University of Toronto
Department of Economics

Working Paper 438

Bayesian Adaptive Hamiltonian Monte Carlo with an Application to High-Dimensional BEKK GARCH Models

By Martin Burda and John Maheu

June 21, 2011

Bayesian Adaptive Hamiltonian Monte Carlo with an Application to High-Dimensional BEKK GARCH Models[*]

Martin Burda,[†] John M. Maheu [‡]

June 21, 2011

Abstract

Hamiltonian Monte Carlo (HMC) is a recent statistical procedure to sample from complex distributions. Distant proposal draws are taken in a sequence of steps following the Hamiltonian dynamics of the underlying parameter space, often yielding superior mixing properties of the resulting Markov chain. However, its performance can deteriorate sharply with the degree of irregularity of the underlying likelihood due to its lack of local adaptability in the parameter space. Riemann Manifold HMC (RMHMC), a locally adaptive version of HMC, alleviates this problem, but at a substantially increased computational cost that can become prohibitive in high-dimensional scenarios. In this paper we propose the Adaptive HMC (AHMC), an alternative inferential method based on HMC that is both fast and locally adaptive, combining the advantages of both HMC and RMHMC. The benefits become more pronounced with higher dimensionality of the parameter space and with the degree of irregularity of the underlying likelihood surface. We show that AHMC satisfies detailed balance for a valid MCMC scheme and provide a comparison with RMHMC in terms of effective sample size, highlighting substantial efficiency gains of AHMC. Simulation examples and an application of the BEKK GARCH model show the usefulness of the new posterior sampler.

JEL codes: C01; C11; C15; C32
Keywords: High-dimensional joint sampling, Markov chain Monte Carlo

[*] We would like to thank Ben Calderhead, Mark Girolami, and Radford Neal for helpful discussions. We would also like to thank the participants of the Seminar on Bayesian Inference in Econometrics and Statistics 2011 at WUSTL, the Workshop on High-Dimensional Econometric Modelling 2010 in London, UK, and Brown University seminar for their insightful comments and suggestions. Both Burda and Maheu thank SSHRC for financial support. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca).

[†] Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-4479; Email: martin.burda@utoronto.ca

[‡] Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-1495; Email: jmaheu@chass.utoronto.ca and RCEA, Italy.

1. **Introduction**

Hamiltonian dynamics have been traditionally used to describe the laws of motion in molecular systems in physics. Following the recent advances in Markov chain Monte Carlo (MCMC) fuelled by increasing availability of fast computation, inferential methods based on Hamiltonian dynamic systems are becoming increasingly popular in the statistics literature (Neal, 1993, 2010; Ishwaran, 1999; Liu, 2004; Girolami and Calderhead, 2011).[1] Hamiltonian Monte Carlo, also called Hybrid Mote Carlo, (HMC) uses Hamiltonian dynamics in constructing distant proposal draws in a sequence of steps and hence concurrently yields relatively low correlation among draws and high acceptance probabilities. Methods based on HMC have been shown to improve sampling of ill-behaved posteriors, and enabled the solution of otherwise intractable high dimensional inference problems (Neal, 2010; Girolami and Calderhead, 2011). These methods are particularly useful for the kind of problems where it is difficult to accurately approximate the surface of the (posterior) log-likelihood around the current parameter draw or the mode in real time needed for obtaining sufficiently high acceptance probabilities in importance sampling (IS) or accept-reject methods. Perpetual re-fitting of a local posterior approximating density around newly accepted draws during the MCMC run may become too costly for methods based on such mechanism to be practical. These types of problems typically arise when the log-likelihood is costly to evaluate and is near-ill-conditioned around the mode.

Even if on a small scale, with a few parameters and small sample size, such problems can be handled by standard procedures, these can become prohibitive in higher parameter dimensions and sample sizes. Examples include recursive models in finance, such as the BEKK GARCH that we treat in our application, state-space models or point process models. In such situations one would typically resort to Random walk (RW) style sampling that is fast to run and does not require the knowledge of the properties of the underlying log-likelihood. However, RW mechanisms can lead to very slow exploration of the parameter space with high autocorrelations among draws which would require a prohibitively large size of the Markov chain to be obtained in implementation to achieve satisfactory mixing and convergence. HMC combines the advantages of sampling that is relatively cheap with RW-like intensity but superior parameter space exploration.

Nonetheless, HMC uses a mechanism whose form is fixed over the parameter space, lacking adaptability to local features of the likelihood. The Riemann Manifold HMC, or RMHMC (Girolami and Calderhead, 2011), alleviates this problem and renders HMC locally adaptable which results in improved convergence and mixing properties. However, relative to HMC, RMHMC implementation requires a substantially increased computational burden with a large number of fixed point evaluations within every MC step which can render its performance inadequate in high-dimensional problems where the likelihood is expensive to evaluate. Indeed, it is precisely this type of problems for which HMC-type methods are most useful relative to other existing methods.

[1] The discussion section of the recent Girolami and Calderhead (2011) article contains over 60 discussion pieces by prominent statisticians expressing their overwhelmingly supportive views.

In this paper we propose an alternative inferential method, the Adaptive HMC (AHMC), that is both relatively fast and locally adaptive. AHMC is based on proposal dynamics generalizing HMC with only minimal additional functional evaluations, yet closely approximating the local adaptability properties of RMHMC. Unlike the RMHMC, AHMC does not attempt to construct a completely locally adaptive proposal sequence, but rather a fast local approximation to the fully adaptive case. This enables AHMC to bypass multiple fixed point evaluations in every step in the proposal sequence within every MC parameter draw that RMHMC needs to take. As a result, AHMC features a substantial speed gain and only a small loss of the degree of adaptability relative to RMHMC.

From the end-user perspective AHMC is easier to code than RMHMC, while the additional elements over HMC are simple to implement. AHMC is not a special case of RMHMC as their dynamic systems are non-nested, while HMC can be obtained as a special case of AHMC by imposing restrictions on the dynamics of the latter.

We lay out a set of necessary and sufficient conditions under which AHMC yields a valid MCMC scheme with a tractable form of its acceptance probability. In particular, these include a reversibility condition and a contraction mapping condition. As Girolami and Calderhead (2011) provide a detailed comparison of RMHMC to a number of alternative samplers including RW, MALA, and HMC on numerous examples showing overall supremacy of RMHMC in terms of effective sample size (ESS), it is sufficient for our purpose to take RMHMC as the benchmark of comparison. In order to uncover any potential trends in performance, we compare the ESS of AHMC to RMHMC on two simulated examples: one with increasing dimensionality of the parameter space and fixed sample size (multivariate Normal posterior) and one with increasing sample size and fixed dimensionality (GARCH(1,1)). Both examples reveal increasing efficiency gains of AHMC in dimensionality and sample size.

Bayesian estimation of multivariate GARCH models is relatively scarce (Dellaportas and Vrontos (2007), Hudson and Gerlach (2008) and Osiewalski and Pipien (2004)). Coming up with a good proposal density inside a Metropolis-Hasting procedure can be challenging. Therefore, we apply our procedure to a high-dimensional BEKK GARCH model with its highly complex likelihood. We show that AHMC facilitates inference on the model in higher dimensions than previously considered practical. The importance of full BEKK inference is highlighted by a marginal likelihood comparison that clearly favors the full model version over its restricted alternatives.

AHMC is related to but distinct from the adaptive radial-based direction sampling (ARDS) method of Bauwens, Bos, van Dijk, and van Oest (2004). While AHMC utilizes deterministic directional derivatives (numerical or analytical) of a Hamiltonian system in order to move within hypersurfaces of approximately equal functional value, ARDS is based on a transformation into radial coordinates, stochastic sampling of directional vectors, and then applying the inverse transformation. The acceptance probability of the Metropolis-Hastings version of ARDS (Proposition 1) is a function of a numerical quadrature over the posterior in a given direction. The importance sampling version of

ARDS relies on a directional approximation of the posterior. In either case, each MC draw of ARDS requires a certain type of relatively detailed posterior approximation which AHMC seeks to avoid in order to be applicable in problems where quadrature evaluation or importance sampling may become computationally prohibitive, as described above. Each method thus focuses on different types of applied problems.

Our work also complements other existing tailored proposal methods for posterior sampling in difficult situations such as Chib and Greenberg (1995), Chib and Ramamurthy (2010), Liesenfeld and Richard (2006) and Pitt and Shephard (1997). The AHMC is a useful addition to the applied econometrician's toolkit and can be applied to the full block of parameters as in our examples or a sub-block of parameters in conjunction with other Gibbs and Metropolis-Hasting steps.

The paper is organized as follows: Section 2 provides an overview of useful statistical background including the detailed balance condition of the Metropolis-Hastings principle. Section 3 reviews HMC and RMHMC, and Section 4 introduces AHMC. Section 5 explores the properties of AHMC on simulated examples and Section 6 details the application of AHMC to a high-dimensional BEKK GARCH model. Section 7 concludes.

2. Statistical Background

Consider an economic model parametrized by a Euclidean vector $\theta \in \Theta$ for which all information in the sample is contained in the model posterior $\pi(\theta; \cdot)$ that we denote by $\pi(\theta)$ which is assumed known up to an unknown integrating constant. Formally, a general class of such models can be characterized by a family $\mathcal{P}_\theta$ of probability measures on a measurable space $(\Theta, \mathcal{B})$ where $\mathcal{B}$ is the Borel $\sigma-$algebra.

The purpose of Markov Chain Monte Carlo (MCMC) methods is to formulate a Markov chain on the parameter space Θ for which, under certain conditions, $\pi(\theta) \in \mathcal{P}_\theta$ is the invariant (also called 'equilibrium' or 'long-run') distribution. The Markov chain of draws of θ can be used to construct simulation-based estimates of the required integrals and functionals $h(\theta)$ of θ that are expressed as integrals. These functionals include objects of interest for inference on θ such as quantiles of $\pi(\theta)$.

The Markov chain sampling mechanism specifies a method for generating a sequence of random variables $\{\theta_r\}_{r=1}^{R}$, starting from an initial point θ_0, in the form of conditional distributions for the draws $\theta_{r+1}|\theta_r \sim G(\theta_r)$. Under relatively weak regularity conditions (Robert and Casella, 2004), the average of the Markov chain converges to the expectation under the stationary distribution:

$$\lim_{R\to\infty} \frac{1}{R}\sum_{r=1}^{R} h(\theta_r) = E_\pi[h(\theta)]$$

A Markov chain with this property is called ergodic. As a means of approximation we rely on large but finite $R \in \mathbb{N}$ which the analyst has the discretion to select in applications.

The Metropolis-Hastings (M-H) principle has been the cornerstone of constructing Markov chains by sampling $\theta_{r+1}|\theta_r$ from $G(\theta_r)$; see Chib and Greenberg (1995) for a detailed overview. $G(\theta_r)$ can be obtained from a given (economic) model and its corresponding posterior $\pi(\theta)$, parametrized by θ, known up to a constant of proportionality.

However, $\pi(\theta)$ typically has a complicated form which precludes direct sampling. Then the goal is to find a transition kernel $P(\theta, d\theta)$ whose nth iterate converges to $\pi(\theta)$ for large n. After this large number, the distribution of the observations generated from the Markov chain simulation is approximately the target distribution. The transition kernel $P(\theta, A)$ for $\theta \in \Theta$ and $A \subset \Theta$ is an unknown conditional distribution function that represents the probability of moving from θ to a point in the set A. Suppose we have a proposal-generating density $q(\theta^*_{r+1}|\theta_r)$ where θ^*_{r+1} is a proposed state given the current state θ_r of the Markov chain. The Metropolis-Hastings (M-H) principle stipulates that θ^*_{r+1} be accepted as the next state θ_{r+1} with the acceptance probability

$$\alpha(\theta_r, \theta^*_{r+1}) = \min\left[\frac{\pi(\theta^*_{r+1})q(\theta_r|\theta^*_{r+1})}{\pi(\theta_r)q(\theta^*_{r+1}|\theta_r)}, 1\right] \tag{2.1}$$

otherwise $\theta_{r+1} = \theta_r$. Then the Markov chain satisfies the so-called detailed balance condition

$$\pi(\theta_r)q(\theta^*_{r+1}|\theta_r)\alpha(\theta_r, \theta^*_{r+1}) = \pi(\theta^*_{r+1})q(\theta_r|\theta^*_{r+1})\alpha(\theta^*_{r+1}, \theta_r)$$

which is sufficient for ergodicity. $\alpha(\theta^*_{r+1}, \theta_r)$ is the probability of the move $\theta_r|\theta^*_{r+1}$ if the dynamics of the proposal generating mechanism were to be reversed. While $\pi(\theta)$ may be difficult or expensive to sample from, the proposal-generating density $q(\theta^*_{r+1}|\theta_r)$ can be chosen to be sampled easily. The popular Gibbs sampler arises as a special case when the M-H sampler is factored into conditional densities.

A variation on (2.1) arises when the parameter space Θ is augmented with a set of independent auxiliary stochastic parameters $\gamma \in \Gamma$ that fulfill a supplementary role in the proposal algorithm, such as facilitating the directional guidance of the proposal mechanism. The detailed balance is then satisfied using the acceptance probability

$$\alpha(\theta_r, \gamma_r; \theta^*_{r+1}, \gamma^*_{r+1}) = \min\left[\frac{\pi(\theta^*_{r+1}, \gamma^*_{r+1})q(\theta_r, \gamma_r|\theta^*_{r+1}, \gamma^*_{r+1})}{\pi(\theta_r, \gamma_r)q(\theta^*_{r+1}, \gamma^*_{r+1}|\theta_r, \gamma_r)}, 1\right] \tag{2.2}$$

The desired posterior can be obtained by marginalizing out γ.

3. Hamiltonian Monte Carlo

The Hamiltonian (or Hybrid) Monte Carlo (HMC) algorithm originates from the physics literature where it was introduced as a fast method for simulating molecular dynamics (Duane, Kennedy, Pendleton, and Roweth, 1987). It has since become popular in a number of application areas including statistical physics (Akhmatskaya, Bou-Rabee, and Reich, 2009; Gupta, Kilcup, and Sharpe, 1988), computational chemistry (Tuckerman, Berne, Martyna, and Klein, 1993), or a generic tool for Bayesian statistical inference (Neal, 1993, 2010; Ishwaran, 1999; Liu, 2004; Beskos, Pillai, Roberts, Sanz-Serna, and Stuart, 2010). A separate stream of literature has developed around the Langevin diffusion mechanisms which use related proposal dynamics but utilize one-step proposals only (Roberts and Rosenthal, 1998; Roberts and Stramer, 2003).

In this Section we provide the stochastic background for HMC. Our synthesis is based on previously published material, but unlike the bulk of literature presenting HMC in terms of the physical laws of motion based on preservation of total energy in the phase-space, we take a fully stochastic perspective familiar to the applied Bayesian econometrician. The HMC principle is thus presented in terms of the joint density over the augmented parameter space leading to a Metropolis acceptance probability update. We hope that our synthesis of the probabilistic perspective on HMC will provide a useful point of reference, in particular to econometricians who wish to further explore the HMC principles.

3.1. **HMC Principle**

Consider a vector of parameters of interest $\theta \in \mathbb{R}^d$ distributed according to the posterior density $\pi(\theta)$. Let $\gamma \in \mathbb{R}^d$ denote a vector of auxiliary parameters with $\gamma \sim \Phi(\gamma; 0, M)$ where Φ denotes the Gaussian distribution with mean vector 0 and covariance matrix M, independent of θ. Denote the joint density of (θ, γ) by $\pi(\theta, \gamma)$. Then the *negative of the logarithm of the joint density of* (θ, γ) is given by the Hamiltonian equation[2]

$$H(\theta, \gamma) = -\ln \pi(\theta) + \frac{1}{2} \ln \left((2\pi)^d \, |M| \right) + \frac{1}{2} \gamma' M^{-1} \gamma \tag{3.1}$$

Hamiltonian Monte Carlo (HMC) is formulated in the following three steps that we will describe in detail further below:

1. Draw an initial auxiliary parameter vector $\gamma_r^0 \sim \Phi(\gamma; 0, M)$;
2. Transition from (θ_r, γ_r) to $(\theta_r^L, \gamma_r^L) = (\theta_{r+1}^*, \gamma_{r+1}^*)$ according to the Hamiltonian dynamics;
3. Accept $(\theta_{r+1}^*, \gamma_{r+1}^*)$ with probability $\alpha(\theta_r, \gamma_r; \theta_{r+1}^*, \gamma_{r+1}^*)$, otherwise keep (θ_r, γ_r) as the next MC draw.

Step 1 provides a stochastic initialization of the system akin to a RW draw. This step is necessary in order to make the resulting Markov chain $\{(\theta_r, \gamma_r)\}_{r=1}^R$ irreducible and aperiodic (Ishwaran, 1999). In contrast to RW, this so-called refreshment move is performed on the auxiliary variable γ as opposed to the original parameter of interest θ, setting $\theta_r^0 = \theta_r$. In terms of the HMC sampling algorithm, the initial refreshment draw of γ_r^0 forms a Gibbs step on the parameter space of (θ, γ) accepted with probability 1. Since it only applies to γ, it will leave the target joint distribution of (θ, γ) invariant and subsequent steps can be performed conditional on γ_r^0 (Neal, 2010).

[2] In the physics literature, θ denotes the position (or state) variable and $-\ln \pi(\theta)$ describes its potential energy, while γ is the momentum variable with kinetic energy $\gamma' M^{-1} \gamma / 2$, yielding the total energy $H(\theta, \gamma)$ of the system, up to a constant of proportionality. M is a constant, symmetric, positive-definite "mass" matrix which is often set as a scalar multiple of the identity matrix.

Step 2 constructs a sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$ according to the Hamiltonian dynamics starting from the current state (θ_r^0, γ_r^0) and setting the last member of the sequence as the HMC new state proposal $(\theta_{r+1}^*, \gamma_{r+1}^*) = (\theta_r^L, \gamma_r^L)$. The role of the Hamiltonian dynamics is to ensure that the M-H acceptance probability (2.2) for $(\theta_{r+1}^*, \gamma_{r+1}^*)$ is kept close to 1. As will become clear shortly, this corresponds to maintaining the difference $-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r^0, \gamma_r^0)$ close to zero throughout the sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$. This property of the transition from (θ_r, γ_r) to $(\theta_{r+1}^*, \gamma_{r+1}^*)$ can be achieved by conceptualizing θ and γ as functions of continuous time t and specifying their evolution using the Hamiltonian dynamics equations[3]

$$\frac{d\theta_i}{dt} = \frac{\partial H(\theta, \gamma)}{\partial \gamma_i} = \left[M^{-1}\gamma\right]_i \tag{3.2}$$

$$\frac{d\gamma_i}{dt} = -\frac{\partial H(\theta, \gamma)}{\partial \theta_i} = \nabla_{\theta_i} \ln \pi(\theta) \tag{3.3}$$

for $i = 1, \ldots, d$. For any discrete time interval of duration s, (3.2)–(3.3) define a mapping T_s from the state of the system at time t to the state at time $t + s$. For practical applications of interest these differential equations (3.2)–(3.3) in general cannot be solved analytically and instead numerical methods are required. The Stormer-Verlet (or leapfrog) numerical integrator (Leimkuhler and Reich, 2004) is one such popular method, discretizing the Hamiltonian dynamics as

$$\gamma(t + \varepsilon/2) = \gamma(t) + (\varepsilon/2)\nabla_\theta \ln \pi(\theta(t)) \tag{3.4}$$

$$\theta(t + \varepsilon) = \theta(t) + \varepsilon M^{-1}\gamma(t + \varepsilon/2) \tag{3.5}$$

$$\gamma(t + \varepsilon) = \gamma(t + \varepsilon/2) + (\varepsilon/2)\nabla_\theta \ln \pi(\theta(t + \varepsilon)) \tag{3.6}$$

for some small $\varepsilon \in \mathbb{R}$. From this perspective, γ plays the role of an auxiliary variable that parametrizes (a functional of) $\pi(\theta, \cdot)$ providing it with an additional degree of flexibility to maintain the acceptance probability close to one for every k. Even though $\ln \pi(\theta_r^k)$ can deviate substantially from $\ln \pi(\theta_r^0)$, resulting in favorable mixing for θ, the additional terms in γ in (3.1) compensate for this deviation maintaining the overall level of $H(\theta_r^k, \gamma_r^k)$ close to constant over $k = 1, \ldots, L$ when used in accordance with (3.4)–(3.6), since $\frac{\partial H(\theta,\gamma)}{\partial \gamma_i}$ and $\frac{\partial H(\theta,\gamma)}{\partial \theta_i}$ enter with the opposite signs in (3.2)–(3.3). In contrast, without the additional parametrization with γ, if only $\ln \pi(\theta_r^k)$ were to be used in the proposal mechanism as is the case in RW style samplers, the M-H acceptance probability would often drop to zero relatively quickly.

Step 3 applies a Metropolis correction to the proposal $(\theta_{r+1}^*, \gamma_{r+1}^*)$. In continuous time, or for $\varepsilon \to 0$, (3.2)–(3.3) would keep $-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r, \gamma_r) = 0$ exactly resulting in $\alpha(\theta_r, \theta_{r+1}^*) = 1$ but for discrete $\varepsilon > 0$, in general, $-H(\theta^*, \gamma^*) + H(\theta, \gamma) \neq 0$ necessitating the Metropolis step. A key feature of HMC is that the generic M-H acceptance probability (2.2) can be expressed in a simple tractable form using only the posterior density $\pi(\theta)$ and the auxiliary parameter Gaussian density $\phi(\gamma; 0, M)$. The transition from (θ_r^0, γ_r^0) to (θ_r^L, γ_r^L) via the proposal sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$ taken according to the

[3] In the physics literature, the Hamiltonian dynamics describe the evolution of (θ, γ) that keeps the total energy $H(\theta, \gamma)$ constant.

discretized Hamiltonian dynamics (3.4)–(3.6) is fully deterministic proposal, placing a Dirac delta probability mass $\delta(\theta_r^k, \gamma_r^k) = 1$ on each (θ_r^k, γ_r^k) conditional on (θ_r^0, γ_r^0). The system (3.4)–(3.6) is time reversible and symmetric in (θ, γ), which implies that the forward and reverse transition probabilities $q(\theta_r^L, \gamma_r^L | \theta_r^0, \gamma_r^0)$ and $q(\theta_r^0, \gamma_r^0 | \theta_r^L, \gamma_r^L)$ are equal: this simplifies the Metropolis-Hastings acceptance ratio in (2.2) to the Metropolis form $\pi(\theta_{r+1}^*, \gamma_{r+1}^*)/\pi(\theta_r^0, \gamma_r^0)$. From the definition of the Hamiltonian $H(\theta, \gamma)$ in (3.1) as the negative of the log-joint densities, the joint density of (θ, π) is given by

$$\pi(\theta, \gamma) = \exp\left[-H(\theta, \gamma)\right] = \pi(\theta) \left((2\pi)^d |M|\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma' M^{-1} \gamma\right) \tag{3.7}$$

Hence, the Metropolis acceptance probability takes the form

$$\begin{aligned}
\alpha(\theta_r, \gamma_r; \theta_{r+1}^*, \gamma_{r+1}^*) &= \min\left[\frac{\pi(\theta_{r+1}^*, \gamma_{r+1}^*)}{\pi(\theta_r^0, \gamma_r^0)}, 1\right] \\
&= \min\left[\exp\left(-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r^0, \gamma_r^0)\right), 1\right] \\
&= \min\left[\exp\left(\ln \pi(\theta_{r+1}^*) - \ln \pi(\theta_r^0) + \ln \phi(\gamma_{r+1}^*; 0, M) - \ln \phi(\gamma_r^0; 0, M)\right), 1\right]
\end{aligned}$$

The expression for $\alpha(\theta_r, \gamma_r; \theta_{r+1}^*, \gamma_{r+1}^*)$ shows, as noted above, that the HMC acceptance probability is given in terms of the difference of the Hamiltonian equations $H(\theta_r^0, \gamma_r^0) - H(\theta_{r+1}^*, \gamma_{r+1}^*)$. The closer can we keep this difference to zero, the closer the acceptance probability is to one. A key feature of the Hamiltonian dynamics (3.2)–(3.3) in Step 2 is that they maintain $H(\theta, \gamma)$ constant over the parameter space in continuous time conditional on $H(\theta_r^0, \gamma_r^0)$ obtained in Step 1, while their discretization (3.4)–(3.6) closely approximates this property for discrete time steps $\varepsilon > 0$ with a global error of order ε^2 corrected by the Metropolis update in Step 3.

3.2. Metropolis adjusted Langevin (MALA)

The Langevin algorithm is equivalent to a special case of HMC when the number of leapfrog steps $L = 1$. In this case, the proposal from the current state θ_r to the proposal θ_{r+1}^* can be expressed as

$$\theta_{r+1}^* = \theta_r + (\varepsilon^2/2)\nabla_\theta \ln \pi(\theta(t)) + \varepsilon z_r$$

where $z \sim N(z; 0, I_d)$ (Ishwaran, 1999). Using a preconditioning matrix M such that

$$\theta_{r+1}^* = \theta_r + (\varepsilon^2/2) M \nabla_\theta \ln \pi(\theta(t)) + \varepsilon U z_r$$

with U obtained via Cholesky decomposition satisfying $M = UU^T$ can further improve the Langevin sampling properties (Roberts and Stramer, 2003).

4. Adaptive Hamiltonian Monte Carlo

Compared to HMC, the ratio in the acceptance probability (2.2) can be kept closer to 1 for higher dimensions of Θ and farther proposals θ_{r+1}^* if the mechanism generating the proposal sequence is allowed to adapt to the curvature of $\pi(\theta)$ with the proposal sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$. Hence, the goal of local adaptivity of HMC is to render the proposal mechanism responsive to the local curvature of the log-likelihood function $\ln \pi(\theta)$.

4.1. Non-separable Hamiltonian Systems

In the adaptive case the Hamiltonian equation takes the form

$$H(\theta,\gamma) = -\ln \pi(\theta) - \ln q(\gamma|\theta) \tag{4.1}$$

where

$$q(\gamma|\theta) = (2\pi)^{-d/2} \left|M(\theta)\right|^{-1/2} \exp\left(-\frac{1}{2}\gamma' M(\theta)^{-1}\gamma\right) \tag{4.2}$$

renders the auxiliary parameter quadratic term $\gamma' M(\theta)^{-1}\gamma/2$ as an explicit function of θ. This property leads to local adaptability of the proposal sequence but also complicates subsequent analysis substantially. The associated Hamiltonian dynamics equations are in general given by

$$\frac{d\theta_i}{dt} = \frac{\partial H(\theta,\gamma)}{\partial \gamma_i} = \left[M(\theta)^{-1}\gamma\right]_i \tag{4.3}$$

$$\begin{aligned}\frac{d\gamma_i}{dt} &= -\frac{\partial H(\theta,\gamma)}{\partial \theta_i} = \nabla_\theta \ln \pi(\theta) - \frac{1}{2}\mathrm{Tr}\left(M(\theta)^{-1}\frac{\partial M(\theta)}{\partial \theta_i}\right)\\ &\qquad + \frac{1}{2}\gamma' M(\theta)^{-1}\frac{\partial M(\theta)}{\partial \theta_i}M(\theta)^{-1}\gamma\end{aligned} \tag{4.4}$$

A number of numerical methods have been devised in the physics and molecular dynamics literature to solve the differential equations (4.3)–(4.4) in order to accurately determine the position of $\theta(t+s)$ and $\gamma(t+s)$ at the next instant $t+s$ given their current position at time t in the state space. These solutions include the generalized Euler and Stormer-Verlet (generalized leapfrog) methods (Hairer, Lubich, and Wanner, 2003; Leimkuhler and Reich, 2004).

The choice of the metric $M(\theta)$ in (4.2) critically influences the properties of the resulting sampler. In general, any non-degenerate form of $M(\theta)$ can lead to an HMC-based method, with $M(\theta) = I_d$, the identity matrix, resulting in HMC as a special case. In a seminal paper, Girolami and Calderhead (2011) provide theoretical justification for combining the generalized leapfrog integrator with the choice of the Fisher information matrix of $\pi(\theta)$ for $M(\theta)$, yielding RMHMC. However, the generalized leapfrog necessitates finding one implicit fixed point in γ and another one in θ at every proposal step k in each dimension i of γ and θ (see Appendix E for details). These fixed points need to be obtained numerically which can lead to prohibitive computational burden in cases where evaluation of the likelihood is expensive, such as in high-dimensional or recursive problems.

4.2. AHMC

In this paper we propose the Adaptive HMC (AHMC), an alternative HMC-based method featuring distant proposals that is locally adaptable and yet avoids determination of two fixed points at every step k of the proposal sequence. Similarly to HMC, the M-H acceptance probability (2.2) can be expressed in the tractable Metropolis form. Showing that AHMC satisfies the conditions for a valid

MCMC scheme is a challenging task that we undertake in Theorem 1 below. Results of this type have been obtained for HMC and RMHMC in the literature, but the AHMC is a non-nested distinct alternative to either of these methods and hence needs to be validated separately. Theorem 2 then lays out the set of conditions on the (posterior) likelihood that are sufficient for satisfying the assumptions made in Theorem 1. These conditions can be easily verified in a given application.

The starting point for AHMC is the non-separable Hamiltonian (4.1)-(4.2). Instead of $M(\theta)$, for each MCMC update r, we use the matrix $M(\overline{\theta_r, \theta^*_{r+1}})$ that is *fixed constant* for the entire leapfrog multi-step proposal sequence $\{\theta^k_r, \gamma^k_r\}^L_{k=1}$, i.e. between θ_r and θ^*_{r+1} inclusive.

A key feature of this approach is that $M(\overline{\theta_r, \theta^*_{r+1}})$ is permitted to change for each MCMC iteration. For a given r, $M(\overline{\theta_r, \theta^*_{r+1}})$ is not a function of θ. Hence the formation of the proposal sequence $\{\theta^k_r, \gamma^k_r\}^L_{k=0}$ can use the standard leapfrog integrator of HMC with the mapping given by (3.4)–(3.6) with M replaced by $M(\overline{\theta_r, \theta^*_{r+1}})$ written as,

$$\gamma_r^{k+1/2} = \gamma_r^k - \frac{\varepsilon}{2}\nabla_\theta \ln \pi(\theta_r^k) \tag{4.5}$$

$$\theta_r^{k+1} = \theta_r^k + \varepsilon\left[M(\overline{\theta_r, \theta^*_{r+1}})^{-1}\gamma_r^{k+1/2}\right] \tag{4.6}$$

$$\gamma_r^{k+1} = \gamma_r^{k+1/2} - \frac{\varepsilon}{2}\nabla_\theta \ln \pi(\theta_r^{k+1}) \tag{4.7}$$

This is in contrast to Girolami and Calderhead (2011) which used a different numerical integrator and requires finding two fixed points for each k. Our approach requires one fixed point for all $k = 1, ..., L$.

In summary, we expect our approach to be more computationally efficient relative to Girolami and Calderhead (2011). First we use (4.5)-(4.7) instead of the more complex integrator associated with (4.3)-(4.4). Second, we require one fixed point in contrast to many fixed points along the proposal path.

The following assumption states a sufficient condition for AHMC to yield a valid MCMC scheme satisfying detailed balance.

ASSUMPTION 1. *$M(\overline{\theta_r, \theta^*_{r+1}})$ is symmetric in its arguments, satisfying*

$$M(\overline{\theta_r, \theta^*_{r+1}}) = M(\overline{\theta^*_{r+1}, \theta_r})$$

This Assumption guarantees the symmetry between the forward mapping sequence T_k and the reversed mapping sequence applying T_k with reversed signs starting at (θ^L_r, γ^L_r). This ensures that the forward proposal sequence $\{\theta^k_r, \gamma^k_r\}^L_{k=0}$ follows the *same path* as the reverse proposal sequence. This symmetry of the proposal mapping sequence fulfills the same role as in HMC, resulting in compliance of AHMC with detailed balance, as shown in Theorem 1 below.

There are many potential ways of setting $M(\overline{\theta_r, \theta^*_{r+1}})$. We take a user-friendly approach with light computational burden and specify

$$M(\overline{\theta_r, \theta^*_{r+1}}) = \frac{1}{2}\left[F(\theta_r) + F(\theta^*_{r+1})\right] \tag{4.8}$$

where $F(\theta)$ is the Fisher information matrix evaluated at θ. The value of $M(\overline{\theta_r, \theta^*_{r+1}})$ that complies with Assumption 1 is then obtained as one fixed point in $\{T_k\}_{k=1}^L$ per proposal draw $(\theta^*_{r+1}, \gamma^*_{r+1})$. Given θ_r, $F(\theta_r)$, and an initial guess $F(\theta^*_{r+1}) = F(\theta_r)$, we take L steps of (4.5)-(4.7) with $k = 1, \ldots, L$, then update $F(\theta^*_{r+1})$ and (4.8), and keep iterating until convergence of $(\theta^*_{r+1}, \gamma^*_{r+1})$ and hence $M(\overline{\theta_r, \theta^*_{r+1}})$ to a fixed point. Conditional on this $M(\overline{\theta_r, \theta^*_{r+1}})$, the actual proposal sequence $\{\theta^k_r, \gamma^k_r\}_{k=0}^L$ with $(\theta^*_{r+1}, \gamma^*_{r+1}) = (\theta^L_r, \gamma^L_r)$ is then drawn by applying (4.5)-(4.7). The conditions for a contraction mapping discussed below ensure the existence and uniqueness of the fixed point. In our experiments we found that only a few iterations were necessary to obtain the fixed point, resulting in relatively rapid speed of the MCMC updates. The exact AHMC algorithm is given in Appendix C.

The HMC results as a special case of AHMC for a globally constant matrix M over the entire parameter space of (θ, γ). As another special case when $\ln \pi(\theta)$ has a globally constant curvature with respect to θ, such as when $\theta = \mu$ for data $y \sim \mathcal{N}(\mu, I)$, the AHMC produces draws equivalent to the HMC. In general, however, when the curvature of $\ln \pi(\theta)$ changes as a function of θ, such as in $\theta = (\mu, \Sigma)$ for data $y \sim \mathcal{N}(\mu, \Sigma)$, AHMC exploits the shape of $\ln \pi(\theta)$ by locally adapting the proposal dynamics to the curvature of $\ln \pi(\theta)$.

A key feature of AHMC, in line with other HMC-based schemes, is that simplifies the acceptance probability (2.2) to the Metropolis form containing only the ratio of the joint densities of (θ, γ). This feature provides for a relatively user-friendly implementation of the algorithm. Theorem 1 shows both detailed balance and the Metropolis property for AHMC.

THEOREM 1. *Conditional on $M(\overline{\theta_r, \theta^*_{r+1}})$ given by Assumption 1, the AHMC satisfies detailed balance, with the acceptance probability*

$$\begin{aligned} \alpha(\theta_r, \gamma_r; \theta^*_{r+1}, \gamma^*_{r+1}) &= \min\left[\frac{\pi(\theta^*_{r+1}, \gamma^*_{r+1})}{\pi(\theta^0_r, \gamma^0_r)}, 1\right] \\ &= \min\left[\exp\left(\widetilde{\alpha}_r\right), 1\right] \end{aligned} \tag{4.9}$$

where

$$\widetilde{\alpha}_r = \ln \pi(\theta^*_{r+1}) - \ln \pi(\theta^0_r) + \ln \phi(\gamma^*_{r+1}; 0, M(\overline{\theta_r, \theta^*_{r+1}})) - \ln \phi(\gamma^0_r; 0, M(\overline{\theta_r, \theta^*_{r+1}}))$$

The proof is given in the Appendix A. Theorem 1 is stated as conditional on $M(\overline{\theta_r, \theta^*_{r+1}})$ obtained as a fixed point whose existence is not guaranteed to hold in general. Here we state a set of sufficient conditions for when this is the case.

ASSUMPTION 2. *$\nabla_\theta \ln \pi(\theta)$ is globally bounded and Lipschitz continuous in θ.*

ASSUMPTION 3. *The parameter space Θ is compact.*

THEOREM 2. *Under the Assumptions 2–3, the fixed point defining $M(\overline{\theta_r, \theta^*_{r+1}})$ exists and is unique for any given θ_r. In particular, for any $\delta \in (0, 1)$ there exists $\varepsilon(\delta) > 0$ dependent on δ only, such that $\forall \varepsilon^* < \varepsilon(\delta)$, $\{T_k\}_{k=1}^L$ is a contraction mapping uniquely determining $M(\overline{\theta_r, \theta^*_{r+1}})$.*

The proof is provided in the Appendix B.

5. Illustrative Examples

In this Section we assess the performance of AHMC on two stylized illustrative examples. Girolami and Calderhead (2011) provide an excellent exposition of a series of problems that highlight the performance edge of RMHMC relative to the non-adaptive HMC, MALA, and RW. Hence, to establish the performance merit of AHMC it is sufficient to take RMHMC as the benchmark of comparison. We first examine sampling of the parameters in multivariate Normal density in Example 1, and then sampling of the parameters in a univariate GARCH(1,1) model in Example 2. In order to uncover any potential trends, in Example 1 we fix the sample size and increase the parameter dimensionality; in Example 2 we fix the dimensionality and increase the sample size.

We compare the relative efficiency of AHMC and RMHMC by using the same approach as Girolami and Calderhead (2011) and Holmes and Held (2006) in making their comparisons. For each example and method, we calculate the effective sample size (ESS) using the posterior samples for each parameter obtained in 10,000 iterations with 5,000 burnin section. The ESS is the number of effectively independent draws from the posterior distribution that the Markov chain is equivalent to. The ESS thus serves as an estimate of the number of independent samples needed to obtain a parameter estimate with the same precision as the MCMC estimate considered based on a given number of dependent samples. The nominal ESS is calculated as $ESS^* = R\left[1 + 2\sum_j \gamma(j)\right]^{-1}$ where R is the number of posterior samples, and $\gamma(j)$ is the monotone sample autocorrelation (Geyer, 1992). The nominal ESS is then normalized for CPU run time required to obtain the given Markov chain of posterior draws, yielding $ESS = 100 \times ESS^*/S$ where S is the number of seconds of CPU run time. The MCMC chains were obtained on a 2.6 GHz unix workstation with the Intel fortran 95 compiler. For obtaining ESS^* from the MCMC output chains we used the R package coda. All results reported are the averages of 10 different runs.

The results for the following examples are given in Tables 1 and 2 and Figures 1 and 2 below. We report the mean, standard deviation, minimum, and maximum ESS for the sampled parameter vector for each simulation setup. We also report the nominal (unnormalized) numbers along with the CPU run time as the ESS inputs. In the tables Ratio denotes the ratio of AHMC to RMHMC of the respective statistics. Values greater than 1 indicate better performance of AHMC. Figures 1 and 2 plot the relative efficiency gain of AHMC over RMHMC, calculated as the ESS means ratio for the two methods. Figure 1 shows the AHMC relative efficiency gain for increasing dimensionality and fixed sample size in Example 1, and Figure 1 for fixed dimensionality and increasing sample size in Example 2. In each Figure, the horizontal dotted line at y-value 1 marks theoretical equivalence of both methods, while the region above 1 represents efficiency gains of AHMC.

5.1. Example 1: Joint Sampling of Parameters of a Multivariate Normal Density

Let $\mathbf{y}_t \sim \mathcal{N}(\mathbf{y}|\mu, \Sigma)$ for $t = 1, \ldots, T$ with

$$\ln \pi(\theta) = -\frac{Td}{2} \ln(2\pi) - \frac{T}{2} \ln |\Sigma| - \frac{1}{2} \sum_{t=1}^{T} (\mathbf{y}_t - \mu)' \Sigma^{-1} (\mathbf{y}_t - \mu)$$

and $\theta \equiv (\mu', vech(\Sigma)')'$. Naturally, a convenient factorization of this problem is readily available, but this stylized example is meant to serve for joint sampling comparison purposes on a familiar and analytically tractable case. In general applications, a conditional factorization of the joint density $\ln \pi(\theta)$ may not be available or practical to implement (this is for instance the case of the BEKK GARCH model analyzed in the next Section). In the simulation study of Example 1, we vary $\dim(\mathbf{y})$ from 2 to 6, which corresponds to the parameter dimensionality $\dim(\theta)$ varying from 3 to 27. The true parameter values were set to $\mu_{\mathbf{0}} = 0$, and Σ to equal the covariance matrix of a first-order autoregressive process with correlation 0.5. Our prior restricts Σ to be positive definite. Each chain was initialized at the true parameter values, with $L = 100$ leapfrog steps, and ϵ tuned to achieve acceptance rates between 0.7 and 0.9. The ESS statistics are reported in Table 1 and Figure 1.

5.2. Example 2: Joint Sampling of GARCH (1,1) Parameters

Let $y_t \sim \mathcal{N}(y|0, \sigma_t^2)$ with $\sigma_t^2 = \gamma + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2$ for $t = 1, \ldots, T$ and $\theta \equiv (\gamma, \alpha, \beta)$ where

$$\begin{aligned}
\ln \pi(\theta) &= -\frac{T}{2} \ln(2\pi) - \frac{1}{2} \sum_{t=1}^{T} \ln(\sigma_t^2(\theta)) - \frac{1}{2} \sum_{t=1}^{T} y_t \sigma_t^{-2}(\theta) \\
\frac{\partial}{\partial \theta} \ln \pi(\theta) &= -\frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_t^2} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} + \frac{1}{2} \sum_{t=1}^{T} e_t^2 \sigma_t^{-4} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} \\
\frac{\partial \sigma_t^2(\theta)}{\partial \gamma} &= 1 + \beta \frac{\partial \sigma_{t-1}^2(\theta)}{\partial \gamma} \\
\frac{\partial \sigma_t^2(\theta)}{\partial \alpha} &= e_{t-1}^2 + \beta \frac{\partial \sigma_{t-1}^2(\theta)}{\partial \alpha} \\
\frac{\partial \sigma_t^2(\theta)}{\partial \beta} &= \sigma_{t-1}^2 + \beta \frac{\partial \sigma_{t-1}^2(\theta)}{\partial \beta}
\end{aligned}$$

and $F(\theta)$ is consistently estimated using the average of the outer products of the scores. In this simulation study we vary the sample size T from 200 to 600. The dimensionality of the parameter space of θ is kept constant at 3. The true parameter values were set to $\gamma_0 = 0.1$, $\alpha_0 = 0.05$ and $\beta_0 = 0.9$. Each chain was initialized at the true parameter values, with $L = 100$ leapfrog steps, and ϵ tuned to achieve acceptance rates between 0.7 and 0.9. The ESS statistics are reported in Table 2 and Figure 2.

In summary, the improvement of AHMC over RMHMC is substantial, with up to 17-fold efficiency gain in Example 1 and up to 10-fold efficiency gain in Example 2. In both examples, the improvement keeps increasing with increasing dimensionality and sample size, indicating sustained efficiency gain of AHMC for more complex and sizeable problems.

Table 1: Simulation Results for Example 1

Variable dimension		2	3	4	5	6
Parameter dimension		5	9	14	20	27
CPU Time (s)	AHMC	40.65	140.74	440.69	2282.66	5653.66
	RMHMC	36.1	214.47	2139.97	24148.21	93067.35
	Ratio	1.13	0.66	0.21	0.08	0.06
Nominal ESS mean	AHMC	543.26	363.52	253.52	189.68	223.65
	RMHMC	495.25	390.87	267.69	184.79	214.29
ESS mean	AHMC	1336.88	258.34	57.54	6.65	3.96
	RMHMC	1372.46	182.25	12.51	0.61	0.23
	Ratio	0.98	1.42	4.61	8.7	17.22
Nominal ESS s.d.	AHMC	100.51	71.95	50.83	40.58	53.66
	RMHMC	150.53	22.78	18.18	11.78	24.69
ESS s.d.	AHMC	247.31	51.09	11.53	1.42	0.95
	RMHMC	416.77	10.63	0.85	0.04	0.03
Nominal ESS min	AHMC	434.33	265.55	175.8	125.54	132.88
	RMHMC	324.08	369.8	249.24	174.42	197.37
ESS min	AHMC	1068.93	188.96	39.91	4.4	2.35
	RMHMC	898.56	172.41	11.65	0.58	0.21
	Ratio	1.28	1.1	3.44	6.11	11.13
Nominal ESS max	AHMC	667.48	466.48	344.34	260.78	314.3
	RMHMC	600.76	412.25	284.33	196.77	242.52
ESS max	AHMC	1642.71	331.48	78.13	9.14	5.56
	RMHMC	1664.81	192.23	13.29	0.65	0.26
	Ratio	0.99	1.73	5.89	11.23	21.5

Figure 1: AHMC Efficiency Gain in Example 1

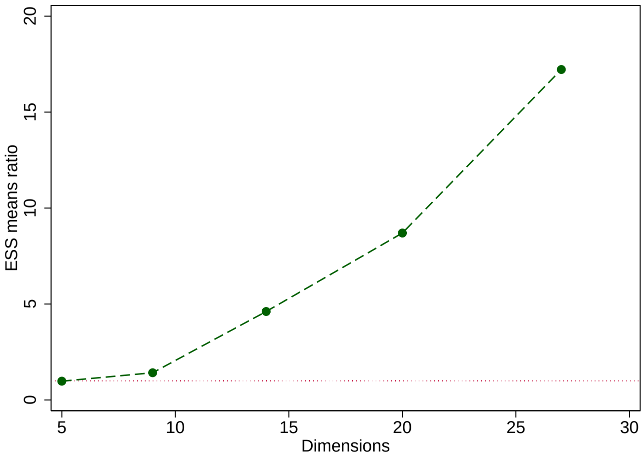
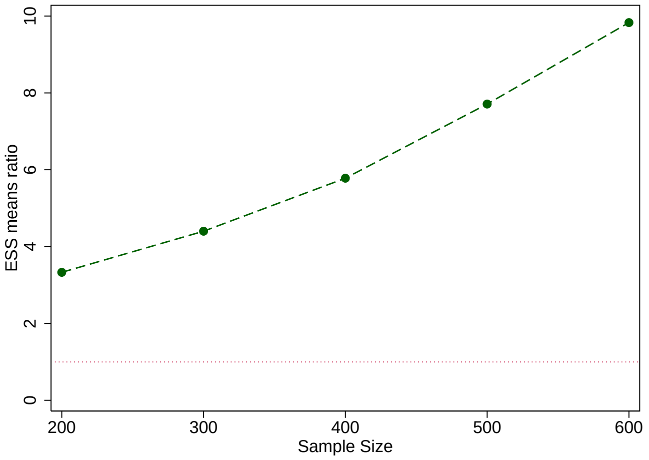


Table 2: Simulation Results for Example 2

Sample size T		200	300	400	500	600
Parameter dimension		3	3	3	3	3
CPU Time (s)	AHMC	50.81	71.27	98.39	124.7	152.69
	RMHMC	212.53	312.29	423.42	519.86	623.52
	Ratio	0.24	0.23	0.23	0.24	0.24
Nominal ESS mean	AHMC	15.7	21.22	20.49	17.54	24.76
	RMHMC	22.26	21.82	17.56	14.53	14.45
ESS mean	AHMC	30.34	29.36	20.87	13.85	16.25
	RMHMC	10.47	6.98	4.15	2.79	2.32
	Ratio	3.33	4.4	5.78	7.71	9.83
Nominal ESS s.d.	AHMC	17.66	28.36	27.69	23.08	35.13
	RMHMC	28.37	30.21	23.98	17.01	18.5
ESS s.d.	AHMC	34.18	39.15	28.2	18.11	23.06
	RMHMC	13.35	9.67	5.67	3.27	2.97
Nominal ESS min	AHMC	4.5	3.25	3.14	3.07	3.74
	RMHMC	4.28	3.04	2.89	2.99	2.85
ESS min	AHMC	8.54	4.52	3.17	2.47	2.46
	RMHMC	2.01	0.97	0.68	0.58	0.46
	Ratio	4.29	4.58	5	4.96	6.03
Nominal ESS max	AHMC	36.03	53.84	52.33	44.06	65.3
	RMHMC	54.91	56.62	45.23	34.02	35.75
ESS max	AHMC	69.67	74.38	53.3	34.66	42.86
	RMHMC	25.83	18.12	10.68	6.54	5.73
	Ratio	3.2	4.36	5.82	8.89	12.37

Figure 2: AHMC Efficiency Gain in Example 2



5.3. **Discussion**

The results show that AHMC clearly outperforms RMHMC in terms of the ESS in all simulation scenarios, except in Example 1 for the smallest dimension where both methods are comparable. The performance edge of AHMC over RMHMC increases rapidly with higher parameter dimensions in Example 1 and with larger sample size for the same parameter vector in Example 2. Both increasing the dimensionality and sample size add additional heavy computational load to the RMHMC in its fixed point iterations that AHMC avoids. These examples highlight the benefits of AHMC on interesting cases in order to motivate its use in applications.

6. **BEKK GARCH Application**

Interest in modeling the volatility dynamics of time-series data continues to grow and be important in many areas of empirical economics and finance. Generally, the literature on multivariate asset return modeling has moved to using more parsimonious models such as Engle (2002), Engle, Shephard, and Sheppard (2009) and Ding and Engle (2001). These approaches put restrictions on the volatility dynamics and feature two-step estimation and approximations to the likelihood. This makes estimation and inference feasible for a larger class of assets. However, it is desirable to consider more flexible models such as the BEKK model of Engle and Kroner (1995) and to perform full likelihood based inference. The BEKK model is one of the most flexible GARCH models that maintain positive definite conditional covariances at the expense of a large number of parameters. Although inference of the model with 2 or 3 assets have appeared in the literature we are not aware of anything beyond this asset dimension. An important question is how much do we lose in terms of statistical fit in moving from a BEKK model to a restricted model with less parameters to estimate. The extension to HMC discussed above provides an approach that can deal with the larger dimensions in the parameter space and jointly estimate the BEKK model in one run and compare the model to restricted versions.

Let r_t be a $N \times 1$ vector of asset returns with $t = 1, \dots, T$ and denote the information set as $\mathcal{F}_{t-1} = \{r_1, \dots, r_{t-1}\}$. We assume returns follow

$$r_t|\mathcal{F}_{t-1} \sim NID(0, H_t) \tag{6.1}$$

$$H_t = CC' + F' r_{t-1} r_{t-1}' F + G' H_{t-1} G. \tag{6.2}$$

H_t is a positive definite $N \times N$ conditional covariance matrix of r_t given information at time $t-1$, C is a lower triangular matrix and F and G are $N \times N$ matrices. Since our main focus is on sampling a complex posterior with many parameters we maintain a Gaussian assumption and a zero intercept for simplicity.[4] The total number of parameters in this model is $N(N+1)/2 + 2N^2$.

[4] Although not estimated, we expect our method could be extended to other innovation distributions such as multivariate Student-t with little modification.

In the following we focus on the full BEKK model in (6.2) but also consider some restricted versions. The first imposes F and G to be diagonal matrices which results in $N(N+1)/2 + 2N$ parameters. The second imposes diagonal matrices on all parameter matrices C, F and G and has $3N$ parameters.

The data is percent log-differences of foreign exchange spot rates for AUD/USD, GBP/USD, CAD/USD, EUR/USD, and JPY/USD from 2000/01/05 - 2006/10/11, (1700 observations). A time series plot of the five ($N = 5$) series is in Figure 3 and summary statistics are in Table 3. The sample mean for all series is close to 0 and excess kurtosis is fairly small. The sample correlations indicate all series tend to move together.

With $N = 5$ there are 65 model parameters in the full BEKK model while there are 25 and 15 parameters, respectively, in the two restricted models. To start the GARCH recursion H_1 is set to the sample covariance of the first 20 observations. The priors are set to independent N(0,100). For identification, the diagonal elements of C and the first element of both F and G are restricted to be positive (Engle and Kroner, 1995). These restrictions are enforced by dropping any parameter draw that violates this. We utilize the analytical expressions for the gradient from Hafner and Herwartz (2008). Starting from a point of high posterior mass we collect a total of 30,000 posterior draws for inference, with 10,000 burnin section. These computations took on the order of 2 days.

Collecting the parameters in $\theta = (vech(C)', vech(F)', vech(G)')'$, Figure 4 displays the conditional log-posterior $\log p(\theta_i|\theta_{-i}, \mathcal{F}_T)$ where θ_{-i} is set to a high probability mass point. Some of the conditional densities are approximately quadratic while others display a more complicated structure. The flat regions in the log-posterior will present challenges to maximizing this function or to obtaining a hessian estimate to compute standard errors in a classical approach.

Figure 5 displays the posterior mean of the conditional correlations for the full BEKK model and the two restricted versions. The BEKK model being the most flexible displays differences with the other models most notably the version that enforces diagonal matrices on C, G, F. That restriction implies unconditional correlations of 0 between assets and is at odds with the sample correlations in Table 3.

These differences in the models are confirmed by the marginal likelihoods reported in Table 4. The marginal likelihoods are estimated following Gelfend and Dey (1994) using a thin tailed truncated normal following Geweke (2005). The evidence is strongly against both of the restricted diagonal models. For example, the log-Bayes factor in favor of the full BEKK model is about 35 compared to the model with diagonal F, G.

In conclusion, our results support the use of the most flexible BEKK model and the AHMC sampler provides a feasible method to sample from a highly complex posterior density effectively.

7. Conclusion

Hamiltonian Monte Carlo (HMC) uses Hamiltonian dynamics in constructing distant proposal draws in a sequence of steps, yielding relatively low correlation among draws and high acceptance probabilities at the same time. In this paper we propose a local adaptation of HMC, the Adaptive Hamiltonian Monte Carlo (AHMC), whereby the proposal sequence follows the local evolution of the parameter space. We show that AHMC yields a valid MCMC procedure satisfying detailed balance. We show that AHMC outperforms in terms of effective sample size the existing locally adaptable HMC-based method – RMHMC – which has been shown elsewhere to outperform HMC and other alternative samplers on a number of cases. In our simulations, the relative performance improvement of AHMC over RMHMC becomes more pronounced with higher dimensionality of the parameter space and the sample size. We apply AHMC to a high-dimensional BEKK GARCH model in 56 parameter dimensions, which substantially exceeds the dimensionality utilized in previous work. Model comparison via marginal likelihood further reveals that the full BEKK model is preferable to its restricted versions with constraints placed on various covariance components, motivating the full high-dimensional implementation of the model.

8. Appendix A: Proof of Theorem 1

Denote by $q(\theta^*_{r+1},\gamma^*_{r+1};\theta^0_r,\gamma^0_r)$ the proposal density and by $q(\theta^0_r,\gamma^0_r;\theta^*_{r+1},\gamma^*_{r+1})$ the reverse proposal density. Given (θ^0_r,γ^0_r), $q(\theta^*_{r+1},\gamma^*_{r+1};\theta^0_r,\gamma^0_r)$ is constructed by the method of change of variables based on the sequence of steps given by the mapping T_k for $k=1,\dots,L$. Since T_k is deterministic, placing the Dirac delta $\delta(\cdot,\cdot)=1$ unit probability mass at each (θ^k_r,γ^k_r), applying successive transformations T_k yields

$$\begin{aligned} q(\theta^*_{r+1},\gamma^*_{r+1};\theta^0_r,\gamma^0_r) &= \left|\det \nabla T(\theta^L_r,\gamma^L_r;\theta^{L-1}_r,\gamma^{L-1}_r)\right|^{-1} \times \left|\det \nabla T(\theta^{L-1}_r,\gamma^{L-1}_r;\theta^{L-2}_r,\gamma^{L-2}_r)\right|^{-1} \times \dots \\ &\quad \dots \times \left|\det \nabla T(\theta^2_r,\gamma^2_r;\theta^1_r,\gamma^1_r)\right|^{-1} \left|\det \nabla T(\theta^1_r,\gamma^1_r;\theta^0_r,\gamma^0_r)\right|^{-1} \delta(\gamma^0_r,\theta^0_r) \end{aligned} \tag{8.1}$$

where $\nabla T(\theta^k_r,\gamma^k_r;\theta^{k-1}_r,\gamma^{k-1}_r)$ denotes the Jacobian matrix of the transformation T_k with respect to θ^k_r and γ^k_r for each $k=1,\dots,L$.

Denote by $\widetilde{T}_k$ the reverse mapping obtained from T_k by reversing the signs in the Hamiltonian proposal dynamics. Then

$$\begin{aligned} q(\theta^0_r,\gamma^0_r;\theta^*_{r+1},\gamma^*_{r+1}) &= \left|\det \nabla \widetilde{T}(\widetilde{\theta}^L_r,\widetilde{\gamma}^L_r;\widetilde{\theta}^{L-1}_r,\widetilde{\gamma}^{L-1}_r)\right|^{-1} \times \left|\det \nabla \widetilde{T}(\widetilde{\theta}^{L-1}_r,\widetilde{\gamma}^{L-1}_r;\widetilde{\theta}^{L-2}_r,\widetilde{\gamma}^{L-2}_r)\right|^{-1} \times \dots \\ &\quad \dots \times \left|\det \nabla \widetilde{T}(\widetilde{\theta}^2_r,\widetilde{\gamma}^2_r;\widetilde{\theta}^1_r,\widetilde{\gamma}^1_r)\right|^{-1} \left|\det \nabla \widetilde{T}(\widetilde{\theta}^1_r,\widetilde{\gamma}^1_r;\widetilde{\theta}^0_r,\widetilde{\gamma}^0_r)\right|^{-1} \delta(\widetilde{\gamma}^0_r,\theta^*_{r+1}) \end{aligned} \tag{8.2}$$

with $(\widetilde{\theta}^0_r,\widetilde{\gamma}^0_r)=(\theta^*_{r+1},\gamma^*_{r+1})$. Conditional on $M(\overline{\theta_r,\theta^*_{r+1}})$ satisfying Assumption1, the leapfrog transformation defined by (4.5)-(4.7) satisfies

$$(\theta^k_r,\gamma^k_r)=(\widetilde{\theta}^{L-k+1}_r,\widetilde{\gamma}^{L-k+1}_r) \quad \text{for each } k=1,\dots,L \tag{8.3}$$

Then

$$(8.4)\quad \left|\det \nabla T(\theta_r^k, \gamma_r^k; \theta_r^{k-1}, \gamma_r^{k-1})\right|^{-1} = \left|\det \nabla \widetilde{T}(\widetilde{\theta}_r^{L-k+1}, \widetilde{\gamma}_r^{L-k+1}; \widetilde{\theta}_r^{L-k}, \widetilde{\gamma}_r^{L-k})\right| \text{ for each } k = 1, \ldots, L$$

and hence $q(\theta^*_{r+1}, \gamma^*_{r+1}; \theta^0_r, \gamma^0_r) = q(\theta^*_{r+1}, \gamma^*_{r+1}; \theta^0_r, \gamma^0_r)$.

The ratio in the acceptance probability (2.2) then satisfies detailed balance in the Metropolis form

$$(8.5)\qquad \frac{\pi(\theta^*_{r+1}, \gamma^*_{r+1}) q(\theta^0_r, \gamma^0_r; \theta^*_{r+1}, \gamma^*_{r+1})}{\pi(\theta_r, \gamma_r) q(\theta^*_{r+1}, \gamma^*_{r+1}; \theta^0_r, \gamma^0_r)} = \frac{\pi(\theta^*_{r+1}, \gamma^*_{r+1})}{\pi(\theta_r, \gamma_r)}$$

since all the Jacobian terms cancel out due to (8.4). By definition of the Hamiltonian equation in (4.1), the ratio in (8.5) is then equivalent to

$$\ln \pi(\theta^*_{r+1}) - \ln \pi(\theta^0_r) + \ln \phi(\gamma^*_{r+1}; 0, M(\overline{\theta_r, \theta^*_{r+1}})) - \ln \phi(\gamma^0_r; 0, M(\overline{\theta_r, \theta^*_{r+1}}))$$

9. **Appendix B: Proof of Theorem 2**

The AHMC mapping is a special case of the general class of s-stage implicit Runge Kutta methods (Leimkuhler and Reich, 2004, p. 150) defined in our notation for each MC step r by the format

$$\begin{aligned} Q_i &= \theta^k + \Delta t \sum_{j=1}^{s} a_{ij} F_j, \quad i = 1, \ldots, s \\ P_i &= \gamma^k + \Delta t \sum_{j=1}^{s} a_{ij} G_j, \quad i = 1, \ldots, s \\ \theta^{k+1} &= \theta^k + \Delta t \sum_{i=1}^{s} b_i F_i \\ \gamma^{k+1} &= \gamma^k + \Delta t \sum_{i=1}^{s} b_i G_i \end{aligned}$$

with $s \geq 1$ the number of stages, (Q_i, P_i), $i = 1, \ldots, s$, the internal stage variables, and the abbreviations

$$\begin{aligned} F_i &= \nabla_\gamma H(Q_i, P_i), \quad i = 1, \ldots, s \\ G_i &= -\nabla_\theta H(Q_i, P_i), \quad i = 1, \ldots, s \end{aligned}$$

where $a_{ij} = b_i = 1$ for all i, j. The proof of Theorem 2 then directly follows from the proof of existence of a unique solution of the s-stage implicit Runge Kutta methods, given by Theorem 7.2 of Hairer, Nørsett, and Wanner (1993, p. 206). Specifically, if

$$(9.1)\qquad \varepsilon(\delta) < \frac{1}{\ell \max_i \sum_j |a_{ij}|}$$

where ℓ is the Lipschitz constant, then there exists a unique solution to T_k defined by (4.5)-(4.7) which can be obtained by iteration resulting in the repeated use of the triangle inequality that results from the Lipschitz condition and the contraction mapping property of (9.1).

10. Appendix C: The AHMC Algorithm

```
Initialize current $\theta$

for $r = 1$ to $R$ {

    draw $\gamma_r^0 \sim q(\gamma_r^0|\theta_r)$

    initialize $\theta_r^0 = \theta_r$, $j = 0$

    (j loop) do while $\left(\left(\left\|\theta_r^{L,j} - \theta_r^{L,j-1}\right\| > \delta_1\right)\right.$ or $\left.\left(\left\|\gamma_r^{L,j} - \gamma_r^{L,j-1}\right\| > \delta_2\right)\right)$ {

        $j = j + 1$

        (k loop) for $k = 1$ to $L$ {

            $\gamma_r^{k+1/2,j} = \gamma_r^{k,j} - \frac{\varepsilon}{2}\nabla_\theta \ln \pi(\theta_r^{k,j})$

            $\theta_r^{k+1,j} = \theta_r^{k,j} + \varepsilon \left[M(\overline{\theta_r, \theta_{r+1}^*})^{-1}\gamma_r^{k+1/2,j}\right]$

            $\gamma_r^{k+1,j} = \gamma_r^{k+1/2,j} - \frac{\varepsilon}{2}\nabla_\theta \ln \pi(\theta_r^{k+1,j})$

        }

        $M(\overline{\theta_r, \theta_{r+1}^*}) = \frac{1}{2}\left[F(\theta_r) + F(\theta_r^{L,j})\right]$

    }

    $\alpha^* = \frac{\pi(\theta_{r+1}^*)q(\widetilde{\gamma}_r^0|\theta_{r+1}^*)}{\pi(\theta_r)q(\gamma_r^0|\theta_r)}$

    draw $u \sim U[0,1]$

    if $(\alpha^* < u)$ then $\{\theta_{r+1} = \theta_r^{L,j}\}$ else $\{\theta_{r+1} = \theta_r\}$

}
```

11. Appendix D: Fisher Information for the Multivariate Normal Density

For the univariate case,

$$F(\theta) = N \begin{bmatrix} \sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{bmatrix}$$

and for the multivariate case

$$F(\theta) = N \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & \frac{1}{2}D_m' \left(\Sigma^{-1} \otimes \Sigma^{-1}\right) D_m \end{bmatrix}$$

where D_m is the duplication matrix (Magnus and Neudecker, 2007). In our empirical application we used the numerical approximation to the diagonal of $F(\theta)$ instead of the full matrix for faster speed of the MC runs.

12. Appendix E: The Generalized Stormer-Verlet Integrator

The RMHMC (Girolami and Calderhead, 2011) utilizes the following Stormer-Verlet numerical integrator (Hairer, Lubich, and Wanner, 2003; Leimkuhler and Reich, 2004) in conjunction with the Fisher information matrix for $M(\theta)$:

$$\begin{aligned}
\gamma_{ri}^{k+1/2} &= \gamma_{ri}^{k} - \frac{\varepsilon}{2}\frac{\partial H(\theta_r^k,\gamma_r^{k+1/2})}{\partial\theta_i} \\
&= \gamma_{ri}^{k} + \frac{\varepsilon}{2}\frac{\partial \ln\pi(\theta_r^k)}{\partial\theta_i} - \frac{\varepsilon}{2}\mathrm{Tr}\left[M(\theta_r^k)^{-1}\frac{\partial M(\theta_r^k)}{\partial\theta}\right]_i \\
&\quad -\frac{\varepsilon}{2}\gamma_r^{k+1/2\prime}M(\theta_r^k)^{-1}\frac{\partial M(\theta_r^k)}{\partial\theta}M(\theta_r^k)^{-1}\gamma_r^{k+1/2} \qquad (12.1)
\end{aligned}$$

$$\begin{aligned}
\theta_{ri}^{k+1} &= \theta_{ri}^{k} + \frac{\varepsilon}{2}\left[\nabla_\gamma H(\theta_r^k,\gamma_r^{k+1/2}) + \nabla_\gamma H(\theta_r^{k+1},\gamma_r^{k+1/2})\right]_i \\
&= \theta_{ri}^{k} + \frac{\varepsilon}{2}\left[M(\theta_r^k)^{-1}\gamma_r^{k+1/2} + M(\theta_r^{k+1})^{-1}\gamma_r^{k+1/2}\right]_i \qquad (12.2)
\end{aligned}$$

$$\begin{aligned}
\gamma_{ri}^{k+1} &= \gamma_{ri}^{k} - \frac{\varepsilon}{2}\frac{\partial H(\theta_r^{k+1},\gamma_r^{k+1/2})}{\partial\theta_i} \\
&= \gamma_{ri}^{k+1/2} + \frac{\varepsilon}{2}\frac{\partial \ln\pi(\theta_r^{k+1})}{\partial\theta_i} - \frac{\varepsilon}{2}\mathrm{Tr}\left[M(\theta_r^{k+1})^{-1}\frac{\partial M(\theta_r^{k+1})}{\partial\theta}\right]_i \\
&\quad -\frac{\varepsilon}{2}\gamma_r^{k+1/2\prime}M(\theta_r^{k+1})^{-1}\frac{\partial M(\theta_r^{k+1})}{\partial\theta}M(\theta_r^{k+1})^{-1}\gamma_r^{k+1/2} \qquad (12.3)
\end{aligned}$$

for $i = 1,\ldots,d$. At every leapfrog step $k = 1,\ldots L$ in the multi-step proposal sequence $\{\theta_r^k\}_{k=1}^L$ for each dimension $i = 1,\ldots,d$ of θ the value of $\gamma_{ri}^{k+1/2}$ is determined numerically as an implicit fixed point of (12.1) and the value of θ_{ri}^{k+1} as an implicit fixed point of (12.2).

References

AKHMATSKAYA, E., N. BOU-RABEE, AND S. REICH (2009): "A comparison of generalized hybrid monte carlo methods with and without momentum flip," *Journal of Computational Physics*, 228(6), 2256–2265.

BAUWENS, L., C. S. BOS, H. K. VAN DIJK, AND R. D. VAN OEST (2004): "Adaptive radial-based direction sampling: some flexible and robust Monte Carlo integration methods," *Journal of Econometrics*, 123, 201–225.

BESKOS, A., N. S. PILLAI, G. O. ROBERTS, J. M. SANZ-SERNA, AND A. M. STUART (2010): "Optimal tuning of the Hybrid Monte-Carlo Algorithm," Working paper, arxiv:1001.4460v1 [math.pr].

CHIB, S., AND E. GREENBERG (1995): "Understanding the Metropolis-Hastings Algorithm," *American Statistician*, 49(4), 327–335.

CHIB, S., AND S. RAMAMURTHY (2010): "Tailored randomized block MCMC methods with application to DSGE models," *Journal of Econometrics*, 155(1), 19 – 38.

DELLAPORTAS, P., AND I. D. VRONTOS (2007): "Modelling volatility asymmetries: a Bayesian analysis of a class of tree structured multivariate GARCH models," *Econometrics Journal*, 10(3), 503520.

DING, Z., AND R. ENGLE (2001): "Large Scale Conditional Covariance Matrix Modeling, Estimation and Testing," *Academia Economic Papers*, 29, 157184.

DUANE, S., A. KENNEDY, B. PENDLETON, AND D. ROWETH (1987): "Hybrid Monte Carlo," *Physics Letters B*, 195(2), 216–222.

ENGLE, R. F. (2002): "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business and Economic Statistics*, 20, 339–350.

ENGLE, R. F., AND K. F. KRONER (1995): "Multivariate Simultaneous Generalized ARCH," *Econometric Theory*, 11(1), 122–150.

ENGLE, R. F., N. SHEPHARD, AND K. SHEPPARD (2009): "Fitting Vast Dimensional Time-Varying Covariance Models," Available at SSRN: http://ssrn.com/abstract=1354497.

GEWEKE, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley, Hoboken, New Jersey.

GEYER, C. J. (1992): "Practical Markov Chain Monte Carlo," *Statistal Science*, 7, 473483.

GIROLAMI, M., AND B. CALDERHEAD (2011): "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods (with Discussion)," *J. R. Stat. Soc. B*, 73(2), 123–214.

GUPTA, R., G. KILCUP, AND S. SHARPE (1988): "Tuning the hybrid monte carlo algorithm," *Physical Review D*, 38(4), 1278–1287.

HAFNER, C. M., AND H. HERWARTZ (2008): "Analytical quasi Maximum Likelihood Inference in Multivariate Volatility Models," *Metrika*, 67, 219–239.

HAIRER, E., C. LUBICH, AND G. WANNER (2003): "Geometric numerical integration illustrated by the Störmer–Verlet method," *Acta Numerica*, pp. 399–450.

HAIRER, E., S. NØRSETT, AND G. WANNER (1993): *Solving Ordinary Differential Equations I*. Springer-Verlag, 2 edn.

HOLMES, C. C., AND L. HELD (2006): "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1(1), 145–168.

HUDSON, B., AND R. GERLACH (2008): "A Bayesian approach to relaxing parameter restrictions in multivariate GARCH models," *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 17(3), 606–627.

ISHWARAN, H. (1999): "Applications of Hybrid Monte Carlo to Generalized Linear Models: Quasi-complete Separation and Neural Networks," *Journal of Computational and Graphical Statistics*, 8, 779–799.

LEIMKUHLER, B., AND S. REICH (2004): *Simulating Hamiltonian Dynamics*. Cambridge University Press.

LIESENFELD, R., AND J.-F. RICHARD (2006): "Classical and Bayesian Analysis of Univariate and Multivariate Stochastic Volatility Models," *Econometric Reviews*, 25(2-3), 335–360.

LIU, J. S. (2004): *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics.

MAGNUS, J., AND H. NEUDECKER (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons.

NEAL, R. M. (1993): "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical report crg-tr-93-1, Dept. of Computer Science, University of Toronto.

——— (2010): "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, ed. by S. Brooks, A. Gelman, G. Jones, , and X.-L. Meng. Chapman & Hall / CRC Press.

OSIEWALSKI, J., AND M. PIPIEN (2004): "Bayesian comparison of bivariate ARCH-type models for the main exchange rates in Poland," *Journal of Econometrics*, 123(2), 371 – 391.

PITT, M. K., AND N. SHEPHARD (1997): "Likelihood Analysis of Non-Gaussian Measurement Time Series," *Biometrika*, 84, 653–667.

ROBERT, C. P., AND G. CASELLA (2004): *Monte Carlo statistical methods*. Springer, New York, second edn.

ROBERTS, G., AND O. STRAMER (2003): "Langevin diffusions and Metropolis-Hastings algorithms," *Methodology and Computing in Applied Probability*, 4, 337–358.

ROBERTS, G. O., AND J. S. ROSENTHAL (1998): "Optimal Scaling of Discrete Approximations to Langevin Diffusions," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60(1), 255–268.

TUCKERMAN, M., B. BERNE, G. MARTYNA, AND M. KLEIN (1993): "Efficient molecular dynamics and hybrid monte carlo algorithms for path integrals," *The Journal of Chemical Physics*, 99(4), 2796–2808.

	Mean	Stdev	Skewness	Ex Kurtosis	Sample Correlation				
AUD/USD	-0.0074	0.7019	0.4444	1.8868	1	0.4814	0.4776	0.5454	0.3614
GBP/USD	-0.0074	0.5281	0.0517	0.6827		1	0.3233	0.7123	0.3787
CAD/USD	-0.0144	0.4657	0.0039	0.7231			1	0.3874	0.2698
EUR/USD	-0.0115	0.6268	0.0640	0.6330				1	0.3995
JPY/USD	0.0087	0.6031	-0.2978	1.5496					1

Table 3: Summary Statistics

χ^2 *quantile*	0.75	0.90	0.99
BEKK	-99549.8	-99549.6	-99549.6
Diagonal BEKK (full C)	-99584.9	-99584.7	-99584.6
Diagonal BEKK (diagonal C)	-100209.9	-100209.7	-100209.6

Table 4: Log-marginal likelihoods

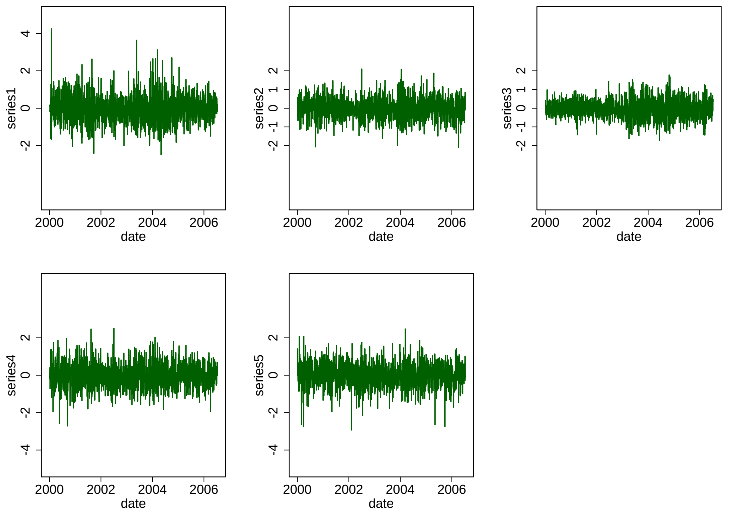


Figure 3: Time-series of log-differences in foreign exchange rates

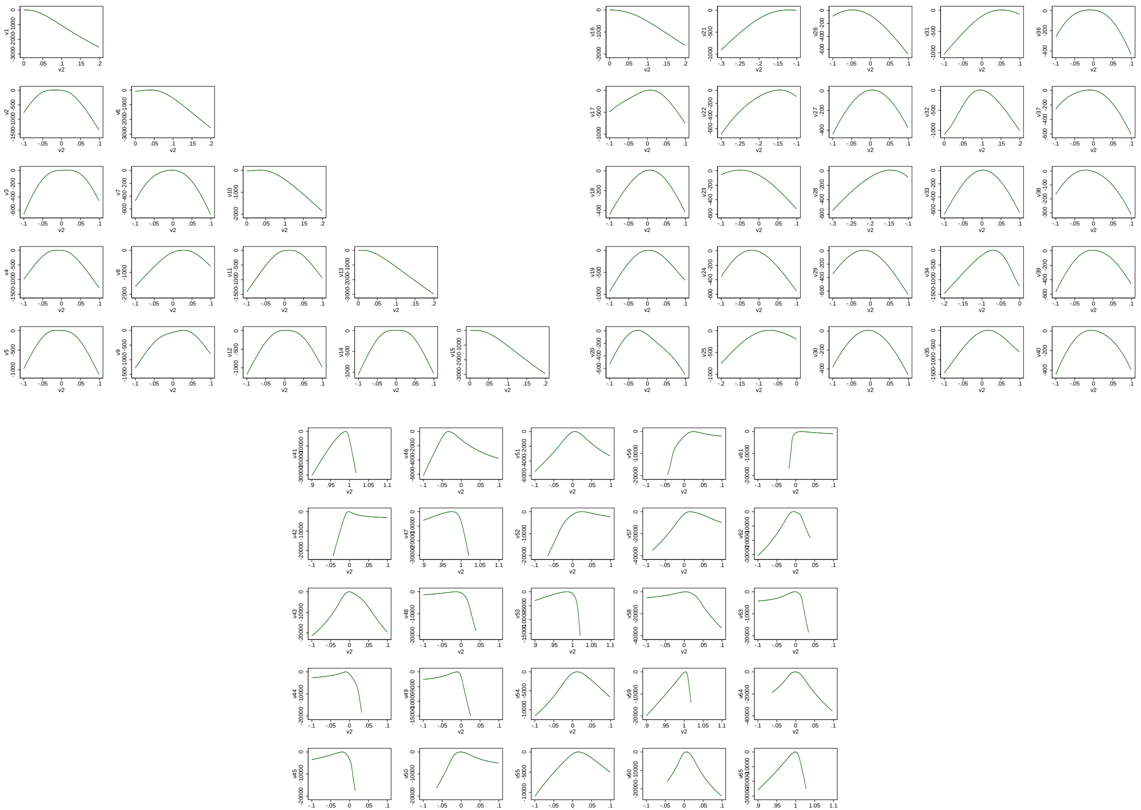


Figure 4: Conditional Log-Posterior Kernels for parameter matrices C, F and G from the BEKK model. Each parameter is plotted conditional on the other parameters being fixed at a point of high mass in the posterior density.

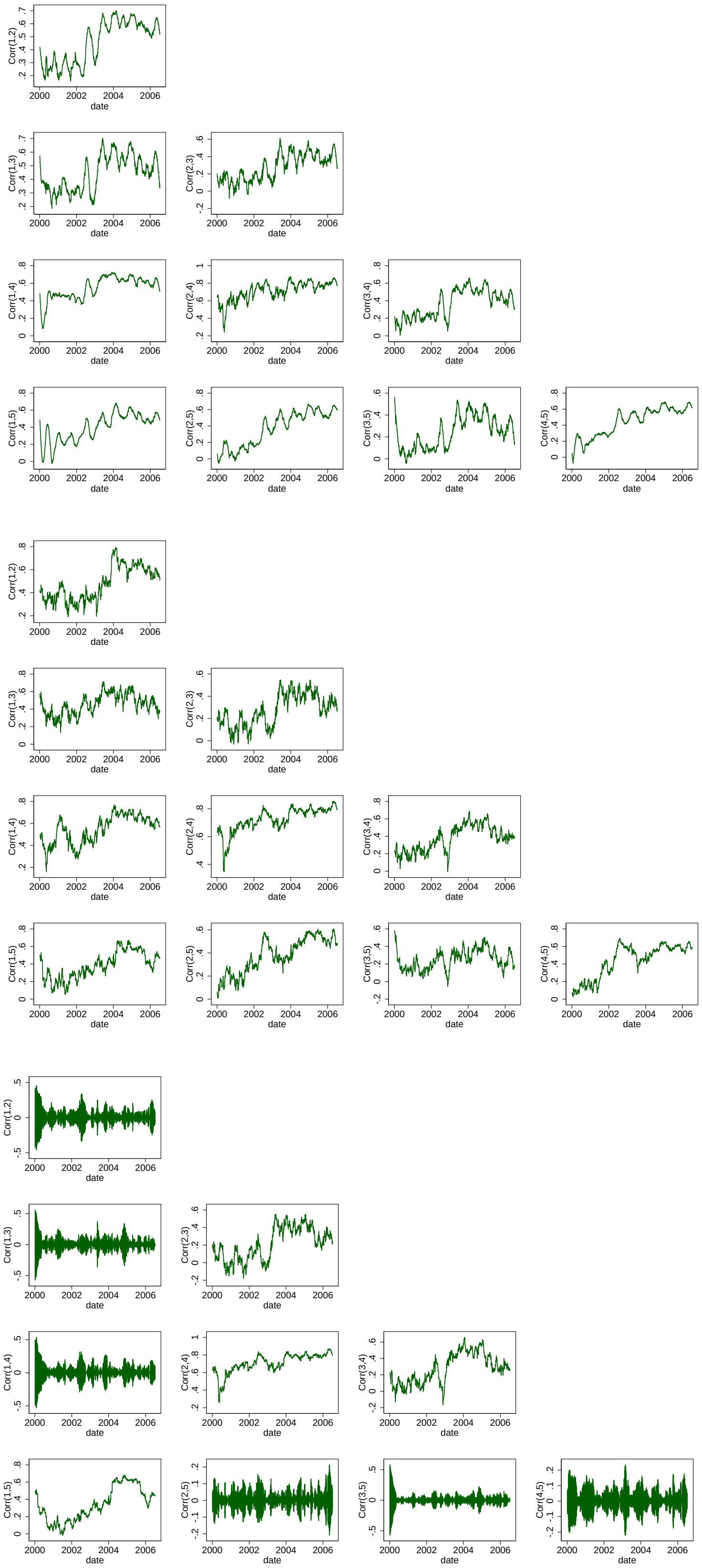


Figure 5: Conditional Correlations: BEKK, Diagonal F, G BEKK, Diagonal C, F, G BEKK