Chapter 3

Generalised autoregressive Method of Moments

3.1. Introduction

This chapter introduces a new estimation framework which extends the Generalised 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 Generalised autoregressive Method of Moments (GaMM) and provide several empirical applications that illustrate its usefulness.

This chapter is based on Creal, Koopman, Lucas, and Zamojski (2015).
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 emphasised 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 stable 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. Appendix 3.D 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 specialised maximum likelihood framework with the generalised autoregressive score dynamics of Creal, Koopman, and Lucas (2011, 2013) and Harvey (2013); see Blasques, Koopman, and Lucas (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 generalised autoregressive conditional heteroscedasticity (GARCH) models. Estimation and inference is relatively straightforward in the GaMM framework and does not require simulation.

This chapter 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, Monfort, and Renault (1993), and the recent extension of GMM to latent variables by Gallant, Giacomini, and Ragusa (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
Generalised autoregressive Method of Moments (1998) in conjunction with their efficient method of moments (EMM) and the approach by Gallant, Giacomini, and Ragusa (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, this chapter 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 generalises 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 generalised autoregressive score approach of Creal, Koopman, and Lucas (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, Koopman, and Lucas (2011, 2013) as well Harvey and Luati (2014), Lucas, Schwaab, and Zhang (2014), and Creal, Schwaab, Koopman, and Lucas (2014). In addition, our framework gives rise to new time-varying parameter models that have not been studied before.

The remainder of this chapter is organised as follows. In Section 3.2, we describe the basic methodology. Section 3.3 contains examples to illustrate the relevance of GaMM for applied work. In Section 3.4, we discuss penalised extensions of the methodology that allow for improved finite sample properties of the estimator. Section 3.5 concludes.
3.2. Methodology

A motivating example

We start with a motivating example. Consider the problem of estimating the mean \( \mu \) 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 \( t = 1, \ldots, \tau \), and \( E[y_t] = \mu_1 \) for \( t = \tau + 1, \ldots, T \), where \( \mu_0 < \mu_1 \). Using the same (unconditional) moment condition to estimate \( \mu \) 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 \( \mu_0 \) is substantially below \( \mu_1 \), then \( (y_t - \hat{\mu}) \) is negative on average for \( t = 1, \ldots, \tau \), and positive for \( t = \tau + 1, \ldots, T \). 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 \( \mu \) 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 \( \mu \) 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 Generalised 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 \( \mu \), 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

$$s_t = -2(y_t - f_t).$$  \hspace{1cm} (3.1)$$

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

$$f_{t+1} = \omega \cdot (I - B_1) + B_1 f_t + A_1 s_t,$$  \hspace{1cm} (3.2)$$

where $\omega$, $A_1$, and $B_1$ are static parameters that need to be estimated and that describe the dynamic behaviour of $f_t$. It is easy to generalise this specification to include more lags of $f_t$ and $s_t$ (see Creal, Koopman, and Lucas, 2013), non-linearity, structural time series dynamics (see Harvey and Luati, 2014), or fractional integration (see Janus, Koopman, and Lucas, 2014).

It is evident that $f_t$ in eq. (3.2) has observation-driven dynamics. Given information up to time $t$, the parameter $f_{t+1}$ is known as it only depends on $y_t, y_{t-1}, \ldots$ which is similar to a GARCH model. This makes the proposed GaMM methodology computationally fast and renders parameter estimation and inference straightforward. We emphasise that the GaMM factor dynamics are not arbitrary. In Section 3.2, we describe the optimality properties of these dynamics.

From the example above, define the vector of static parameters as $\theta = (\omega, A_1, B_1)^T$.

To estimate $\theta$, we need unconditional moment conditions which we obtain these by instrumenting the conditional moment condition $E_{t-1}[y_t - f_t]$. We propose the unconditional moment conditions

$$E \left[ (y_t - f_t) \otimes (1, f_{t-1}, s_{t-1})^T \right] = 0,$$  \hspace{1cm} (3.3)$$

with $\otimes$ denoting the Kronecker product. Due to the presence of the constant term in the vector of instruments, the original (conditional) moment condition $E[y_t - f_t]$ 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 $E_{t-1}[y_t - f_t] = 0$ by including in eq. (3.3) the autocorrelation of the misfit of the cross-product between $s_t = -2(y_t - f_t)$ and $s_{t-1}$. The fact that the unconditional expectation of this cross-product needs to be zero in eq. (3.3) forces the dynamic scheme in eq. (3.2) to remove as much autocorrelation in $(y_t - f_t)$ as possible.

The distributional properties of the GMM estimator for $\theta$ in a framework with GaMM dynamics turn out to be straightforward. Equation (3.3) fits into the GMM framework of Hansen (1982) under standard assumptions. Therefore, consistency and asymptotic normality of the estimator for $\theta$ follow easily, as do the optimal weighting matrices for two-stage feasible GMM estimation of $\theta$. 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.

**GaMM dynamics**

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

$$E [g_t (w_t; f, \theta)] = 0,$$

where $g_t : \mathbb{R}^m \times F \times \Theta \rightarrow \mathbb{R}^K$, $w_t \in \mathbb{R}^m$ is observed, and $f \in F$ and $\theta \in \Theta$ denote parameter vectors that lie in the parameter spaces $F$ and $\Theta$, 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 $\hat{f}_t$. As $f_t$ is observation-driven, it can be written as a function of past observations $w_{t-1}, w_{t-2}, \ldots$

We now replace the unconditional moment condition in eq. (3.4) by its conditional counterpart

$$E_{t-1} [g_t (w_t; f_t, \theta)] = 0.$$  

The dynamic specification for $f_t$ starts by considering the GMM objective function at
time $t$, 

$$
E_{t-1} \left[ g_t(w_t; f_t, \theta) \right] \Omega_t E_{t-1} \left[ g_t(w_t; f_t, \theta) \right]^\top.
$$  \hfill (3.6)

where $\Omega_t$ is a time $t-1$ measurable weighting matrix that is positive semi-definite almost surely. The expectations in eq. (3.6) are computed under the true time $t-1$ conditional measure $F_{w_t}$, for which eq. (3.5) holds. To propagate $f_t$ forward to $f_{t+1}$ given the realisation of $w_t$, we take a scaled steepest descent step of eq. (3.6) using an appropriate derivative of the function in eq. (3.6) evaluated at $f_t$. The derivative concept used should account for the specific value of $w_t$ realised 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 $f_t$ given the objective function in eq. (3.6); see, e.g., Hampel, Ronchetti, Rousseeuw, and Stahel (2011).

To define the Fréchet derivative, consider a contaminated measure $F^\epsilon_{w_t} = (1-\epsilon)F_{w_t} + \epsilon \delta_{w_t}$, where $\delta_{w_t}$ is the Dirac measure that puts unit assumption on the realised value of $w_t$. By considering $F^\epsilon_{w_t}$ instead of $F_{w_t}$, we account for the appropriate information in $w_t$ when updating $f_t$ to $f_{t+1}$. The first order condition corresponding to eq. (3.6) evaluated at the measure $F^\epsilon_{w_t}$ is

$$
(G^\epsilon_t)^\top \Omega_t E^\epsilon_{t-1} \left[ g_t(w_t; f_t, \theta) \right] = 0,
G^\epsilon_t = E^\epsilon_{t-1} \left[ \partial g_t(w_t; f_t, \theta) / \partial f_t \right],
$$  \hfill (3.7)

where $E^\epsilon_{t-1}$ and $E_t \equiv E^0_{t-1}$ denote the expectations operators taken with respect to $F^\epsilon_{w_t}$ and $F^0_{w_t} \equiv F_{w_t}$, respectively. The influence function of $f_t$ for the criterion function in eq. (3.6) is the Fréchet derivative of eq. (3.7) in the direction $\delta_{w_t}$, evaluated at $\epsilon = 0$. After some minor algebra, it is easy to show that the influence function equals $G^1_t \Omega_t g_t(w_t; f_t, \theta)$, with $G_t = G^0_t$ for $\epsilon = 0$; see again Hampel, Ronchetti, Rousseeuw, and Stahel (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 $f_t$ to $f_{t+1}$. The resulting scaled step is

$$
\nabla_t = -(G^1_t \Omega_t G)^* G^1_t \Omega_t g_t(w_t; f_t, \theta),
$$  \hfill (3.8)
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 eq. (3.7) 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 $G_t$ does not depend on either the data or the time-varying parameters; see the case of the stable distribution in Section 3.3, where most of the relevant matrices have rank one.

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

$$ f_{t+1} = \omega + \sum_{j=1}^p B_j \left( f_{t+1-j} - \omega \right) + \sum_{i=1}^q A_i s_{t+1-i}, \tag{3.9} $$

which we call GaMM($p, q$) dynamics, where $\omega = \omega(\theta)$, $B_j = B_j(\theta)$, and $A_i = A_i(\theta)$ are appropriately sized vectors and matrices that depend on the static parameter vector $\theta$, and

$$ s_t = -S_t \nabla_t, \tag{3.10} $$

for some time $t - 1$ measurable (almost surely) positive semi-definite scaling matrix $S_t$. The scaling matrix $S_t$ adjusts the direction of the step, for example, if one wants to annihilate the effect of $(G_t^T \Omega_t G_t)^*$ in eq. (3.8) to fully concentrate on the steepest descent direction. Additionally, we can use an efficient conditional weighting matrix $\Omega_t = E_t^{-1} [g_t (w_t; f_t, \theta) g_t^T (w_t; f_t, \theta)]^*$ in eq. (3.6). However, as $\Omega_t$ needs to be computed at each time $t$, this is computationally more demanding. In this chapter, we set $\Omega_t = I$ and $S_t = I$, which provides good results in the empirical examples considered later on; see Section 3.3 and Appendix 3.D.

Under the assumption that eq. (3.5) 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 eq. (3.9) is $E[f_t] = \omega$ if $f_t$ is first-order stationary. To start up the recursion in eq. (3.9), we can set $f_t = \omega$ for $t = 0, -1, \ldots, -p+1$ and $s_t = 0$ for $t = 0, -1, \ldots, -q+1$. 
Alternatively, the initial values can be estimated together with the static parameters in \( \theta \), 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 \( \theta = (\theta^T_f, \theta^T_c) \) where \( \theta_f \) contains the parameters governing the dynamics of \( f_t \), i.e., \( \omega \), \( A \), and \( B \), and \( \theta_c \) includes the remaining static parameters.

The joint GaMM(1,1) dynamics for the vector \( \tilde{f}_{t+1} = (f^T_t, \theta^T_c) \) are

\[
\tilde{f}_{t+1} = \begin{bmatrix} f_t^T \\ \theta_c \end{bmatrix} = \omega + B \left( \tilde{f}_{t-j+1} - \bar{\omega} \right) + A \tilde{s}_{t-i+1},
\]

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 \( \theta_c \). This can be useful, for example, when modelling 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 \( \sigma^2_t \) and constant mean \( \mu \), 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

\[
g_t = \begin{bmatrix} y_t - \mu \\ (y_t - \mu)^2 - \sigma^2_t \end{bmatrix} = \begin{bmatrix} \varepsilon_t \\ \varepsilon_t^2 - \sigma^2_t \end{bmatrix}, \quad G_t = -I.
\]

Fixing \( C^B = 0 \), the GaMM(1,1) recursion for \( f_t = \sigma^2_t \) is

\[
\sigma^2_{t+1} = (\sigma^2_t - \omega) B + \omega + C^A \varepsilon_t + A \varepsilon^2_t - A \sigma^2_t
= \omega (1 - B) + A \varepsilon^2_t + (B - A) \sigma^2_t + C^A \varepsilon_t.
\]

This equation is known as the GARCH(1,1) model with Leverage.
Equation (3.13) coincides with the familiar GARCH(1,1) model of Engle (1982) and Bollerslev (1986) with an additional leverage effect $C^A \epsilon_t$. 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 generalises the score driven approach from Creal, Koopman, and Lucas (2013).

**Local optimality of GaMM**

In this section, we derive generic local optimality properties for the GaMM dynamics introduced in Section 3.2. Similar optimality properties were derived in Blasques, Koopman, and Lucas (2015) for the generalised autoregressive score model of Creal, Koopman, and Lucas (2011, 2013). In particular, Blasques, Koopman, and Lucas show that generalised 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 generalised autoregressive score framework. In our current GMM context, optimality instead centres around the quadratic objective function of the moment conditions. As a result, the concepts and results in Blasques, Koopman, and Lucas need to be adapted accordingly. Interestingly, the optimality results in Blasques, Koopman, and Lucas (2015) hold whether or not the statistical model is correctly specified. Similarly, our results hold whether or not the moment conditions $E[g(w_t; f, \theta)] = 0$ are correctly specified.

We introduce the local GMM objective function

$$C(t, f_t, W) = E_{t-1} [g_t (w_t; f, \theta) | w_t \in W] \Omega_{t-1}^{-1} E_{t-1} [g_t (w_t; f, \theta) | w_t \in W]. \tag{3.14}$$

Equation (3.14) considers the behaviour of the GMM objective function in eq. (3.6) for a restricted set $W \subseteq \mathbb{R}^m$ of realisations of the random variable $w_t$. If $W = \mathbb{R}^m$, then
eq. (3.14) coincides with eq. (3.6). For concreteness, let \( w_t \) denote the realisation of the random variable \( \mathbf{w} \) at time \( t \). Our results are local in nature in that they focus on sets \( \mathcal{W} \) that contain the empirical realisation \( w_t \), 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 \( \mathbf{w} \). 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 realisations of the data. Note that we do not require the moment conditions to be correctly specified, i.e., we do not require \( E_{t-1} [g_i(\mathbf{w}; f_t, \theta)] = 0 \). 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 3.1.** Conditional on a realisation \( \mathbf{w}_t = w_t \in \mathcal{W} \) for a neighbourhood \( \mathcal{W} \) of \( w_t \), an update from \( f_t \) to \( f_{t+1} \) is called Realized Locally (RL) optimal if and only if

\[
C(t, f_t, \mathcal{W}) - C(t, f_{t+1}, \mathcal{W}) \geq 0,
\]

(3.15)

for every \( (w_t, f_t) \). The update is called Conditionally Locally (CL) optimal if and only if

\[
\int_{\mathcal{W}} [C(t, f_t, \mathcal{W}) - C(t, f_{t+1}, \mathcal{W})] \, dF_{\mathbf{w}}(w_t) \geq 0,
\]

(3.16)

for every \( f_t \), where \( F_{\mathbf{w}} \) 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, \( f_{t+1} \) can be seen as a function of \( w_t \). Note that \( f_t \) does not depend on \( w_t \). The concept of RL optimality considers updates from \( f_t \) to \( f_{t+1} \) that improve the local criterion \( C(t, f_t, \mathcal{W}) \) in a neighbourhood \( \mathcal{W} \) of the realisation \( w_t \) for given values of \( \mathbf{w}_t = w_t \) and \( f_t \). It, thus, treats the realised data value \( w_t \) as given. CL optimality goes one step further and no longer conditions on the realisation \( w_t \), but rather takes the expectation over the entire neighbourhood \( \mathbf{w}_t \in \mathcal{W} \). This accounts for the fact that \( \mathbf{w}_t \) has an impact on \( f_{t+1} \) through the update equation. Note that Definition 3.1 generalises
the definitions in Blasques, Koopman, and Lucas (2015) from a fully parametric setting to the semi-parametric setting of GMM. The proof of the following proposition can be found in Appendix 3.A.

**Proposition 3.1.** Let $g_t(w; f, \theta)$ and $G_t = G_t(w; f, \theta)$ be continuous in all their arguments, and let $G_t^\top \Omega_{t-1} G_t$ be positive definite for all $w_t \in W$ and given $f_t$. Then there is a GaMM(1,1) update with $\omega = 0$, $A = a \cdot I$ for $a \in \mathbb{R}^+$, and $B = I$ that is both RL-optimal and CL-optimal.

Proposition 3.1 ensures that the score driven GaMM(1,1) factor dynamics improve the local GMM objective function in eq. (3.14) 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 $E_t^{-1} [g(w_t; f_t, \theta)] = 0$ are correctly specified: the GaMM dynamics still operate to minimise the local deviations from the moment conditions laid down by the econometrician by efficiently processing the new observations and updating $f_t$ to $f_{t+1}$. Note that the positive semi-definiteness of the matrix $G_t^\top \Omega_{t-1} G_t$ may be easily satisfied, depending on the specific setting. For example, if $G_t$ does not depend on $w_t$ such as in our example of stable distributions with time-varying scale, the positive semi-definiteness follows directly for small enough $W$.

Along the same lines as the results in Blasques, Koopman, and Lucas (2015) for generalised 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 eq. (3.9) rather than the restrictive setting of $\omega = 0$, $A = a \cdot I$, and $B = I$. Each of these properties continues to hold under model misspecification.

**Choice of instruments**

To estimate $\theta^\top = (\theta_c^\top, \theta_f^\top)$, we augment the conditional moment conditions in eq. (3.5) by the matrix of instruments $W_t \in \mathbb{R}^{L \times K}$ to arrive at the unconditional moment condi-
Generalised autoregressive Method of Moments

g_{t}(w_{t}; f_{t}, \theta) = 0,

(3.17)

and the corresponding GMM objective function

\[
\min_{\theta \in \Theta} \bar{g}_{T}^{\top} \bar{\Omega}_{T} \bar{g}_{T}, \quad \bar{g}_{T} = \frac{1}{T} \sum_{t=1}^{T} W_{t} g_{t}(w_{t}; f_{t}, \theta),
\]

(3.18)

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

The matrix of instruments we propose equals \( W_{t} = (1 \ f_{t-1} \ s_{t-1})^{\top} \otimes I \), such that

\[
E[(1 \ f_{t-1} \ s_{t-1})^{\top} \otimes g_{t}(w_{t}; f_{t}, \theta)] = E[(1 \ f_{t-1} \ s_{t-1})^{\top} \otimes g_{t}(w_{t}; f_{t}, \theta)] = 0.
\]

(3.19)

The instrumented moment conditions \( E[1 \otimes g_{t}(w_{t}; f_{t}, \theta)] = 0, E[s_{t-1} \otimes g_{t}(w_{t}; f_{t}, \theta)] = 0, \) and \( E[f_{t-1} \otimes g_{t}(w_{t}; f_{t}, \theta)] = 0 \) 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, \( (g_{t}(w_{t}; f_{t}, \theta) \otimes s_{t-1}) \) holds the cross products of the moment conditions \( g_{t} \) and their lags \( g_{t-1} \) via the lagged scores \( s_{t-1} \). If there is autocorrelation in \( g_{t} \), GMM adjusts \( A \) and \( B \) to push this autocorrelation closer to zero in line with the intuition underlying the GaMM dynamics.

A different way to motivate eq. (3.19) is to consider what optimal instruments would be for estimating \( \theta \); 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

\[
E[g_{s}(w_{s}; f_{s}, \theta) g_{t}^{\top}(w_{t}; f_{t}, \theta)] = 0.
\]

(3.20)

The optimal instruments are then given by

\[
W_{t}^{\top} = \Omega_{t} E_{t-1} \left[ \frac{d g_{t}(w_{t}; f_{t}, \theta)}{d \theta^{\top}} \right],
\]

(3.21)
with $\Omega_t = E_{t-1}[g_t(w_t; f_t, \theta) g_t^\top(w_t; f_t, \theta)]^\top$. Note that $f_t$ depends on $\theta$ through the GaMM dynamics in eq. (3.11). We make this explicit by writing $f_t = f_t(w_{t-1}; f_{t-1}, \theta)$, to obtain

$$\frac{d g_t(w_t; f_t, \theta)}{d \theta^\top} = \frac{\partial g_t(w_t; f_t, \theta)}{\partial \theta^\top} + \frac{\partial g_t(w_t; f_t, \theta)}{\partial f_t^\top} \frac{d f_t(w_{t-1}; f_{t-1}, \theta)}{d \theta^\top},$$

such that

$$W_t^\top = \Omega_t E_{t-1} \left[ \frac{\partial g_t(w_t; f_t, \theta)}{\partial \theta^\top} \right] + \Omega_t G_t \frac{d f_t(w_{t-1}; f_{t-1}, \theta)}{d \theta^\top}, \quad (3.22)$$

with $G_t = G_t^0$ as defined in eq. (3.7). Let $(\omega^\top, \text{vec } [B]^\top, \text{vec } [A]^\top)^\top$ constitute the lower part of $\theta$. Then the last derivative in eq. (3.22) follows the recursion

$$\frac{d f_{t+1}(w_{t+1}; f_{t+1}, \theta)}{d \theta^\top} = \frac{d (I - B) \omega}{d \theta^\top} + (f_t(w_{t-1}; f_{t-1}, \theta)^\top \otimes I) \frac{d \text{vec } [B]}{d \theta^\top} + (s_t(w_t; f_t, \theta)^\top \otimes I) \frac{d \text{vec } [A]}{d \theta^\top} + B \frac{d f_t(w_{t-1}; f_{t-1}, \theta)}{d \theta^\top} + A \frac{\partial s_t(w_t; f_t, \theta)}{\partial f_t^\top} \frac{d f_t(w_{t-1}; f_{t-1}, \theta)}{d \theta^\top} +$$

$$= \begin{pmatrix} 0 
 0 
 I - B 
 f_t^\top \otimes I 
 s_t^\top \otimes I 
 \end{pmatrix} + B \frac{d f_t}{d \theta^\top} + A \frac{\partial s_t}{\partial \theta^\top} + A \frac{\partial s_t}{\partial f_t} \frac{d f_t}{d \theta^\top}. \quad (3.23)$$

In empirical work, we expect GaMM dynamics to work well if time-variation in parameters is persistent and relatively slow. This implies $A$ will often be estimated close to zero and $B$ close to the identity matrix. Consequently, the last two terms in eq. (3.23) are typically small relative to the first two terms. As long as $I - B$ is non-singular, the first term in eq. (3.23) is equivalent to the matrix of instruments proposed earlier in eq. (3.19). The instruments in eq. (3.19) therefore account for the dominant sources of variation in the optimal instruments in eq. (3.22), while avoiding the use of a second recursion for the derivatives in eq. (3.23) during estimation. The main difference between eq. (3.19) and eq. (3.23) is the presence of $B \frac{d f_t}{d \theta^\top}$, which induces additional smoothing of the instruments in $W_t$ as proposed in eq. (3.19).

The instruments in eq. (3.19) have computational advantages over optimal instruments. Components of eq. (3.22) can be hard to compute for specific models, particularly
given the need to compute conditional expectations. The instruments in eq. (3.19) simplify the computational challenges because the time-varying parameter $f_t$ as well as its derivatives are always time $(t - 1)$-measurable by construction. We demonstrate the usefulness of the instruments in eq. (3.19) for different models in Section 3.3 using both simulated and empirical data.

**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 $\hat{\theta}$, under standard high-level regularity conditions. We make the following assumptions.

**Assumption 3.1.** Let $\Theta$ denote the compact parameter space with interior $\text{int}(\Theta)$. The GMM objective function in eq. (3.18) is almost surely twice continuously differentiable and has a unique minimum at $\theta_0 \in \text{int}(\Theta)$.

**Assumption 3.2.** (i) A central limit theorem holds for $T^{1/2} \bar{g}_T$, i.e., $T^{1/2} \bar{g}_T \overset{d}{\to} N(0, \bar{V})$, with $\bar{g}_T$ as defined in eq. (3.18), and $T \cdot \text{Var}[\bar{g}_T] \to \bar{V}$. (ii) Let $\bar{G}_t = d(W_t g_t)/d\theta$, then $E[\bar{G}_t]$ exists and a uniform law of large numbers result holds for $T^{-1} \sum_{t=1}^T \bar{G}_t \to \bar{G}$. (iii) The matrix $\bar{\Omega}_T$ converges in probability to a positive semi-definite matrix $\bar{\Omega}$. (iv) The matrix $\bar{H} = \bar{G}^T \bar{\Omega} \bar{G}$ 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 3.2.** Under Assumptions 3.1 and 3.2, the GMM estimator for the static parameters $\theta$, including the parameters $\omega$, $A$ and $B$ specifying the GaMM dynamics, is consistent and asymptotically normal, i.e.,

$$T^{1/2} \left( \hat{\theta} - \theta_0 \right) \overset{d}{\to} N \left( 0, H^{-1} \bar{D} H^{-1} \right), \quad \bar{D} = \bar{G}^T \bar{\Omega} \bar{V} \bar{\Omega} \bar{G}. \quad (3.24)$$
The efficient weighting matrix is \( \Omega = V^{-1} \), in which case the asymptotic covariance matrix collapses to \((G^T \Omega G)^{-1}\).

Though the differentiability conditions in Assumption 3.1 are typically straightforward to verify and the uniqueness of the optimum is typically imposed by assumption, Assumption 3.2 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 \( \{W_t\} \) and \( \{g_t\} \), which in turn depend on the stationarity and ergodicity of the underlying data, and that of the time-varying parameter \( f_t \) and of its derivatives with respect to \( \theta \).

A key difference here is the non-linearity of the GaMM transition equation describing the dynamics of \( f_t \) as a function of the data, i.e., the filtering equations. Even if the data \( w_t \) are stationary, ergodic, near epoch dependent, and have the appropriate moments, these properties are not necessarily inherited by the ‘filtered’ time-varying parameter \( f_t \) and its derivatives. In order to obtain stationarity and ergodicity results for \( f_t \), the transition equation (3.2) 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 3.2. For particular models or model classes, lower level conditions are obtainable; see, e.g., Straumann and Mikosch (2006) for generalised autoregressive conditional heteroscedasticity models, or Blasques, Koopman, and Lucas (2014) for generalised autoregressive score models.

**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 \( f_t \).
The GaMM methodology can solve these problems by introducing a penalised 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, Commenges, and Joly, 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 penalised 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 penalised procedure then weights a sequence of smaller departures as much as it does one single large departure. The penalised criterion function takes the form

$$
\min_{\theta \in \Theta} \bar{g}^T \Omega_T \bar{g}_T + \frac{\lambda_T}{T} \sum_{t=1}^{T} \iota^T |s_t|, 
$$

(3.25)

where $\iota$ denotes a vector of ones, and $\lambda_T \geq 0$ is a smoothing parameter. The impact of the penalty term $T^{-1} \lambda_T \sum_{t=1}^{T} \iota^T |s_t|$ is twofold. First, the penalty on $|s_t|$ favours 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 $s_t$, the penalty favours 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 $E_t[\iota^T |s_t|] < \infty$, and furthermore that $T \cdot \lambda_T \to 0$. 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 $T \cdot \tilde{g}_T^{\top} \Omega_T \tilde{g}_T$ converges to a non-degenerate random variable under Assumptions 3.1 and 3.2. The above assumptions on $E_{t-1}[\epsilon^{\top} | s_t]$ and $\lambda_T$ then ensure that the penalty term in eq. (3.25) becomes negligible asymptotically compared to the first term. Further assumptions need to be developed to ensure consistency and asymptotic normality of penalised GaMM estimation. We leave this for future work.

3.3. 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 modelling frameworks, such as for example the generalised autoregressive score framework of Creal, Koopman, and Lucas (2013). An additional illustration of GaMM for time-varying linear regression models with endogeneity problems is provided in Appendix 3.D.

Stable distributions

Model

The use of $\alpha$-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, Renault, and Veredas (2011), and can be particularly convenient for modelling highly erratic data such as changes in electricity or energy prices.

A challenging aspect of $\alpha$-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 generalised autoregressive
score model as in Creal, Koopman, and Lucas (2013). Similarly, using standard volatility models such as generalised autoregressive conditional heteroscedasticity (GARCH) models does not seem to be appropriate either: as the stable distribution allows for the realisation 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 $\alpha$-stable distributions is fully characterised by four parameters: the stability parameter $\alpha \in (0, 2]$ (which corresponds with a tail index of $\alpha^{-1}$), the skewness parameter $\beta \in [-1, 1]$, the scale parameter $\sigma \in \mathbb{R}^+$, and the location parameter $\mu \in \mathbb{R}$. 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 $\alpha$-stable distributions are not known. Although the densities are not known in closed-form, the characteristic function of $\alpha$-stable distributions is (i) readily available and has a simple and intuitive form. The $\alpha$-stable model with time-varying scale provides a good example of how GaMM dynamics can be operationalised 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:

$$\log c(u; \alpha, \beta, \sigma_t, \mu) = \log E_t \left[ e^{iuX_t} \right]$$

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

Given eq. (3.2), we can therefore formulate the conditional moment condition

$$E_{t-1} [g_i^c] = 0, \quad g_i^c = \exp(iu_iX_t) - c(u_i; \alpha, \beta, \sigma_t, \mu), \quad i = 1, \ldots, K,$$
Example applications

where \( m \) denotes the number of grid points \( u_1, \ldots, u_K \). 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 \( \alpha, \beta, \sigma_t, \) and \( \mu \); for example, for large values of \( \sigma_t \) the grid should be much denser around zero than for smaller values of \( \sigma_t \); see also the discussion in Carrasco and Florens (2000).

For a given grid \( \{u_1, \ldots, u_K\} \), the moment conditions are constructed by stacking the real and imaginary parts of \( g^c_t \) for different values of \( u_i \),

\[
\mathbf{g}^T_t = (\text{Re}(\mathbf{g}^c_t)^T, \text{Im}(\mathbf{g}^c_t)^T).
\]  

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

Given eqs. (3.2) to (3.3), the matrix \( \mathbf{G}_t \) from eq. (3.7) is

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

(3.5)

where \( \mathbf{c}_t \equiv c(u; \alpha, \beta, \sigma_t, \mu) \). 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 3.2 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 Appendix 3.B for details about the sampling procedure. We estimate the static parameters $\theta^T = (\alpha, \beta, \mu, \omega, B, A)$ as described in Section 3.2 using GaMM(1,1) dynamics. We guarantee positivity of the estimated scale $\sigma_t$ by defining $f_t = \log \sigma_t$.

Results based on 10,000 replications are shown in Panels A–D of Figure 3.1. 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 $\alpha$-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 $\sigma_t$ 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 summarise the bias and root mean squared error (RMSE) of the time-varying scale into a single summary statistic,

$$\text{Average Bias} = \frac{1}{T} \sum_{t=1}^{T} (\hat{\sigma}_t - \sigma_{t,0}), \quad \text{Average RMSE} = \left[ \frac{1}{T} \sum_{t=1}^{T} (\hat{\sigma}_t - \sigma_{t,0})^2 \right]^{\frac{1}{2}}. \quad (3.6)$$

We also compute the bias and RMSE for the static parameters $\alpha$, $\beta$, and $\mu$. Box plots for the bias and RMSE across all simulations are presented to the right of each of the different panels of Figure 3.1. GaMM(1,1) overall produces unbiased estimates of the static parameters $\alpha$, $\beta$, and $\mu$. These parameters are typically also estimated with a low RMSE. The estimated path for the scale parameter also appears to be unbiased, though
**Figure 3.1**

**Estimating Time-Varying Scale of a Stable Distribution**

This figure contains summary figures for a series of simulations in which returns are drawn from a stable distribution. We let the scale parameter, $\sigma$, 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 $\sigma$, 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.

Panel A: Sigma has structural breaks; $\alpha_0 = 1.5$, $\beta_0 = -0.5$, and $\mu_0 = -0.5$

Panel B: Sigma is threshold sinusoidal (high periodicity); $\alpha_0 = 1.5$, $\beta_0 = -0.5$, and $\mu_0 = -0.5$

Panel C: Sigma follows an AR(1) process; $\alpha_0 = 1.5$, $\beta_0 = -0.5$, and $\mu_0 = -0.5$
it tends to oscillate around the true value for any given replication. This results in a somewhat higher RMSE statistic. The oscillating behaviour 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.

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 3.2. We provide two benchmark estimates, namely a GJR-GARCH(1,1,1) model of Glosten, Jagannathan, and Runkle (1993a), and the Student’s $t$-GAS(1,1) model of Creal, Koopman, and Lucas (2011) and Harvey (2013).

Overall, the estimated paths of $\sigma_t$ 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 $t$
Example applications

Figure 3.2

S&P 500 Returns as Draws From a Stable Distribution

This figure contains estimated time-varying scale of daily returns for 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).

Panel A: Stable distribution fitted with GaMM(1,1)

Panel B: t-GAS(1,1)

Panel C: GJR-GARCH(1,1,1)

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 realisation 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 $\sigma_t$ 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 $\sigma_t$ 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.

**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:

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

(3.7)
where $R_{t+1}^g$ denotes the vector of gross asset returns, $\beta$ is a subjective discount factor, and $\gamma$ 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 (3.7) imply a stochastic discount factor $M_t = \beta (C_{t+1}/C_t)^{-\gamma}$. Empirical estimates of the risk aversion parameter $\gamma$ are sensitive to the particular sample period, starting values, and instruments employed. Results of many studies suggest that time series estimates of $\gamma$ are typically too high compared to risk aversion estimates obtained from experimental data and that the simple model in eq. (3.7) 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 $\gamma$ and $\beta$.

Poor performance of the standard model in eq. (3.7) 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 quantity. Furthermore, stability of deep parameters in the simple structural model in eq. (3.7) is already questioned by Ghysels and Hall (1990). Ghysels and Hall introduce a structural break test for $\gamma$, 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 eq. (3.7). The relative risk aversion, $\gamma_t$, is allowed vary in time and we filter it out by endowing the

\footnote{Allowing for habit formation implies that relative risk aversion depends on $\gamma$ and on the current values of consumption and consumption habit.}
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 $\gamma_t$
to remain fixed forever when making their decision at time $t$. This results in the Euler
equation:

$$E_t \left[ \beta \left( \frac{C_{t+1}}{C_t} \right)^{-\gamma_t} R^r_{t+1} \right] = 1.$$  (3.8)

Using eq. (3.8), we can directly employ the GaMM framework by taking $f_t = \gamma_t$.

One complication in our current context concerns the estimation of $G_t$, which in this
case depends both on the consumption growth data, the return data, and on the time
varying parameter $f_t$. 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 realisations from a distribution
characterised by a different value of $\gamma_t$. Still, if $f_t$ varies sufficiently slowly, observations
in the recent past can be informative about the curvature $G_t$ of the moment conditions
now. We therefore estimate $G_t$ as an exponentially weighted moving average

$$G_t = \lambda G_{t-1} + (1 - \lambda) \frac{\partial g_t(w_t; f_t, \theta)}{\partial f_t},$$  (3.9)

where we choose $\lambda$ in the range (0.98, 1.0). If $f_t$ varies sufficiently slowly, such an exponenti-
ally weighted moving average estimates the local curvature of the score accurately
enough to provide an adequate form of scaling in the GaMM transition dynamics from
eq. (3.9). This holds even though $\hat{G}_t$ may not be a consistent estimate of $G_t$. See
Appendix 3.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, Heaton, and Li (2008) or Savov (2011). Furthermore, we endow
Figure 3.3

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 \( \gamma_t \). 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% empirical confidence bands constructed based on the simulation results.

Panel A: risk aversion parameter exhibits a structural break in the middle of the sample

Panel B: risk aversion parameter follows an AR(1) process

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:

\[ \Delta c_t = \mu_c + \varepsilon_{ct} \quad \varepsilon_{ct} \sim N \left( 0, \sigma^2 \right) \]  

(3.10)
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 3.3. Note that the classical full sample GMM estimates of $\gamma$ 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 $\gamma_t$ observations actually lie below the full sample GMM estimate. If there is time-variation in $\gamma_t$, the full sample GMM estimates are thus severely biased towards the high-end realisations 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.

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 $\gamma_t$.

**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 the U.S. Bureau of Economic Analysis. Apart from the six Fama-French portfolios we obtain all remaining data from the FRED database.\(^2\) 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

\(^2\)The 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).
Figure 3.4

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 \( \gamma_t \). 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 labelled separately.

The second step only. Standard errors for parameter \( \theta \) 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 \( se_{(p)}^{(p)} \). 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 \( \gamma \) and \( \beta \). The discount rate is estimated at \( \hat{\beta} = 1.42 \) with
\( \text{se}_{\beta}^{(7)} = 0.14 \) while the curvature parameter is \( \hat{\gamma} = 133.06 \) with \( \text{se}_{\gamma}^{(7)} = 38.92 \). 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 \( \beta \) 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 upwards.\(^3\)

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 \( \text{se}_{\beta}^{(4)} = 0.001 \). Figure 3.4 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 \( \bar{\hat{\gamma}}_t = 1.42 \) (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 \( \text{se}_{\omega}^{(4)} = 0.75 \), \( \hat{B} = 1.00 \) with \( \text{se}_{B}^{(4)} = 0.006 \), and finally \( \hat{A} = 0.02 \) with \( \text{se}_{A}^{(4)} = 0.005 \).

There are two components in the time-varying risk aversion parameter \( \gamma_t \): 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 \( \gamma_t \). After the recession ends, we observe risk aversion returning to its long-term path. This

---

\(^3\)It is worth noting what impact do \( \beta \) and \( \gamma \) have on moment conditions in a neighbourhood of time \( t \). \textit{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.
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 behaviour in the loss domain and risk aversion in the gain domain.

More interesting than the short-term cyclical behaviour 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 realised 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.

### 3.4. Example results for penalised GaMM

To illustrate the usefulness of the penalised version of GaMM from Section 3.2, we consider a simple linear regression model with an endogenous regressor $x_t$ and an instrument $z_t$,

\[
y_t = \beta_{0,t} + \beta_{1,t}x_t + \varepsilon_{yt}, \quad \varepsilon_{yt} \sim N\left(0, \sigma_{\varepsilon_y}^2\right), \quad E[\varepsilon_{yt}\varepsilon_{xt}] = \rho, \quad (3.1)
\]

\[
x_t = \pi_{0,x} + \pi_{1,x}z_t + \varepsilon_{xt}, \quad \varepsilon_{xt} \sim N\left(0, \sigma_{\varepsilon_x}^2\right), \quad z_t \sim N\left(\mu_z, \sigma_{\varepsilon_z}^2\right), \quad (3.2)
\]

\[
\beta_{i,t+1} = \pi_{\beta}\beta_{i,t} + \eta_{i,t}, \quad \eta_{i,t} \sim N\left(0, \sigma_{\eta_i}^2\right), \quad i = 0, 1. \quad (3.3)
\]
If $\rho \neq 0$ and $\sigma_{e_x}^2 > 0$, we have a standard endogenous regressor problem. Lower values of $\sigma_{e_x}^2$ 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 $\mathbf{\beta}_t^\top = (\beta_{0,t}, \beta_{1,t})$, and let $\mathbf{\Pi}_x$ be such that $\mathbf{x}_t^\top = \mathbf{z}_t^\top \mathbf{\Pi}_x + (0, \varepsilon_{xt})$. In Appendix 3.D, 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 $\mathbf{\beta}_t$. Given the availability of an instrument variable $z_t$, the obvious way to estimate $\mathbf{\beta}_t$ is via the conditional moment condition

$$E_{t-1} [z_t (y_t - \mathbf{x}_t^\top \mathbf{\beta}_t)] = 0. \tag{3.4}$$

The complication here is that the parameters $\mathbf{\beta}_t$ are time-varying. The moment condition in eq. (3.4), however, lends itself directly to the GaMM framework by setting $\mathbf{f}_t = \mathbf{\beta}_t$.

We consider a setting where $\beta_{1,t}$ has a structural break at $t = t^\star$. The result without the use of the penalty is presented in Panel A of Figure 3.1. A perfect fit to the structural break would require $s_{t^\star} \neq 0$, while $s_t = 0$ for all $t \neq t^\star$. The GaMM criterion function in eq. (3.18) does not incorporate this prior knowledge about the properties of $\{s_t\}$. Panel A of Figure 3.1 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 minimised. 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 $\bar{g}_n$ in eq. (3.18) 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 3.1. 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.
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 biassed 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.

Panel A: Path estimated without the penalised criterion function

Panel B: Path estimated with the penalised criterion function

To obtain further insight, we run a Monte-Carlo simulation study for a variety of parameter settings using the penalised criterion function in eq. (3.25) with $\lambda_T = \{10, 5, 1, 0.6, 0.2, 0\}$, a variety 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 Appendix 3.D. 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 $\hat{\beta}_i$ in Figure 3.2. The figure contains the average bias and RMSE for the constant (top) and for the slope parameter (bottom). Within
Figure 3.2

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.

Panel A: Bias (left) and RMSE (right) for the exogenous variable

Panel A: Bias (left) and RMSE (right) for the endogenous variable

each panel, we see 6 boxplots for the different values of the smoothing parameter $\lambda$. 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 substantial. The GaMM methods perform much better in this respect. Biases are visible for very high values of $\lambda = 10$, but quickly vanish for smaller values of $\lambda$. Even for $\lambda = 10$, the average bias is smaller than for the state space models estimated with Kalman Filter, which does not account at all for the en-
dogeneity problem. The penalised estimates also come with substantially lower RMSEs though this difference becomes smaller if the endogeneity problem is increased; see the additional results in Appendix 3.D.

We conclude that already small values of the penalty parameter $\lambda_T$ 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 $\theta$ is first estimated using the penalised GaMM criterion, e.g., with $\lambda_T = 1$. The resulting estimates can then be used as starting values for $\theta$ 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.

3.5. Conclusions

In this chapter we proposed a new approach for modelling time-varying parameters in linear and non-linear econometric models identified through moment conditions. We call the approach the Generalised 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 generalised autoregressive score approach of Creal, Koopman, and Lucas (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 penalised version of the GaMM criterion function. We showed how penalisation 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 chapter 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.
3.A: Proofs

Proof of Proposition 3.1. Define

\[
\tilde{g}_t(f) = \int_{\mathcal{W}} g_t(w; f, \theta) \, dF_w(w), \tag{3.A.1}
\]

\[
\tilde{G}_t(f) = \frac{\partial}{\partial f} \int_{\mathcal{W}} g_t(w; f, \theta) \, dF_w(w), \tag{3.A.2}
\]

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

\[
\tilde{g}_t(f_t)\mathbf{\Omega}_{t-1}\tilde{g}_t(f_t) - \tilde{g}_t(f_{t+1})\mathbf{\Omega}_{t-1}\tilde{g}_t(f_{t+1})
\]

\[
= -2(f_{t+1} - f_t)\mathbf{\Omega}_{t-1}\tilde{G}_t(f_{t+1})\mathbf{\Omega}_{t-1}\tilde{g}_t(f_{t+1})
\]

\[
= 2g_t(w_t; f_{t+1}, \theta)\mathbf{\Omega}_{t-1}\mathbf{G}_t\mathbf{V}^\top \mathbf{A}^\top \tilde{G}_t(f_{t+1})\mathbf{\Omega}_{t-1}\tilde{g}_t(f_{t+1}),
\]

\[
= 2a \int_{\mathcal{W}} g_t(w_t; f_{t+1}, \theta)\mathbf{\Omega}_{t-1} \cdot \mathbf{G}_t\mathbf{V}^\top \tilde{G}_t(f_{t+1})\mathbf{\Omega}_{t-1} g_t(w; f_{t+1}, \theta) \, dF_w(w), \tag{3.A.3}
\]

where \( f_{t+1} \) is a point between \( f_t \) and \( f_{t+1} \). For \( \mathcal{W} = \{w_t\} \), the integrand is strictly positive. Given the assumed continuity of \( g_t(w; f, \theta) \) and \( G_t(w; f, \theta) \) in \( w \) and \( f \), and the subsequent continuity of \( f_{t+1} \) as a function of \( w_t \), the result follows immediately for a small enough ball \( \mathcal{W} \) around \( \{w_t\} \). \( \square \)

3.B: Sampling from stable distributions

To simulate draws from a stable distribution, we follow the generalised 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 \( v = \pi \left(v_u - \frac{1}{2}\right) \) and \( w = -\log w_u \). For given \( \alpha, \beta, \sigma, \) and \( \mu \), we obtain a random vector...
\[ Z \sim S(\alpha, \beta, \sigma, \mu) \] from the transformations

\[ B = \frac{1}{\alpha} \tan \left( \beta \tan \left( \frac{\pi}{2} \right) \right), \quad (3.B.1) \]

\[ S = \left[ 1 + \beta^2 \tan^2 \left( \frac{1}{2} \alpha \pi \right) \right]^{1/2}, \quad (3.B.2) \]

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

for \( \alpha \neq 1 \), and

\[ B = \frac{\pi}{2} + \beta v, \quad (3.B.4) \]

\[ Z = \frac{2}{\pi} B \sigma \tan (v) - \frac{2}{\pi} \sigma \beta \log \left( \frac{\pi w \cos (v)}{B} \right) + \mu + \frac{2}{\pi} \beta \sigma \log (\sigma), \quad (3.B.5) \]

for \( \alpha \equiv 1 \).

### 3.C: Estimating procedure for the derivative matrix

Two considerations need to be made when it is necessary to estimate \( G_t \) and \( \theta \) jointly; see the example in Section 3.3. First of all, since in these models \( G_t \) depends on the true values of \( \theta \) and \( f_t \) at time \( t \); if \( f_{0,t=i} \neq f_{0,t=j} \) then \( G_{t=i} \neq G_{t=j} \) (parameter-value inconsistency). It follows that estimates of \( G_t \) need to take into account in which neighbourhood of the parameter space \( \mathcal{F} \) the filtered parameter \( \hat{f}_{t-1} \) is located. More importantly, especially at the beginning of the estimation procedure for \( \theta \) it may very well be that while \( \hat{f}_{t=i} = \hat{f}_{t=j} \), it is still true that \( f_{0,t=i} \neq f_{0,t=j} \) (time-period inconsistency) and thus using the same value of \( G_t \) for \( t = i \) and \( t = j \) should be avoided. If \( \{f_t\} \) is relatively slowly-varying, we can account for both issues by estimating \( G_t \) as an exponentially weighted moving average. Let for some \( \lambda \):

\[ G_t^{(i)} = \lambda \hat{G}_t^{(i)} = \lambda \hat{G}_t^{(i)}_{t-1} + (1 - \lambda) \frac{\partial g_t(w_t; f_t, \theta)}{\partial f_t}, \quad (3.C.1) \]

where the superscript \( (i) \) denotes the \( i \)-th iteration of the optimising algorithm and the arrow denotes direction in the time domain in which we apply the EWMA. In practice we find that \( \lambda \) can be chosen to be in the \((0.98, 1.0)\) interval. For \( i = 0 \), we initialise the
recursion in eq. (3.C.1) by setting $\mathcal{G}_{t=0}^{(0)} = I$. In order to update this initial guess for $\mathcal{G}_{t=0}$ we propose to apply the EWMA in reverse, i.e. let:

$$\mathcal{G}_{t}^{(i)} = \lambda \mathcal{G}_{t+1}^{(i)} + (1 - \lambda) \frac{\partial g_t(w_t; f_t, \theta)}{\partial f_t},$$

(3.C.2)

$$\mathcal{G}_{t=0}^{(i+1)} = \mathcal{G}_{t=0}^{(i)}.$$  

(3.C.3)

Note that estimating $\mathcal{G}_t$ 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 $\mathcal{G}_t$. This is because given the estimate of $\hat{\theta}$ form the first-step, the path $\{\hat{f}_t\}$ 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 $\mathcal{G}^{(i)}_t$ based on previously obtained $\{\mathcal{G}^{(i-1)}_t\}$ and $\{f^{(i-1)}_t\}$. For $i = 0$, we initialise the kernel density smoother with $\{\mathcal{G}_t\}$ and $\{f_t\}$ obtained at the first stage of the estimation.

3.D: 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.

$$y_t = x_t^T \beta_t + \varepsilon_t$$

Given the availability of an instrument variable $z_t$, the obvious way to estimate $\beta_t$ is via the conditional moment condition

$$E_{t-1} [z_t (y_t - x_t^T \beta_t)] = 0.$$  

(3.D.1)
The complication here is that the parameters $\beta_t$ are time-varying. The moment condition in eq. (3.D.1), however, lends itself directly to the GaMM framework by setting $f_t = \beta_t$.

In contrast to the setting with stable distributions in Section 3.3 where $G_t$ did not depend on the data at time $t$, $G_t = -E_{t-1}[z_t|x_t^T]$ 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 $G_t$ in a straightforward way by noting from eq. (3.D.3) that $E_{t-1}[z_t|x_t^T] = z_t^T \Pi_x$. Define $X$ and $Z$ as the $T \times 2$ matrices with $t$th row equal to $x_t^T$ and $z_t^T$, respectively. Given the exogeneity of the instruments $z_t$, we can estimate $G_t$ in the standard way by $\hat{G}_t^T = z_t z_t^T \hat{\Pi}_x = z_t z_t^T (Z^T Z)^{-1} Z^T X$. The fact that we can use all observations to estimate $\hat{\Pi}_x$ rather than only the observations up to time $t-1$ follows from the exogeneity assumption for $z_t$ and the assumption that $\Pi_x$ is static.

**Simulation results**

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

$$y_t = \beta_{0,t} + \beta_{1,t} x_t + \varepsilon_{yt}, \quad \varepsilon_{yt} \sim N\left(0, \sigma^2_{\varepsilon_y}\right), \quad E[\varepsilon_{yt}\varepsilon_{xt}] = \rho, \quad (3.D.2)$$

$$x_t = \pi_{0,x} + \pi_{1,x} z_t + \varepsilon_{xt}, \quad \varepsilon_{xt} \sim N\left(0, \sigma^2_{\varepsilon_x}\right), \quad z_t \sim N\left(\mu_z, \sigma^2_z\right), \quad (3.D.3)$$

$$\beta_{i,t+1} = \pi_{\beta} \beta_{i,t} + \eta_{i,t}, \quad \eta_{i,t} \sim N\left(0, \sigma^2_{\eta_i}\right), \quad i = 0, 1. \quad (3.D.4)$$

If $\rho \neq 0$ and $\sigma^2_{\varepsilon_x} > 0$, we have a standard endogenous regressor problem. For lower values of $\sigma^2_{\varepsilon_x}$, the size of the bias when estimating $\beta_{1,t}$ using standard least-squares based methods, is larger. Define $x_t^T = (1, x_t)$, $z_t^T = (1, z_t)$, and $\beta_t^T = (\beta_{0,t}, \beta_{1,t})$, and let $\Pi_x$ be such that $x_t^T = z_t^T \Pi_x + (0, \varepsilon_{xt})$. As a benchmark model we use an ordinary state space model consisting of eq. (3.D.2) and eq. (3.D.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.
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:

\[ y_t = \beta_{0,t} + x_t \beta_{1,t} + \varepsilon_t \]
\[ x_t = 0.5 z_t + \phi_t \]
\[ \beta_{i,t+1} = 0.98 \beta_{i,t} + \eta_{i,t} \]

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

We consider moderate \((\sigma_{\varepsilon}^2 = 4.0)\) and high \((\sigma_{\varepsilon}^2 = 0.5)\) degree of the endogeneity problem. For both the moderate and the high endogeneity bias case, we consider different degrees of time-variation in \(\beta_t\), 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.

Panel A: Average bias of coefficients for the exogenous variable

Panel B: Average bias of coefficients for the endogenous variable

Our data generating process uses eqs. (3.D.2) to (3.D.4). The parameters we selected
are $\rho = 0.5$, $\sigma^2_{\epsilon_x} = 1$, $\pi_x = 0.5$, $\mu_z = 0$, $\sigma^2_z = 1$, and $\pi^\top_\beta = [0.98, 0.98]^\top$. We vary $\sigma_{\epsilon_x}$ and $\sigma_\eta$ to study the effect of different magnitudes of the endogeneity problem and of the time-variation of the coefficients. We use $\sigma^2_{\epsilon_x} = \{0.5, 4\}$ and $\sigma^2_\eta = \{0.01, 0.25, 0.75\}$, where lower values of $\sigma_{\epsilon_x}$ 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 $\beta_t$ than in eq. (3.6.4), including structural breaks and slowly varying sinusoid waves.

The results of 10,000 replications are presented in Figure 3.D.1 and Figure 3.D.2. The figures summarise a substantial amount of information. In the top panel (Panel A), we summarise 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 eq. (3.6). In the left-half of the plot, the endogeneity problem is moderate with a high value of $\sigma^2_{\epsilon_x} = 4.0$. In the right-hand half of the graph, the endogeneity problem is more severe with $\sigma^2_{\epsilon_x} = 0.5$. This becomes particularly 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 $\beta_t$, from low ($\sigma^2_\eta = 0.01$) to high ($\sigma^2_\eta = 0.75$). 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 $\beta_t$, and severity of the endogeneity problem. 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 3.D.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.
Figure 3.D.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:

\[
\begin{align*}
    y_t &= \beta_{0,t} + x_t \beta_{1,t} + \varepsilon_t \\
    x_t &= 0.5 z_t + \phi_t \\
    \beta_{i,t+1} &= 0.98 \beta_{i,t} + \eta_{i,t}
\end{align*}
\]

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

We consider moderate (\(\sigma_\varepsilon^2 = 4.0\)) and high (\(\sigma_\varepsilon^2 = 0.5\)) degree of the endogeneity problem. For both the moderate and the high endogeneity bias case, we consider different degrees of time-variation in \(\beta_t\), 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.

Panel A: Average RMSE of coefficients for the exogenous variable

Panel B: Average RMSE of coefficients for the endogenous variable

In Panel A, we see that in most cases both the GaMM and the Kalman Filter esti-
mates 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 3.D.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 biassed (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 $\sigma^2_{\varepsilon x} = 0.5$ than for $\sigma^2_{\varepsilon x} = 4.0$. 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 3.D.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 behaviour 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.

Summarising, 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 penalised criterion function. Partial results are
**Figure 3.D.3**

**Impact of Penalizing the Criterion Function on Bias**

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 Figure 3.D.1 and Figure 3.D.2. 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 GaMM without the penalised criterion function was initialised with estimates obtained after running GaMM with penalty weight of 1.

Panel A: Average bias of coefficients for the exogenous variable

Panel B: Average bias of coefficients for the endogenous variable

discussed in Section 3.4. Figure 3.D.3 and Figure 3.D.4 contain results for other sample sizes, and different values of both $\sigma_{e_x}$ and $\sigma_{\eta}$. 
Figure 3.D.4
Impact of Penalizing the Criterion Function on RMSE

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 Figure 3.D.1 and Figure 3.D.2. 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 GaMM without the penalised criterion function was initialised with estimates obtained after running GaMM with penalty weight of 1.

Panel A: Average RMSE of coefficients for the exogenous variable

Panel B: Average RMSE of coefficients for the endogenous variable