Tractable estimation and smoothing of highly non-linear dynamic state-space models.

Tom D. Holden[1]*, School of Economics, University of Surrey*

Abstract: We present an algorithm for tractably estimating non-linear dynamic models, such as DSGE models with occasionally binding constraints, or stochastic volatility models. The algorithm presents an extended skew-t, augmented-state, version of the Cubature Kalman Filter of Arasaratnam and Haykin (2009) with dynamic state space reduction, to give adequate speed, and to ensure that it can handle the large state spaces generated, for example, by pruned perturbation solutions to medium-scale DSGE models. The use of an extended skew-t approximation to the state's distribution allows the filter to also track the distribution's third and fourth moments. We extend the base algorithm to allow for alternative cubature procedures to further improve the tracking of non-linearities. We illustrate that the method can solve some of the identification problems that plague linearized DSGE models, and show that the method can readily handle the estimation of stochastic volatility models with time varying correlation between level and volatility innovations. We go on to extend the algorithm to produce smoothed estimates of states, and we use this to assess which shocks caused the great recession in the model of Christiano, Motto, and Rostagno (2014).

Keywords: *non-linear, state-space, estimation, filtering, smoothing, cubature Kalman filter, occasionally binding constraints, zero lower bound, DSGE, stochastic volatility*

JEL Classification: *C13, C32, E3, E4, E5*

This version: 9 February 2017

The latest version of this paper may be downloaded from:
https://github.com/tholden/dynareOBC/raw/master/EstimationPaper.pdf

[1] Contact address: School of Economics, FASS, University of Surrey, Guildford, GU2 7XH, England
Contact email: thomas.holden@gmail.com Web-site: http://www.tholden.org/
Financial support provided to the author by the ESRC and the EC is greatly appreciated. Furthermore, the author gratefully acknowledges the use of the University of Surrey FASS cluster for the numerical computations performed in this paper.
The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement "Integrated Macro-Financial Modelling for Robust Policy Design" (MACFINROBODS, grant no. 612796).

1. Introduction

Traditionally, DSGE models have been linearized prior to solution and estimation. Assuming that the driving shocks are normally distributed, the resulting solution is a linear-Gaussian state space model that may be estimated without additional approximation error via the standard Kalman filter. Thanks to linearity, and the special properties of the normal distribution, this is remarkably tractable, and has enabled the estimation of rich macroeconomic models, as popularized by Smets and Wouters (2003). However, any small departure from linearity or Gaussianity means that tracking the distribution of the state without approximation error would require keeping track of the full distribution of the state, which is an infinite dimensional object. As a result, any non-linear filter will inevitably introduce additional approximation error.

One important departure from linearity is the presence of occasionally binding constraints (OBCs). Such constraints are ubiquitous in modern macroeconomic models, appearing, variously, in the zero lower bound in nominal interest rates, in the irreversibility of investment, and in borrowing and collateral constraints. If they are ignored during estimation, the resulting parameters can be severely biased, as lower bounded variables will tend to have higher mean in an accurate approximation than they do under a first order one. Furthermore, when occasionally binding constraints are ignored, there can be failures of identification, as some of the model's parameters may only alter behaviour when at the bound.

Models with occasionally binding constraints present a variety of difficulties. As established by Holden (2016a), they may possess multiple solutions in some states, and no solutions at all in others, even when a terminal condition is specified. For estimation, this means both that the likelihood may be minus infinity for some parameters, and that unless care is taken, the likelihood may be discontinuous in the parameters, invalidating standard inference. Holden (2016b) is the only current simulation algorithm for models with occasionally binding constraints that is always guaranteed to select the same solution in the presence of multiplicity, restoring continuity in the parameters, and hence enabling conventional statistical inference in models with occasionally binding constraints. Furthermore, Holden (2016b) establishes that computing solutions to models with occasionally binding constraints is a computationally difficult problem, in a sense which that paper makes precise.

Indeed, Holden (2016b) is the first algorithm for simulating such models that is guaranteed to complete in finite time. Since almost all non-linear estimation methods are based upon repeated simulation, this in turn implies that estimating models with occasionally binding constraints is likely to be particularly difficult, and essentially impossible in general if the Holden (2016b) simulation algorithm is not used.

Some progress has recently been made in estimating non-linear DSGE models using direct inversion of shocks (see e.g. Holden 2014; Guerrieri and Iacoviello 2015; Kollmann 2016), or using relatively standard particle filters (see e.g. Fernández-Villaverde and Rubio-Ramírez 2007; Fernández-Villaverde, Guerrón-Quintana, and Rubio-Ramírez 2015; Gust et al. 2016). The former approach is potentially quite computationally intensive in models with many shocks, as it requires the inversion of a large system of non-linear equations in each period for each likelihood evaluation. Additionally, this system of equations may have multiple solutions, and selecting a particular one will give an incorrect value for the likelihood. Finally, the direct inversion approach requires that there are no more shocks than observables, which precludes the inclusion of econometrically reasonable measurement errors on each equation.

The standard particle filter has the advantage that given an exact simulation procedure, it produces an estimate of the likelihood is unbiased (though noisy). This has been shown to be sufficient for correct Bayesian inference at least by Andrieu, Doucet, and Holenstein (2010). However, this advantage is less compelling for non-linear DSGE models, since their simulation is impossible without approximation, which itself introduces biases into the likelihood. Furthermore, these methods require too many simulation-steps to be computationally tractable in the presence of occasionally binding constraints, at least on a current desktop machine. In addition, one of the chief reasons macroeconomists were led towards Bayesian methods was due to the weak identification of many linearized DSGE models. Since using non-linear filters potentially solves the weak identification issue, we would ideally like to be able to return to maximum likelihood estimation, which is less econometrically divisive. However, due to the noise in the standard particle filter's estimate of the likelihood, numerically maximising the likelihood is almost impossible in this case. While techniques such as the smooth particle filter of Pitt (2002) may remedy this, they come at the cost of introducing substantial sampling noise into the final parameter estimates.

An alternative track of the literature has sought to exploit the properties of pruned perturbation solutions (Kim et al. 2008) to enable the use of the standard Kalman filter, despite the non-linearity, albeit with some additional approximation error. One prominent example of this approach is that of Kollmann (2013), who exploits the

existence of an augmented state-space in which the pruned perturbation solution is linear, though non-Gaussian. Another is that of Meyer-Gohde (2014), who exploits the ability to get a closed form first order approximation around the mean of an underlying pruned perturbation approximation. However, both of these methods are restricted to models which are everywhere differentiable, ruling out occasionally binding constraints.

In this paper, we take an intermediary route between these two extremes, based on the cubature Kalman filter of Arasaratnam and Haykin (2009) and the smoother extension of it of Arasaratnam and Haykin (2011). Like the standard particle filter approach, we will rely on approximating the distribution of the state via simulation at a collection of points. Indeed, the approach we take is sometimes referred to as a type of particle filter (see e.g. Adjemian et al. 2016). Like the alternative approach though, we will maintain a parameteric approximation to the state of the model, and, in the presence of OBCs, we use an underlying simulation algorithm from Holden (2016b) that exploits the pruned perturbation structure.

Relative to Arasaratnam and Haykin (2009), our approach expands and improves along several lines. Firstly, rather than approximating the distribution of the state by a Gaussian, our version approximates the state's distribution by an extended skew-t distribution, which enables us to capture the skewness and kurtosis that is generated by non-linear models.

Secondly, we present a version exploiting an augmented-state representation that both reduces the number of sampling procedures required per step from two to one, and which enables the use of the filter on models without additive shocks. While Wan and Van Der Merwe (2000) presented an augmented version of the related unscented Kalman filter for models with non-additive noise, and Li et al. (2009) presented an augmented-state version of the cubature Kalman filter for models with additive shocks, we present an augmented-state version of the cubature Kalman filter for non-additive models.

Thirdly, we present a technique that ensures the dimensionality of both this initial state, and subsequent states, are as low as possible, without introducing major inaccuracies. Given that pruned perturbation solutions produce very large state spaces, this is essential for medium-scale DSGE applications. Fourthly, we give an algorithm for the initialization of the state distribution. Fifthly, we introduce a choice of alternative cubature methods, which will produce better estimates at some additional computational cost. Sixthly, we discuss techniques to ensure consistency. Seventhly, we discuss techniques for the numerical maximisation of the likelihood, which needs considerable care given the potential multi-modality. Such difficulties in numerical maximisation were another key reason why the profession shifted towards

Bayesian methods, so addressing them is crucial if we are to argue for using maximum likelihood estimates. Finally, we discuss the computation of standard errors, given the non-normality of the actual surprise content of observations.

Our paper is structured as follows. In the following section, we present our estimation algorithm, and discuss assorted extensions and econometric issues. In section 3, we then test the algorithm's performance, and illustrate how non-linear estimation can address the weak identification of linearized DSGE models. Section 4 discusses smoothing and provides an application to the Christiano, Motto, and Rostagno (2014) model. Section 5 discusses our implementation of these algorithms in the author's DynareOBC add-on for Dynare (Adjemian et al. 2011), which is freely available from http://github.org/tholden/dynareOBC. Finally, section 6 concludes. All files needed for the replication of this paper's numerical results are included in the "Examples" directory of the aforementioned toolkit.[2]

2. Our algorithm for estimating non-linear models

In this section, we first describe the extended skew t-distribution, of which we will make heavy use. We then give the core algorithm, leaving unspecified for the time being how the required integrals may be numerically evaluated. We then go on to examine alternative methods of performing the required integration. The following sub-section discusses the computation of the initial distribution of the state. This is followed by material on the practical numerical maximisation of the likelihood, and the computation of standard errors.

2.1. The extended skew t-distribution

We say that Z has the extended skew t-distribution with location $\xi \in \mathbb{R}^n$, scale-matrix $\Omega \in \mathbb{R}^{n\times n}$, skew direction $\delta \in \mathbb{R}^n$, shape parameter $\tau \in \mathbb{R}$ and "degrees of freedom" parameter $\nu \in \mathbb{R}^+$, and write $Z \sim \mathrm{EST}(\xi, \Omega, \delta, \tau, \nu)$ if the density of Z at $z \in \mathbb{R}^n$ is given by:

$$f_{\mathrm{EST}_{\xi,\Omega,\delta,\tau,\nu}}(z) =$$
$$\left|\breve{\Omega}\right|^{-\frac{1}{2}} f_{T_{\nu,n}}\left(\breve{\Omega}^{-\frac{1}{2}}(z-\xi)\right)\frac{1}{F_{T_{\nu,1}}(\tau)}F_{T_{\nu+n,1}}\left(\frac{\delta'\breve{\Omega}^{-1}(z-\xi)+\tau}{\sqrt{1-\delta'\breve{\Omega}^{-1}\delta}}\sqrt{\frac{\nu+n}{\nu+(z-\xi)'\breve{\Omega}^{-1}(z-\xi)}}\right),$$

where:

$$\breve{\Omega} := \Omega + \delta\delta',$$

$f_{T_{\nu,n}}$ is the p.d.f. of a standard n-dimensional t-distribution, with "degrees of freedom" parameter ν, and where $F_{T_{\nu,n}}$ is the c.d.f. of the same distribution. This distribution

[2] These files may be viewed online at https://github.com/tholden/dynareOBC/tree/master/Examples.

was introduced by Arellano-Valle and Genton (2010) who used a different, though equivalent parameterisation.[3]

Proposition 2 of Arellano-Valle and Genton (2010) implies[4] that if $X_0 \sim \mathrm{T}(0,1,\nu)$ (i.e. X_0 is a draw from a standard t-distribution with degrees of freedom parameter ν) and

$$X_1 \sim \mathrm{T}(0,\Omega,\nu+1)$$

(i.e. X_1 is a draw from a t-distribution located at 0 with scale matrix Ω and degrees of freedom parameter $\nu+1$), with X_0 and X_1 independent, and if:

$$\overline{X}_0 \overset{d}{=} X_0 | X_0 + \tau > 0,$$

then:

$$\xi + X_1 \sqrt{\frac{\nu + \overline{X}_0^2}{\nu+1}} + \delta \overline{X}_0 \overset{d}{=} Z \sim \mathrm{EST}(\xi, \Omega, \delta, \tau, \nu).$$

Furthermore, by the results of Arellano-Valle and Genton (2010), the first two moments of Z are given by:

$$\mathbb{E}Z = \xi + \delta \mathbb{E}\overline{X}_0,$$
$$\operatorname{var} Z = \left(\frac{\nu + \mathbb{E}\overline{X}_0^2}{\nu-1}\right)\Omega + \operatorname{var}\overline{X}_0\, \delta\delta',$$

where closed form expressions for the moments of $\overline{X}_0$ are given in Arellano-Valle and Genton (2010).

We can also calculate a natural multivariate pseudo-median of Z, that we shall denote by λ. In particular, if we view Z as a function of the two random variables $\overline{X}_0$ and X_1, then a natural pseudo-median may be obtained by applying the same function to the median of $\overline{X}_0$ and the median of X_1. We take this definition since it will be particularly easy to calculate such a pseudo-median of the dynamic model, but we must note that potentially there may be other (less natural) representations of Z as functions of other shocks which might lead to differing pseudo-medians. Given this definition, and the fact that X_1 is elliptically symmetric, and hence median 0 by any reasonable definition, we have that $\lambda = \xi + \delta \operatorname{med}\overline{X}_0$ where med is the standard univariate median operator. Consequently, if $\mathbb{E}Z = \mu$ and $\operatorname{var} Z = \Sigma$, then:

$$\delta = \frac{\mu - \lambda}{\mathbb{E}\overline{X}_0 - \operatorname{med}\overline{X}_0},$$
$$\xi = \mu - \delta \mathbb{E}\overline{X}_0,$$
$$\Omega = \left[\!\!\left[\left(\frac{\nu-1}{\nu + \mathbb{E}\overline{X}_0^2}\right)(\Sigma - \operatorname{var}\overline{X}_0\, \delta\delta') \right]\!\!\right],$$

[3] Writing EST-AVG for the extended skew-t distribution under their parametrization, here we have that $Z \sim \text{EST-AVG}\left(\xi, \widetilde{\Omega}, \frac{\operatorname{diag}(\operatorname{diag}\widetilde{\Omega})^{\frac{1}{2}}\widetilde{\Omega}^{-1}\delta}{\sqrt{1-\delta'\widetilde{\Omega}^{-1}\delta}}, \nu, \frac{\tau}{\sqrt{1-\delta'\widetilde{\Omega}^{-1}\delta}}\right)$, where the diag operator maps matrices to a vector containing their diagonal, and vectors to a diagonal matrix with the vector on the diagonal.

[4] Note that this proposition contains a typo in the published version of the paper. As may be clearly seen from the proof, the bar over the τ at the very end of the proposition's statement should not be there.

where throughout this paper, $[\![A]\!]$ will denote the nearest symmetric positive semi-definite matrix to A, under the Frobenius norm,[5] plus the smallest multiple of the identity matrix needed to make $[\![A]\!]$ pass a standard numerical test for positive definiteness[6]. This enables us to choose ξ, δ and Ω to fit the desired location, mean and covariance, given ν and τ. To calibrate ν and τ, note that if we define:

$$\check{Z} := \frac{(\mu-\lambda)'(Z-\mu)}{\sqrt{(\mu-\lambda)'\Sigma(\mu-\lambda)}} = \frac{\delta'(Z-\mu)}{\sqrt{\delta'\Sigma\delta}} \in \mathbb{R}$$

then by Proposition 5 and Proposition 6 of Arellano-Valle and Genton (2010):

$$\mathbb{E}\check{Z} = 0, \qquad \mathbb{E}\check{Z}^2 = 1,$$

$$\mathbb{E}\check{Z}^3 = \frac{\delta'\breve{\Omega}\delta}{\delta'\Sigma\delta}\sqrt{\frac{\delta'\breve{\Omega}\delta}{\delta'\Sigma\delta}}\left[\omega_3 - 3\omega_2\frac{\delta'\delta}{\sqrt{\delta'\breve{\Omega}\delta}}\mathbb{E}\overline{X}_0 + 3\omega_1\frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}(\mathbb{E}\overline{X}_0)^2 - \frac{(\delta'\delta)^3}{\delta'\breve{\Omega}\delta\sqrt{\delta'\breve{\Omega}\delta}}(\mathbb{E}\overline{X}_0)^3\right],$$

$$\mathbb{E}\check{Z}^4 = \frac{\delta'\breve{\Omega}\delta}{\delta'\Sigma\delta}\frac{\delta'\breve{\Omega}\delta}{\delta'\Sigma\delta}\left[\omega_4 - 4\omega_3\frac{\delta'\delta}{\sqrt{\delta'\breve{\Omega}\delta}}\mathbb{E}\overline{X}_0 + 6\omega_2\frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}(\mathbb{E}\overline{X}_0)^2 - 4\omega_1\frac{(\delta'\delta)^3}{\delta'\breve{\Omega}\delta\sqrt{\delta'\breve{\Omega}\delta}}(\mathbb{E}\overline{X}_0)^3 + \frac{(\delta'\delta)^4}{\left(\delta'\breve{\Omega}\delta\right)^2}(\mathbb{E}\overline{X}_0)^4\right],$$

where:

$$\omega_1 = \frac{\delta'\delta}{\sqrt{\delta'\breve{\Omega}\delta}}\mathbb{E}\overline{X}_0$$

$$\omega_2 = \frac{\delta'\Sigma\delta}{\delta'\breve{\Omega}\delta} + \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}(\mathbb{E}\overline{X}_0)^2 = \left(1 - \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\right)\left(\frac{\nu + \mathbb{E}\overline{X}_0^2}{\nu-1}\right) + \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\mathbb{E}\overline{X}_0^2$$

$$\omega_3 = \frac{3\nu}{\nu-1}\frac{\delta'\delta}{\sqrt{\delta'\breve{\Omega}\delta}}\left(1 - \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\right)\left(\mathbb{E}\overline{X}_0 + \frac{1}{\nu}\mathbb{E}\overline{X}_0^3\right) + \frac{(\delta'\delta)^3}{\delta'\breve{\Omega}\delta\sqrt{\delta'\breve{\Omega}\delta}}\mathbb{E}\overline{X}_0^3$$

$$\omega_4 = \frac{3\nu^2}{(\nu-1)(\nu-3)}\left(1 - \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\right)^2\left(1 + \frac{2}{\nu}\mathbb{E}\overline{X}_0^2 + \frac{1}{\nu^2}\mathbb{E}\overline{X}_0^4\right) + \frac{6\nu}{\nu-1}\frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\left(1 - \frac{(\delta'\delta)^2}{\delta'\breve{\Omega}\delta}\right)\left(\mathbb{E}\overline{X}_0^2 + \frac{1}{\nu}\mathbb{E}\overline{X}_0^4\right) + \frac{(\delta'\delta)^4}{\left(\delta'\breve{\Omega}\delta\right)^2}\mathbb{E}\overline{X}_0^4$$

By numerically inverting the equations for $\mathbb{E}\check{Z}^3$ and $\mathbb{E}\check{Z}^4$, we can choose τ and ν to hit any desired skewness and kurtosis for $\check{Z}$, prior to calculating ξ, δ and Ω.

2.2. Core algorithm

The solution to a general non-linear DSGE model, or other non-linear state space model, may always be written in the following form for all $t \in \mathbb{N}^+$:

$$z_t := \begin{bmatrix} x_t \\ y_t \\ \varepsilon_t \end{bmatrix} = g_t(x_{t-1}, \varepsilon_t),$$

where $x_t \in \mathbb{R}^{n_x}$, $g_t \colon \mathbb{R}^{n_x} \times \mathbb{R}^{n_\varepsilon} \to \mathbb{R}^{n_x+n_y+n_\varepsilon}$, and $\varepsilon_t \sim \text{NIID}\left(0_{n_\varepsilon}, I_{n_\varepsilon \times n_\varepsilon}\right)$. The restriction to normal shocks is without loss of generality, since if $\varepsilon \sim \text{N}(0,1)$, we can generate

[5] Computed using the algorithm of Higham (1988).

[6] If $[\![A]\!]$ is large enough, the Cholesky decomposition of $[\![A]\!]$ should complete successfully.

shocks from a distribution with cumulative distribution function H by evaluating $H^{-1}(\Phi(\varepsilon))$, where Φ is the cumulative distribution function of the standard normal. Given that the shocks may enter non-linearly, such an expression may be incorporated into our g.

We force g to "pass-through" ε_t to its output to simplify notation in the below. We also allow g to return some of the model's control variables in y_t, as often it is impossible (or at least inconvenient) to substitute control variables out of a model's equations. Furthermore, y_t may be of independent interest, or may feature in observation equations.

We suppose that rather than observing x_t, y_t and/or ε_t, for all $t \in \mathbb{N}^+$, we instead observe:

$$m_t = h_t(z_t) + \zeta_t,$$

where $m_t \in \mathbb{R}^{n_{m,t}}$, $h_t \colon \mathbb{R}^{n_x+n_y+n_\varepsilon} \to n_{m,t}$ and $\zeta_t \sim \text{NIID}\left(0_{n_m}, \Lambda_t\right)$, where $\Lambda_t \in \mathbb{R}^{n_{m,t} \times n_{m,t}}$ is diagonal and full rank. Restricting ourselves to additive, uncorrelated, Gaussian measurement error is again without loss of generality, as richer measurement error processes can be directly incorporated into z_t. It is important that we do always allow for some additional measurement error though, since our state space dimension reduction procedure may possibly induce stochastic singularity even in models with as many shocks as observables, if some shocks make a small enough contribution.

Now, suppose that we believe that:

$$x_{t-1}|\mathcal{F}_{t-1} \sim \text{EST}\left(\hat{x}_{t-1|t-1}, P^*_{t-1|t-1}, \delta^*_{t-1|t-1}, \tau_{t-1|t-1}, \nu_{t-1|t-1}\right),$$

where $P^*_{t-1|t-1} \coloneqq S^*_{t-1|t-1} S^{*\prime}_{t-1|t-1}$, $\mathcal{F}_{t-1} \coloneqq \{m_1, \dots, m_{t-1}\}$ is the period $t-1$ information set, and where $S^*_{t-1|t-1}$ is not necessarily square. As in the standard Kalman filter, we wish to calculate the approximate distribution of $x_t|\mathcal{F}_t$, i.e. $x_t|(m_t, \mathcal{F}_{t-1})$, for which it suffices to have an extended skew t-distribution approximation to $\begin{bmatrix} x_t \\ m_t \end{bmatrix}|\mathcal{F}_{t-1}$. However, rather than splitting the calculation up into separate integrations for the predict and update step, as in both the cubature Kalman filter of Arasaratnam and Haykin (2009) and an older version of this paper (Holden 2016c), we combine both steps as in the augmented extended Kalman filter of Wan and Van Der Merwe (2000). Let $N_0, N_{10} \in \mathbb{R}$ be draws from $\text{N}(0,1)$ and $N_{11} \in \mathbb{R}^{k_{t-1|t-1}}$ be a draw from $\text{N}\left(0_{k^*_{t-1|t-1}}, I_{k^*_{t-1|t-1}}\right)$, where $k^*_{t-1|t-1} \coloneqq \operatorname{cols} S^*_{t-1|t-1}$, then note that by standard properties of the multivariate t-distribution, and the results of the previous section, the distribution of $x_{t-1}|\mathcal{F}_{t-1}$ is equal to that of:

$$\begin{aligned} x_{t-1|t-1}(N) \coloneqq \hat{x}_{t-1|t-1} &+ S^*_{t-1|t-1} N_{11} F^{-1}_{\sqrt{\frac{\nu+1}{\lambda^2_{\nu+1}}}}\left(\Phi_1(N_{10})\right) \sqrt{\frac{\nu + F^{-1}_{E_{\tau,\nu}}\left(\Phi_1(N_0)\right)^2}{\nu+1}} \\ &+ \delta^*_{t-1|t-1} F^{-1}_{E_{\tau,\nu}}\left(\Phi_1(N_0)\right), \end{aligned}$$

where $\tau = \tau_{t-1|t-1}$, $\nu = \nu_{t-1|t-1}$, $N := \begin{bmatrix} N_0 \\ N_{10} \\ N_{11} \end{bmatrix}$, Φ_1 is the c.d.f. of a standard univariate normal, $F_{\sqrt{\frac{\nu+1}{\chi^2_{\nu+1}}}}$ is the c.d.f. of $\sqrt{\frac{\nu+1}{Q}}$ where $Q \sim \chi^2_{\nu+1}$,[7] and $F_{E_{\tau,\nu}}$ is the c.d.f. of EST(0,0,1,τ,ν) i.e. the c.d.f. of $\overline{X}_0$ in the notation of the previous section. Hence, if we define $w_t := \begin{bmatrix} z_t \\ \zeta_t \end{bmatrix}$:

$$\mathbb{E}\left[\begin{bmatrix} w_t \\ m_t \end{bmatrix} \middle| \mathcal{F}_{t-1}\right] = \int_{\mathbb{R}^{\left(2+k^*_{t-1|t-1}+n_\varepsilon\right)}} \begin{bmatrix} g_t(x_{t-1|t-1}(N),\varepsilon) \\ 0 \\ h_t\Big(g(x_{t-1|t-1}(N),\varepsilon)\Big) \end{bmatrix} \phi_{2+k^*_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix} N \\ \varepsilon \end{bmatrix}\right) d\begin{bmatrix} N \\ \varepsilon \end{bmatrix},$$

$$\begin{aligned} &\operatorname{var}\left[\begin{bmatrix} w_t \\ m_t \end{bmatrix} \middle| \mathcal{F}_{t-1}\right] \\ &= \left[\!\!\left[\int_{\mathbb{R}^{\left(2+k^*_{t-1|t-1}+n_\varepsilon\right)}} a_{t|t-1}(N,\varepsilon) a_{t|t-1}(N,\varepsilon)' \phi_{2+k^*_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix} N \\ \varepsilon \end{bmatrix}\right) d\begin{bmatrix} N \\ \varepsilon \end{bmatrix} \right]\!\!\right] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Lambda_t & \Lambda_t \\ 0 & \Lambda_t & \Lambda_t \end{bmatrix}, \end{aligned}$$

where:

$$a_{t|t-1}(N,\varepsilon) := \begin{bmatrix} g_t(x_{t-1|t-1}(N),\varepsilon) \\ 0 \\ h_t\Big(g(x_{t-1|t-1}(N),\varepsilon)\Big) \end{bmatrix} - \mathbb{E}\left[\begin{bmatrix} w_t \\ m_t \end{bmatrix} \middle| \mathcal{F}_{t-1}\right],$$

and where $\phi_k : \mathbb{R}^k \to \mathbb{R}^+$ is the probability density function of a standard k-dimensional normal variable. Thus, to derive a Gaussian approximation to the distribution of $\begin{bmatrix} w_t \\ m_t \end{bmatrix} | \mathcal{F}_{t-1}$, we just need to evaluate a pair of $2 + k^*_{t-1|t-1} + n_\varepsilon$-dimensional standard Gaussian integrals, for the expectation and variance above. The co-variance term is not needed for filtering, but will be needed for smoothing, so we note its formula here. Of course, we in fact want an extended skew t approximation to $\begin{bmatrix} w_t \\ m_t \end{bmatrix} | \mathcal{F}_{t-1}$. The obvious choice for the pseudo-median (λ in the notation of the previous section) is $\begin{bmatrix} g_t(x_{t-1|t-1}(0),0) \\ 0 \\ h_t\Big(g(x_{t-1|t-1}(0),0)\Big) \end{bmatrix}$, since $x_{t-1|t-1}(0)$ is the pseudo-median of $x_{t-1}|\mathcal{F}_{t-1}$. From this, we can then calculate the quantity we called $\check{Z}$ in the previous section, and so we can represent the third and fourth moments of "$\check{Z}$" also as Gaussian integrals with respect to N and ε.

Methods for approximately evaluating these integrals will be discussed in the following section, but we observe here that any integration rule will take the form of a weighted sum over a set of sample points. Hence, the necessary integrals may be

[7] Hence: $F_{\sqrt{\frac{\nu+1}{\chi^2_{\nu+1}}}}(d) = \frac{\Gamma\left(\frac{\nu+1}{2}, \frac{\nu+1}{2d^2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}$, where $\Gamma(a,z) = \int_z^\infty e^{-t} t^{a-1}\, dt$.

evaluated simultaneously by first evaluating $\begin{bmatrix} g_t(x_{t-1|t-1}(N), \varepsilon) \\ 0 \\ h_t\big(g(x_{t-1|t-1}(N), \varepsilon)\big) \end{bmatrix}$ at the sample points, then using the integration rule to calculate the approximation to $\mathbb{E}\left[\begin{bmatrix} w_t \\ m_t \end{bmatrix} \middle| \mathcal{F}_{t-1}\right]$, which then enables us to calculate $a_{t|t-1}(N, \varepsilon)$ without further evaluations of g or h. Providing that the sample points include the zero point, we obtain the pseudo-median "for free", and using this we can calculate the value of "$\check{Z}$" at the sample points as well. As detailed in the previous section, we can then derive an approximation to the distribution of $\begin{bmatrix} w_t \\ m_t \end{bmatrix} | \mathcal{F}_{t-1}$ that matches the approximate mean, covariance and pseudo-median, along with the skewness and kurtosis of the projection onto the line through the mean and pseudo-median. I.e., we can find $\widehat{w}_{t|t-1} = \begin{bmatrix} \hat{z}_{t|t-1} \\ 0_{n_{m,t}} \end{bmatrix} = \begin{bmatrix} \hat{x}_{t|t-1} \\ \hat{y}_{t|t-1} \\ 0_{n_\varepsilon} \\ 0_{n_{m,t}} \end{bmatrix}$, $\widehat{m}_{t|t-1}$, $P_{t|t-1}$, $Q_{t|t-1}$, $R_{t|t-1}$, $\delta_{t|t-1}$, $\eta_{t|t-1}$, $\tau_{t|t-1}$ and $\nu_{t|t-1}$ such that:

$$\begin{bmatrix} w_t \\ m_t \end{bmatrix} | \mathcal{F}_{t-1} \overset{\text{approx}}{\sim} \text{EST}\left(\begin{bmatrix} \widehat{w}_{t|t-1} \\ \widehat{m}_{t|t-1} \end{bmatrix}, \begin{bmatrix} P_{t|t-1} & R_{t|t-1} \\ R'_{t|t-1} & Q_{t|t-1} \end{bmatrix}, \begin{bmatrix} \delta_{t|t-1} \\ \eta_{t|t-1} \end{bmatrix}, \tau_{t|t-1}, \nu_{t|t-1}\right),$$

where $\overset{\text{approx}}{\sim}$ is shorthand for "is approximately distributed as". To ensure that scale parameters have standard asymptotics, we always impose that $\nu_{t|t-1} > 4$. Note that although ε_t and ζ_t are actually normally distributed, their thin tails will not push upwards our estimates of $\nu_{t|t-1}$. This is because the median of ε_t and ζ_t are equal to their mean, so, by construction, "$\check{Z}$" will be independent of them. This is desirable, as we are really only interested in the tails of x_t and m_t. We also note that if we let $P^*_{t|t-1} := P_{t|t-1,11}$ be the upper left $n_x \times n_x$ block of $P_{t|t-1}$ and $\hat{x}_{t|t-1} := \widehat{w}_{t|t-1,1}$ & $\delta^*_{t|t-1} := \delta_{t|t-1,1}$ be the top n_x rows of $\widehat{w}_{t|t-1}$ & $\delta_{t|t-1}$, respectively, then by Proposition 3 of Arellano-Valle and Genton (2010):

$$x_t | \mathcal{F}_{t-1} \overset{\text{approx}}{\sim} \text{EST}\big(\hat{x}_{t|t-1}, P^*_{t|t-1}, \delta^*_{t|t-1}, \tau_{t|t-1}, \nu_{t|t-1}\big).$$

While the extended skew t approximation will not be exact, in practice it will often beat particle approximations unless implausibly large numbers of particles are used. We note one caveat to this procedure though. If a low order cubature method is used, then it will be impossible to arrive at a reliable estimate of $\nu_{t|t-1}$, since this requires fourth moments. In this case, we suggest that the modeller assume that for all t, $\nu_{t|t-1} = \bar{\nu}$, where $\bar{\nu}$ is an additional free parameter to be estimated by the modeller. The estimation of this additional non-structural parameter potentially reduces efficiency, but this cost may be worthwhile if the resulting improved approximation to the likelihood sufficiently reduces bias.

We want to know the distribution of $w_t|\mathcal{F}_t$, but by the definition of $\mathcal{F}_t$, this is just equal to the distribution of $w_t|(m_t, \mathcal{F}_{t-1})$. Hence, by Proposition 4 of Arellano-Valle and Genton (2010):

$$w_t|\mathcal{F}_t \overset{\text{approx}}{\sim} \text{EST}(\widehat{w}_{t|t}, P_{t|t}, \delta_{t|t}, \tau_{t|t}, \nu_{t|t}),$$

where:

$$\begin{aligned}
\widehat{w}_{t|t} &:= \check{R}_{t|t-1}\check{Q}_{t|t-1}^{-1}(m_t - \widehat{m}_{t|t-1}),\\
P_{t|t} &:= \frac{\nu_{t|t-1} + (m_t - \widehat{m}_{t|t-1})'\check{Q}_{t|t-1}^{-1}(m_t - \widehat{m}_{t|t-1})}{\nu_{t|t-1} + n_{m,t}}\left[\!\!\left[\tilde{P}_{t|t-1} - \frac{\tilde{\delta}_{t|t-1}\tilde{\delta}_{t|t-1}'}{1 - \eta_{t|t-1}'\check{Q}_{t|t-1}^{-1}\eta_{t|t-1}}\right]\!\!\right],\\
\delta_{t|t} &:= \sqrt{\frac{\nu_{t|t-1} + (m_t - \widehat{m}_{t|t-1})'\check{Q}_{t|t-1}^{-1}(m_t - \widehat{m}_{t|t-1})}{(\nu_{t|t-1} + n_{m,t})(1 - \eta_{t|t-1}'\check{Q}_{t|t-1}^{-1}\eta_{t|t-1})}}\tilde{\delta}_{t|t-1},\\
\tau_{t|t} &:= \sqrt{\frac{\nu_{t|t-1} + n_{m,t}}{\nu_{t|t-1} + (m_t - \widehat{m}_{t|t-1})'\check{Q}_{t|t-1}^{-1}(m_t - \widehat{m}_{t|t-1})}}\frac{\eta_{t|t-1}'\check{Q}_{t|t-1}^{-1}(m_t - \widehat{m}_{t|t-1}) + \tau_{t|t-1}}{\sqrt{1 - \eta_{t|t-1}'\check{Q}_{t|t-1}^{-1}\eta_{t|t-1}}},\\
\nu_{t|t} &:= \nu_{t|t-1} + n_{m,t},
\end{aligned}$$

and:

$$\begin{aligned}
\tilde{P}_{t|t-1} &:= \check{P}_{t|t-1} - \check{R}_{t|t-1}\check{Q}_{t|t-1}^{-1}\check{R}_{t|t-1}',\\
\tilde{\delta}_{t|t-1} &:= \delta_{t|t-1} - \check{R}_{t|t-1}\check{Q}_{t|t-1}^{-1}\eta_{t|t-1},\\
\check{P}_{t|t-1} &:= P_{t|t-1} + \delta_{t|t-1}\delta_{t|t-1}',\\
\check{R}_{t|t-1} &:= R_{t|t-1} + \delta_{t|t-1}\eta_{t|t-1}',\\
\check{Q}_{t|t-1} &:= Q_{t|t-1} + \eta_{t|t-1}\eta_{t|t-1}',
\end{aligned}$$

so:

$$\check{Q}_{t|t-1}^{-1} = \left(Q_{t|t-1} + \eta_{t|t-1}\eta_{t|t-1}'\right)^{-1} = Q_{t|t-1}^{-1} + \frac{Q_{t|t-1}^{-1}\eta_{t|t-1}\eta_{t|t-1}'Q_{t|t-1}^{-1}}{1 - \eta_{t|t-1}'Q_{t|t-1}^{-1}\eta_{t|t-1}}.$$ [8]

So, as before, if we let $P_{t|t}^* := P_{t|t,11}$ be the upper left $n_x \times n_x$ block of $P_{t|t}$ and $\hat{x}_{t|t} := \widehat{w}_{t|t,1}$ & $\delta_{t|t}^* := \delta_{t|t,1}$ be the top n_x rows of $\widehat{w}_{t|t}$ & $\delta_{t|t}$, respectively, then by Proposition 3 of Arellano-Valle and Genton (2010):

$$x_t|\mathcal{F}_t \overset{\text{approx}}{\sim} \text{EST}\left(\hat{x}_{t|t}, P_{t|t}^*, \delta_{t|t}^*, \tau_{t|t}, \nu_{t|t}\right).$$

To complete the inductive step, we just have to find $S_{t|t}^*$ such that $S_{t|t}^* S_{t|t}^{*\prime} \approx P_{t|t}^*$. We give a general procedure for the reduced rank factorization of any symmetric positive semi-definite (henceforth, p.s.d.) matrix M. First, let $U_M D_M U_M'$ be the Schur decomposition of M, where U_M is orthogonal, and D_M is diagonal and weakly positive, since M is p.s.d.. Let $d_M := \operatorname{diag} D_M$, where the diag operator maps matrices to a vector containing their diagonal, and vectors to diagonal matrices with the given vector on their diagonal. Now choose a threshold κ. Then, without loss of generality, we may suppose that only the $k_{M,\kappa}$ first elements of d_M are strictly greater than κ. Now

[8] Note that $Q_{t|t-1}$ is invertible as Λ is full rank, by assumption.

let $U_{M,\kappa,\cdot 1}$ be the first $k_{M,\kappa}$ columns of U_M, and $d_{M,\kappa,1}$ be the first $k_{M,\kappa}$ rows of d_M. Consequently, if we define:

$$S_{M,\kappa} := U_{M,\kappa,\cdot 1} \operatorname{diag} d_{M,\kappa,1}^{\frac{1}{2}},$$

where exponentiation operates element-wise on vectors, then, providing κ is small:

$$M \approx U_{M,\kappa,\cdot 1} \operatorname{diag} d_{M,\kappa,1} U'_{M,\kappa,\cdot 1} = S_{M,\kappa} S'_{M,\kappa},$$

giving a reduced rank factorisation of M. Since the Schur decomposition coincides with the singular value decomposition for p.s.d. matrices, the result is the optimal rank $k_{M,\kappa}$ factorisation of M under the Frobenius norm, by the Eckart–Young theorem.

If we now fix $\kappa^* > 0$, and define $S^*_{t|t} := S_{P^*_{t|t},\kappa^*}$, then:

$$x_t|\mathcal{F}_t \overset{\text{approx}}{\sim} \operatorname{EST}\left(\hat{x}_{t|t}, S^*_{t|t} S^{*\prime}_{t|t}, \delta^*_{t|t}, \tau_{t|t}, \nu_{t|t}\right),$$

which completes one time-step. In all the examples below we set $\kappa^* := 10^{-12}$, which is large enough to provide reasonable dimension reduction without overly affecting accuracy. There are two benefits to taking reduced rank approximations. Firstly, by reducing the dimensionality of the space over which we must integrate, it will greatly speed up the computation of integrals. Secondly, integration rules are often much better behaved in lower dimensions. This may mean they evaluate less far from the centre, avoiding distortions caused by extreme tail non-linearities, or it may mean that they have more uniform weights, avoiding e.g. failures of positive semi-definiteness caused by a negative weight.

We close this section by noting that we can use these calculations and approximations to obtain the approximate likelihood, in the standard way. In particular:

$$\begin{aligned}
\log f(\mathcal{F}_t) &= \log f(\mathcal{F}_{t-1}) + \log f(m_t|\mathcal{F}_{t-1}) \\
&\approx \log f(\mathcal{F}_{t-1}) - \frac{1}{2}\log|\check{Q}_{t|t-1}| + \log f_{T_{\nu_{t|t-1},n_{m,t}}}\left(\check{Q}_{t|t-1}^{-\frac{1}{2}}(m_t - \hat{m}_{t|t-1})\right) \\
&\quad - \log F_{T_{\nu_{t|t-1},1}}(\tau_{t|t-1}) + \log F_{T_{\nu_{t|t},1}}(\tau_{t|t}),
\end{aligned}$$

This gives an iterative formula for progressively calculating the approximate log-likelihood.[9]

2.3. Cubature methods

Degree 3 monomial rule Arasaratnam and Haykin (2009) suggest approximating the Gaussian integrals via degree three monomial cubature. The degree three rule they advocate is based upon the following approximation:

$$\int_{\mathbb{R}^k} q(z)\phi_k(z)\,dz \approx \frac{1}{2k}\sum_{j=1}^{k}\left[q\left(e_j\sqrt{k}\right) + q\left(-e_j\sqrt{k}\right)\right],$$

which is exact when the arbitrary function q is in fact a sum of monomials of at most degree 3 (where $e_{j,i}$ is 0 for $i \neq j$ and 1 for $i = j$). Indeed, this is the degree 3 rule

[9] In practice, we use the Cholesky factor of $\check{Q}_{t|t-1}$, rather than a matrix square root, and we rewrite all of the quadratic forms here and in the previously given update formulas in terms of this Cholesky factor.

requiring the minimum possible number of function evaluations, making it highly computationally efficient. However, since our likelihood requires the evaluation of $q(0)$, it makes sense to include this additional point in the cubature rule, since its evaluation has no additional cost. The equally weighted degree 3 cubature rule including the zero point, with the minimum number of other points, takes the form:

$$\int_{\mathbb{R}^k} q(z)\phi_k(z)\,dz \approx \frac{1}{2k+1}\left[q(0) + \sum_{j=1}^{k}\left[q\left(\frac{e_j}{2}\sqrt{2+4k}\right) + q\left(-\frac{e_j}{2}\sqrt{2+4k}\right)\right]\right],$$

which just adds one additional point.

In a model without occasionally binding constraints, under a first order approximation either rule will give the exact mean and variance, and under a second or third order approximation, it will give the exact mean, but only an approximate variance. In the presence of occasionally binding constraints, it will only give approximate means and variances, whatever the order of approximation.

However, in practice these rules tends to perform remarkably well. As discussed in Holden (2016b), these degree 3 monomial cubature rules are particularly robust since they have positive, equal weights. All known higher degree integration rules that do not use more than polynomial in k nodes also feature negative weights on at least some nodes (Cools 2003), which means that their result is not guaranteed to lie within the convex hull of the source evaluations, and it means that the approximated covariance matrix need not be p.s.d..

A downside of either rule is that when k is large, the rule evaluates a long distance from the mean (approximately the same distance for either rule). When the true integrand is actually a sum of monomials of at most degree 3, this obviously does not matter, but in reality q often has substantially more curvature. Indeed, in the presence of e.g. occasionally binding constraints, q may feature extreme behaviour in the tails. Hence, when q is evaluated far into the tails, we are likely to obtain biased estimates of the integral. Our dimension reduction algorithm obviously helps with this, but still in large models this may be problematic.

Additionally, since these rules only integrate degree 3 monomials exactly, they are never going to be able to give reasonable approximations to $\mathbb{E}\check{Z}^4$. Hence, if we are using these rules, then we will have to assume that for all t, $\nu_{t|t-1} = \bar{\nu}$, as previously suggested, meaning that the fatness of the tails of the approximating distribution will not be dynamic.

Genz and Keister (1996) rules By way of motivation, note that with the standard cubature Kalman filter, if we obtained a non-p.s.d. covariance matrix at one step, it would produce a catastrophic failure of the likelihood evaluation. We avoid such problems for two reasons. Firstly, our use of the $[\![\quad]\!]$ operator ensures our $P_{t|t}$ is positive definite. Secondly, even if we did not force $P_{t|t}$ to be positive definite, given our dimension reduction algorithm, we would still not have problems. To see this, suppose that $P_{t|t}$ were not p.s.d.. It would nonetheless be symmetric though, as it is being approximated by a weighted sum of symmetric matrices. Hence, by the spectral theorem for real symmetric matrices, the Schur decomposition would enable us to calculate $S_{t|t}$ just as before, still selecting the eigenvalues that are greater than κ^* to give a p.s.d. approximation to $P_{t|t}$. This would be a reasonable approximation to the true covariance of interest provided (plausibly enough) that the reason we ended up with a non-p.s.d. matrix was that the true covariance has some very small magnitude eigenvalues.

In light of this discussion, it is natural for us to reappraise the use of higher degree cubature rules in our setting. Holden (2016b) found that the rules of Genz and Keister (1996) performed very well in a different context, with not excessively high computational cost, so these rules seem a natural thing to try here too. These rules rules allow one to choose the maximum degree of monomial that should be integrated exactly, up to a maximum order of 51. The number of points used is $\mathrm{O}\left(k^d\right)$, where $2d + 1$ is the degree of monomial that is integrated exactly. When $d > 0$ and $k > 1$, the rule features negative weights on at least one node, but this enables it to ensure that the maximum over the absolute vectors of integration points is independent of k. This contrasts with the aforementioned rule in which the higher is k, the further into the tails of the distribution one has to evaluate the integrand, which may lead to poor performance as discussed previously. Furthermore, by using a higher degree rule, we can generally obtain a better approximation to the integrand, leading to improved estimates. This is particularly important when it comes to the evaluation of $\mathbb{E}\check{Z}^4$, which we use to calibrate $\nu_{t|t-1}$. Numerical experiments suggest that providing n is not too small, using a rule that exactly integrates monomials of up to degree 9 performs well.

2.4. Selecting the initial distribution of the state

For models in which g_t varies over time, there is no choice but to maximise the likelihood conditional on the first observation, but for models in which $g_t \equiv g$, we would like to be able to maximise the unconditional likelihood. In order to evaluate this, we need to be able to calculate the unconditional distribution of m_1, which in turn requires the unconditional distribution of x_0. Now, if we have a model without occasionally binding constraints, then it is possible to get the moments of x_0 without simulation if we are using a pruned solution. Even if we are not using a pruned

solution, we can at least derive reasonable approximations to these quantities without resorting to simulation. However, in the presence of occasionally binding constraints, or with more general non-linearities, we will not be able to calculate a reasonable approximation to the stationary distribution of x_t without resorting to simulation.

One approach then would be to simulate a long run from the model, discard a burn-in period, and then take the moments of the remainder. This has two drawbacks. Firstly, due to the high degree of auto-correlation in many models, removing all sampling variation from the estimate would require a prohibitively long simulation run. Even if the same random seed was used for each run, so the objective was still continuous in the parameters even in finite samples, the result would just be that the sampling variation was transmitted to the final parameter estimates. This is essentially the same problem as is encountered by the particle filter in maximum likelihood contexts.

Secondly, it is not clear that the stationary distribution of the model is actually what is needed here. In general, thanks to the approximations intrinsic in any variant of the cubature Kalman filter, including ours, the value to which e.g. $\hat{x}_{t|t}$ and $P^*_{t|t}$ would converge given an infinite string of completely uninformative observations will not agree with the stationary distribution of the model. If it is agreed that parameter estimates should not change when the data set is augmented by a run of initial missing observations, then rather than trying to evaluate the stationary distribution of x_t, we should be trying to evaluate the limit of $\hat{x}_{t|t}$, $P^*_{t|t}$, etc. when no information arrives.

This is the approach we take here. In particular, we imagine that we were tasked with running the cubature Kalman filter on an infinite run of missing observations. Since this works in "pseudo-time", to distinguish "pseudo-times" from real times, we place $\tilde{}$ over all "pseudo-times" in the following. We start by initializing $\hat{x}_{\tilde{0}|\tilde{0}}$, $P_{\tilde{0}|\tilde{0}}$ $\delta^*_{\tilde{0}|\tilde{0}}$, $\tau_{\tilde{0}|\tilde{0}}$ and $\nu_{\tilde{0}|\tilde{0}}$ with some easily computable approximation, such as that derived from the pruned perturbation approximation to the model, omitting occasionally binding constraints. We then run our cubature Kalman filter forward, with $n_{m,\tilde{t}} = 0$, meaning that $\hat{x}_{\tilde{t}|\tilde{t}} = \hat{x}_{\tilde{t}|\widetilde{t-1}}$, $P^*_{\tilde{t}|\tilde{t}} = P^*_{\tilde{t}|\widetilde{t-1}}$, $\delta^*_{\tilde{t}|\tilde{t}} = \delta^*_{\tilde{t}|\widetilde{t-1}}$, $\tau_{\tilde{t}|\tilde{t}} = \tau_{\tilde{t}|\widetilde{t-1}}$ and $\nu_{\tilde{t}|\tilde{t}} = \nu_{\tilde{t}|\widetilde{t-1}}$ for all $\tilde{t} \in \mathbb{N}^+$. We continue until the change in these quantities is sufficiently small (e.g. on the order of 10^{-8}). For some models, this procedure may not converge exactly, in which case rather than making a full step from $\hat{x}_{\widetilde{t-1}|\widetilde{t-1}}$ to $\hat{x}_{\tilde{t}|\tilde{t}}$, $P^*_{\widetilde{t-1}|\widetilde{t-1}}$ to $P^*_{\tilde{t}|\tilde{t}}$, etc., we instead make a partial step to a weighted average of the old and new points. DynareOBC contains code for dynamically adjusting the weight which works well in practice, ensuring convergence. When $\hat{x}_{\tilde{t}|\tilde{t}}$, $P^*_{\tilde{t}|\tilde{t}}$, $\delta^*_{\tilde{t}|\tilde{t}}$, $\tau_{\tilde{t}|\tilde{t}}$ and $\nu_{\tilde{t}|\tilde{t}}$ have converged, we set $\hat{x}_{0|0}$, $P^*_{0|0}$, $\delta^*_{0|0}$, $\tau_{0|0}$ and $\nu_{0|0}$ to the found limiting values.

2.5. Maximising the likelihood and computing standard errors

Traditionally, DSGE models have chiefly been estimated by Bayesian methods. This has apparently been for two reasons. Firstly, many parameters in linearized DSGE models are either unidentified or just weakly identified. By placing a prior over the parameter space, although the likelihood may be flat in places, the posterior density will not be, ensuring that any numerical maximisation algorithm will return the same maximum a posterior estimate. Nonetheless, the prior does not solve the underlying non-identification. Instead, when the likelihood is flat, then the prior becomes highly "informative". Of course, if the prior reflects true external information (e.g. from panel micro-data), then it is completely appropriate to incorporate this information into the final estimate. However, often the prior used in macroeconomic modelling is instead a product (albeit indirect) of the same data on which the model is now being estimated. It is preferable then to attempt to fix the underlying weak identification problem, which is what non-linear estimation potentially permits. This is thanks to the fact that two parameters may have identical effects on dynamics very close to the steady-state, but quite different effects further away.

The second reason people have not traditionally pursued maximum likelihood estimation of DSGE models is because the likelihoods tend to be highly multi-modal. The hope is that with a strong enough prior, the posterior density might be unimodal, even though the likelihood is not. This hope is somewhat naïve though, since at least asymptotically the likelihood will asymptotically dominate the prior along any dimensions in which there is identification. Evidence for the practical relevance of this is provided by Herbst and Schorfheide (2014) who show that a range of popular DSGE models actually possess multimodal posteriors, though this was previously missed due to the difficulties of integrating over high dimensional spaces.

Consequently, the only way that a Bayesian approach could have a computational advantage over a classical approach would be if somehow it was at least easier to integrate over high dimensional spaces than it was to maximise in them. But this cannot be true. If one has an algorithm that can sample from a distribution in a high dimensional space, then by starting standard local maximisation procedures from these draws, one will eventually find the global maximum. Furthermore, since local maximisation requires far fewer evaluations than (say) MCMC would to explore a mode, this will be faster. Of course, one can never guarantee that a maximisation procedure has found the true global maximum, but neither can one guarantee that an integration procedure has explored every mode. The fact that most MCMC implementations start by using conventional methods to search for a global mode provides further evidence that integration must be harder than maximisation.

In short then, the standard arguments for a Bayesian approach do not seem relevant when the model is estimated non-linearly. To make a maximum likelihood approach practical though, we must provide a global search algorithm with decent performance. By default, DynareOBC uses a customised version of the CMA-ES algorithm of Hansen et al. (1995; 2006). While first order (and first generation) evolutionary algorithms evolve a population of parameter vectors by combining parameters from multiple "parents" and adding independent noise to each component, the second order (and second generation) CMA-ES algorithm draws noise from a covariance matrix which mirrors the shape of the objective. This covariance matrix is dynamically updated over time in an entirely parameter-free way, and the result is an algorithm which is almost competitive with local algorithms on unimodal objectives,[10] but which also possesses good global search properties. The modified version in DynareOBC is designed to exploit parallel computing environments, further speeding up the search.

Once we have found the location of the maximum of the likelihood function, we then need to compute standard errors. Now, recall that the likelihood is coming from approximations to the distributions of $m_t|\mathcal{F}_{t-1}$ for $t \in \{1, \dots, T\}$. In non-linear models, these distributions may be quite a way from the true distribution, hence, what we have here is in fact a quasi-maximum likelihood estimate. Quasi-maximum likelihood estimates are consistent with respect to the pseudo-true parameters (the ones minimising the Kullback–Leibler divergence to the true distribution) and asymptotically normal (White 1982), but they require standard errors to be computed using a sandwich formula, as detailed in e.g. Canova (2007).

In fact, any "maximum-likelihood" estimation of an approximation (first order or otherwise) to a model is actually a quasi-maximum likelihood procedure. The true distribution of the measurement will be non-normal in general, so the result of, for example, linearization followed by the Kalman filter will still not be the true distribution. While Gaussian quasi-maximum likelihood appears to be consistent with respect to the true parameters under milder conditions than in the non-Gaussian case (Bollerslev and Wooldridge 1992), at a minimum this requires that the mean and covariance of the approximating Gaussian model are equal to the true mean and covariance, which is certainly not the case following an approximation to the source model. Thus, for example, the maximum a posterior estimate of the parameters of a linearized DSGE model is inconsistent in general for the true (un-approximated) model's parameters. Given that the profession does not seem to be overly concerned with this inconsistency, there does not seem to be any good reason to be concerned with the inconsistency coming from our explicit approximations to the true distribution of the measurement. Indeed, it seems reasonable to hope that by better

[10] CMA-ES requires around ten times more function evaluations than local methods on quadratic objectives.

fitting the true distribution, thanks to higher order approximations to the source model, and more flexible distributional approximations, we ought to reduce the magnitude of this inconsistency if anything. Nonetheless, in some circumstances, being assured of consistency is desirable. In the next section, we derive a modified estimator for these circumstances.

2.6. Alternative objective delivering a consistent estimator

We now describe an alternative to the standard quasi-maximum likelihood estimator that is consistent for the true parameters of the model, or at least for the parameters of the best approximation to the model from which one can simulate. For any θ, let $m^*_{\theta,1}, m^*_{\theta,2}, \ldots$ be a series of simulated "measurements" generated by simulating the model and its measurement equation, and define $\mathcal{F}^*_{\theta,t^*} := \{m^*_{\theta,1}, \ldots, m^*_{\theta,t^*}\}$. Furthermore, let $\tilde{f}(m_t|\mathcal{F}_{t-1}, \theta)$ be the approximate likelihood function derived in section 2.2, with $\tilde{f}(m^*_{\theta,t^*}|\mathcal{F}^*_{t^*-1}, \theta)$ the approximate likelihood of the simulated data, and let $\tilde{s}(\theta; m_t, \mathcal{F}_{t-1}) := \frac{\partial \tilde{f}(m_t|\mathcal{F}_{t-1}, \theta)}{\partial \theta}$ be the corresponding score. Then define:

$$\bar{s}_T(\theta) := \frac{1}{T}\sum_{t=1}^{T} \tilde{s}(\theta; m_t, \mathcal{F}_{t-1}) - \frac{1}{T^*(T)}\sum_{t^*=1}^{T^*(T)} \tilde{s}(\theta; m^*_{\theta,t^*}, \mathcal{F}^*_{t^*-1})$$

where $\frac{T^*(T)}{T} \to \infty$ as $T \to \infty$. Finally define our estimate as:

$$\hat{\theta}_T := \arg\min_{\theta} \bar{s}_T(\theta)' \bar{s}_T(\theta).$$

Note that if $\tilde{f}$ were the true density, then as $\frac{T^*(T)}{T} \to \infty$, the right-hand term in $\bar{s}_T(\theta)$ would tend to zero, so the first order conditions of this problem would be asymptotically identical to those for the quasi-maximum likelihood estimator. However, in general this will not be the case, and $\hat{\theta}_T$ will differ from the standard quasi-maximum likelihood estimator. That $\hat{\theta}_T$ is consistent and asymptotically normal (given the standard technical assumptions) follows immediately from the consistency of the simulated method of moments (Duffie and Singleton 1990). Since this is a just identified case, and we are assuming that $\frac{T^*(T)}{T} \to \infty$ as $T \to \infty$, this estimator is also asymptotically efficient in the class of asymptotically normal estimators using these moment conditions.[11]

3. A test of the performance of our algorithm

In this section, we give some indications of the accuracy of our approach, by applying it to three simple non-linear models, with one from finance, and two simple DSGE models, one of which contains occasionally binding constraints. The finance

[11] An alternative procedure would have been to find, by simulation, the value of the true parameters at which the expected score at the QMLE parameters was equal to zero. This two-step procedure was proposed by Tauchen and Gallant (1995) and would deliver similar results, but is likely to be computationally more expensive due to the need to maximise twice.

model is a generalization of the standard discrete time stochastic volatility (SV) model, generalized to allow for time variation in the parameters of the standard SV model. One of the DSGE models is a minimal example of a DSGE model with occasionally binding constraints, the other merely removes the constraint from the former. We restrict ourselves to OBC models for which the results of Holden (2016a) imply there is a unique solution in all states of the world, to avoid additional uncertainty coming from equilibrium selection, though the algorithm of Holden (2016b) does give a natural procedure for doing this. Likewise, we restrict ourselves to models for which an exact solution is available, to enable us to assess the impact of approximation error on the final parameter estimates.

We start by presenting results for the stochastic volatility model. We then describe the two DSGE models, and their identification properties, before presenting the estimation results.

3.1. Results for a stochastic volatility model with time varying models

TODO

3.2. A simple "DSGE" model without occasionally binding constraints

Suppose the representative household in an economy chooses consumption C_t and zero net supply bond holdings B_t to maximise:

$$\mathbb{E}_t \sum_{k=0}^{\infty} \beta^k \frac{C_{t+k}^{1-\gamma} - 1}{1-\gamma},$$

subject to the restriction that:

$$A_t + R_{t-1} B_{t-1} = C_t + B_t$$

for all $t \in \mathbb{Z}$, where A_t's evolution is given by: $\log A_t = \log A_{t-1} + g_t$, where

$$g_t = (1-\rho)\bar{g} + \rho g_{t-1} + \sigma \varepsilon_t$$

and $\varepsilon_t \sim \mathrm{N}(0,1)$. Market clearing implies $A_t = C_t$ and $B_t = 0$ for all $t \in \mathbb{Z}$, implying that the first order condition may be written as:

$$1 = \beta R_t \mathbb{E}_t \exp(-\gamma g_{t+1}) = R_t \mathbb{E}_t \exp(\log \beta - \gamma g_{t+1}),$$

and from this, it is easy to see that in the exact solution:

$$R_t = \left[\beta \exp\left(\frac{\gamma^2 \sigma^2}{2} - \gamma \mu_t \right) \right]^{-1},$$

where $\mu_t = (1-\rho)\bar{g} + \rho g_t$.

To make our estimation task more challenging, we suppose that the econometrician only observes $\log R_t$, not g_t. The model has five parameters, but with this limited information set, only three of them can be identified. To see this, first note that since γ just scales $-\gamma g_{t+1}$, and $\log \beta$ just shifts $\log \beta - \gamma g_{t+1}$, the parameter vector $(\beta, \gamma, \bar{g}, \rho, \sigma)$ must be observationally equivalent to the parameter vector $(1,1, \gamma\bar{g} - \log \beta\,, \rho, \gamma\sigma)$. In light of this, when we estimate the model, we fix β and γ at their true values. These estimates will be based on an 1000 period artificial data-set constructed

from the exact solution using the following parameters: $\beta := 0.99$, $\gamma := 5$, $\bar{g} := 0.005$, $\rho := 0.95$ and $\sigma := 0.007$ (after discarding 100 periods of burn-in).

3.3. A simple "DSGE" model with occasionally binding constraints

To produce a model with occasionally binding constraints, we modify the previous model, changing the law of motion for g_t. In particular, we suppose that for all $t \in \mathbb{Z}$:

$$g_t = \max\{0, (1-\rho)\bar{g} + \rho g_{t-1} + \sigma\varepsilon_t\}$$

where $\varepsilon_t \sim \mathrm{N}(0,1)$. This specification may be thought of as capturing the fact that technologies cannot be un-invented. The first order condition of this model is as before, but now, in the exact solution, slightly more onerous calculation gives us that:

$$R_t = \left[\beta\left[\left(1 - \Phi\left(\frac{\mu_t}{\sigma}\right)\right) + \left(1 - \Phi\left(\frac{\gamma\sigma^2 - \mu_t}{\sigma}\right)\right)\exp\left(\frac{\gamma^2\sigma^2}{2} - \gamma\mu_t\right)\right]\right]^{-1},$$

where μ_t is as before.

This model also contains five parameters, but thanks to the additional non-linearity, four of them can be identified despite only $\log R_t$ being observed. While γ still just scales the $-\gamma g_{t+1}$ term in the first order condition, varying $\log\beta$ is no longer equivalent to shifting $\bar{g}$, due to the zero lower bound on productivity. Thus, when we estimate this model, we only need to fix γ at its true value. For comparison though, we also perform runs with β fixed too. These estimates will use an artificial data set constructed from the exact solution exactly as before, with identical parameters.

3.4. "DSGE" Estimation results

TODO: UPDATE THE RESULTS BELOW WHICH WERE CREATED USING A GAUSSIAN APPROXIMATION TO THE DISTRIBUTION OF THE STATE

Results from estimating both models are contained in Table 1 below. We also include the results of estimating the unbounded model on data from the bounded model, to illustrate the biases that can occur when bounds are ignored. Furthermore, to illustrate the potential costs of different types of approximation error, we include both results where the simulations in the filter predict step are performed using the exact solution, and results where they are performed using the approximate solution algorithm from Holden (2016b), either without cubature or with degree 3 monomial cubature in the internal simulation algorithm (referred to as "fast cubature" in the below). If cubature is not incorporated into the inner simulation algorithm, then expectations are not fully rational, as agents are continually surprised by the presence of the bound. Using cubature in the inner solution algorithm fixes this, producing more accurate simulations, and, hopefully, better estimates.

When using this approximate solution algorithm without cubature, we try both with an order one approximation to the underlying model, and with an order three one. The latter illustrates our algorithm's dimension reduction method, since it turns out

that in these models, order two and order three solutions agree. Finally, when we use the exact solution algorithm, we try both with degree 3 cubature for the integrals in the filter step, and with the degree 51 Genz and Keister (1996) rules, which are essentially exact. All other integrals are performed with the degree 3 monomial rule.

Bound in model	*Bound in d.g.p.*	*Simulation approxi-mation*	*Filter int. degree*	$log\, p(\mathcal{Y}_t)$ *true params.*	$log\, p(\mathcal{Y}_t)$ *estimated params.*	β	$\bar{g}$	ρ	σ	$\sqrt{\Lambda}$
No	No	Order 1	3	1991.00	2005.96	*Fixed*	2.72E-03	9.24E-01	7.04E-03	0.00E+00
							(NaN)	*(8.09E-03)*	*(2.95E-05)*	*(5.53E-11)*
No	No	Order 3	3	1991.02	2005.96	*Fixed*	2.84E-03	9.24E-01	7.04E-03	0.00E+00
							(NaN)	*(1.07E-02)*	*(5.26E-05)*	*(1.13E-10)*
No	No	Exact	3	1991.02	2005.96	*Fixed*	2.84E-03	9.24E-01	7.04E-03	0.00E+00
							(1.57E-01)	*(2.51E+01)*	*(4.14E-02)*	*(1.76E-07)*
No	No	Exact	51	1991.02	2005.96	*Fixed*	2.84E-03	9.24E-01	7.04E-03	0.00E+00
							(2.60E-05)	(4.89E-03)	(4.46E-05)	(1.35E-10)
No	Yes	Order 1	3	2116.99	2193.82	*Fixed*	1.51E-02	8.98E-01	6.00E-03	0.00E+00
							(NaN)	*(1.27E-02)*	*(9.19E-05)*	*(3.22E-10)*
No	Yes	Order 3	3	2116.92	2193.82	*Fixed*	1.52E-02	8.98E-01	6.00E-03	0.00E+00
							(NaN)	*(2.31E-02)*	*(1.11E-04)*	*(5.06E-10)*
No	Yes	Exact	3	2116.92	2193.82	*Fixed*	1.52E-02	8.98E-01	6.00E-03	0.00E+00
							(1.21E-04)	(2.57E-03)	(5.09E-05)	(1.92E-10)
No	Yes	Exact	51	2116.92	2193.82	*Fixed*	1.52E-02	8.98E-01	6.00E-03	0.00E+00
							(6.71E-05)	*(1.70E-02)*	*(2.85E-06)*	*(9.73E-11)*
Yes	Yes	Order 1 No cub.	3	2181.32	2239.61	*Fixed*	1.02E-02	9.21E-01	6.25E-03	2.12E-11
							(NaN)	*(7.45E-02)*	*(9.36E-05)*	*(1.68E-16)*
Yes	Yes	Order 3 No cub.	3	2187.98	2242.91	*Fixed*	1.06E-02	9.20E-01	6.31E-03	4.25E-11
							(NaN)	*(5.27E-03)*	*(6.43E-05)*	*(4.53E-15)*
Yes	Yes	Order 1 Fast cub.	3	2232.11	2258.29	*Fixed*	3.13E-03	9.44E-01	7.02E-03	5.23E-09
							(NaN)	*(7.47E-03)*	*(1.65E-12)*	*(1.05E-10)*
Yes	Yes	Exact	3	2233.50	2254.06	*Fixed*	2.89E-03	9.46E-01	6.59E-03	3.80E-05
							(2.78E-05)	(5.51E-03)	(6.23E-05)	(2.38E-07)
Yes	Yes	Exact	51	2236.48	2255.92	*Fixed*	1.36E-03	9.49E-01	6.69E-03	3.73E-05
							(3.79E-05)	(2.39E-02)	(1.60E-04)	(1.34E-07)
Yes	Yes	Order 1 No cub.	3	2181.32	2253.06	9.75E-01	1.00E-06	9.44E-01	6.37E-03	1.40E-09
						(NaN)	*(NaN)*	*(2.55E-02)*	*(4.97E-04)*	*(6.27E-10)*
Yes	Yes	Order 3 No cub.	3	2187.98	2253.07	9.75E-01	1.00E-06	9.44E-01	6.37E-03	0.00E+00
						(NaN)	*(NaN)*	*(4.02E-03)*	*(5.57E-05)*	*(1.11E-10)*
Yes	Yes	Order 1 Fast cub.	3	2232.11	2262.63	9.93E-01	6.53E-03	9.33E-01	6.91E-03	1.38E-11
						(NaN)	*(NaN)*	*(8.46E-03)*	*(5.58E-05)*	*(1.58E-11)*
Yes	Yes	Exact	3	2233.50	2257.95	9.96E-01	9.75E-03	9.23E-01	6.85E-03	2.11E-11
						(1.94E-03)	*(1.37E-04)*	*(3.92E-03)*	*(6.77E-05)*	*(3.94E-16)*
Yes	Yes	Exact	51	2236.48	2259.46	9.96E-01	9.58E-03	9.24E-01	6.95E-03	0.00E+00
						(2.50E-08)	(2.10E-10)	(2.03E-08)	(1.52E-10)	(1.56E-14)

Table 1: Estimates (Standard errors in brackets. ***Standard errors in italics may be unreliable due to reported numerical difficulties inverting the Hessian. NaN stands for "not a number" and is indicative of near zero Eigenvalues in both the Hessian and the Fisher information matrix, resulting in 0/0 expressions.*****)**

All numerical maximisation was performed with the CMA-ES algorithm, with the results polished by MATLAB's "fmincon". We start CMA-ES from the true parameters, with the measurement error standard deviation, $\sqrt{\Lambda}$, set to 0.0001, but thanks to the initial broad search undertaken by the CMA-ES algorithm, it soon moves

away from this point, so identical results would be derived with a different initial point. We constraint β and ρ to lie in $[10^{-6}, 1-10^{-6}]$, $\bar{g}$ to lie in $[10^{-6}, \infty)$, and σ and $\sqrt{\Lambda}$ to lie in $[0, \infty)$. Estimation times range from a few minutes for the models without bounds, to around four hours for the runs with "fast cubature" in the internal simulation algorithm, running on either a 12 or 20 core machine. We stress though that thanks to our dimension reduction techniques, and the fact that simulation speed is almost independent of model size, running times should be of a similar order of magnitude even with medium-scale models.

Turning to the results, we first note that the estimates of $\bar{g}$ are all quite poor. For example, in the model without occasionally binding constraints, $\bar{g}$ is roughly half its true value, even with the exact simulation algorithm, which turns this into a standard linear filtering problem. However, in all cases this is just driven by sampling error. Given the high persistence in g_t, $\bar{g}$ cannot be precisely estimated. To provide further intuition, note that in the model without occasionally binding constraints:

$$\mathbb{E} \log R_t = \gamma \bar{g} - \frac{\gamma^2 \sigma^2}{2} - \log \beta.$$

Hence, conditional on the other parameters, we can estimate $\bar{g}$ by evaluating:

$$\frac{1}{\gamma}\left[\frac{1}{T}\sum_{t=1}^{T} \log R_t + \frac{\gamma^2 \sigma^2}{2} + \log \beta\right],$$

which, on our generated data set without occasionally binding constraints, gives a value of 2.58×10^{-3}, broadly in line (though somewhat worse) than our estimates.

Next, note that without occasionally binding constraints, three of our methods give identical results. This is due to the fact that the third order approximation is exact in this case, and to the fact that the exact measurement equation is linear. The standard errors do not perfectly agree here though, essentially due to a blow up in numerical errors coming from the poor-conditioning of the Hessian, which itself comes from the weak identification of $\bar{g}$.

We now examine the estimates generated by running the model without occasionally binding constraints on the data generated with an occasionally binding constraint. Here, we see a big upward bias in $\bar{g}$, and a big downwards bias in ρ and σ. The former is due to the fact that the mean of a zero lower bounded process is higher than the mean of the process without bound. The latter is due to the fact that hitting the bound prevents g_t from getting any lower, both compressing its range, and reducing its mean half-life, as returning from e.g. $g_t = -0.01$ takes longer than returning from $g_t = 0$. This illustrates the severe biases that may accompany an attempt to estimate a model without occasionally binding constraints when there clearly are such constraints in the data generating process.

The estimates using a model with occasionally binding constraints unsurprisingly fare much better. With "fast cubature", we get results very close to those using the

exact solution, suggesting that even in models for which an exact solution is not available, we are likely to get good results using "fast cubature". The estimates of ρ seem particularly impressive, which is most likely explained by the additional information coming from the zero lower bound, as ρ will now have a non-linear effect.

When we attempt to estimate β as well, understandably, the estimates of other parameters suffer somewhat. For example, β is biased downwards, and $\bar{g}$ is driven to its lower bound of 10^{-6} when we do not incorporate cubature in the simulation algorithm. To understand this, observe that in the absence of cubature, our simulation algorithm essentially ignores the bound when computing expectations. In particular, if d is a dummy which equals 0 under a first order approximation, and 1 under a second or higher order approximation, then our solution's approximation in the absence of cubature, which we shall denote $\tilde{R}_t^{(d)}$ solves:

$$1 = \beta \tilde{R}_t^{(d)} \exp\left(d\frac{\gamma^2\sigma^2}{2} - \gamma\mu_t \right).$$

Hence:

$$\begin{aligned}
\log \tilde{R}_t^{(d)} &= \gamma\mu_t - d\frac{\gamma^2\sigma^2}{2} - \log\beta \\
&= \gamma\big((1-\rho)\bar{g} + \rho g_t\big) - d\frac{\gamma^2\sigma^2}{2} - \log\beta \\
&\geq \gamma(1-\rho)\bar{g} - d\frac{\gamma^2\sigma^2}{2} - \log\beta,
\end{aligned}$$

as $g_t \geq 0$. Thus, using this approximation, a reasonable estimate of $\log\beta$ conditional on the other parameters would be:

$$\gamma(1-\rho)\bar{g} - d\frac{\gamma^2\sigma^2}{2} - \min_{t=1,\ldots,T} \log \tilde{R}_t.$$

Note though, that this is likely to be substantially biased down, since our approximation to $\mathbb{E}_t \exp(-\gamma g_{t+1})$ is biased upwards, as it ignores the impact of the bound. Now suppose we are trying to simultaneously estimate $\bar{g}$ as well. Given our approximation, the natural estimate of $\bar{g}$ is the solution for $\bar{g}$ to:

$$\gamma\big((1-\rho)\bar{g} + \rho\mathbb{E}g_t\big) = \frac{1}{T}\sum_{t=1}^{T} \log \tilde{R}_t^{(d)} + d\frac{\gamma^2\sigma^2}{2} + \log\beta,$$

which, using our estimate of $\log\beta$ is the solution for $\bar{g}$ to:

$$\gamma\rho\mathbb{E}g_t = \frac{1}{T}\sum_{t=1}^{T} \log \tilde{R}_t^{(d)} - \min_{t=1,\ldots,T} \log \tilde{R}_t\,. \tag{1}$$

Since $\mathbb{E}g_t$ depends positively on $\bar{g}$, and our estimate of $\log\beta$ is biased down, so too will be our estimate of $\bar{g}$. This in turn causes further downwards bias in the estimate of $\log\beta$, which further pushes down $\bar{g}$, and so on.

To see the magnitude of these expected biases, note that our estimation algorithm approximates the stationary distribution of g_t by a Gaussian, say $\mathrm{N}(\psi, \omega^2)$, hence, if $g \sim \mathrm{N}(\psi, \omega^2)$ and $\varepsilon \sim \mathrm{N}(0,1)$:

$$(1-\rho)\bar{g} + \rho g + \sigma\varepsilon \sim \mathrm{N}(\tilde{\psi}, \tilde{\omega}^2),$$

where $\tilde{\psi} := (1-\rho)\bar{g} + \rho\psi$ and $\tilde{\omega} := \sqrt{\rho^2\omega^2 + \sigma^2}$, so:

$$\psi = \mathbb{E}[\max\{0, (1-\rho)\bar{g} + \rho g + \sigma\varepsilon\}] = \tilde{\psi}\Phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right) + \tilde{\omega}\phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right),$$

and:

$$\begin{aligned}\omega^2 &= \mathbb{E}[(\max\{0, (1-\rho)\bar{g} + \rho g + \sigma\varepsilon\} - \psi)^2] \\ &= \tilde{\psi}^2\Phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right)\left[1 - \Phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right)\right] + \tilde{\omega}^2\left[\Phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right) - \phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right)^2\right] + \tilde{\psi}\tilde{\omega}\phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right)\left[1 - 2\Phi\left(\frac{\tilde{\psi}}{\tilde{\omega}}\right)\right],\end{aligned}$$

which implicitly define ψ and ω in terms of $\bar{g}$. Substituting $\psi \approx \mathbb{E}g_t$ into equation (1), and solving the resulting non-linear equations gives an estimate of $\bar{g}$ on our artificial data of -0.0080.[12] Since we constrain $\bar{g}$ to be greater than 10^{-6} when we estimate, this means that we should be unsurprised that we hit this bound. Given this, the implied estimate for β is:

$$\exp\left[10^{-6}\gamma(1-\rho) - d\frac{\gamma^2\sigma^2}{2} - \min_{t=1,\ldots,T}\log\tilde{R}_t\right],$$

which, on our data, equals 0.9759 under a first order approximation, and 0.9753 under a third order approximation. These estimates are closely in line with the estimates produced by our procedure, fully explaining our results.

Luckily, when we include cubature in the underlying simulation algorithm, ensuring that agent's expectations are truly rational, all of these biases disappear. Indeed, our estimates with "fast cubature" including the bound are the most accurate estimates of both β and $\bar{g}$ from any of our estimation runs. Given that "fast cubature" is computationally tractable even on large models, this suggests that we should be able to use our estimation procedure to get reliable estimates even in the medium scale DSGE models used by policy makers.

4. Performing smoothing

4.1. Algorithm

The smoother contains a forward pass that follows the algorithm given in section 2.2, followed by a backwards pass that uses information saved from the forward pass. Since a smoother is generally only run once, we anticipate users wishing to use the Genz and Keister (1996) rules for the forward pass, to ensure high accuracy in the presence of strong non-linearities.

A natural strategy for smoothing were we approximating the distribution of the state by a Gaussian would be as follows.[13] We would first calculate an approximation to

[12] Replication code for this is contained in the "NaturalEstimate.m" file within the "Examples\BoundedProductivityEstimation" folder of DynareOBC.

[13] This broadly follows both the derivation of the standard Rauch Tung Striebel (1965) smoother, and the smoother for the non-augmented Gaussian cubature Kalman filter of Arasaratnam and Haykin (2011).

$\mathrm{cov}[x_{t-1}, x_t|\mathcal{F}_{t-1}]$ just as we calculated the other moments in section 2.2. We could then calculate a Gaussian approximation to $\begin{bmatrix} x_t \\ w_{t-1} \end{bmatrix}|\mathcal{F}_{t-1}$, with marginals agreeing with the results of section 2.2, using standard properties of the Gaussian distribution along with the fact that $x_t|(w_{t-1}, \mathcal{F}_{t-1}) \overset{d}{=} x_t|(x_{t-1}, \mathcal{F}_{t-1})$. This in turn would give us $w_{t-1}|(x_t, \mathcal{F}_T) \overset{d}{=} w_{t-1}|(x_t, \mathcal{F}_{t-1})$, as x_t is Markov. However, such a strategy cannot work with an extended skew t-distribution. To see this, note that for $\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}|\mathcal{F}_{t-1}$ or $\begin{bmatrix} x_t \\ w_{t-1} \end{bmatrix}|\mathcal{F}_{t-1}$ to be extended skew t-distributed, $x_t|\mathcal{F}_{t-1}$ and $x_{t-1}|\mathcal{F}_{t-1}$ would have to have identical "degrees of freedom" parameters, i.e. we would need $\nu_{t|t-1} = \nu_{t-1|t-1}$. Aside from being a probability zero event in a sufficiently general model, this would also imply that $\nu_{t|t} = \nu_{t|t-1} + n_{m,t} = \nu_{t-1|t-1} + n_{m,t}$, so $\nu_{t|t} \to \infty$ as $t \to \infty$, giving an unsatisfactory degree of non-stationarity to the filter.

Instead, we rely on the standard forward-backward smoothing identity:

$$\begin{aligned} f(w_{t-1}|\mathcal{F}_T) &= \int f(w_{t-1}, x_t|\mathcal{F}_T)\, dx_t = \int f(w_{t-1}|x_t, \mathcal{F}_T) f(x_t|\mathcal{F}_T)\, dx_t \\ &= \int f(w_{t-1}|x_t, \mathcal{F}_{t-1}) f(x_t|\mathcal{F}_T)\, dx_t = \int \frac{f(x_t|w_{t-1}, \mathcal{F}_{t-1}) f(w_{t-1}|\mathcal{F}_{t-1})}{f(x_t|\mathcal{F}_{t-1})} f(x_t|\mathcal{F}_T)\, dx_t \\ &= \int \frac{f(x_t|x_{t-1}) f(w_{t-1}|\mathcal{F}_{t-1})}{f(x_t|\mathcal{F}_{t-1})} f(x_t|\mathcal{F}_T)\, dx_t = \mathbb{E}\left[\frac{f(x_t|\mathcal{F}_T)}{f(x_t|\mathcal{F}_{t-1})} \middle| x_{t-1} \right] f(w_{t-1}|\mathcal{F}_{t-1}) \end{aligned}$$

where we have used the Markov property of x_t at the start of the second and third lines. Now, suppose that:

$$x_t|\mathcal{F}_T \overset{\text{approx}}{\sim} \mathrm{EST}\big(\hat{x}_{t|T}, P^*_{t|T}, \delta^*_{t|T}, \tau_{t|T}, \nu_{t|T}\big),$$

and recall that:

$$w_{t-1}|\mathcal{F}_{t-1} \overset{\text{approx}}{\sim} \mathrm{EST}(\widehat{w}_{t-1|t-1}, P_{t-1|t-1}, \delta_{t-1|t-1}, \tau_{t-1|t-1}, \nu_{t-1|t-1}).$$

Then, much as in section 2.2, define $S_{t-1|t-1} := S_{P_{t-1|t-1}, \kappa^*}$, and let $N_0, N_{10} \in \mathbb{R}$ be draws from $\mathrm{N}(0,1)$ and $N_{11} \in \mathbb{R}^{k_{t-1|t-1}}$ be a draw from $\mathrm{N}\big(0_{k_{t-1|t-1}}, I_{k_{t-1|t-1}}\big)$, where $k_{t-1|t-1} := \mathrm{cols}\, S_{t-1|t-1}$, so the distribution of $w_{t-1}|\mathcal{F}_{t-1}$ is equal to that of:

$$\begin{aligned} w_{t-1|t-1}(N) := \widehat{w}_{t-1|t-1} &+ S_{t-1|t-1} N_{11} F^{-1}_{\sqrt{\frac{\nu+1}{\chi^2_{\nu+1}}}}\big(\Phi_1(N_{10})\big) \sqrt{\frac{\nu + F^{-1}_{E_{\nu,\tau}}\big(\Phi_1(N_0)\big)^2}{\nu+1}} \\ &+ \delta_{t-1|t-1} F^{-1}_{E_{\nu,\tau}}\big(\Phi_1(N_0)\big), \end{aligned}$$

where $\nu = \nu_{t-1|t-1}$, $\tau = \tau_{t-1|t-1}$ and $N := \begin{bmatrix} N_0 \\ N_{10} \\ N_{11} \end{bmatrix}$. Then, by the derived forward-backward smoothing identity, for any function q:

$$\mathbb{E}[q(w_{t-1})|\mathcal{F}_T] \approx \int_{\mathbb{R}^{(2+k_{t-1|t-1}+n_\varepsilon)}} q\big(w_{t-1|t-1}(N)\big) r_t(N, \varepsilon) \phi_{2+k_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix} N \\ \varepsilon \end{bmatrix}\right) d\begin{bmatrix} N \\ \varepsilon \end{bmatrix},$$

where:

$$r_t(N,\varepsilon) = \frac{f_{\text{EST}_{\hat{x}_{t|T},\llbracket P^*_{t|T}\rrbracket,\delta^*_{t|T},\tau_{t|T},\nu_{t|T}}}\Big(g_{t,1}(w_{t-1|t-1,1}(N),\varepsilon)\Big)}{f_{\text{EST}_{\hat{x}_{t|t-1},\llbracket P^*_{t|t-1}\rrbracket,\delta^*_{t|t-1},\tau_{t|t-1},\nu_{t|t-1}}}\Big(g_{t,1}(w_{t-1|t-1,1}(N),\varepsilon)\Big)},$$

and where $g_{t,1}$ and $w_{t-1|t-1,1}$ give the first n_x outputs of g_t and $w_{t-1|t-1}$ respectively. The required integral here is again just a Gaussian one, and so may be efficiently evaluated using the methods detailed in section 2.3.

We note that:

$$1 = \mathbb{E}[1|\mathcal{F}_T] \approx \int_{\mathbb{R}^{(2+k_{t-1|t-1}+n_\varepsilon)}} r_t(N,\varepsilon)\phi_{2+k_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix}N\\ \varepsilon\end{bmatrix}\right) d\begin{bmatrix}N\\ \varepsilon\end{bmatrix},$$

so the term $r_t(N,\varepsilon)$ is effectively a weight. Thus, for increased numerical robustness, it is better to represent $\mathbb{E}[q(w_{t-1})|\mathcal{F}_T]$ in our numerical calculations as:

$$\mathbb{E}[q(w_{t-1})|\mathcal{F}_T] \approx \frac{\int_{\mathbb{R}^{(2+k_{t-1|t-1}+n_\varepsilon)}} q\big(w_{t-1|t-1}(N)\big) r_t(N,\varepsilon)\phi_{2+k_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix}N\\ \varepsilon\end{bmatrix}\right) d\begin{bmatrix}N\\ \varepsilon\end{bmatrix}}{\int_{\mathbb{R}^{(2+k_{t-1|t-1}+n_\varepsilon)}} r_t(N,\varepsilon)\phi_{2+k_{t-1|t-1}+n_\varepsilon}\left(\begin{bmatrix}N\\ \varepsilon\end{bmatrix}\right) d\begin{bmatrix}N\\ \varepsilon\end{bmatrix}}.$$

Using this result, just as in section 2.2, we may evaluate the integrals necessary to find $\widehat{w}_{t-1|T}$, $P_{t-1|T}$, $\delta_{t-1|T}$, $\tau_{t-1|T}$ and $\nu_{t-1|T}$ such that:

$$w_{t-1}|\mathcal{F}_T \overset{\text{approx}}{\sim} \text{EST}(\widehat{w}_{t-1|T}, P_{t-1|T}, \delta_{t-1|T}, \tau_{t-1|T}, \nu_{t-1|T}).$$

Defining, as usual , $P^*_{t-1|T} := P_{t-1|T,11}$ to be the upper left $n_x \times n_x$ block of $P_{t-1|T}$ and $\hat{x}_{t-1|T} := \widehat{w}_{t-1|T,1}$ & $\delta^*_{t-1|T} := \delta_{t-1|T,1}$ to be the top n_x rows of $\widehat{w}_{t-1|T}$ & $\delta_{t-1|T}$, respectively, then by Proposition 3 of Arellano-Valle and Genton (2010):

$$x_{t-1}|\mathcal{F}_T \overset{\text{approx}}{\sim} \text{EST}\big(\hat{x}_{t-1|T}, P^*_{t-1|T}, \delta^*_{t-1|T}, \tau_{t-1|T}, \nu_{t-1|T}\big).$$

Therefore, by induction, we can calculate an approximation to the distribution of $x_t|\mathcal{F}_T$ and $w_t|\mathcal{F}_T$ for all $t \in \{1, \dots, T\}$, giving our smoother.

4.2. Application: Which shocks caused the great recessions?

TODO in Christiano, Motto, and Rostagno (2014)

5. Further details on the DynareOBC toolkit

Code implementing the estimation algorithm discussed here is contained in the author's "DynareOBC" toolkit which is a suite of MATLAB files designed to augment the abilities of Dynare (Adjemian et al. 2011). The toolkit may be freely downloaded from http://github.org/tholden/dynareOBC, and this site also contains complete documentation for its assorted options.[14] To use it for simulation, one merely has to include a "max", "min" or "abs" in the MOD file describing the DSGE model to be simulated, and then to invoke DynareOBC with the MATLAB command "dynareOBC ModFileName.MOD". Using it for estimation is almost as easy, and the examples in the "Examples\BoundedProductivityEstimation" sub-folder should make it clear to the user how to proceed.

[14] A PDF of the toolkit's documentation is available from: https://github.com/tholden/dynareOBC/raw/master/ReadMe.pdf.

While base Dynare now supports using the cubature Kalman filter for estimating second order approximations to models (Adjemian et al. 2016), it does not implement either the state initialization, or the state space reduction technique developed here; it does not support third order approximations; it does not calculate quasi-Maximum likelihood standard errors; and it cannot handle occasionally binding constraints. In all these regards then, the DynareOBC estimation procedure is an improvement on that already contained in base Dynare.

6. Conclusion

This paper has presented an efficient algorithm for estimating non-linear models, including those with occasionally binding constraints. Thanks to the algorithm's dimension reduction techniques, the algorithm keeps the costs of forming predictive distributions manageable, allowing it to be used even on models for which simulation is expensive, such as those with occasionally binding constraints. We went on to show that identification is easier in non-linear models, and that estimating ignoring occasionally binding constraints may introduce substantial biases. The latter is particularly relevant for the zero lower bound on nominal interest rates, given the amount of time many economies have now spent at the bound. Macro-econometricians thus have no choice but to include the occasionally binding constraint in their model when they estimate medium-scale DSGE models, if they want to use up to date data. Luckily, the algorithms presented in this paper will readily scale up to handle such models.

Code implementing all of the algorithms discussed here is contained in the author's "DynareOBC" toolkit which augments the abilities of Dynare (Adjemian et al. 2011) with the ability to solve and estimate models with occasionally binding constraints.

7. References

Adjemian, Stéphane, Houtan Bastani, Fréderic Karamé, Michel Juillard, Junior Maih, Ferhat Mihoubi, George Perendia, Johannes Pfeifer, Marco Ratto, and Sébastien Villemot. 2011. *Dynare: Reference Manual Version 4*. CEPREMAP.

Adjemian, Stéphane, Michel Juillard, Fréderic Karamé, and Ferhat Mihoubi. 2016. *New Methods for Estimating Nonlinear Dynamic General Equilibrium Models with Particle Filters: A New Software Tool to Be Distributed via Dynare*. MACFINROBODS.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein. 2010. 'Particle Markov Chain Monte Carlo Methods'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (3): 269–342.

Arasaratnam, Ienkaran, and Simon Haykin. 2009. 'Cubature Kalman Filters'. *Automatic Control, IEEE Transactions on* 54 (6): 1254–1269.

———. 2011. 'Cubature Kalman Smoothers'. *Automatica* 47 (10): 2245–2250.

Arellano-Valle, Reinaldo B., and Marc G. Genton. 2010. 'Multivariate Extended Skew-T Distributions and Related Families'. *Metron - International Journal of Statistics* LXVIII (3): 201–234.

Bollerslev, Tim, and Jeffrey M Wooldridge. 1992. 'Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances'. *Econometric Reviews* 11 (2): 143–172.

Canova, F. 2007. *Methods for Applied Macroeconomic Research*. Princeton University Press.

Christiano, Lawrence J., Roberto Motto, and Massimo Rostagno. 2014. 'Risk Shocks'. *American Economic Review* 104 (1): 27–65.

Cools, Ronald. 2003. 'An Encyclopaedia of Cubature Formulas'. *Oberwolfach Special Issue* 19 (3) (June): 445–453.

Duffie, Darrell, and Kenneth J Singleton. 1990. *Simulated Moments Estimation of Markov Models of Asset Prices*. National Bureau of Economic Research Cambridge, Mass., USA.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, and Juan F. Rubio-Ramírez. 2015. 'Estimating Dynamic Equilibrium Models with Stochastic Volatility'. *Journal of Econometrics* 185 (1): 216–229.

Fernández-Villaverde, Jesús, and Juan F. Rubio-Ramírez. 2007. 'Estimating Macroeconomic Models: A Likelihood Approach'. *The Review of Economic Studies* 74 (4): 1059–1087.

Genz, Alan, and Bradley D. Keister. 1996. 'Fully Symmetric Interpolatory Rules for Multiple Integrals over Infinite Regions with Gaussian Weight'. *Journal of Computational and Applied Mathematics* 71 (2): 299–309.

Guerrieri, Luca, and Matteo Iacoviello. 2015. *Collateral Constraints and Macroeconomic Asymmetries*. National Bank of Poland Working Papers. National Bank of Poland, Economic Institute.

Gust, Christopher J., Edward Herbst, Matthew E. Smith, and David Lopez-Salido. 2016. 'The Empirical Implications of the Interest-Rate Lower Bound'. In . Banca d'Italia.

Hansen, Nikolaus. 2006. 'The CMA Evolution Strategy: A Comparing Review'. In *Towards a New Evolutionary Computation*, 75–102. Springer.

Hansen, Nikolaus, Andreas Ostermeier, and Andreas Gawelczyk. 1995. 'On the Adaptation of Arbitrary Normal Mutation Distributions in Evolution Strategies: The Generating Set Adaptation.' In *ICGA*, 57–64. Citeseer.

Herbst, Edward, and Frank Schorfheide. 2014. 'Sequential Monte Carlo Sampling For Dsge Models'. *Journal of Applied Econometrics* 29 (7) (November): 1073–1098.

Higham, Nicholas J. 1988. 'Computing a Nearest Symmetric Positive Semidefinite Matrix'. *Linear Algebra and Its Applications* 103 (May 1): 103–118.

Holden, Tom D. 2014. 'Chapter 5: Estimating Non-Linear Models'. In *An Advanced Course On The Science and Art of DSGE Modelling*, by Szabolcs Deak, Tom D. Holden, and Antonio Mele. University of Surrey.

———. 2016a. 'Existence and Uniqueness of Solutions to Dynamic Models with Occasionally Binding Constraints.' Unpublished.

———. 2016b. 'Computation of Solutions to Dynamic Models with Occasionally Binding Constraints.' Unpublished.

———. 2016c. 'Estimation of Dynamic Models with Occasionally Binding Constraints.' Zenodo.

Kim, Sunghyun Henry, Jinill Kim, Ernst Schaumburg, and Christopher A. Sims. 2008. 'Calculating and Using Second Order Accurate Solutions of Discrete Time Dynamic Equilibrium Models'. *Journal of Economic Dynamics and Control* 32 (11): 3397–3414.

Kollmann, Robert. 2013. *Tractable Latent State Filtering for Non-Linear DSGE Models Using a Second-Order Approximation*. Working Papers ECARES. ULB – Universite Libre de Bruxelles.

———. 2016. *Tractable Likelihood-Based Estimation of Non-Linear DSGE Models Using Higher-Order Approximations*. MPRA Paper. University Library of Munich, Germany.

Li, Pengfei, Jianping Yu, Mingjie Wan, Jianjun Huang, and Jingxiong Huang. 2009. 'The Augmented Form of Cubature Kalman Filter and Quadrature Kalman Filter for Additive Noise'. In *Information, Computing and Telecommunication, 2009. YC-ICT '09. IEEE Youth Conference on*, 295–298.

Meyer-Gohde, Alexander. 2014. *Risky Linear Approximations*. SFB 649 Discussion Papers. Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Pitt, Michael K. 2002. *Smooth Particle Filters for Likelihood Evaluation and Maximisation*. The Warwick Economics Research Paper Series (TWERPS). University of Warwick, Department of Economics.

Rauch, Herbert E., C. T. Striebel, and F. Tung. 1965. 'Maximum Likelihood Estimates of Linear Dynamic Systems'. *AIAA Journal* 3 (8): 1445–1450.

Smets, Frank, and Rafael Wouters. 2003. 'An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area'. *Journal of the European Economic Association* 1 (5): 1123–1175.

Tauchen, George E, and A Ronald Gallant. 1995. *Which Moments to Match*.

Wan, Eric A., and Rudolph Van Der Merwe. 2000. 'The Unscented Kalman Filter for Nonlinear Estimation'. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, 153–158. Ieee.

White, Halbert. 1982. 'Maximum Likelihood Estimation of Misspecified Models'. *Econometrica* 50: 1–25.