Discussion Paper Series – CRC TR 224

Discussion Paper No. 355
Project C 05

Ensemble MCMC Sampling for DSGE Models

Gregor Boehl[1]

June 2022

[1] University of Bonn, Email: gboehl@uni-bonn.de

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

Ensemble MCMC Sampling for DSGE Models

Latest version: http://gregorboehl.com/live/ademc_boehl.pdf

Gregor Boehl

University of Bonn

June 14, 2022

Abstract

This paper develops an *adaptive differential evolution Markov chain Monte Carlo* (ADEMC) sampler. The sampler satisfies five requirements that make it suitable especially for the estimation of models with high-dimensional posterior distributions and which are computationally expensive to evaluate: (i) A large number of chains (the "ensemble") where the number of chains scales inversely (nearly one-to-one) with the number of necessary ensemble iterations until convergence, (ii) fast burn-in and convergence (thereby superseding the need for numerical optimization), (iii) good performance for bimodal distributions, (iv) an endogenous proposal density generated from the state of the full ensemble, which (v) respects the bounds of prior distribution. Consequently, ADEMC is straightforward to parallelize. I use the sampler to estimate a heterogeneous agent New Keynesian (HANK) model including the micro parameters linked to the stationary distribution of the model.

Keywords: Bayesian Estimation, Monte Carlo Methods, DSGE Models, Heterogeneous Agents
JEL: C11, C13, C15, E10

1 Introduction

Since the pioneering work of Geweke (1999) and Schorfheide (2000), Bayesian estimations have found their way into the toolboxes of macroeconomic researchers at universities and central banks alike. They are used extensively to bring modern New Keynesian-type DSGE models to the data, to evaluate their empirical performance, and to asses the

*Address: Institute for Macroeconomics and Econometrics, University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany. I am grateful to Alexander Meyer-Gohde, Frank Schorfheide, Felix Strobel and participants of several conferences and seminars for discussions and helpful comments on the contents of this paper. Part of the research leading to the results in this paper has received financial support from the Alfred P. Sloan Foundation under the grant agreement G-2016-7176 for the Macroeconomic Model Comparison Initiative (MMCI) at the Institute for Monetary and Financial Stability. I also gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG) under CRC-TR 224 (project C05) and under project number 441540692.
Email address: gboehl@uni-bonn.de
URL: https://gregorboehl.com

effects of macroeconomic policy. Bayesian estimations allow to combine all available prior information on the model parameters with the likelihood of these parameters, and provide means to sensibly quantify the uncertainty of the resulting parameter estimates. They are, however, challenging on a technical domain: the estimation of DSGE model requires sampling from the posterior distribution, which is an high-dimensional object with ex-ante unknown properties. The quality, and thereby the economic usefulness of these estimations crucially depends on our ability to precisely approximate the posterior distribution.

This paper seeks to substantiate this ability for particularly complex models by introducing a novel sampling algorithm that is inspired by contemporary methods in the field of astrophysics: the *adaptive differential-evolution Markov chain Monte Carlo* (ADEMC) method. This method has three serious advantages over conventional methods. First, ADEMC allows to sample from high-dimensional, ill-conditioned, and potentially bimodal distributions. Second, ADEMC is "embarrassingly parallelizable", allowing the estimation of models whose likelihood is computationally expensive to evaluate. Third, ADEMC is equally efficient for burn-in – that is, during convergence to the high-density region of the posterior – and for posterior sampling. The first point is important because DSGE models are especially hard to sample from, e.g. because of indeterminacy regions where the likelihood is zero.[1] The second point is very relevant because current DSGE models are becoming more and more expensive to evaluate, either because of nonlinearities or because of massive heterogeneities across multiple agents.[2] The third point is crucial because often, a significant amount of computational resources is spend not during posterior sampling but for mode finding. Such mode finding methods are difficult to parallelize and tend to converge to local maxima only.

ADEMC is a member of the broader class of ensemble MCMC methods. Instead of using a single or small number of state-dependent chains (such as e.g. the Metropolis algorithm), ensemble samplers use a large number of chains (the "ensemble"). For each iteration, proposals are generated based on the current state of the full ensemble and, as the ensemble evolves over time, proposal steps naturally adapt direction and scale of the posterior distribution. These methods are essentially parameter-free, self-tuning and do – if at all – only require decisions on hyper parameters. This represents a fourth, major advantage of ADEMC. Straightforward parallelization then comes naturally from running a large number of chains as long as these are, in each iteration, independent. Hence, ensemble MCMC methods form a simple, yet potentially extremely powerful tool. I build on earlier work of Ter Braak (2006) and Nelson et al. (2013) by adapting their differential evolution algorithm to DSGE models. Two central features are added: and adaptation stage, which I find to greatly reduce burn-in times, and the distinction between parameter and proposal space, which boosts acceptance ratios, thereby increasing robustness and sampling efficiency.

[1] An additional problem with nonlinear models constitutes trough the process of nonlinear filtering, which is generally also based on sampling. For this reason, likelihood estimates will, to some degree, be noisy. Neither "classic" samplers, nor local optimization methods are well suited to deal with such problems reliably, in particular given that economic models can have quite high dimensional parameter spaces.

[2] See, e.g. Boehl and Strobel (2020), for the estimation of medium scale DSGE models with the zero-lower bound on nominal interest rates as an example for nonlinear estimation, or Bayer et al. (2020) for the estimation of heterogeneous agent models.

In this paper I asses the performance of ADEMC on three typical problems that are often encountered when estimating DSGE models. I first evaluate the algorithm's capability to deal with high dimensional odd-shaped, bimodal distributions. I document that ADEMC performs well as long as the two modes are not fully disconnected. I then test the performance of the sampler on the estimation exercise from Smets and Wouters (2007). Independently of the number of chains used, ADEMC indeed returns the original parameter estimates. For the given example, I find that convergence times roughly scale one-to-one with the number of chains, which suggests that the losses through parallelization mainly amount to the computational overhead of the serialization. Finally, I estimate a heterogeneous agents New Keynesian model, including the micro parameters governing its steady state distribution. This exercise is important because it was so far deemed impossible due to the large computational costs associated with one single likelihood evaluation of the model.

Literature

The workhorse of Bayesian estimation is, to date, still the random walk Metropolis Hastings (RWMH) algorithm. The shortcomings of RWMH are manifold and well documented (e.g. Chib and Ramamurthy, 2010; Herbst and Schorfheide, 2015). The main issue is that convergence of RWMH to the posterior distribution can be extremely slow, and sampling from ill-shaped or bimodal distributions is either very difficult, or requires additional tweaking. To circumvent the first problem, numerical optimization routines are frequently used to find a good initial guess for RWMH. These routines are often slow as well, and not very robust when applied to more complicated posterior distributions.[3] With RWMH going back to the seminal work of Metropolis et al. (1953) and Hastings (1970), it is arguably time to take a sidestep and to acknowledge (and lever) process in this area in other fields. In particular, research in the field of astrophysics has since made considerable progress on this frontier.[4]

The ADEMC sampler introduced here waives the need for explicit numerical mode finding. Rather, it can by itself be seen as an stochastic optimizer since it is based on the concept of differential evolution (DE). This concept was initially suggested by Storn and Price (1997) as a heuristic global optimization method. A *population* of candidate solutions evolve and mutate iteratively, thereby exchanging candidate solutions. The method is shown to perform well on ill-behaved, constrained or high dimensional optimization problems for which no gradient information is available. Strens et al. (2002) first propose to use proposals based on differential evolution in the context of Monte Carlo sampling. Independently, Ter Braak (2006) develops the idea of combining a DE-update proposal with Metropolis acceptance steps, which yields a Differential evolution MCMC algorithm. Nelson et al. (2013) iterate on this work and provide general recommendations for the use of the algorithm.

Goodman and Weare (2010) introduce *Ensemble MCMC* coneptionally. The authors develop the idea of an ensemble of Markov chains iterating on a distribution, thereby

[3] E.g., they are apt to "getting stuck" at local maxima. Another problem is that they are often iterative and parallelization is hence not straightforward. Another issue is that the routine may run into regions where the likelihood is zero due to issues with the model solution.

[4] A notable exception from withing economics is Herbst and Schorfheide (2014) which, however, at its core still relies on RWMH chains.

using proposals similar to those from the numerical optimization method of Nelder and Mead (1965). They show that such sampler can be affine invariant and hence "be uniformly effective over all the convex bodies of a given dimension regardless of their shape." The authors provide examples where ensemble MCMC significantly outperforms RWMH. While Goodman and Weare (2010) indeed performs well in terms of sampling efficiency, I find that it is rather slow to converge to the posterior distribution when starting off from a bad initial distribution, and does not perform well for bimodal distributions. The recent success of of Ensemble MCMC methods may also come from its excellent implementation in the free and open source packet *emcee* (Foreman-Mackey et al., 2013), which also provides routines for efficient parallelization.

The recent rise in the implementation of *automatic differentiation* (AD, e.g. in the Python packages *JAX* or *TensorFlow*, or in the new *Julia* programming language) have renewed interest in the Hamiltonian Monte Carlo (HMC) method (Duane et al., 1987). HMC requires the derivatives of the likelihood function, which are normally hard to evaluate since its calculation via finite difference methods is computationally very expensive. However, AD provides computationally efficient means to calculate these derivatives. Deviates of HMC are for example implemented in the *STAN* framework. While HMC indeed clearly outperforms RWMH in terms of sampling efficiency and also in its capability to sample from more complex distributions, it comes with two disadvantages. First, it is required to implement the likelihood function – and hence the complete model and filtering routines – in a framework that allows for AD. Secondly, HMC is a single-particle method which is not straightforward to parallelize, which rules out its application to models whose likelihood is already expensive to evaluate.

A prominent exception from single-particle samplers in the spirit of RWMH *in econometrics* is the Sequential Monte Carlo (SMC) method suggested in Herbst and Schorfheide (2014). The authors propose to run many RWMH in parallel, interrupted by several resampling and re-weighting steps. They additionally develop a tempering scheme to allow for the smooth transition the group of chains from the prior towards the posterior. The tempering schedule is necessary because the RWMH chains do not perform well in exploring the parameter space. While this approach successfully tackles many of the flaws of RWMH, it is yet relatively time consuming. Each individual chain will still have the convergence speed of an individual RWMH, which is further slowed down by the tempering scheme. In contrast, the proposal density of ADEMC is endogenous and, through the adaptation extensions, chains converge more quickly.

The rest of the paper is structured as follows. Section 2 explains the basic ADEMC algorithm. Section 3 studies the performance of the algorithm on a high dimensional bimodal distribution. In section 4 the sampler is used on the Smets-Wouters model and in section 5 it is applied to the estimation of a large-scale HANK model. Section 6 concludes.

2 Sampling

Let $\pi(x)$ be the probably density of a target distribution with $x \in \mathbb{R}^n$. In practice, $\pi(x)$ is the posterior density $\pi(x) = p(x|Y)$, which is given by

$$p(x|Y) = \frac{p(Y|x)p(x)}{p(Y)}, \tag{1}$$

where $p(Y|x)$ is the likelihood function, $p(x)$ the prior density, and

$$p(Y) = \int p(Y|x)p(x)dx \tag{2}$$

is an unknown constant for given data. The prior $p(x)$ is commonly specified such that it is straightforward to evaluate and $p(Y|x)$ can be calculated using various Bayesian filtering techniques.

The classic random walk Metropolis-Hastings algorithm (RWMH) goes as follows. Start with a *single* parameter vector X_h. A new proposal is generated by $\hat{X}_h = X_h + \varepsilon_h$ where $\varepsilon_h \sim \mathcal{N}(0, \Sigma)$ is called the *proposal distribution*. The proposal $\hat{X}_h$ is then accepted with the Metropolis acceptance probability which is calculated as defined further below in (8). If it is accepted, set $X_{h+1} = \hat{X}_h$. Otherwise, set $X_{h+1} = X_h$. A large literature discusses the properties of RWMH, see e.g. Sokal (1997) or Roberts and Rosenthal (2001).

Clearly, the performance of the algorithm crucially depends on the choice of the covariance matrix Σ. This is problematic since Σ has more than $n^2/2$ degrees of freedom and it is very challenging to determine ex-ante which choice of Σ will maintain a high acceptance ratio while still exploring the posterior distribution to a satisfactory degree. To maintain a sufficiently large acceptance ratio, Σ is often scaled down to relatively small values. Consequently, RWMH is very slow to converge to the high density region of the posterior (so-called "burn-in"). To speed up computation, RWMH is thus often used subsequent to a numerical optimization routine in the aspiration to obtain a good starting value.

ADEMC aims to supersede conventional samplers in five dimensions:

i) Convergence speed scales with the number of chains (almost) one-to-one.
ii) Fast burn-in to the high density region of the posterior.
iii) Good performance for bimodel distributions.
iv) The proposal density is generated endogenously from the state of all chains.
v) The proposal density respects the bounds of prior distribution.

While each of these points addresses a generic sampling problem, point v) is of particular importance for the context of DSGE models. The reason is that DSGE models often feature exogenous AR(1) processes which are estimated to be close to a unit root. However, since the prior of the AR(1)-coefficient is bounded above by one, parameter values close to unit roots will often cause poor sampling performance because any proposals with values of the AR-coefficient larger one will be rejected.

2.1 Proposal space

To circumvent this problem define the *prior space* $\mathbb{X} : x \in \mathbb{X} \Leftrightarrow p(x) > 0$ to be the space of all parameter combinations for which the prior density is positive and let the *proposal space* $\mathbb{Z} = \mathbb{R}^n$ be unbounded. Let f_b be a *bijective map*

$$f_b : \mathbb{R}^n \to \mathbb{X} \tag{3}$$

such that for any $x \in \mathbb{X}$ there exists a unique $z \in \mathbb{R}^n$ for which $f_b(z) = x$. This implies that f_b is always uniquely invertible, and that by definition, f_b maps within the bounds of the prior distribution whereas its domain is unbounded. In the following, f_b will be

used to map samples from the proposal space into prior space, thereby maintaining that every sample has a positive prior density.

At its core, and in the spirit of Ter Braak (2006), ADEMC replaces the random-walk proposal distribution of MH with a proposal that follows the *differential evolution* concept. To that end, ADEMC holds an *ensemble* of chains at each iteration s (or, using the evolutionary terminology, at each "generation" s). Define

$$\mathbf{X}_s = (X_{s,1}, \ldots, X_{s,n_c}), \tag{4}$$

to be an ensemble of n_c *chains* (or "particles", in SMC terminology), with individual chains $X_{s,i}$ indexed by $i = 1, 2, \ldots, n_c$. While $\mathbf{X}_s$ holds the ensemble in prior space, let

$$\mathbf{Z}_s = (Z_{s,1}, \ldots, Z_{s,n_c}) = (f_b^{-1}(X_{s,1}), \ldots, f_b^{-1}(X_{s,n_c})) \tag{5}$$

be its representation in proposal space.

2.2 Adaptation stage

Before introducing the differential evolution proposal, another step will help to increase burn-in speed (point ii) from above): the adaptation stage. In this stage, ensemble members that are sufficiently far away from the current mode are substituted for better candidates. For this purpose, define a threshold value δ very close to zero. Divide the ensemble in three subsets $\mathbf{X}_s = \{\mathbf{X}_{1,s}, \mathbf{X}_{2,s}, \mathbf{X}_{3,s}\}$ of (almost) equal sizes, and randomly assign to each chain i another chain $j \neq i$ of the same subset. The *candidate chain* $\tilde{X}_{s,i}$ is then generated by comparing $X_{s,i}$ and $X_{s,j}$:

$$\tilde{X}_{s,i} = \begin{cases} X_{s,j} & \text{if } \pi(X_{s,i})/\pi(X_{s,j}) < \delta, \\ X_{s,i} & \text{else.} \end{cases} \tag{6}$$

$X_{s,j}$ is hence *adapted* if the Metropolis acceptance probability of $X_{s,i}$ contra $X_{s,j}$ is below δ. Thereby, the value of δ must be chosen sufficiently low to maintain that adaptations only occur during burn-in but not once the sampler is converging to the final posterior distribution.

2.3 Evolution stage

For each chain i in iteration s, create a displacement vector from two chains k and $l \neq k$ from *different* subset than i, and add this displacement vector to chain i. More formally, the proposed replacement for chain $X_{s,i}$ is

$$\hat{Z}_{s,i} = \tilde{Z}_{s,i} + \gamma(Z_{s,j} - Z_{s,k}) + \epsilon_{s,i} \qquad \forall i = 1, 2, \ldots, n_c, \tag{7}$$

where γ is a scaling factor, $\epsilon_{s,i}$ is some (very) small noise, and $\tilde{Z}_{s,i} = f_b^{-1}(\tilde{X}_{s,i})$ is the bijective transform of the candidate chain associated with $X_{i,s}$, which results from the adaptation stage. Generate ensemble $\mathbf{X}_{s+1}$ by accepting the proposals $\hat{X}_{s,i} = f_b(\hat{Z}_{s,i})$ with a Metropolis acceptance probability of

$$P(X_{s+1,i} = \hat{X}_{s,i}) = \min\left\{1, \frac{\pi(\hat{X}_{s,i})}{\pi(X_{s,i})}\right\} \qquad \forall i = 1, 2, \ldots, n_c, \tag{8}$$

or reject $\hat{X}_{s,i}$ and set $X_{s+1,i} = X_{s,i}$ with probability

$$P(X_{s+1,i} = X_{s,i}) = 1 - P(X_{s+1,i} = \hat{X}_{s,i}). \tag{9}$$

Note that in the case of a rejection of the proposal, $X_{s+1,i}$ remains independent of the candidate chain from the adaptation stage. As the ensemble evolves over time, proposal steps naturally adapt direction and scale of the posterior distribution.

2.4 Parameterization

Theorem 1 in Ter Braak (2006) shows that the unique stationary distribution of DE-MCMC has the PDF $\pi(x)$. This result applies one-to-one to the stationary distribution of $\mathbb{X}$ (in the stationary distribution, no more adaptations are proposed for a reasonably chosen δ). If $\pi(x)$ follows a Gaussian distribution, after convergence each individual proposal $\hat{X}_{s,i}$ is of the same form as an RWMH proposal.[5] Respectively, if the target distribution is near-Gaussian, the optimal choice is $\gamma = \frac{2.38}{\sqrt{2n}}$ from the RWMH literature (e.g. Roberts and Rosenthal, 2001), which is expected to give an acceptance probability of 23% for large n.

The value of the adaptation threshold should be chosen to be close to zero, and be low enough to maintain that adaptations are proposed (and accepted) during burn-in, but never during posterior sampling. A larger value of δ can speed-up burn-in, but potentially makes it more likely to end up in a local maximum. For all exercises in this paper I set $\delta = \exp(-100)$ and find that results are robust to variations to this number.

Sections 4 and Appendix B investigate the question of the optimal number of chains n_c vs. the number of iterations S in detail. As I document there, a value of $n_c \geq 4n$ is often sufficient to guarantee that the ensemble have full rank after the adaptation stage. However, using a larger number of chains does neither have severe advantages nor drawbacks. In other words, as long as $n_c \geq 4n$, in terms of pure function evaluations, a larger number of chains compensates one-to-one for fewer ensemble iterations.

A straightforward choice for the functional form of the bijective transform f_b I chose $x_p = \exp(z_p)$ for priors following a gamma and inverse gamma distribution (hence for distributions bounded below by zero), and the standard logistic function $x_p = 1/\left(1 + \exp(-z_p)\right)$ for beta priors.

2.5 Initialization

Ensemble initialization can be done in several ways. Goodman and Weare (2010) suggest to initialize the ensemble as a small ball around some initial value. However, if the posterior is oddly or irregularly shaped – e.g., if the posterior is bimodal –, this bears the risk that the ensemble can not fully unfold. Further, it is unclear which initial value to choose, in particular since we seek to avoid mode finding methods for the reasons discussed above. In the following I initialize the ensemble with the prior distribution $p(\theta)$. This will put equal initial weight to each region of the prior space:

$$\mathbf{X}_0 = (X_{0,1}, \ldots, X_{0,n_c}) \quad \text{with} \quad P(X_{0,i} = x) = p(x) \; \forall i = 1, \ldots, n_c. \tag{10}$$

[5] This can be seen by acknowledging that, if $\pi(x)$ is Gaussian, each draw $X_{s,i}$ is also Gaussian, and the difference between two chains hence also follows a Gaussian distribution.

Initializing the ensemble with the prior distribution ensures that the full relevant parameter space is considered, independently of multi-modality or possible discontinuities. It hence combines the strength of multi-start algorithms (see, e.g. Arnoud et al., 2019) with an efficient MCMC sampler.

3 A high dimensional bimodal toy distribution

This section studies the performance of ADEMC on a distribution with known properties. I focus on a class of high dimensional bimodal distributions potentially different maximum densities on each mode. Such densities are known to pose a challenge to MCMC samplers. Let the probability density of the random variable T be given by the multivariate Gaussian mixture

$$\pi(x) = \lambda P(X = x) + (1 - \lambda)P(Y = x) \tag{11}$$

where $X \sim \mathcal{N}_n(\mu_X, \Sigma)$ and $Y \sim \mathcal{N}_n(\mu_Y, \Sigma)$ are both n-dimensional Gaussian distributions with the same covariance, and $\lambda \in (0, 1)$ is a weighting parameter. Set $\Sigma = \sigma I_n$ to the identity matrix scaled by the scalar $\sigma > 0$. $\mu_X = (m/2, 0, \cdots, 0)'$ and $\mu_Y = (-m/2, 0, \cdots, 0)'$ are both vectors of zeroes apart from the first entries, which are $m/2$ and $-m/2$ respectively. The distribution of T is then bimodal whenever $m \neq 0$ and the distance between the two modes is given by $|m|$. When keeping σ fix, increasing m imposes a greater challenges to Monte Carlo sampling because the modes are less connected. I chose T to be in $n = 35$ dimensions since this is typical size of a target distribution when estimating medium-scale DSGE models (the dimensionality of the posterior of the HANK model estimated in section 5 is 36).

Figure 1 illustrates the results from this exercise graphically. For each of the histograms, I run 210 chains (correspondingly, $n_c/n = 6$) for 4000 iterations with $\delta = \exp(-100)$ and then marginalize over the first dimension. Each ensemble is initialized with a sample from $\mathcal{N}_n(\mathbf{0}_n, \sqrt{2}I_n)$. The initial ensemble is hence distributed across the domain of T, with relatively more chains closer to the origin. Results are shown for $\sigma = 0.05$ and distances of $m \in \{1, 2, 3\}$ (the columns of figure 1).[6] The first row shows results for $\lambda = 0.5$ where, hence, both modes peak at the same maximum density. For $m = 1$ both modes are connected, meaning that for any point between the modes the density is still reasonably large (that is, larger than 0.1 for the cases considered here). For $m = 2$ the trough between the models is relatively short in distance, but the minimum density is already close to zero. The gap for which the density is zero again increases considerably when setting $m = 3$, for which the modes are fully disconnected. The challenge to MCMC sampling lies in the fact that the chains must be able to bridge this gap, which is unlikely once the density is close to zero.

As the plots suggest, ADEMC performs very well on the first example. Even when the target distribution is fully disconnected ($m = 3$), the sampling error only increases marginally. Note that, since the initial chains are spread evenly over the domain of T the exercise for $m = 3$ is not very informative: it may as well be that chains initialized close to the first mode converge to the first mode while those initialized close to the second

[6] These values are chosen to demonstrate the frontier of what is possible with ADEMC, without additional adjustments of the algorithm.

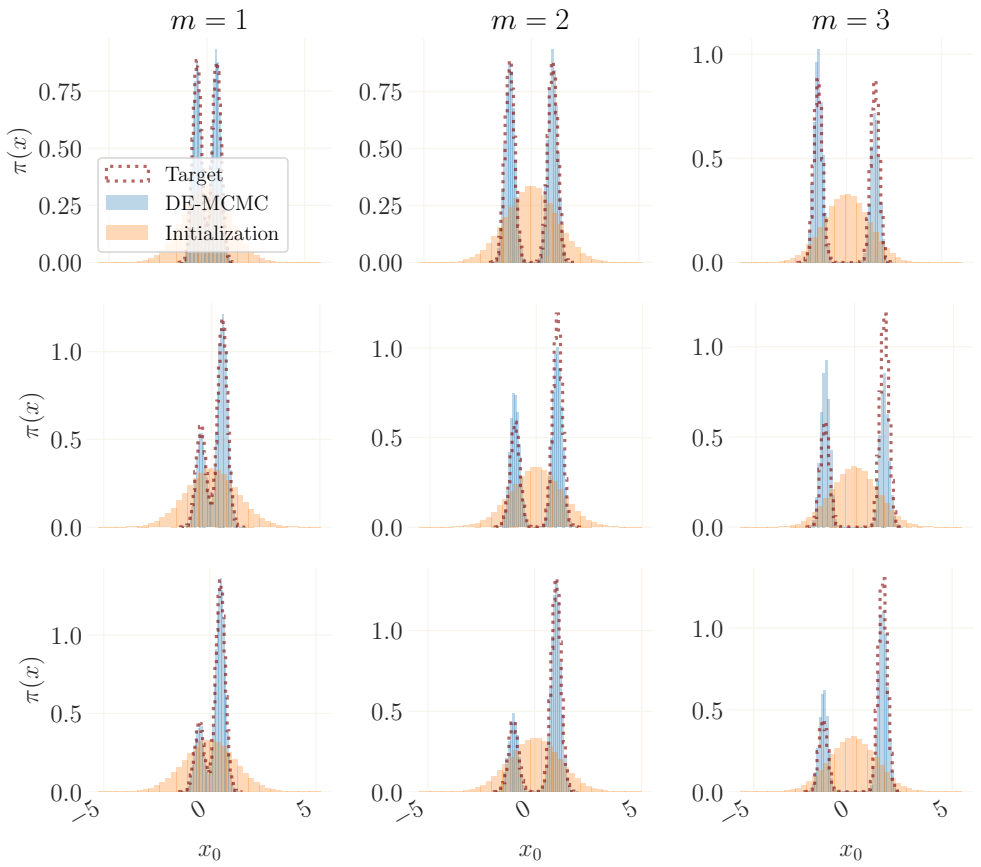

Figure 1: A 35 dimensional multivariate Gaussian mixture, marginalized over the first dimension. The dashed line depicts the target distribution. The orange region is the initialization of the ensemble, and the blue region the posterior after convergence.

mode converge to the latter. Given that initial chains are distributed evenly, convergence of a similar number of chains in each mode can be expected.

For the simulations in the second row of figure 1, λ is set to 0.33. This example is more challenging because some chains must "jump" modes in order to correctly reflect their different densities. This may cause problems in practice, and a single-particle sampler is likely to "get stuck" in the mode with the higher density, thereby ultimately misrepresent the posterior distribution. However, the figure suggests that this is not the case for ADEMC as long as both modes are still well connected ($m = 1$). For $m = 2$ (mild disconnected) both modes show some small sampling error at the peak. This sampling error gets larger as the size of the trough increases ($m = 3$). This suggests that there is still a good chance that chains may switch modes, but the distance between modes plays a crucial role on sampling performance.

For the last row λ is set to 0.25, implying that one mode has a considerably larger

posterior density. Even for $m = 1$ convergence to the posterior distribution for any single-particle sample would be very slow because changing from one mode to the other is relatively unlikely. As for the simulations for $\lambda = 0.33$, the figure suggests that ADEMC does perform well on the case in which both modes are connected, while the quality of the sample deteriorates when m increases. As above, if chains are equally distributed across the parameter space, each mode is expected to attract the same number of chains. The scaling constant is $\gamma = \frac{2.38}{\sqrt{n}} \approx 0.402 < 1$. If now two reference chains $X_{t,j}$ and $X_{t,k}$ get drawn from different modes, $\hat{X}_{t,i}$ is likely to lie between both modes and will probably be rejected. Correspondingly, the exchange of chains between the two modes is rather limited once the trough between them is too deep.

From the exercises where the distance $|m|$ between the modes is large (and the density at the trough is low), I conclude that ADEMC fails to reliably sample from distributions which feature *disconnected* modes. This problem can be expected to be more severe once the peak densities of the modes are sufficiently different. There are different potential resolutions to this problem. For example, sampling quality improves when letting some chains randomly draw a value of $\gamma = 1$. If then, two reference chains are drawn from different modes, the proposal will suggest a candidate which lies on the respectively other mode. This helps to balance between the two modes. The disadvantage of this approach is a relatively lower acceptance fraction. Another solution would be a tempering scheme as in Herbst and Schorfheide (2014). When sampling given a lower temperature the modes are not disconnected and chains can balance between the two modes during the early stages. However, in this case the acceptance fraction is also assumed to be lower. Lastly, another viable option would be to introduce resampling stages. To summarize, ADEMC performs surprisingly well if the target distribution is bimodal and both modes are connected.

4 The Smets-Wouters model

A common benchmark case for the Bayesian estimation of DSGE models is the model of (Smets and Wouters, 2007, SW, henceforth). I use this this famous reference in two exercises. First, I show that ADEMC is able to exactly recover the posterior distribution from the original paper in significantly less time. Secondly, I use it to benchmark ADEMC against the differential evolution MCMC (DE-MCMC) method of Ter Braak (2006) in the implementation of (Foreman-Mackey et al., 2013). Lastly, I use this model to study the trade-off between more chains and longer chains. For each of the exercises I use exactly the same model specification, priors, data and data treatment as in the original paper. All estimations are run on on a workstation with 40 Intel Xeon CPUs with 3.1GHz each and a total or 32GB RAM. I use the freely available pydsge package[7] to parse and solve the model, and to calculate the likelihood using the standard Kalman filter.

Table A.3 in Appendix A shows summary statistics over the posterior distribution of the estimation. For the estimation, 200 chains are used and run for 3500 iterations, of which 500 are kept as the posterior. The original estimation relies on 250.000 samples (of which 50.000 are discarded) after running a optimization procedure. Overall, the ADEMC estimates and the posterior values from the original estimation of SW are very

[7]The toolbox is developed and maintained at GitHub: https://github.com/gboehl/pydsge.

closely aligned. The ADEMC estimate of $\beta_{tpr} = 0.136$ is slightly below the SW estimate (0.16). Yet, both estimates are mutually contained in the respectively other standard error (0.052 for ADEMC). The ADEMC mean for $\bar{l}$ is 0.969 versus 0.53 in SW. Since the parameter has a large standard deviation in both estimates and considerable differences between mean and mode, it can be assumed that it is not well identified. In summary, the estimates indicates that the ADEMC can fully recover the original values of SW.

ρ_u and σ_u

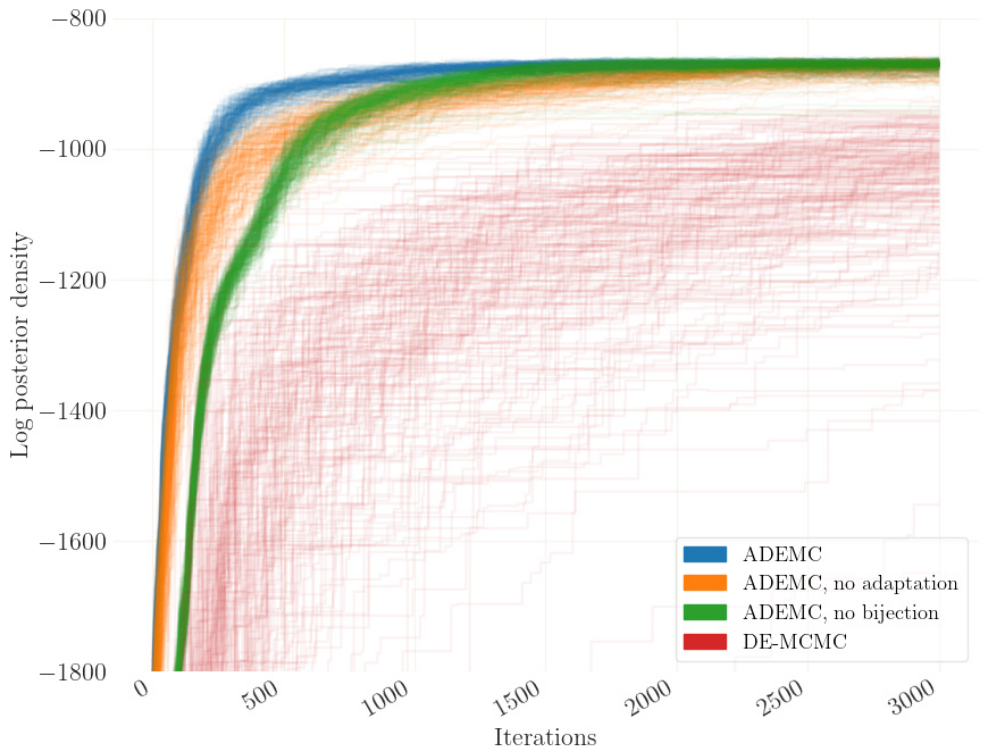

Figure 2: The progress of the likelihood of several ensembles with different setups. Note that for each ensemble each individual chain is plotted.

Let us now turn to the comparison of the performance of ADEMC with the DE-MCMC sampler from Ter Braak (2006). Figure 2 shows the likelihood of each chain during convergence for four different setups: ADEMC with $\delta = \exp(-100)$ (blue), ADEMC without the bijective map between prior and proposal space (green), ADEMC with bijective transform (orange) but without adaptation ($\delta = 0$) and plain DE-MCMC. The figure suggests that DE-MCMC converges slowly and does not converge to the prior distribution within the given 3000 iterations (red lines). Relative to this, the extension of prior and proposal space is quite powerful, and, through the significant increase in the acceptance ratio, speeds up convergence considerably (orange lines). However, although burn-in of ADEMC without adaptation is faster than ADEMC without bijective transform (green lines), final convergence to the posterior distribution seems relatively

slow. As to be expected, adding the bijective transform additionally boosts sampling performance (blue lines).

Appendix B includes a benchmark of the sensitivity of the estimation results with respect to the number of chains n_c. Figure B.3 shows estimations with different multiples of the number of parameters n as the number of chains n_c. I start with $n_c = 2n$, which is the minimum number of chains suggested by Foreman-Mackey et al. (2013). The exercise reveals two core-findings: first, ADEMC requires a larger number of chains as e.g. Goodman and Weare (2010) or Ter Braak (2006). This is because with each successful adaptation step, the accepted proposal is an explicit linear combination of existing chains. This is not problematic if only few adaptations are successful, since for $n_c > n$ the rank of the ensemble is n. If however many chains are accepted after adaptation, it may occur that the rank of the ensemble falls below n, which is undesirable. Consequently, after each adaptation a check on the condition number of the ensemble is performed, and I exclude estimations for which the ensemble rank falls below n. The results documented in Appendix B suggests that a choice of $n_c \geq 4n$ is advisable to maintain that the ensemble has full rank.

The second finding in Appendix B is that the number of chains and the number of iterations are close substitutes. Figure B.3 illustrates burn-in speed and convergence in terms of the number of total function evaluations. For the chosen range of $n_c \in (4n, 10n)$ it seems that no setup emerges which is to be strongly preferred. A larger number of chains approximately compensates one-to-one for fewer iterations. This hints that not the number of iterations is central, but the total number of function evaluations across chains. This is an important finding because it suggests that estimations can be scaled very well when parallelizing chains using computers with a larger number of processing units.

5 Full estimation of HANK

To fully unfold the potential of ADEMC I use the sampler to estimate a full-blown Heterogeneous-Agent New Keynesian (HANK) model with idiosyncratic portfolio choice. Thereby I also include those parameters that change the models steady state and the stationary distribution. Such HANK models are a relatively new class of models (see. Gornemann et al. (2012) and Kaplan et al. (2018a)) that combine the New Keynesian paradigm with household heterogeneity and incomplete markets. For example, this allows to study the impact of economic inequality on macroeconomic aggregates and vice versa.

While the estimation of HANK models is pioneered by Winberry (2018), Bayer et al. (2020, henceforth BBL) and Auclert et al. (2021), neither of these papers include the micro parameters that change the model's steady state into their estimation. The steady state forms central attributes of the ergodic distribution of assets and income, and qualitatively and quantitatively determines the relevance of the novel channels exposed by this class of models. Hence, these parameters are likely to crucially affect the macroeconomic dynamics of this model. Additionally, all three papers use a first-order approximation approach around the steady state, which lends another reason why the parameters that affect the latter are of central interest. The reasons for excluding the steady state in BBL and Auclert et al. (2021) is that calculating it involves finding a stationary distribution such that all equilibrium conditions are satisfied, which is computationally expensive.

Instead, both opt to calibrate the steady state and only estimate those parameters that are invariant to the steady state.

Since finding the steady state and the stationary distribution for HANK is economically important, but computationally expensive – about 10 seconds for the implementation considered here – it is a perfect use case for ADEMC. To remain consistent with the literature, and to keep the estimation results comparable, I only add a minor technical innovation to the rest of the estimation procedure.

5.1 Model and Data

The model shares many features with the two-asset HANK model of Auclert et al. (2021) and Kaplan et al. (2018b). I allow for two types of extensions: first, to ease comparison I use the priors of Smets and Wouters (2007). Accordingly, some of the functional forms (e.g. capital adjustment costs and Calvo pricing) are adapted from there. Secondly, and also in the sprit of Smets and Wouters (2007), I extend the model by several forms of inertia to allow for additional endogenous persistence in response to aggregate shocks.[8] I here discuss only those equations that deviate from Auclert et al. (2021) and refer the interested reader to Appendix C for further details on the model.

Households supply labor and have access to a liquid and an illiquid asset. Importantly, they face borrowing constraints on both assets, and adjustment costs on the illiquid asset. Firms accumulate capital, and staggered price setting results in a conventional Phillips curve. Adding price indexation with parameter ι_p, inflation π_t (in percentage points) is determined by

$$\pi_t - \bar{\pi} = \frac{\beta}{1+\beta\iota_p}\left(E_t\pi_{t+1} - \bar{\pi}\right) + \frac{\iota_p}{1+\beta\iota_p}(\pi_{t-1} - \bar{\pi}) + \kappa_p\left(\widehat{MC}_t - \frac{1}{\mu}\right) + \epsilon_{p,t}, \tag{12}$$

where $\bar{\pi}$ is the steady state inflation. $\epsilon_{p,t}$ is assumed to follow an AR(1) process around its zero and the slope of the Phillips curve is given by $\kappa_p = \frac{1-\zeta_p\beta}{1+\iota_p\beta}\frac{1-\zeta_p}{\zeta_p}$. Labor unions also set nominal wages subject Calvo frictions, which gives rise to a Phillips curve for wages. Adding wage indexation with parameter ι_w, this yields

$$\begin{aligned}\pi_t^w - \bar{\pi} = \frac{\beta}{1+\beta\iota_w}\left(E_t\pi_{t+1}^w - \bar{\pi}\right) + \frac{\iota_w}{1+\beta\iota_w}(\pi_{t-1}^w - \bar{\pi}) \\ + \kappa_w\left(\varphi N_t^{1+\nu} - \frac{(1-\tau_t)w_tN_t}{\mu_t^w}\int e_{it}c_it^{-\sigma}di\right) + \epsilon_{w,t},\end{aligned} \tag{13}$$

where $\epsilon_{w,t}$ as well follows an AR(1) process and $\kappa_w = \frac{1-\zeta_w\beta}{1+\iota_w\beta}\frac{1-\zeta_w}{\zeta_w}$. Monetary policy follows a conventional monetary policy rule with inflation indexing when setting the nominal interest rate r_t:

$$r_t^n - r^n = \rho\left(r_{t-1}^n - r^n\right) + (1-\rho)\left[\phi_\pi\left(\pi_t - \bar{\pi}\right) + \phi_y\Delta Y_t\right] + \epsilon_{r,t}, \tag{14}$$

with $\epsilon_{r,t}$ as an exogenous AR(1) process for monetary policy surprises. Note that in order to remain agnostic about the central bank's welfare objective, a traditional measure of

[8] It is well known that endogenous persistence is a crucial feature to replicate the hump-shaped empirical responses that are reported in the VAR literature.

output gap is absent in this equation. The setup of capital adjustment costs is as in Smets and Wouters (2007) and yields the following expressions for Tobin's Q and the firm's investment decisions

$$R_{t+1}q_t = (1-\delta)E_t q_{t+1} + \alpha E_t \left\{ Z_{t+1} \frac{N_{t+1}}{K_t}^{1-\alpha} \widehat{MC}_{t+1} \right\} \tag{15}$$

$$1 = \exp(\epsilon_{i,t}) q_t \left[1 - S\left(\frac{I_t}{I_{t-1}}\right) - S'\left(\frac{I_t}{I_{t-1}}\right)\frac{I_t}{I_{t-1}} \right] + E_t \left\{ \exp(\epsilon_{i,t+1}) \frac{q_{t+1}}{R_{t+1}} S'\left(\frac{I_{t+1}}{I_t}\right)\left(\frac{I_{t+1}}{I_t}\right)^2 \right\} \tag{16}$$

where R_t is the gross real interest rate, $S(x) = \frac{1}{2S''}(x-1)^2$ is a quadratic adjustment cost function and $\epsilon_{i,t}$ is an exogenous AR(1) process on the marginal productivity of investment. Finally, labor income taxation is progressive with parameter Ξ, such that after-tax labor income y_{jt} is given by

$$y_{jt} = {y_{jt}^p}^{1-\Xi} + \int p(e_{jt}) \left(y_{jt}^p - {y_{jt}^p}^{1-\Xi} \right), \tag{17}$$

with pretax income $y_{jt}^p = (1-\tau_t) w_t N_t e_{it}$.

For the estimation, I use a subset of the data used in Boehl et al. (forthcoming), which includes a standard setup for medium scale models: growth rates of consumption, investment, output and wages, together with inflation, labor hours and the federal funds rate. The data is at quarterly frequency and ranges from 1983:I to 2008:IV. Investment and consumption time series are adjusted such that investment also includes durables consumption as in Justiniano et al. (2010). Those seven observables are matched by seven economic shocks: the two markup shocks, the monetary policy shock, a government spending shock on G_t, a discount factor shock on β_t and $\epsilon_{i,t}$, which is a shock on the marginal efficiency of investment. Further details can be found in Appendix D.

5.2 Estimation methodology

Model solution and likelihood inference is done following the methodology introduced in Auclert et al. (2021).[9] In brief, let y_t be the time t vector of model variables (including disaggregated variables) and let the sequence of first-order conditions and marking clearing conditions, up to some distant point T periods in the future, be

$$F = \{ f(y_{t-1}, y_t, E_t y_{t+1}; x) \}_{t=0}^{T} = 0, \tag{18}$$

which depends on the parameter vector x. Denote by $Y_t \subset y_t$ only the aggregated variables and by $Z_t \subset y_t$ those variables that are purely exogenous. The authors propose a novel and computationally efficient procedure of finding the steady state Jacobian matrix of F with respect to $\{Y_t\}_{t=0}^T$ and $\{Z_t\}_{t=0}^T$. These sequence-space Jacobians (SSJ) can then be used to calculate impulse responses to aggregate shocks up to a first order approximation. Notably, this works for the broad class of models for which it is not required to explicitly keep track of any of the disaggregated distribution variables on a

[9] The authors provide their set of methods as a Python toolbox maintained at GitHub: https://github.com/shade-econ/sequence-jacobian.

global domain. Simulations are based on the sequence space rather than, as in BBL, the *state space*. The authors show that the first-order sequence space representation can be used directly for likelihood inference, without the need for using the Kalman filter (which would require a state space representation). In their application, the authors are able to re-use (parts of) the Jacobians depending on the types of parameters to be estimated. In my application, each Jacobian has to be calculated from scratch due to the re-calculation of the steady state.

The steady state $\bar{y}$ must satisfy

$$f(\bar{y}, \bar{y}, \bar{y}; x) = 0. \tag{19}$$

Given a guess for the steady state values of aggregated variables $\bar{Y}$, the stationary distribution of idiosyncratic variables can be found by finding the stationary decision rules via backward iteration, and finding the stationary distribution via forward iteration. Hence, there exists a known mapping $\bar{Y} \to \bar{y}$, and finding $\bar{Y}$ can be done using conventional root finding methods. Often, the size of this root finding problem can further be reduced to only searching a subset $\bar{K} \subset \bar{Y}$ since $\bar{Y}$ can be expressed in terms of $\bar{K}$. Still, finding $\bar{y}$ is relatively time consuming and must be repeated for any parameter draw x.[10]

5.3 Estimation results

As usual, some parameters are fixed prior to the estimation. Those parameters, most of which configuring the technical setup of the estimation (e.g. the number of grid points), can be found in table C.4 in Appendix C. All other parameters are estimated using the priors presented in the first three columns in tables 1 and 2, which follow the specification of Smets and Wouters (2007). Exceptions are the portfolio adjustment cost parameter χ_0, tax progressively parameter Ξ, and the standard deviation of the AR(1) process for idiosyncratic labor productivity σ^e, which are specific to the HANK model. I opt for flat priors for these parameters.

For the estimation I run an ADEMC ensemble with a total of 288 chains[11] for 2500 iterations. The ensemble converges to the high-density region of the posterior after about 800 iterations. The estimation takes 111 hours on a machine with 48 cores. See figures E.4 to E.9 in Appendix E for graphical illustrations of the convergence of the chains over time, and the posterior distribution. Tables 1 and 2 show summary statistics of the posterior distribution.

This paper focusses on the properties of the ADEMC sampler instead of the particular economic dynamics of the HANK model. Hence, I deem an in-debt analysis of the economic implications of the estimated model out of the scope of this paper and leave it as a promising endeavour for future research. However, the comparison of the parameter estimates from the HANK model with those of Smets and Wouters (2007) – for a somewhat smaller sample – reveals some surprising differences.[12]

[10] For any numerical root finding method a good initial guess is crucial, and so it is for finding the steady state. For bad initial guesses, the root search may either diverge, crash because it is causing numerical errors when solving for the stationary distribution, or simply take up a very long time. The former is problematic because it prohibits to calculate the likelihood also for cases in which a likelihood actually exists. In practice, I use the steady state values for the prior mean as the initial guess, which causes about 2/3 of all parameter vectors sampled from the prior distribution to be accepted.

[11] The number is the number of available CPUs (48) times 6.

[12] The estimation of Smets and Wouters (2007) is replicated in Appendix A.

		Prior			Posterior		
		distribution	mean	std.	mean	std.	mode
σ	intertemporal elasticity of substitution	normal	1.500	0.375	1.809	0.254	2.162
φ	Frisch elasticity	normal	2.000	0.750	1.926	0.781	1.688
ζ_p	Calvo parameter for price setting	beta	0.500	0.100	0.594	0.072	0.562
ζ_w	Calvo parameter of wage setting	beta	0.500	0.100	0.414	0.092	0.457
ι_p	price inertia	beta	0.500	0.150	0.316	0.151	0.368
ι_w	wage inertia	beta	0.500	0.150	0.333	0.174	0.317
S''	derivative capital adjustment costs	gamma	4.000	2.000	3.221	1.842	2.538
ϕ_π	monetary policy coefficient inflation	gamma	1.500	0.250	2.281	0.315	2.122
ϕ_y	monetary policy coefficient output	gamma	0.125	0.050	0.236	0.092	0.238
ρ	monetary policy persistence	beta	0.750	0.100	0.646	0.079	0.644
$\bar{y}$	trend output	normal	0.400	0.100	0.442	0.033	0.445
$\bar{n}$	steady state labor hours	normal	0.000	2.000	-0.914	2.582	-0.580
π^*	inflation target	gamma	0.625	0.100	0.570	0.077	0.614
i^*	steady state nominal interest rate	gamma	1.250	0.100	1.196	0.125	1.281
χ_0	portfolio adjustment costs (scale)	gamma	0.250	0.200	0.110	0.133	0.012
Ξ	tax progressivity	beta	0.200	0.100	0.109	0.079	0.106
σ^e	standard deviation of labor productivity	normal	1.500	0.500	1.407	0.608	0.906

Table 1: Estimation results for HANK: model parameters

In HANK, the inverse elasticity of substitution, σ, is relatively large. This stands in contrast to the estimate of SW, and even stronger contradicts findings documented in Boehl and Strobel (2022b,a) for US data until 2019 of values close-to or below one. Potentially, this could be due to the fact that the precautionary motive is more emphasized in HANK due to the assumption of incomplete financial markets. Similar as in SW, the Frisch elasticity of labor supply φ is estimated significantly above its prior mean.

A remarkable finding is that, in the HANK model both the price and the wage Phillips curve are identified to be comparably steep, i.e. the Calvo adjustment probabilities ζ_p and ζ_w are estimated to be relatively low. This stands in contrast to many more recent estimates which find rather large values for these parameters, which result in a very flat Phillips curve. While this effect may come from different data samples, it calls for further investigation. The relatively lower estimate of S'' may indicate that capital adjustment costs play a smaller role in the HANK model, which may be due to the fact that in the HANK model, portfolio adjustment represent a additional friction that actively influences the capital investment decision. The other parameters in table 1, which govern the monetary policy rule and the steady state values of the observables, are well-aligned with the original estimates in SW. This may likely be the case because the parameters are identified independently of the model's setup of the household sector.

The estimate of χ_0 is below its prior, suggesting a somewhat less accentuated role of the households' portfolio choice problem. The estimate of the standard deviation of the idiosyncratic labor productivity is close to its prior value. Lastly, the estimates of the parameters that govern the exogenous autoregressive processes are much in line with conventional estimates, where technology, government spending and investment specific shocks have a high autocorrelation.

		Prior			Posterior		
		distribution	mean	std.	mean	std.	mode
ρ_z	AR coefficient technology shock	beta	0.500	0.200	0.951	0.046	0.956
ρ_r	AR coefficient MP shock	beta	0.500	0.200	0.599	0.120	0.604
ρ_g	AR coefficient gov. spending shock	beta	0.500	0.200	0.982	0.068	0.999
ρ_w	AR coefficient wage MU shock	beta	0.500	0.200	0.960	0.093	0.986
ρ_p	AR coefficient price MU shock	beta	0.500	0.200	0.903	0.085	0.936
ρ_i	AR coefficient investment shock	beta	0.500	0.200	0.764	0.210	0.860
ρ_β	AR coefficient interest wedge shock	beta	0.500	0.200	0.968	0.054	0.945
σ_z	standard dev. technology shock	inv.gamma	0.100	0.250	0.227	0.070	0.234
σ_r	standard dev. MP shock	inv.gamma	0.100	0.250	0.389	0.221	0.346
σ_g	standard dev. gov. spending shock	inv.gamma	0.100	0.250	0.233	0.037	0.214
σ_w	standard dev. wage MU shock	inv.gamma	0.100	0.250	2.385	0.836	2.404
σ_p	standard dev. price MU shock	inv.gamma	0.100	0.250	0.209	0.071	0.212
σ_i	standard dev. investment shock	inv.gamma	0.100	0.250	1.022	0.483	1.222
σ_β	standard dev. interest wedge shock	inv.gamma	0.100	0.250	0.045	0.073	0.059

Table 2: Estimation results for HANK: parameters of exogenous processes

6 Conclusion

This paper develops an adaptive differential evolution Monte Carlo Marcov chain (ADEMC) method. Such ensemble MCMC methods have in common that they are running a large number of chains, and the proposal density is generated endogenously from the state of all chains, thereby automatically adapting to the shape of the current estimate of the posterior distribution. The separation of parameter space and proposal space guarantees that the proposal density respects the bounds of prior distribution, which results in significantly higher acceptance fractions and, consequently, higher sampling efficiency. An adaptation stage leads to faster burn-in times to the high density region of the posterior. I show that ADEMC is easy to parallelize, where the number of iterations required until convergence decreases almost one-to-one with the number of chains. This makes the method easily scalable for large-scale problems. I further document good performance for bimodal distributions, as long as the two modes are not too far apart.

I apply ADEMC to estimate novel HANK models, thereby being the first to include the micro parameters which govern the households' problem and give rise to the endogenous distribution of assets. To retain focus of this paper, I abstain from a detailed analysis of the estimated model, and do not put the resulting parameter estimates in relation to estimates e.g. from the micro data. This is an promising endeavour for future research.

References

Arnoud, A., Guvenen, F., Kleineberg, T., 2019. Benchmarking global optimizers. Technical Report. National Bureau of Economic Research.

Auclert, A., Bardóczy, B., Rognlie, M., Straub, L., 2021. Using the sequence-space jacobian to solve and estimate heterogeneous-agent models. Econometrica 89, 2375–2408.

Bayer, C., Born, B., Luetticke, R., 2020. Shocks, Frictions, and Inequality in US Business Cycles. CEPR Discussion Papers 14364.
Boehl, G., Goy, G., Strobel, F., forthcoming. A structural investigation of quantitative easing. Review of Economics and Statistics .
Boehl, G., Strobel, F., 2020. US business cycle dynamics at the zero lower bound. Bundesbank Discussion Papers 65/2020. Deutsche Bundesbank.
Boehl, G., Strobel, F., 2022a. Estimation of DSGE Models with the Effective Lower Bound. CRC 224 Discussion Papers. University of Bonn and University of Mannheim, Germany.
Boehl, G., Strobel, F., 2022b. The Empirical Performance of Financial Frictions Since 2008. CRC 224 Discussion Papers. University of Bonn and University of Mannheim, Germany.
Chib, S., Ramamurthy, S., 2010. Tailored randomized block mcmc methods with application to dsge models. Journal of Econometrics 155, 19–38.
Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid monte carlo. Physics letters B 195, 216–222.
Edge, R.M., Gürkaynak, R.S., Kisacikoglu, B., 2013. Judging the DSGE model by its forecast. Technical Report. mimeo.
Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J., 2013. emcee: the mcmc hammer. Publications of the Astronomical Society of the Pacific 125, 306.
Geweke, J., 1999. Using simulation methods for bayesian econometric models: inference, development, and communication. Econometric reviews 18, 1–73.
Goodman, J., Weare, J., 2010. Ensemble samplers with affine invariance. Communications in applied mathematics and computational science 5, 65–80.
Gornemann, N., Kuester, K., Nakajima, M., 2012. Monetary policy with heterogeneous agents. Working Papers 12-21. Federal Reserve Bank of Philadelphia. URL: http://ideas.repec.org/p/fip/fedpwp/12-21.html.
Hastings, W.K., 1970. Monte carlo sampling methods using markov chains and their applications .
Herbst, E., Schorfheide, F., 2014. Sequential monte carlo sampling for dsge models. Journal of Applied Econometrics 29, 1073–1098.
Herbst, E.P., Schorfheide, F., 2015. Bayesian estimation of dsge models, in: Bayesian Estimation of DSGE Models. Princeton University Press.
Justiniano, A., Primiceri, G.E., Tambalotti, A., 2010. Investment shocks and business cycles. Journal of Monetary Economics 57, 132–145. URL: https://ideas.repec.org/a/eee/moneco/v57y2010i2p132-145.html.
Kaplan, G., Moll, B., Violante, G.L., 2018a. Monetary policy according to HANK. NBER Working Papers 3. National Bureau of Economic Research, Inc. URL: https://ideas.repec.org/p/nbr/nberwo/21897.html.
Kaplan, G., Moll, B., Violante, G.L., 2018b. Monetary policy according to hank. American Economic Review 108, 697–743.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21, 1087–1092.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. The computer journal 7, 308–313.
Nelson, B., Ford, E.B., Payne, M.J., 2013. Run dmc: an efficient, parallel code for analyzing radial velocity observations using n-body integrations and differential evolution markov chain monte carlo. The Astrophysical Journal Supplement Series 210, 11.
Roberts, G.O., Rosenthal, J.S., 2001. Optimal scaling for various metropolis-hastings algorithms. Statistical science 16, 351–367.
Schorfheide, F., 2000. Loss function-based evaluation of dsge models. Journal of Applied Econometrics 15, 645–670.
Smets, F., Wouters, R., 2007. Shocks and frictions in us business cycles: A bayesian dsge approach. American Economic Review 97, 586–606.
Sokal, A., 1997. Monte carlo methods in statistical mechanics: foundations and new algorithms, in: Functional integration. Springer, pp. 131–192.
Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization 11, 341–359.
Strens, M.J., Bernhardt, M., Everett, N., 2002. Markov chain monte carlo sampling using direct search optimization, in: ICML, Citeseer. pp. 602–609.
Ter Braak, C.J., 2006. A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. Statistics and Computing 16, 239–249.

Winberry, T., 2018. A method for solving and estimating heterogeneous agent macro models. Quantitative Economics 9, 1123–1151.

Appendix A Posterior distribution of the estimation of the Smets-Wouters model

	Prior			Posterior		
	distribution	mean	std./df	mean	sd.	mode
σ_c	normal	1.500	0.375	1.361	0.136	1.443
σ_l	normal	2.000	0.750	1.958	0.578	1.614
β_{tpr}	gamma	0.250	0.100	0.136	0.052	0.132
h	beta	0.700	0.100	0.706	0.050	0.664
S''	normal	4.000	1.500	5.471	1.093	4.637
ι_p	beta	0.500	0.150	0.229	0.105	0.244
ι_w	beta	0.500	0.150	0.582	0.137	0.556
α	normal	0.300	0.050	0.182	0.018	0.188
ζ_p	beta	0.500	0.100	0.646	0.060	0.647
ζ_w	beta	0.500	0.100	0.725	0.066	0.706
Φ_p	normal	1.250	0.125	1.578	0.075	1.537
ψ	beta	0.500	0.150	0.560	0.120	0.540
ϕ_π	normal	1.500	0.250	2.066	0.174	2.026
ϕ_y	normal	0.125	0.050	0.093	0.023	0.073
ϕ_{dy}	normal	0.125	0.050	0.229	0.027	0.238
ρ	beta	0.750	0.100	0.812	0.025	0.824
ρ_r	beta	0.500	0.200	0.118	0.062	0.080
ρ_g	beta	0.500	0.200	0.983	0.008	0.984
ρ_z	beta	0.500	0.200	0.964	0.011	0.963
ρ_u	beta	0.500	0.200	0.235	0.126	0.166
ρ_p	beta	0.500	0.200	0.898	0.099	0.892
ρ_w	beta	0.500	0.200	0.976	0.027	0.980
ρ_i	beta	0.500	0.200	0.730	0.067	0.774
μ_p	beta	0.500	0.200	0.693	0.141	0.677
μ_w	beta	0.500	0.200	0.883	0.059	0.884
ρ_{gz}	normal	0.500	0.250	0.505	0.089	0.475
σ_g	inv.gamma	0.100	2.000	0.530	0.032	0.517
σ_u	inv.gamma	0.100	2.000	1.905	0.457	1.708
σ_z	inv.gamma	0.100	2.000	0.459	0.029	0.471
σ_r	inv.gamma	0.100	2.000	0.243	0.015	0.228
σ_p	inv.gamma	0.100	2.000	0.139	0.020	0.130
σ_w	inv.gamma	0.100	2.000	0.251	0.023	0.245
σ_i	inv.gamma	0.100	2.000	0.456	0.066	0.432
$\overline{\gamma}$	normal	0.400	0.100	0.418	0.020	0.415
$\overline{l}$	normal	0.000	2.000	0.969	1.141	2.142
$\overline{\pi}$	gamma	0.625	0.100	0.665	0.104	0.722

Table A.3: SW estimation. The inverse gamma distribution is parameterized in terms of degrees of freedom as in dynare.

Appendix B Benchmarking against the number of chains

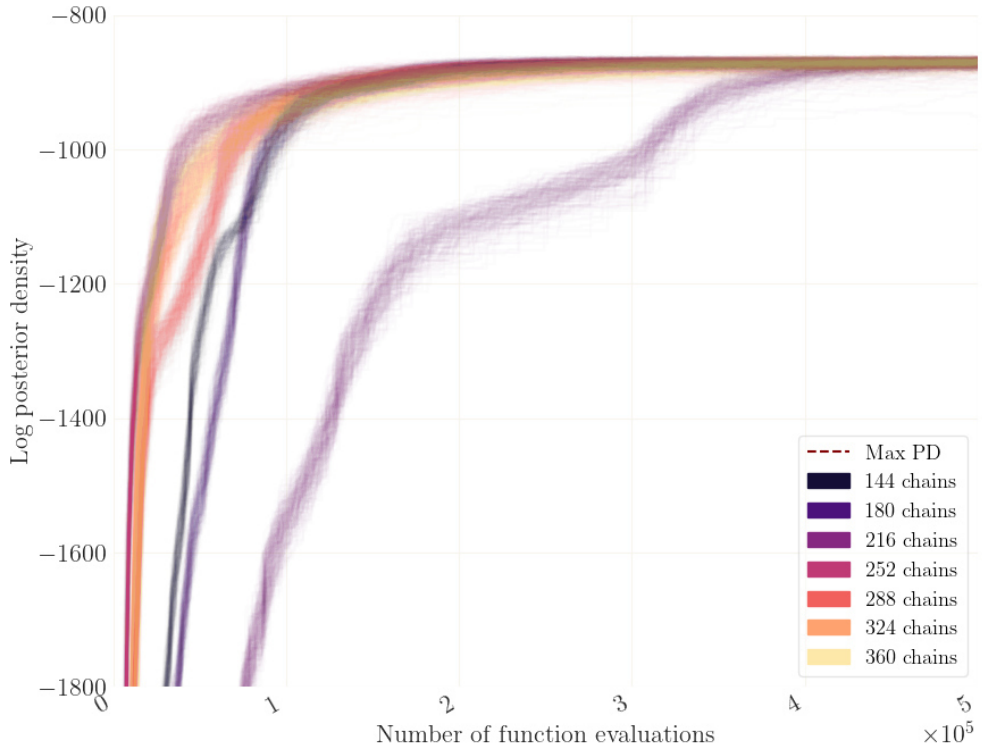

Figure B.3: The progress of the likelihood of several ensembles with different numbers n_c of chains for the model of Smets and Wouters (2007). Note that for each ensemble each individual chain is plottet.

Figure B.3 shows several estimations using different multiples of the number of parameters n_θ as the number of chains, starting with $n_c/n_\theta = 2$, which is the minimum suggested by Foreman-Mackey et al. (2013). The red dashed line marks the highest value of the posterior density that is found across all ensembles and chains. Note that, in order to directly compare computational efficient, the x-axis of the figure does *not* show the number of iterations n_i but the number of total function evaluations, which is $n_i \times n_c$. The figure suggests that no measurable tradeoff exists.

Appendix C Details on the HANK model

This part of the model is by large adopted from Auclert et al. (2021).

Appendix C.1 Households

The Bellman equation of households is given by

$$V_t(e_{it}, l_{it-1}, a_{it-1}) = \max_{c_{it}, b_{it}, a_{it}} \left\{ \frac{c_{it}^{1-\sigma}}{1-\sigma} - \varphi \frac{N_t^{1+\nu}}{1+\nu} + \beta E_t V_{t+1}(e_{it+1}, b_{it+1}, a_{it}) \right\} \quad \text{(C.1)}$$

such that

$$c_{it} + a_{it} + b_{it} = \frac{(1-\tau_t)w_t N_t}{\int P(e_{jt}) e_{jt}^{1-\Xi} dj} e_{it}^{1-\Xi} + (1+r_t^a)a_{it-1} + (1+r_t^b)b_{it-1} - \Phi_t(a_{it}, a_{it-1}), \tag{C.2}$$

$$a_{it} \geq 0, \tag{C.3}$$

$$b_{it} \geq \bar{b}, \tag{C.4}$$

where $\Phi_t(\cdot)$ is the portfolio adjustment cost function

$$\Phi_t(a_{it}, a_{it-1}) = \frac{\chi_1}{\chi_2} \left| \frac{a_{it} - (1+r_t^a)a_{it-1}}{(1+r_t^a)a_{it-1} + \chi_0} \right|^{\chi_2} [(1+r_t^a)a_{it-1} + \chi_0], \tag{C.5}$$

with $\chi_0, chi_1 > 0$ and $\chi_2 > 1$. Individual labor productivity e_{it} is assumed to follow a random walk process with coefficient ρ_e and a standard deviation of the innovations of σ_t^e, which is by itself assumed to follow an exogenous AR(1) process on an aggregate level.

Appendix C.2 Financial market

No arbitrage at the financial market requires that

$$1 + E_t r_{t+1} = \frac{1+i_t}{1+E_t\pi_{t+1}} = \frac{E_t[d_{t+1} + p_{t+1}]}{p_t} = 1 + E_t R_{t+1}^a = 1 + E_t r_{t+1}^b + \omega, \tag{C.6}$$

with ω the parameter governing the cost for liquidity transformation charged by the financial intermediary. Ex-post returns are subject to surprise inflation and capital gains

$$1 + r_t = \frac{1+i_{t-1}}{1+\pi} = 1 + r_t^b + \omega \tag{C.7}$$

and

$$1 + r_t^a = \Theta_p \left(\frac{d_t + p_t}{p_{t-1}} \right) + (1-\Theta_p)(1+r_t), \tag{C.8}$$

where Θ_p denotes the share of equity in the illiquid portfolio.

Appendix C.3 Firms

Firms have a production function

$$y_{jt} = F(k_{jt-1}, n_{jt}) = k_{jt-1}^{\alpha} n_{jt}^{1-\alpha} \tag{C.9}$$

and aggregat marginal costs are given by

$$\widehat{MC}_t = w_t / F_N(\cdot), \tag{C.10}$$

which enter the Phillips curve (12). Aggregate investment is given by

$$I_t = K_t - (1-\delta)K_{t-1} + S\left(\frac{I_t}{I_{t-1}}\right), \tag{C.11}$$

Parameter		Value	Target
β	time preference parameter	–	r^*
χ_1	portfolio adj. cost scale	–	$B = 1.04Y$
$\bar{b}$	borrowing constraint	0	
ρ_e	autocorrelation of earnings	0.966	
ν	disutility of labor	–	N = 1
μ_p	steady state markup	–	$p + B^g = 14Y$
μ_w	steady state wage markup	1.1	
Z	TFP	0.468	Y = 1
α	captial share	0.33	K = 10Y
ω	steady state liquidity premium	0.1	
G	steady state government spending	0.2	
B^g	bond supply	2.8	
n_e	points for Markov chain of e	3	
n_b	points for liquid asset grid	25	
n_a	points for illiquid asset grid	35	

Table C.4: Parameters fixed for the estimation of HANK.

with the quadradic capital adjustment cost function $S(x) = \frac{1}{2S''}(x-1)^2$ as given in the main body, and $\delta > 0$ the parameter for capital depreciation. Dividents are defined as

$$d_t = Y_t - w_t - I_t - \psi_t. \tag{C.12}$$

Tobin's Q and the capital investiment decisions follow equations (15) and (16) from the main body.

Appendix C.4 Market clearning

The optimality condition for labor unions is (13) and the monetary policy rule is given by (14). Balanced budget requires

$$\tau_t w_t N_t = r_t B^g + G_t, \tag{C.13}$$

and market clearing requires

$$Y_t = \int c_{it} di + G_t + I_t + \psi_t + \omega b_{it} di, \tag{C.14}$$

$$p_t + B^g = \int a_{it} + b_{it} di. \tag{C.15}$$

Appendix C.5 Fixed parameters

The parameters that are not estimated are set as in table C.4.

Appendix D Data

The following measurement equations are used for the HANK estimation:

$$\begin{aligned}
\text{Real GDP growth} &= \overline{\gamma} + (y_t - y_{t-1}),\\
\text{Real consumption growth} &= \overline{\gamma} + (c_t - c_{t-1}),\\
\text{Real investment growth} &= \overline{\gamma} + (i_t - i_{t-1}),\\
\text{Real wage growth} &= \overline{\gamma} + (w_t - w_{t-1}),\\
\text{Labor hours} &= \overline{n} + n_t,\\
\text{Inflation} &= \overline{\pi} + \pi_t,\\
\text{Federal funds rate} &= 100\left(\frac{\overline{\pi}}{\beta\gamma^{-\sigma_c}} - 1\right) + r_t,
\end{aligned}$$

The observables are constructed as follows:

- GDP: ln(GDP/GDPDEF/CNP16OV_ma)*100
- CONS: ln((PCEC-PCEDG) / GDPDEF / CNP16OV_ma)*100
- INV: ln((GPDI+PCEDG) / GDPDEF / CNP16OV_ma)*100
- LAB: ln(13*AWHNONAG * CE16OV / CNP16OV_ma)*100
- INFL: ln(GDPDEF)*100
- WAGE: ln(COMPNFB / GDPDEF)*100
- FFR: FEDFUNDS/4

Due to artificial dynamics in the civilian noninstitutional population series that arise from irregular updating (Edge et al., 2013), we use a 4-quarter trailing moving average, denoted CNP16OV_ma, to calculate per capita variables. We take log changes for GDP, CONS, INV and WAGE in our measurement equations. Data for the GZ spread is downloaded from the Federal Reserves Board.[13] Data for the 10-year equivalents of the Fed's SOMA Treasury security holdings are extracted from the Domestic Open Market Reports published by the New York Fed.[14] All other data is downloaded from the FRED database of the St. Louis Fed, with the above mnemonics representing:

- GDP: GDP - Gross Domestic Product, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate, FRED
- GDPDEF: Gross Domestic Product: Implicit Price Deflator , Index 2012=100, Quarterly, Seasonally Adjusted , FRED

[13] See `https://www.federalreserve.gov/econresdata/notes/feds-notes/2016/files/ebp_csv.csv`.

[14] We collected the data for the SOMA Domestic Securities Holdings in Ten-Year Equivalents from the Open Market Reports, downloaded from `https://www.newyorkfed.org/markets/annual_reports`.

- CNP16OV: Civilian noninstitutional population, Thousands of Persons, Quarterly, Seasonally Adjusted, FRED
- CNP16OV_ma: a four-quarter trailing average of CNP16OV
- PCEC: Personal Consumption Expenditures, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate, FRED
- PCEDG: Personal Consumption Expenditures: Durable Goods, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate, FRED
- GPDI: Gross Private Domestic Investment, Billions of Dollars, Quarterly, Seasonally Adjusted Annual Rate, FRED
- AWHNONAG: Average Weekly Hours of Production and Nonsupervisory Employees: Total private, Hours, Quarterly, Seasonally Adjusted, FRED
- CE16OV: Employment Level, Thousands of Persons, Quarterly, Seasonally Adjusted, FRED
- COMPNFB, Nonfarm Business Sector: Compensation Per Hour, Index 2012=100, Quarterly, Seasonally Adjusted, FRED
- FEDFUNDS: Effective Federal Funds Rate, Percent, FRED

Appendix E Details on the estimation of HANK

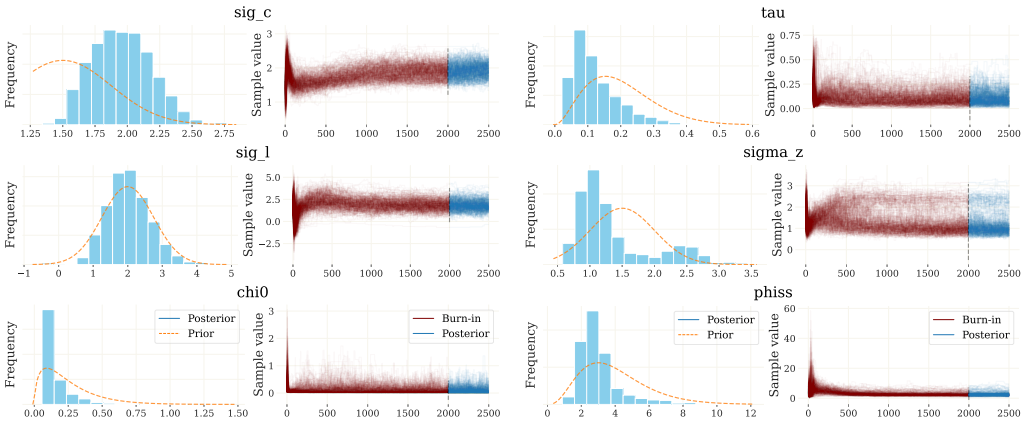

Figure E.4: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.

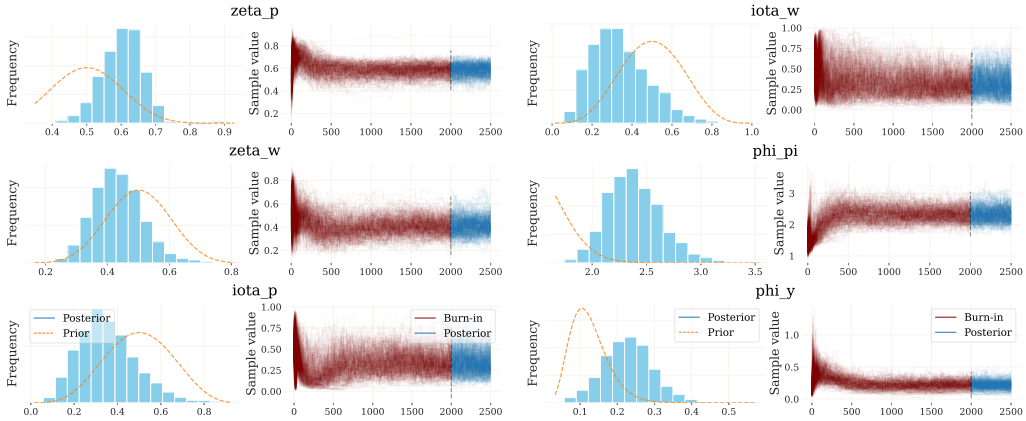

Figure E.5: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.

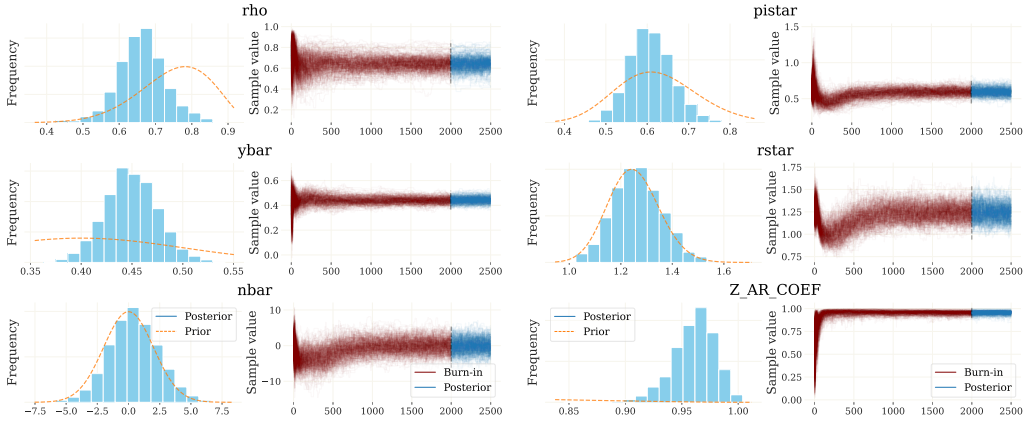

Figure E.6: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.

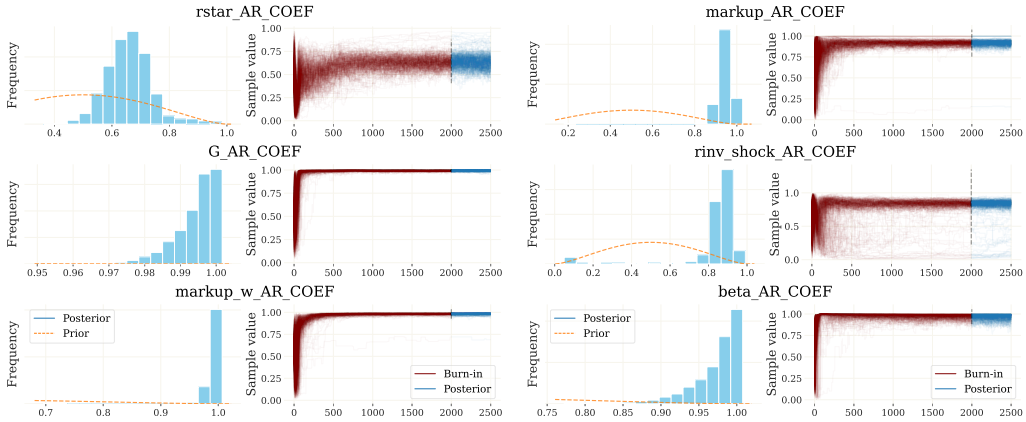

Figure E.7: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.

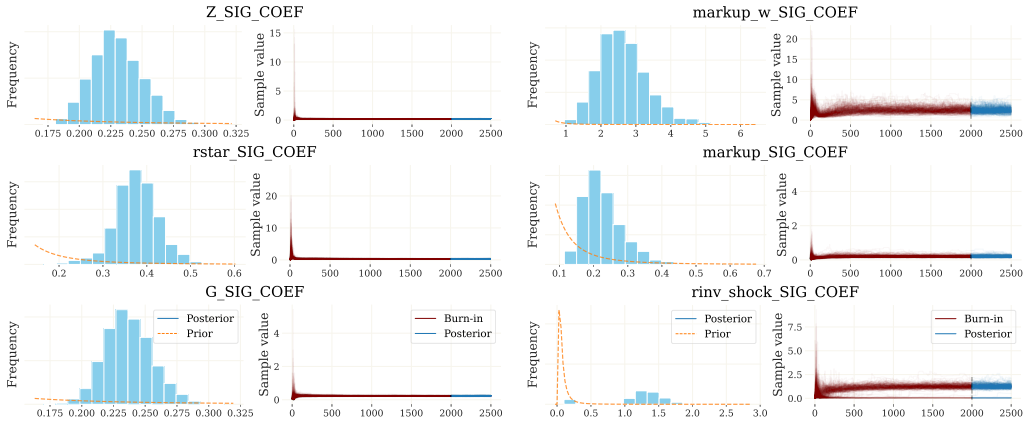

Figure E.8: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.

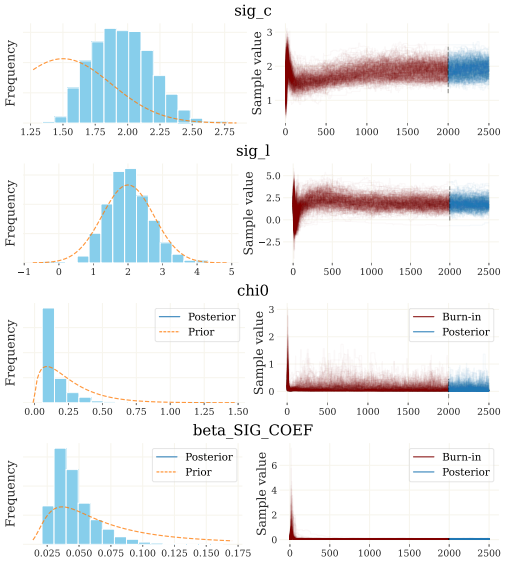

Figure E.9: Traceplots of the 200 ADEMC chains. The left panel shows a KDE of the parameter distribution. The right displays the trace of each of the chains over time.