider moderate ($\sigma_{\varepsilon_x}^2 = 4.0$) and high ($\sigma_{\varepsilon_x}^2 = 0.5$) degree of the endogeneity problem. For both the moderate and the high endogeneity bias case, we consider different degrees of time-variation in β_t , from low ($\sigma_\eta^2 = 0.01$) to high ($\sigma_\eta^2 = 0.75$). Simulations are run for different sample sizes (horizontal axis). The results are based on 10,000 replications.



important for the coefficient $\beta_{1,t}$ in Panel B. For both the moderate and the high endogeneity bias case, we consider different degrees of time-variation in β_t , from low ($\sigma_\eta^2 = 0.01$) to high ($\sigma_\eta^2 = 0.75$). For each of the $2 \times 9 = 18$ combinations, we plot the results for three different simulated sample sizes $T = \{1000, 2500, 5000\}$. Each group of three box-plots corresponds to a combination of sample size, degree of time-variation in β_t , and severity of the endogeneity prob-

lem. The three box-plots correspond to three different models: GaMM(1,1) (black), state space model with random walk dynamics (dark grey), and state space model with autoregressive (AR) dynamics of order 1 (light grey).

Figure SA.2

RMSE of GaMM for a Simple Endogeneity Problem

The figure compares performance of GaMM (black) to Kalman Filter with AR1 (light grey) or Random Walk (dark grey) dynamics for coefficients. We simulate observations and true parameters from a simple linear model where all coefficients are time-varying:

$$\begin{aligned} y_t &= \beta_{0,t} + x_t \beta_{1,t} + \varepsilon_t \\ x_t &= 0.5z_t + \phi_t \\ \beta_{i,t+1} &= 0.98\beta_{i,t} + \eta_{i,t} \end{aligned}$$

$$E[\varepsilon_t \phi_t] = 0.5 \quad \varepsilon_t \sim N(0, 1), \quad \phi_t \sim N(0, \sigma_\phi) \quad z_t \sim N(0, 1) \quad \eta_{i,t} \sim N(0, \sigma_\eta)$$

We consider moderate ($\sigma_{\varepsilon_x}^2 = 4.0$) and high ($\sigma_{\varepsilon_x}^2 = 0.5$) degree of the endogeneity problem. For both the moderate and the high endogeneity bias case, we consider different degrees of time-variation in β_t , from low ($\sigma_\eta^2 = 0.01$) to high ($\sigma_\eta^2 = 0.75$). Simulations are run for different sample sizes (horizontal axis). The results are based on 10,000 replications.

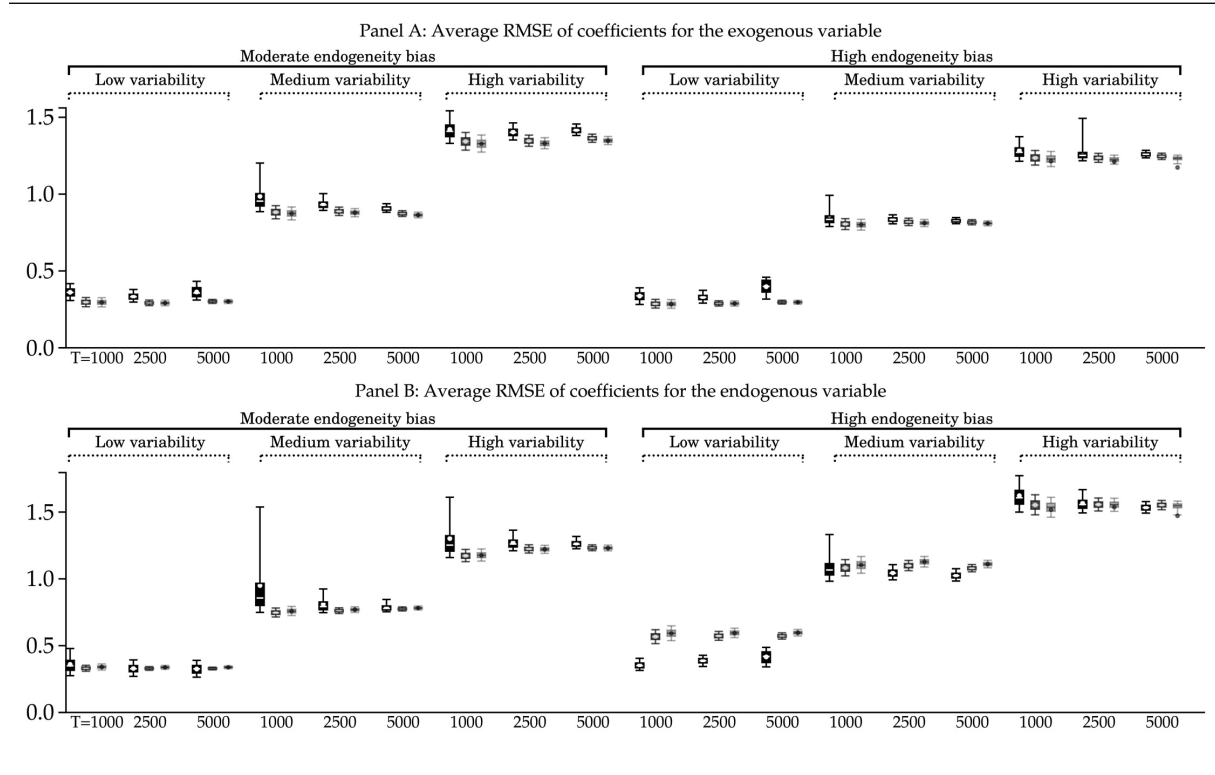


Figure SA.1 shows the in-sample performance in terms of the average bias. Typically, out of the two state space models, the specification with random walk dynamics for the parameters

performs better and we continue with this model as our main benchmark.

In Panel A, we see that in most cases both the GaMM and the Kalman Filter estimates based on random walk dynamics offer a similar bias performance for the exogenous parameter. The average and median biases are close to zero. It is also clear that the distribution of the average bias for the GaMM approach has a higher spread. The relative differences in performance diminish substantially as the sample size increases. For the largest sample sizes the results produced with GaMM are often as accurate as the ones produced with the Kalman Filter.

In Panel B of Figure SA.1 we see that GaMM clearly outperforms the Kalman Filter in terms of bias of the estimator for the parameter corresponding to the endogenous variable. Regardless of sample size, size of the time-variability, or magnitude of the endogeneity bias; the Kalman Filter estimates are clearly biased (as expected), whereas the GaMM approach results in unbiased estimates of the parameter path in almost all cases. As expected, the bias of the Kalman Filter estimates is larger for $\sigma_{\varepsilon_x}^2 = 0.5$ than for $\sigma_{\varepsilon_x}^2 = 4.0$. In fact, the 5% of best results generated by the Kalman Filter are in most cases worse than 95% of the results produced by GaMM. As before, the performance of GaMM improves with the sample size, making the improvement over the biased least-squares based methods even more apparent.

Figure SA.2 focuses on average root-mean square errors produced by both methods. For both the exogenous and endogenous variables, RMSEs produced by either method are comparable on average. However, the distribution of RMSE produced by GaMM has a considerably heavier right tail: in some cases, the RMSE behavior of GaMM can be substantially worse than that of the Kalman Filter methods. The number of simulations with poor RMSE produced by GaMM reduces substantially as the sample size increases.

Summarizing, the trade-off between the Kalman Filter and GaMM approach seems to mirror the differences between OLS and IV estimation in the case of static parameters. The Kalman Filter produces results which are biased but with low sampling variability, whereas paths estimated by GaMM appear to be unbiased, but at the cost of a higher sampling variance.

We also repeat the exercise using the penalized criterion function. Partial results are discussed in Section 4. Figure SA.3 contains results for other sample sizes, and different values of both σ_{ε_x} and σ_{η}

Figure SA.3

Performance of GaMM for a Simple Endogeneity Problem With Penalized Criterion Function

The figure compares performance of GaMM (black) to Kalman Filter with AR1 (dark grey) or Random Walk (light grey) dynamics for coefficients. We simulate observations and true parameters from a simple linear model where all coefficients are time-varying as in Figures 3 and 4. GaMM estimation was performed for different values of the smoothing parameter, $\lambda = \{10, 1, 0.2, 0, 0^*\}$, and are presented in this order. 0^* denotes the case where unpenalized GaMM was initialized with estimates obtained after running GaMM with penalty weight of 1.

