Bayesian Autoregressive Distributed Lag *via* Stochastic Gradient Hamiltonian Monte Carlo

Al-Ahmadgaid B. Asaad

alasaadstat@gmail.com

School of Statistics
University of the Philippines Diliman

April 2017

Background and Motivation

Simple Linear Regression Model

Consider the following model:

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

- In classical statistics, w_0 and w_1 are assumed to be **fixed**.
- In Bayesian statistics, w_0 and w_1 are assumed to be **random variable and follows some distribution**.

Background and Motivation

Simple Linear Regression Model

Consider the following model:

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

- In classical statistics, w_0 and w_1 are assumed to be **fixed**.
- In Bayesian statistics, w_0 and w_1 are assumed to be **random variable and follows some distribution**.

Background and Motivation

Simple Linear Regression Model

Consider the following model:

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

- In classical statistics, w_0 and w_1 are assumed to be **fixed**.
- In Bayesian statistics, w_0 and w_1 are assumed to be **random variable and follows some distribution**.

Background and Motivation

Simple Linear Regression Model

Consider the following model:

$$y_i = w_0 + w_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

- In classical statistics, w_0 and w_1 are estimated using **Maximum Likelihood Estimation**.
- In Bayesian statistics, w_0 and w_1 are estimated using **Bayes' theorem**.

Background and Motivation

Bayes' Theorem

Let $\mathbf{y}$ and $\mathbf{w}$ be the data and the weights, respectively, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\int \mathbb{P}(\mathbf{y}|\mathbf{w})\,\mathrm{d}\,\mathbf{w}} \tag{3}$$

- $\mathbb{P}(\mathbf{w}|\mathbf{y})$ is the **a posteriori**;
- $\mathbb{P}(\mathbf{w})$ is the **a priori**;
- $\mathbb{P}(\mathbf{y}|\mathbf{w})$ is the **likelihood**;
- $\mathbb{P}(\mathbf{y})$ is the **marginal likelihood** or **model evidence**;

Background and Motivation

Bayes' Theorem

Let $\mathbf{y}$ and $\mathbf{w}$ be the data and the weights, respectively, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\int \mathbb{P}(\mathbf{y}|\mathbf{w})\,\mathrm{d}\,\mathbf{w}} \tag{3}$$

- $\mathbb{P}(\mathbf{w}|\mathbf{y})$ is the **a posteriori**;
- $\mathbb{P}(\mathbf{w})$ is the **a priori**;
- $\mathbb{P}(\mathbf{y}|\mathbf{w})$ is the **likelihood**;
- $\mathbb{P}(\mathbf{y})$ is the **marginal likelihood** or **model evidence**;

Background and Motivation

Bayes' Theorem

Let $\mathbf{y}$ and $\mathbf{w}$ be the data and the weights, respectively, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\int \mathbb{P}(\mathbf{y}|\mathbf{w})\, \mathrm{d}\, \mathbf{w}} \tag{3}$$

- $\mathbb{P}(\mathbf{w}|\mathbf{y})$ is the **a posteriori**;
- $\mathbb{P}(\mathbf{w})$ is the **a priori**;
- $\mathbb{P}(\mathbf{y}|\mathbf{w})$ is the **likelihood**;
- $\mathbb{P}(\mathbf{y})$ is the **marginal likelihood** or **model evidence**;

Background and Motivation

Bayes' Theorem

Let $\mathbf{y}$ and $\mathbf{w}$ be the data and the weights, respectively, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\int \mathbb{P}(\mathbf{y}|\mathbf{w})\,\mathrm{d}\,\mathbf{w}} \tag{3}$$

- $\mathbb{P}(\mathbf{w}|\mathbf{y})$ is the **a posteriori**;
- $\mathbb{P}(\mathbf{w})$ is the **a priori**;
- $\mathbb{P}(\mathbf{y}|\mathbf{w})$ is the **likelihood**;
- $\mathbb{P}(\mathbf{y})$ is the **marginal likelihood** or **model evidence**;

Background and Motivation

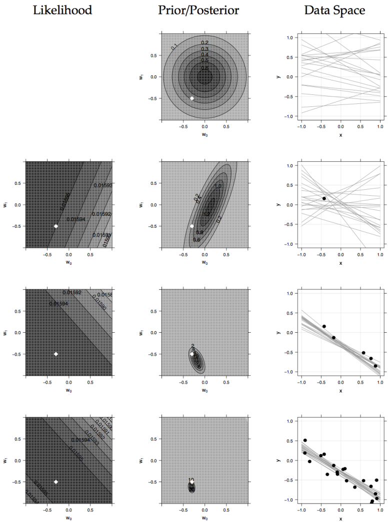
Bayes' Theorem

Let $\mathbf{y}$ and $\mathbf{w}$ be the data and the weights, respectively, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\mathbb{P}(\mathbf{y})} = \frac{\mathbb{P}(\mathbf{y}|\mathbf{w})\mathbb{P}(\mathbf{w})}{\int \mathbb{P}(\mathbf{y}|\mathbf{w})\,\mathrm{d}\,\mathbf{w}} \tag{3}$$

- $\mathbb{P}(\mathbf{w}|\mathbf{y})$ is the **a posteriori**;
- $\mathbb{P}(\mathbf{w})$ is the **a priori**;
- $\mathbb{P}(\mathbf{y}|\mathbf{w})$ is the **likelihood**;
- $\mathbb{P}(\mathbf{y})$ is the **marginal likelihood** or **model evidence**;

Example of Bayesian Modeling



Metropolis-Hasting

Algorithm 3 *Metropolis-Hasting MCMC*

1: Initialize $\mathbf{w}_r \sim \mathbb{G}(\mathbf{w}), r = 0$
2: **for** $r \in \{1, \cdots, r_{\max}\}$ **do**
3: Propose: $\mathbf{w}_{new} \sim \mathbb{G}(\mathbf{w}_{new}|\mathbf{w}_{r-1})$
4: Acceptance: $\alpha(\mathbf{w}_{new}|\mathbf{w}_{r-1}) \triangleq \min\left\{1, \frac{\mathbb{P}(\mathbf{w}_{new}|\mathbf{w}_{r-1})\mathbb{G}(\mathbf{w}_{r-1}|\mathbf{w}_{new})}{\mathbb{P}(\mathbf{w}_{r-1}|\mathbf{w}_{new})\mathbb{G}(\mathbf{w}_{new}|\mathbf{w}_{r-1})}\right\}$
5: Draw $x \sim \text{Unif}(0, 1)$
6: **if** $x < \alpha(\mathbf{w}_{new}|\mathbf{w}_{r-1})$ **then**
7: $\mathbf{w}_r \triangleq \mathbf{w}_{new}$
8: **else**
9: $\mathbf{w}_r \triangleq \mathbf{w}_{r-1}$
10: **end if**
11: **end for**

Metropolis-Hasting

Limitations

- Specification of the proposal distribution, $\mathbb{G}$, is often difficult for high dimensional parameters.
- Autocorrelations of the markov chains is often high in magnitude, hence violates the assumption of IID samples.

Hamiltonian Dynamics for MCMC

Algorithm 5 *Hamiltonian MCMC*

Initialize Leap Frog parameters: γ and τ;
Set initial location $\mathbf{w}^{(r=0)}(t=0)$;
for $r \in \{0, \cdots, r_{\max}\}$ **do**
  Draw initial momentum, $\mathbf{p}^{(r)}(t=0) \sim \frac{\exp[-\mathbb{K}(\mathbf{p})]}{Z}$;
  Compute $\mathbb{H}(\mathbf{w}^{(r)}(0), \mathbf{p}^{(r)}(0)) \triangleq \mathbb{U}(\mathbf{w}^{(r)}(0)) + \mathbb{K}(\mathbf{p}^{(r)}(0))$;
  Simulate Hamiltonian dynamics using Leap Frog:
  for $t \in \{0, \cdots, \tau\}$ **do**

$$\mathbf{p}^{(r)}(t+\gamma/2) \triangleq \mathbf{p}^{(r)}(t) - (\gamma/2)\frac{\partial \mathbb{U}(\mathbf{w}^{(r)}(t))}{\partial \mathbf{w}^{(r)}(t)} \tag{4.7}$$

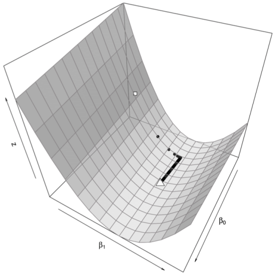
$$\mathbf{w}^{(r)}(t+\gamma) \triangleq \mathbf{w}^{(r)}(t) + \gamma\frac{\partial \mathbb{K}(\mathbf{p}^{(r)}(t+1))}{\partial \mathbf{p}^{(r)}(t+1)}, \tag{4.8}$$

$$\mathbf{p}^{(r)}(t+\gamma) \triangleq \mathbf{p}^{(r)}(t+\gamma/2) - (\gamma/2)\frac{\partial \mathbb{U}(\mathbf{w}^{(r)}(t+\gamma))}{\partial \mathbf{w}^{(r)}(t+\gamma)} \tag{4.9}$$

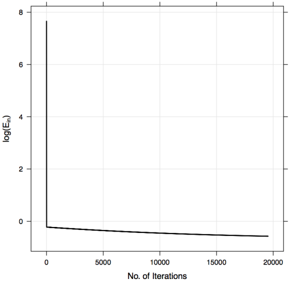
  end for
  if $\Delta\mathbb{H} < 0$ **then**
    $\mathbf{w}^{(r+1)}(0) \triangleq \mathbf{w}^{(r)}(\tau+\gamma)$
  else
    if $a < \exp(\Delta\mathbb{H}), a \sim \text{Unif}(0,1)$ **then**
      $\mathbf{w}^{(r+1)}(0) \triangleq \mathbf{w}^{(r)}(\tau+\gamma)$
    else
      $\mathbf{w}^{(r+1)}(0) \triangleq \mathbf{w}^{(r)}(0)$
    end if
  end if
end for

Batch Gradient Descent

Consider the following error surface function of the simple linear regression model:



(a) BGD on Loss Function Surface.

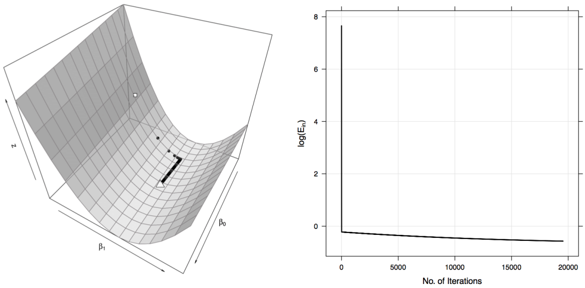


(b) Loss Function under BGD.

Figure: Batch Gradient Descent on SLR Loss Function.

Gradient Descent

Batch gradient descent can be very computationally expensive especially for large dataset.



(a) BGD on Loss Function Surface.

(b) Loss Function under BGD.

Figure: Batch Gradient Descent on SLR Loss Function.

Stochastic Gradient Descent

- One might suggest that instead of using **all observations**, would it be feasible to just use **one** or **sample of observations**?
- The answer is **Yes**! and that is the idea behind **Stochastic Gradient Descent** (SGD).
- SGD updates the parameter using only one observation for every iteration, which is a lot faster.
- Sometimes it is called **Minibatch Gradient Descent** (MGD) if it uses **samples of observations**.
- MGD can take advantage of vectorization in computation and hence even faster than SGD;

Stochastic Gradient Descent

- One might suggest that instead of using **all observations**, would it be feasible to just use **one** or **sample of observations**?
- The answer is **Yes**! and that is the idea behind **Stochastic Gradient Descent** (SGD).
- SGD updates the parameter using only one observation for every iteration, which is a lot faster.
- Sometimes it is called **Minibatch Gradient Descent** (MGD) if it uses **samples of observations**.
- MGD can take advantage of vectorization in computation and hence even faster than SGD;

Stochastic Gradient Descent

- One might suggest that instead of using **all observations**, would it be feasible to just use **one** or **sample of observations**?
- The answer is **Yes**! and that is the idea behind **Stochastic Gradient Descent** (SGD).
- SGD updates the parameter using only one observation for every iteration, which is a lot faster.
- Sometimes it is called **Minibatch Gradient Descent** (MGD) if it uses **samples of observations**.
- MGD can take advantage of vectorization in computation and hence even faster than SGD;

Stochastic Gradient Descent

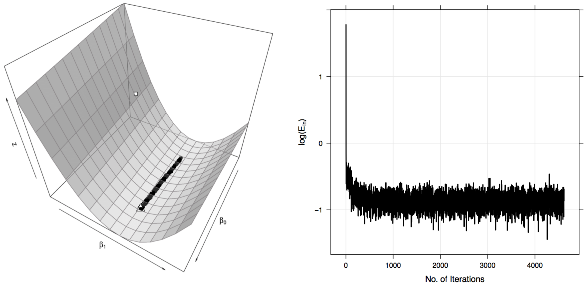
- One might suggest that instead of using **all observations**, would it be feasible to just use **one** or **sample of observations**?
- The answer is **Yes**! and that is the idea behind **Stochastic Gradient Descent** (SGD).
- SGD updates the parameter using only one observation for every iteration, which is a lot faster.
- Sometimes it is called **Minibatch Gradient Descent** (MGD) if it uses **samples of observations**.
- MGD can take advantage of vectorization in computation and hence even faster than SGD;

Stochastic Gradient Descent

- One might suggest that instead of using **all observations**, would it be feasible to just use **one** or **sample of observations**?
- The answer is **Yes**! and that is the idea behind **Stochastic Gradient Descent** (SGD).
- SGD updates the parameter using only one observation for every iteration, which is a lot faster.
- Sometimes it is called **Minibatch Gradient Descent** (MGD) if it uses **samples of observations**.
- MGD can take advantage of vectorization in computation and hence even faster than SGD;

Stochastic Gradient Descent

Consider again the following error surface function of the simple linear regression model:

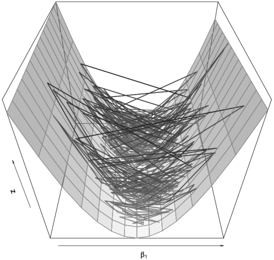


(a) SGD on Loss Function Surface. (b) Loss Function under SGD.

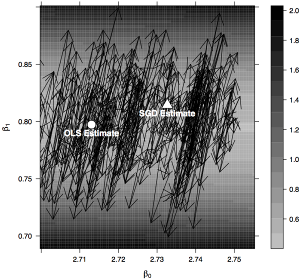
Figure: Stochastic Batch Gradient Descent on SLR Loss Function.

Stochastic Gradient Descent

Zoom into the SGD gradient vectors.



(a) SGD on Loss Function Surface.



(b) SGD on Loss Function Contour.

Figure: A Closer Look at SGD Gradient Vectors.

Stochastic Gradient Hamiltonian Monte Carlo

- To take advantage of the features of SGD, Bayesians decided to apply this algorithm to MCMC methods involving gradient computation.
- Such MCMCs are the **HMC** and **LMC**.
- For HMC in particular, Chen, Fox, and Guestrin 2014 were the pioneers for marrying the two methods.
- Their work has been inspired by Ahn, Korattikara, and Welling 2012; and Welling and Teh 2011.

Stochastic Gradient Hamiltonian Monte Carlo

- To take advantage of the features of SGD, Bayesians decided to apply this algorithm to MCMC methods involving gradient computation.
- Such MCMCs are the **HMC** and **LMC**.
- For HMC in particular, Chen, Fox, and Guestrin 2014 were the pioneers for marrying the two methods.
- Their work has been inspired by Ahn, Korattikara, and Welling 2012; and Welling and Teh 2011.

Stochastic Gradient Hamiltonian Monte Carlo

- To take advantage of the features of SGD, Bayesians decided to apply this algorithm to MCMC methods involving gradient computation.
- Such MCMCs are the **HMC** and **LMC**.
- For HMC in particular, Chen, Fox, and Guestrin 2014 were the pioneers for marrying the two methods.
- Their work has been inspired by Ahn, Korattikara, and Welling 2012; and Welling and Teh 2011.

Stochastic Gradient Hamiltonian Monte Carlo

- To take advantage of the features of SGD, Bayesians decided to apply this algorithm to MCMC methods involving gradient computation.
- Such MCMCs are the **HMC** and **LMC**.
- For HMC in particular, Chen, Fox, and Guestrin 2014 were the pioneers for marrying the two methods.
- Their work has been inspired by Ahn, Korattikara, and Welling 2012; and Welling and Teh 2011.

Langevin Dynamics

- The Stochastic Gradient HMC works by considering Langevin dynamics on its momentum.
- The said dynamics extend the idea of the Newton's second law of motion.
- Originally, the second law proceeds as follows: let $\mathbf{f}$ be the force, $\mathbf{p}$ be the momentum, m be the mass, $\mathbf{v}$ be the veclocity, and $\mathbf{a}$ be the acceleration, then

$$\mathbf{f} = \frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} = \frac{\mathrm{d}(m\mathbf{v})}{\mathrm{d}\,t} = m\frac{\mathrm{d}\,\mathbf{v}}{\mathrm{d}\,t} = m\mathbf{a}. \tag{4}$$

Langevin Dynamics

- The Stochastic Gradient HMC works by considering Langevin dynamics on its momentum.
- The said dynamics extend the idea of the Newton's second law of motion.
- Originally, the second law proceeds as follows: let $\mathbf{f}$ be the force, $\mathbf{p}$ be the momentum, m be the mass, $\mathbf{v}$ be the veclocity, and $\mathbf{a}$ be the acceleration, then

$$\mathbf{f} = \frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} = \frac{\mathrm{d}(m\mathbf{v})}{\mathrm{d}\,t} = m\frac{\mathrm{d}\,\mathbf{v}}{\mathrm{d}\,t} = m\mathbf{a}. \tag{4}$$

Langevin Dynamics

- The Stochastic Gradient HMC works by considering Langevin dynamics on its momentum.
- The said dynamics extend the idea of the Newton's second law of motion.
- Originally, the second law proceeds as follows: let $\mathbf{f}$ be the force, $\mathbf{p}$ be the momentum, m be the mass, $\mathbf{v}$ be the veclocity, and $\mathbf{a}$ be the acceleration, then

$$\mathbf{f} = \frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} = \frac{\mathrm{d}(m\mathbf{v})}{\mathrm{d}\,t} = m\frac{\mathrm{d}\,\mathbf{v}}{\mathrm{d}\,t} = m\mathbf{a}. \tag{4}$$

Langevin Dynamics

- The Stochastic Gradient HMC works by considering Langevin dynamics on its momentum.
- The said dynamics extend the idea of the Newton's second law of motion.
- Originally, the second law proceeds as follows: let $\mathbf{f}$ be the force, $\mathbf{p}$ be the momentum, m be the mass, $\mathbf{v}$ be the veclocity, and $\mathbf{a}$ be the acceleration, then

$$\mathbf{f} = \frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} = \frac{\mathrm{d}(m\mathbf{v})}{\mathrm{d}\,t} = m\frac{\mathrm{d}\,\mathbf{v}}{\mathrm{d}\,t} = m\mathbf{a}. \tag{4}$$

Langevin Dynamics

- The idea of Langevin dynamics is to take into account or at least approximate the effect of neglected degrees of freedom;
- and this is achieved by adding two force terms: one represents the frictional force, $\eta \mathbf{v}^{\bigstar}$; and the other represents the random force, $\mathbf{e}$.
- So that the Langevin equation is given below:

$$\frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} - \eta \mathbf{v}^{\bigstar} + \mathbf{e} = m\mathbf{a}, \tag{5}$$

where the random force is assumed to have zero mean and is uncorrelated, i.e. $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \xi \mathbf{I})$.

Langevin Dynamics

- The idea of Langevin dynamics is to take into account or at least approximate the effect of neglected degrees of freedom;
- and this is achieved by adding two force terms: one represents the frictional force, $\eta \mathbf{v}^{\bigstar}$; and the other represents the random force, $\mathbf{e}$.
- So that the Langevin equation is given below:

$$\frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} - \eta \mathbf{v}^{\bigstar} + \mathbf{e} = m\mathbf{a}, \tag{5}$$

where the random force is assumed to have zero mean and is uncorrelated, i.e. $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \xi \mathbf{I})$.

Langevin Dynamics

- The idea of Langevin dynamics is to take into account or at least approximate the effect of neglected degrees of freedom;
- and this is achieved by adding two force terms: one represents the frictional force, $\eta \mathbf{v}^{\bigstar}$; and the other represents the random force, $\mathbf{e}$.
- So that the Langevin equation is given below:

$$\frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} - \eta \mathbf{v}^{\bigstar} + \mathbf{e} = m\mathbf{a}, \tag{5}$$

where the random force is assumed to have zero mean and is uncorrelated, i.e. $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \xi \mathbf{I})$.

Langevin Dynamics

- The idea of Langevin dynamics is to take into account or at least approximate the effect of neglected degrees of freedom;
- and this is achieved by adding two force terms: one represents the frictional force, $\eta \mathbf{v}^{\blacklozenge}$; and the other represents the random force, $\mathbf{e}$.
- So that the Langevin equation is given below:

$$\frac{\mathrm{d}\,\mathbf{p}}{\mathrm{d}\,t} - \eta \mathbf{v}^{\blacklozenge} + \mathbf{e} = m\mathbf{a}, \tag{5}$$

where the random force is assumed to have zero mean and is uncorrelated, i.e. $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \xi \mathbf{I})$.

Stochastic Gradient Hamiltonian Monte Carlo

Algorithm 6 *Stochastic Hamiltonian MCMC*

1: Initialize Leapfrog parameters: γ and τ;

2: Initialize estimate for $\hat{\mathfrak{B}}(\mathbf{w}) = \frac{\gamma}{2}\hat{\mathfrak{A}}(\mathbf{w})$, and specify the matrix $\mathfrak{C}$;

3: Set initial location $\mathbf{w}^{(r=0)}(t=0)$;

4: **for** $r \in \{0, \cdots, r_{\max}\}$ **do**

5: Draw initial momentum, $\mathbf{p}^{(r)}(t=0) \sim \frac{\exp[-\mathbb{K}(\mathbf{p})]}{Z}$;

6: Simulate Hamiltonian dynamics using Leapfrog:

7: **for** $t \in \{0, \cdots, \tau\}$ **do**

$$\Delta\mathbf{w}^{(r)}(t+\gamma) \triangleq \gamma\nabla_{\mathbf{p}^{(r)}(t+1)}\mathbb{K}(\mathbf{p}^{(r)}(t+1)), \tag{4.31}$$

$$\Delta\mathbf{p}^{(r)}(t+\gamma) \triangleq -\gamma\nabla_{\mathbf{w}^{(r)}(t)}\tilde{\mathbb{U}}(\mathbf{w}^{(r)}(t)) - \mathfrak{C}(\mathbf{w}^{(r)}(t))\boldsymbol{\Sigma}^{-1}\mathbf{p} + \xi^{\star} \tag{4.32}$$

$$\text{where } \xi^{\star} \sim \mathcal{N}(\mathbf{0}, 2\gamma(\mathfrak{C} - \mathfrak{B})) \tag{4.33}$$

8: **end for**

9: $\mathbf{w}^{(r+1)} \triangleq \mathbf{w}^{(r)}$

10: **end for**

Autoregressive Distributed Lag (ADL) Model

- In this thesis, the objective model is the Autoregressive Distributed Lag (ADL) which is a specialized type of dynamic linear models (L. J. Welty et al 2009).
- And to the best knowledge of the author, none has yet integrated the SGHMC to Bayesian ADL (BADL). Thus in this paper, the proposed model is abbreviated as BADL-SGHMC.

Autoregressive Distributed Lag (ADL) Model

- In this thesis, the objective model is the Autoregressive Distributed Lag (ADL) which is a specialized type of dynamic linear models (L. J. Welty et al 2009).
- And to the best knowledge of the author, none has yet integrated the SGHMC to Bayesian ADL (BADL). Thus in this paper, the proposed model is abbreviated as BADL-SGHMC.

Autoregressive Distributed Lag (ADL) Model

The simplest ADL model is of order $p = 1$ and $q = 0$, denoted by ADL$(1, 0)$:

$$y(t) = w_0 + w_1 y(t-1) + w_2 x(t) + \varepsilon(t), \quad \varepsilon \sim \mathcal{N}(0, \sigma). \tag{6}$$

With other m explanatory variables:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{m} w_{2+i} x_i(t) + \varepsilon(t), \quad \varepsilon \sim \mathcal{N}(0, \sigma). \tag{7}$$

Autoregressive Distributed Lag (ADL) Model

The simplest ADL model is of order $p = 1$ and $q = 0$, denoted by ADL$(1, 0)$:

$$y(t) = w_0 + w_1 y(t-1) + w_2 x(t) + \varepsilon(t), \quad \varepsilon \sim \mathcal{N}(0, \sigma). \tag{6}$$

With other m explanatory variables:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{m} w_{2+i} x_i(t) + \varepsilon(t), \quad \varepsilon \sim \mathcal{N}(0, \sigma). \tag{7}$$

Autoregressive Distributed Lag (ADL) Model

ADL$(1,1)$:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{m}\sum_{l=0}^{1} w_{(2+l\cdot m)+i}\, x_i(t-l) + \varepsilon(t). \quad (8)$$

For general ADL(p,q):

$$y(t) = w_0 + \sum_{k=1}^{p} w_k y(t-k) + \sum_{i=1}^{m}\sum_{l=0}^{q} w_{\kappa(p,l,m,i)}\, x_i(t-l) + \varepsilon(t). \quad (9)$$

where $\varepsilon \sim \mathcal{N}(0,\sigma)$.

Autoregressive Distributed Lag (ADL) Model

ADL$(1,1)$:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{m} \sum_{l=0}^{1} w_{(2+l\cdot m)+i} x_i(t-l) + \varepsilon(t). \quad (8)$$

For general ADL(p,q):

$$y(t) = w_0 + \sum_{k=1}^{p} w_k y(t-k) + \sum_{i=1}^{m} \sum_{l=0}^{q} w_{\kappa(p,l,m,i)} x_i(t-l) + \varepsilon(t). \quad (9)$$

where $\varepsilon \sim \mathcal{N}(0,\sigma)$.

Autoregressive Distributed Lag (ADL) Model

And since the error term is centered at 0, then

$$\mathbb{E}[y(t)] = w_0 + \sum_{k=1}^{p} w_k y(t-k) + \sum_{i=1}^{m} \sum_{l=0}^{q} w_{\kappa(p,l,m,i)} x_i(t-l). \quad (10)$$

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
    1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
    2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
    3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
    1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
    2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
    3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
    1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
    2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
    3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
   1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
   2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
   3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
   1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
   2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
   3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
    1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
    2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
    3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

General Objectives

1. derive the necessary theoretical results;
2. compare the performance of the proposed model, BADL-SGHMC, for three cases:
    1. leapfrog step size $\gamma = .09$ for 1,000 iterations;
    2. leapfrog step size $\gamma = .009$ for 10,000 iterations;
    3. leapfrog step size $\gamma = .0009$ for 100,000 iterations.
3. apply the proposed model to forecasting Philippine's economic growth; and
4. create software packages for SGHMC for R and Julia programming languages.

Objectives of the Study

Specific Objectives

1. The derivation of the theoretical results:
    1. given the *a priori* on the parameters of the BADL model, derive the posterior distribution; and
    2. given the *a posteriori* from the preceding objective, derive the stochastic gradient of the potential energy;

Objectives of the Study

Specific Objectives

1. The derivation of the theoretical results:
    1. given the *a priori* on the parameters of the BADL model, derive the posterior distribution; and
    2. given the *a posteriori* from the preceding objective, derive the stochastic gradient of the potential energy;

Objectives of the Study

Specific Objectives

1. The comparison on the performance of the proposed model, BADL-SGHMC, against the performance of the BADL-MH and BADL-HMC. This is done by considering four markov chains for each parameter of the BADL with dispersed random initial values from uniform distribution.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

The following are the statistical methodologies used for assessing the performance of the models:

1. Heidelberger-Welch, for stationarity test on the Markov chains;
2. Gelman-Rubin, for convergence test of averages of the Markov chains;
3. Autocorrelations, for assessing the assumption of the independent and identically distributed (IID) samples from the *a posteriori*; and
4. Root Mean Squared Error (RMSE), for assessing the in-sample and out-of-sample forecast performance of the three models.

The above procedures are performed across three cases of the leapfrog step size mentioned in the General Objectives.

Objectives of the Study

Specific Objective

1. Apply the derived theoretical results to forecasting Philippine's year-over-year economic growth rate. In particular, the comparison of the models detailed in the preceding objective is performed using this data. The following are the time series involved in modeling:
   1. The Response Variable
      - Growth Rate of Gross Domestic Product
   2. The Predictors
      - Growth Rate of Peso/US Dollar Exchange Rate;
      - Growth Rate of Stock Price Index; and
      - Growth Rate of Gross International Reserves.

Objectives of the Study

Specific Objective

1. Apply the derived theoretical results to forecasting Philippine's year-over-year economic growth rate. In particular, the comparison of the models detailed in the preceding objective is performed using this data. The following are the time series involved in modeling:
    1. The Response Variable
        - Growth Rate of Gross Domestic Product
    2. The Predictors
        - Growth Rate of Peso/US Dollar Exchange Rate;
        - Growth Rate of Stock Price Index; and
        - Growth Rate of Gross International Reserves.

Objectives of the Study

Specific Objective

1. Apply the derived theoretical results to forecasting Philippine's year-over-year economic growth rate. In particular, the comparison of the models detailed in the preceding objective is performed using this data. The following are the time series involved in modeling:
    1. The Response Variable
        - Growth Rate of Gross Domestic Product
    2. The Predictors
        - Growth Rate of Peso/US Dollar Exchange Rate;
        - Growth Rate of Stock Price Index; and
        - Growth Rate of Gross International Reserves.

Objectives of the Study

Specific Objective

1. Apply the derived theoretical results to forecasting Philippine's year-over-year economic growth rate. In particular, the comparison of the models detailed in the preceding objective is performed using this data. The following are the time series involved in modeling:
    1. The Response Variable
        - Growth Rate of Gross Domestic Product
    2. The Predictors
        - Growth Rate of Peso/US Dollar Exchange Rate;
        - Growth Rate of Stock Price Index; and
        - Growth Rate of Gross International Reserves.

Objectives of the Study

Specific Objective

1. Create software packages for SGHMC for R and Julia programming languages using Github.com as the repository. That is, the package can be installed from this website.

Autoregressive Distributed Lag (ADL) Model

BADL(1, 1)-SGHMC:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{m} \sum_{l=0}^{1} w_{(2+l\cdot m)+i} x_i(t-l) + \varepsilon(t). \quad (11)$$

- **w** is treated as random vector.
- **w** is estimated using Bayesian MCMC, specifically the SGHMC.

Theoretical Results

Proposition

Let $\mathscr{D} = \{[\mathbf{x}(t), y(t)], \forall t \in \mathbb{Z}_+^{\tau}\}$ be the data such that $y(t)$ is modeled by a Gaussian function with mean given in Equation (10) and constant variance $\alpha^{-1} \in \mathbb{R}_+$. If $\mathbf{w}$ is the vector of coefficients of ADL(p, q) such that $\mathbf{w} \sim \mathcal{N}_d(\mathbf{0}, \beta^{-1}\mathbf{I})$, where $\beta^{-1} \in \mathbb{R}_+$, then the posterior is a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma} = (\alpha \mathfrak{G}^T \mathfrak{G} + \beta \mathbf{I})^{-1}$ and mean vector $\boldsymbol{\mu} = \alpha \boldsymbol{\Sigma} \mathfrak{G}^T \mathbf{y}$, where $\mathfrak{G}$ is the design matrix.

Theoretical Results

proof: Let $\mathbf{w} \triangleq [w_0\ w_1\ \cdots\ w_{\kappa(p,q,m,m)}]^{\mathsf{T}}$, $\kappa(p,l,m,m) \triangleq [(p+1)+l \cdot m]+m$ and let $\mathbf{z}(t) \triangleq [1\ y(t-1)\ \cdots\ x_m(t-q)]^{\mathsf{T}}$, then the ADL$(p,q)$ can be written as

$$y(t) = \mathbf{w}^{\mathsf{T}}\mathbf{z}(t) + \varepsilon(t), \quad \varepsilon(t) \sim \mathcal{N}(0, \alpha^{-1}). \tag{12}$$

The likelihood is therefore given by

$$\mathcal{L}(\mathbf{w}|\mathbf{y}) \triangleq \left(\frac{\alpha}{2\pi}\right)^{\tau/2} \exp\left\{-\sum_{t=1}^{\tau} \frac{\alpha[y(t) - \mathbf{w}^{\mathsf{T}}\mathbf{z}(t)]^2}{2}\right\}. \tag{13}$$

Let $\mathbf{y} \triangleq [y(1)\ y(2)\ \cdots\ y(\tau)]^{\mathsf{T}}$ and let $\mathfrak{G} \triangleq [(\mathbf{z}(t)^{\mathsf{T}})]$, i.e. $\mathfrak{G} \in \mathbb{R}^{\tau} \times \mathbb{R}^{d}$. Thus in matrix form

$$\mathcal{L}(\mathbf{w}|\mathbf{y}) \propto \exp\left[-\frac{\alpha}{2}(\mathbf{y} - \mathfrak{G}\mathbf{w})^{\mathsf{T}}(\mathbf{y} - \mathfrak{G}\mathbf{w})\right]. \tag{14}$$

Theoretical Results

The prior is given by

$$\mathbb{P}(\mathbf{w}) = \frac{1}{\sqrt{(2\pi)^d |\beta^{-1}\mathbf{I}|}} \exp\left[-\frac{1}{2}\mathbf{w}^\mathsf{T}\beta\mathbf{I}\mathbf{w}\right]. \tag{15}$$

So that the posterior would be

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \exp\left[-\frac{\alpha}{2}(\mathbf{y}-\mathfrak{G}\mathbf{w})^\mathsf{T}(\mathbf{y}-\mathfrak{G}\mathbf{w})\right]\exp\left[-\frac{1}{2}\mathbf{w}^\mathsf{T}\beta\mathbf{I}\mathbf{w}\right] \tag{16}$$

$$= \exp\left\{-\frac{1}{2}\left[\alpha(\mathbf{y}-\mathfrak{G}\mathbf{w})^\mathsf{T}(\mathbf{y}-\mathfrak{G}\mathbf{w}) + \mathbf{w}^\mathsf{T}\beta\mathbf{I}\mathbf{w}\right]\right\}. \tag{17}$$

Expanding the terms in the exponential factor becomes

$$\alpha\mathbf{y}^\mathsf{T}\mathbf{y} - 2\alpha\mathbf{w}^\mathsf{T}\mathfrak{G}^\mathsf{T}\mathbf{y} + \mathbf{w}^\mathsf{T}(\alpha\mathfrak{G}^\mathsf{T}\mathfrak{G} + \beta\mathbf{I})\mathbf{w}. \tag{18}$$

Theoretical Results

Hence

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{C} \exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}(\alpha\mathfrak{G}^{\mathsf{T}}\mathfrak{G} + \beta\mathbf{I})\mathbf{w} - 2\alpha\mathbf{w}^{\mathsf{T}}\mathfrak{G}^{\mathsf{T}}\mathbf{y}\right]\right\}. \quad (19)$$

Notice the terms in the exponential factor is of the form $ax^2 - 2bx$. This suggest a quadratic equation and therefore can be factored by completing the square. To do so, let $\mathbf{D} \triangleq \alpha\mathfrak{G}^{\mathsf{T}}\mathfrak{G} + \beta\mathbf{I}$ and $\mathbf{b} \triangleq \alpha\mathfrak{G}^{\mathsf{T}}\mathbf{y}$, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{C} \exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}\mathbf{D}\mathbf{w} - 2\mathbf{w}^{\mathsf{T}}\mathbf{b}\right]\right\} \quad (20)$$

$$= \mathcal{C} \exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}\mathbf{D}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{b} - \mathbf{b}^{\mathsf{T}}\mathbf{w}\right]\right\}. \quad (21)$$

Theoretical Results

Next is to add a term that is not a function of $\mathbf{w}$ which can be assumed to be part of the constant $\mathcal{C}$. Let this term be $\mathbf{b}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{b}$, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{C}\exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}\mathbf{D}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{b} - \mathbf{b}^{\mathsf{T}}\mathbf{w} + \mathbf{b}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{b}\right]\right\}. \quad (22)$$

In order to proceed, the matrix $\mathbf{D}$ must be symmetric and invertible since later this will be the covariance matrix of the posterior which requires such property. If satisfied, then $\mathbf{I} = \mathbf{D}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{D}$. So that

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{C}\exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}\mathbf{D}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{D}\mathbf{D}^{-1}\mathbf{b} - \mathbf{b}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{D}\mathbf{w} + \mathbf{b}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{D}\mathbf{D}^{-1}\mathbf{b}\right]\right\}.$$

Theoretical Results

Finally, let $\mathbf{\Sigma} \triangleq \mathbf{D}^{-1}$ and $\boldsymbol{\mu} \triangleq \mathbf{D}^{-1}\mathbf{b}$, then

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) \propto \mathcal{C} \exp\left\{-\frac{1}{2}\left[\mathbf{w}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{w} - \mathbf{w}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\mathbf{w} + \boldsymbol{\mu}^{\mathsf{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}\right]\right\} \tag{23}$$

$$= \mathcal{C} \exp\left\{-\frac{1}{2}\left[(\mathbf{w} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{w} - \boldsymbol{\mu})\right]\right\}. \tag{24}$$

Thus $\mathcal{C} = \frac{\mathcal{C}_0}{\mathbb{P}(\mathbf{y})}$, where $\mathcal{C}$ is the constant of the Guassian kernel in Equation (24). Therefore,

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \mathcal{N}_d(\mathbf{w}|\boldsymbol{\mu}, \mathbf{\Sigma}), \tag{25}$$

where $\mathbf{\Sigma} = (\alpha\mathfrak{G}^{\mathsf{T}}\mathfrak{G} + \beta\mathbf{I})^{-1}$ and $\boldsymbol{\mu} = \alpha\mathbf{\Sigma}\mathfrak{G}^{\mathsf{T}}\mathbf{y}$.

Theoretical Results

Thus $\mathcal{C} = \frac{\mathcal{C}_0}{\mathbb{P}(\mathbf{y})}$, where $\mathcal{C}$ is the constant of the Guassian kernel in Equation (24). Therefore,

$$\mathbb{P}(\mathbf{w}|\mathbf{y}) = \mathcal{N}_d(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{26}$$

where $\boldsymbol{\Sigma} = (\alpha\boldsymbol{\mathfrak{G}}^{\mathsf{T}}\boldsymbol{\mathfrak{G}} + \beta\mathbf{I})^{-1}$ and $\boldsymbol{\mu} = \alpha\boldsymbol{\Sigma}\boldsymbol{\mathfrak{G}}^{\mathsf{T}}\mathbf{y}$.

Theoretical Results

Proposition

Let the posterior of the parameters be $\mathbb{P}(\mathbf{w}|\mathbf{y})$ *given in Proposition 2.1, with* $\mathbf{y} = [y(1)\ y(2)\ \cdots\ y(\tau)]^{\mathrm{T}}$. *Further, let* $\mathbf{w} \sim \mathcal{N}_d(\mathbf{0}, \beta^{-1}\mathbf{I})$, *then the gradient noise of* $-\log \mathbb{P}(\mathbf{w}|\mathbf{y})$, *needed for SGHMC's computation is given below:*

$$-\alpha \sum_{t=1}^{\tau} (y(t) - \mathbf{w}^T \mathbf{z}(t))\mathbf{z}(t) + \beta \mathbf{w} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathfrak{A}(\mathbf{w})). \quad (27)$$

Theoretical Results

proof: Again, let $\mathbf{w} \triangleq [w_0 \;\; w_1 \;\; \cdots \;\; w_{\kappa(p,q,m,m)}]^{\mathsf{T}}$, $\kappa(p,l,m,m) \triangleq [(p+1)+l\cdot m]+m$ and let $\mathbf{z}(t) \triangleq [1 \;\; y(t-1) \;\; \cdots \;\; x_m(t-q)]^{\mathsf{T}}$, then

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}[-\log\mathbb{P}(\mathbf{w}|\mathbf{y})] = -\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\left[\ell(\mathbf{w}|\mathbf{y}) + \log\mathbb{P}(\mathbf{w}) - \log\mathbb{P}(\mathbf{y})\right] \tag{28}$$

$$= -\left[\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\ell(\mathbf{w}|\mathbf{y}) + \frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\log\mathbb{P}(\mathbf{w})\right] \tag{29}$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\ell(\mathbf{w}|\mathbf{y}) = \frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\log\left\{\left(\frac{\alpha}{2\pi}\right)^{\tau/2}\exp\left[-\sum_{t=1}^{\tau}\frac{\alpha(y(t)-\mathbf{w}^{\mathsf{T}}\mathbf{z}(t))^2}{2}\right]\right\} \tag{30}$$

$$= -\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\sum_{t=1}^{\tau}\frac{\alpha(y(t)-\mathbf{w}^{\mathsf{T}}\mathbf{z}(t))^2}{2} = \alpha\sum_{t=1}^{\tau}(y(t)-\mathbf{w}^{\mathsf{T}}\mathbf{z}(t))\mathbf{z}(t) \tag{31}$$

Theoretical Results

and the derivative of the prior with log transformation is given by

$$\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\log\mathbb{P}(\mathbf{w}) = -\frac{\beta}{2}\frac{\mathrm{d}}{\mathrm{d}\,\mathbf{w}}\mathbf{w}^{\mathsf{T}}\mathbf{w} = -\beta\mathbf{w}. \tag{32}$$

Equation (27) then follows from stochastic gradient noise.

Forecasting Philippine's Economic Growth

1. **Peso/US Dollar Exchange Rate (ERATE)**
   Monthly average of Philippine Peso-US Dollar exchange rate.
2. **Stock Price Index (SPI)**
   Philippine stock price index (SPI) serves as a measure of the changes in, and the movements of, the average prices of company shares of stock traded in the Philippine Stock Exchange (PSE).
3. **Gross International Reserves (GIR)**
   Foreign assets that are readily available to and controlled by the BSP for direct financing of payments imbalances and for managing the magnitude of such imbalances.

Forecasting Philippine's Economic Growth

1. **Peso/US Dollar Exchange Rate (ERATE)**
   Monthly average of Philippine Peso-US Dollar exchange rate.
2. **Stock Price Index (SPI)**
   Philippine stock price index (SPI) serves as a measure of the changes in, and the movements of, the average prices of company shares of stock traded in the Philippine Stock Exchange (PSE).
3. **Gross International Reserves (GIR)**
   Foreign assets that are readily available to and controlled by the BSP for direct financing of payments imbalances and for managing the magnitude of such imbalances.

Forecasting Philippine's Economic Growth

1. **Peso/US Dollar Exchange Rate (ERATE)**
   Monthly average of Philippine Peso-US Dollar exchange rate.
2. **Stock Price Index (SPI)**
   Philippine stock price index (SPI) serves as a measure of the changes in, and the movements of, the average prices of company shares of stock traded in the Philippine Stock Exchange (PSE).
3. **Gross International Reserves (GIR)**
   Foreign assets that are readily available to and controlled by the BSP for direct financing of payments imbalances and for managing the magnitude of such imbalances.

Forecasting Philippine's Economic Growth

4. **Balance of Payments - Current Account (BOP)**
   Consists of the aggregate balance of goods, services, income and current transfers. This account measures the net transfer of real resources between the domestic economy and the rest of the world.

http://www.bsp.gov.ph/statistics/statistics_metadata.asp

1. Quarterly economic series starts from 1999 first quarter to 2016 third quarter.
2. The indicators are known to have correlation with the gross domestic product of the Philippines.

Forecasting Philippine's Economic Growth

4. **Balance of Payments - Current Account (BOP)**
   Consists of the aggregate balance of goods, services, income and current transfers. This account measures the net transfer of real resources between the domestic economy and the rest of the world.

http://www.bsp.gov.ph/statistics/statistics_metadata.asp

1. Quarterly economic series starts from 1999 first quarter to 2016 third quarter.
2. The indicators are known to have correlation with the gross domestic product of the Philippines.

Forecasting Philippine's Economic Growth

4. **Balance of Payments - Current Account (BOP)**
   Consists of the aggregate balance of goods, services, income and current transfers. This account measures the net transfer of real resources between the domestic economy and the rest of the world.

http://www.bsp.gov.ph/statistics/statistics_metadata.asp

1. Quarterly economic series starts from 1999 first quarter to 2016 third quarter.
2. The indicators are known to have correlation with the gross domestic product of the Philippines.

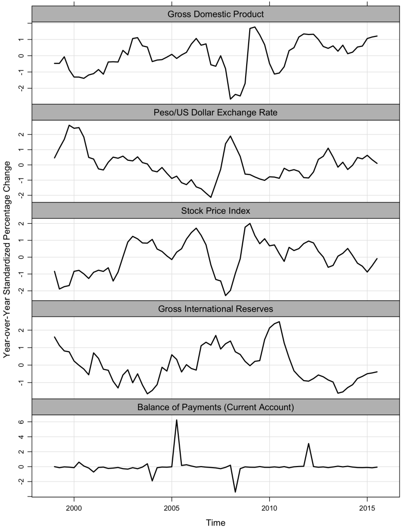
Forecasting Philippine's Economic Growth

4. **Balance of Payments - Current Account (BOP)**
   Consists of the aggregate balance of goods, services, income and current transfers. This account measures the net transfer of real resources between the domestic economy and the rest of the world.

http://www.bsp.gov.ph/statistics/statistics_metadata.asp

1. Quarterly economic series starts from 1999 first quarter to 2016 third quarter.
2. The indicators are known to have correlation with the gross domestic product of the Philippines.

Forecasting Philippine's Economic Growth



Forecasting Philippine's Economic Growth

The ADL(1, 1) model derived from Equation (9) has the following form:

$$y(t) = w_0 + w_1 y(t-1) + \sum_{i=1}^{4}\sum_{l=0}^{1} w_{(2+l\cdot 4)+i} x_i(t) + \varepsilon(t)$$

where $\varepsilon(t) \sim \mathcal{N}(0, 1/\alpha)$. From Proposition 2.2, the gradient noise of the posterior distribution is given by

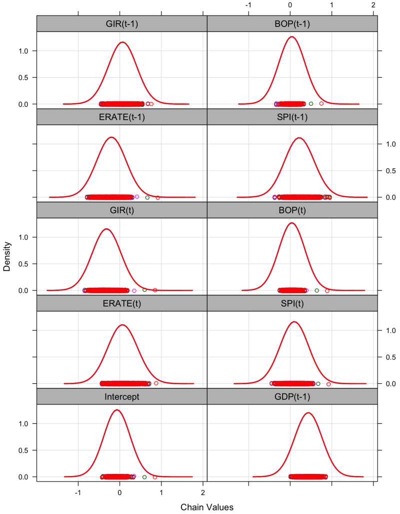
$$-\alpha \sum_{t=1}^{\tau}(y(t) - \mathbf{w}^{\mathsf{T}}\mathbf{z}(t))\mathbf{z}(t) + \beta\mathbf{w} + \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathfrak{A}(\mathbf{w})) \quad (33)$$

Forecasting Philippine's Economic Growth

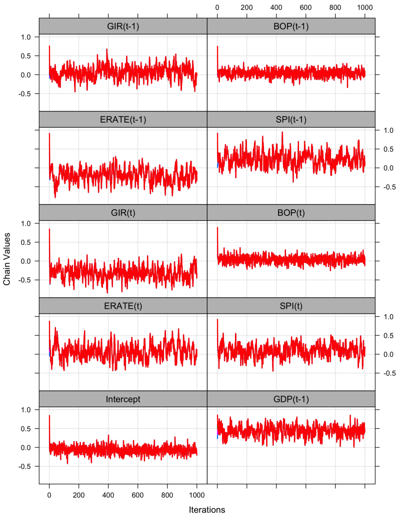
	Variables	BADL(1, 1)-MH		BADL(1, 1)-HMC		BADL(1, 1)-SGHMC	
		Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
1st Case	w_0	-0.127	0.123	-0.066	0.079	-0.068	0.103
	GDP($t-1$)	0.439	0.144	0.261	0.052	0.434	0.14
	ERATE(t)	-0.102	0.087	0.333	0.154	0.072	0.199
	SPI(t)	0.159	0.264	0.155	0.086	0.097	0.165
	GIR(t)	-0.129	0.198	-0.288	0.081	-0.317	0.17
	BOP(t)	0.094	0.108	0	0.031	0.039	0.093
	ERATE($t-1$)	0.323	0.223	-0.399	0.144	-0.204	0.186
	SPI($t-1$)	0.247	0.235	0.359	0.092	0.225	0.192
	GIR($t-1$)	0.133	0.131	0.013	0.084	0.064	0.166
	BOP($t-1$)	0.162	0.085	0.062	0.033	0.04	0.095
2nd Case	w_0	-0.082	0.13	-0.076	0.022	-0.066	0.045
	GDP($t-1$)	0.292	0.105	0.304	0.046	0.43	0.039
	ERATE(t)	0.099	0.091	0.5	0.055	0.08	0.068
	SPI(t)	-0.049	0.122	0.114	0.055	0.11	0.065
	GIR(t)	-0.26	0.102	-0.355	0.039	-0.294	0.071
	BOP(t)	0.03	0.115	-0.019	0.021	0.04	0.042
	ERATE($t-1$)	-0.268	0.166	-0.548	0.077	-0.206	0.074
	SPI($t-1$)	0.459	0.154	0.419	0.07	0.22	0.074
	GIR($t-1$)	-0.033	0.101	0.098	0.023	0.049	0.052
	BOP($t-1$)	-0.006	0.101	0.038	0.023	0.045	0.038
3rd Case	w_0	-0.095	0.094	-0.094	0.022	-0.042	0.098
	GDP($t-1$)	0.273	0.142	0.304	0.027	0.446	0.039
	ERATE(t)	0.693	0.181	0.523	0.043	0.108	0.089
	SPI(t)	0.345	0.09	0.172	0.035	0.148	0.093
	GIR(t)	-0.343	0.196	-0.325	0.038	-0.202	0.144
	BOP(t)	0.018	0.049	-0.004	0.02	0.066	0.078
	ERATE($t-1$)	-0.822	0.169	-0.586	0.056	-0.124	0.158
	SPI($t-1$)	0.173	0.113	0.345	0.041	0.266	0.093
	GIR($t-1$)	0.048	0.17	0.071	0.033	0.045	0.061
	BOP($t-1$)	0.081	0.053	0.062	0.017	0.084	0.091

Table 7.1: *Estimated Coefficients of the BADL(1, 1) using Different MCMCs Across Cases.*

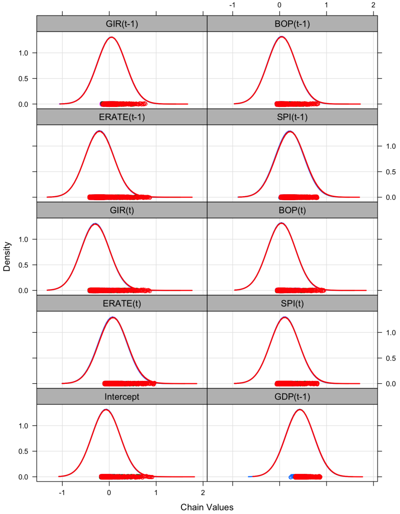
KDE of the Unfiltered Chains using SGHMC Case 1



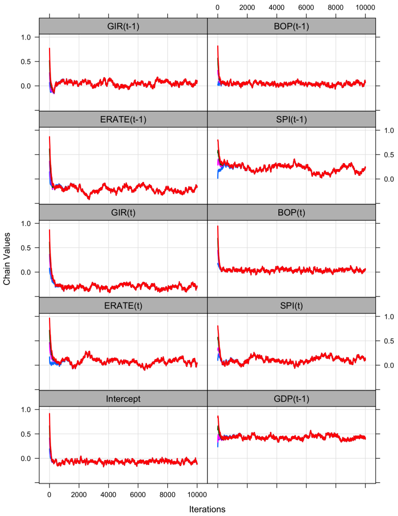
Trace of the Unfiltered Chains using SGHMC Case 1



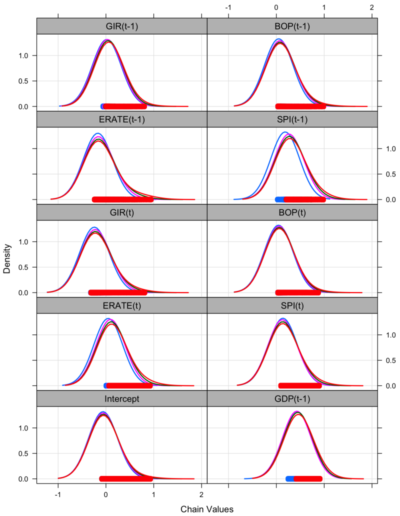
KDE of the Unfiltered Chains using SGHMC Case 2



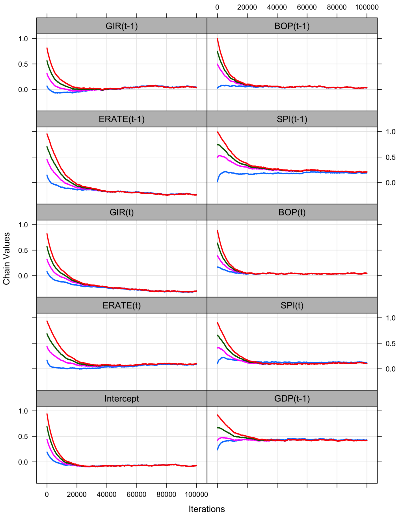
Trace of the Unfiltered Chains using SGHMC Case 2



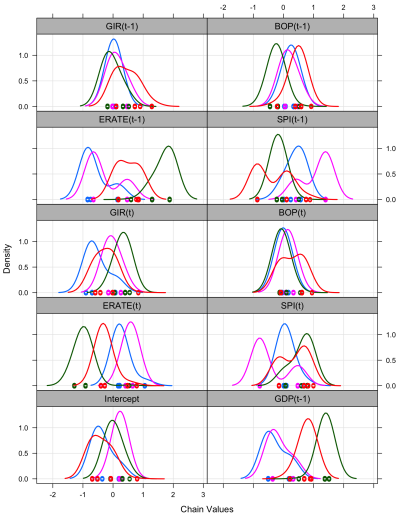
KDE of the Unfiltered Chains using SGHMC Case 3



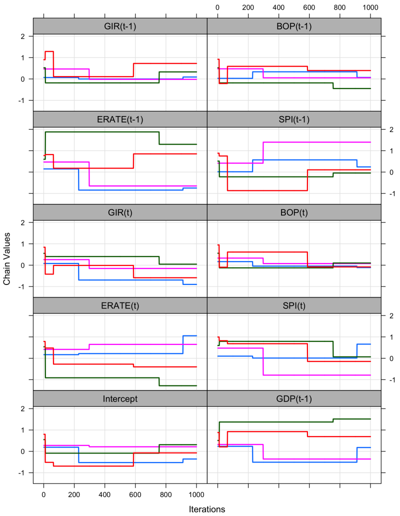
Trace of the Unfiltered Chains using SGHMC Case 3



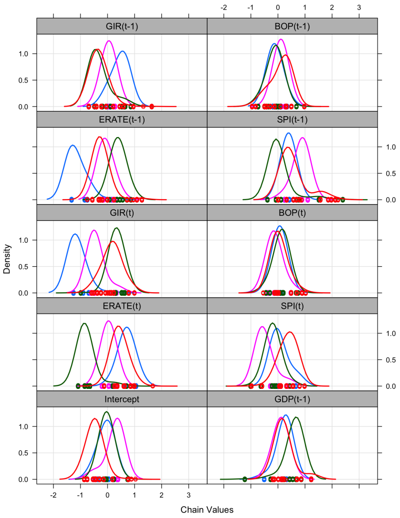
KDE of the Unfiltered Chains using MH Case 1



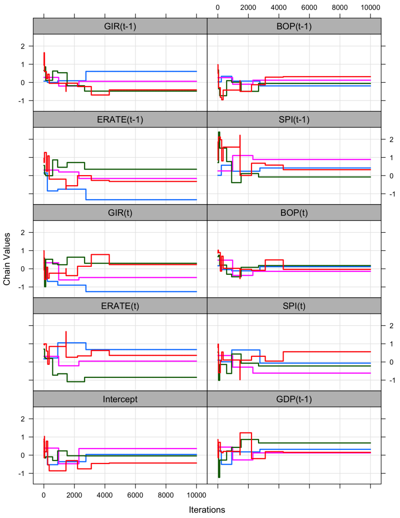
Trace of the Unfiltered Chains using MH Case 1



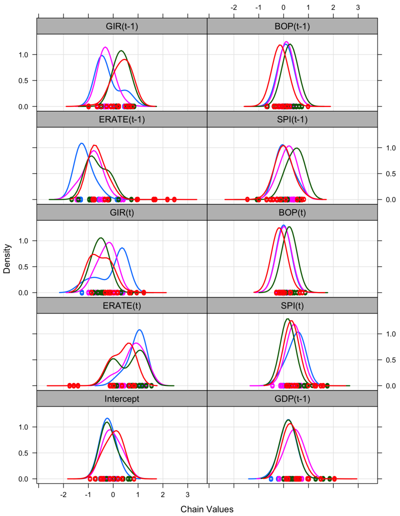
KDE of the Unfiltered Chains using MH Case 2



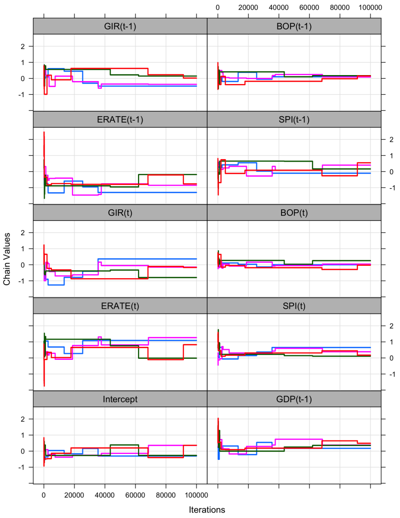
Trace of the Unfiltered Chains using MH Case 2



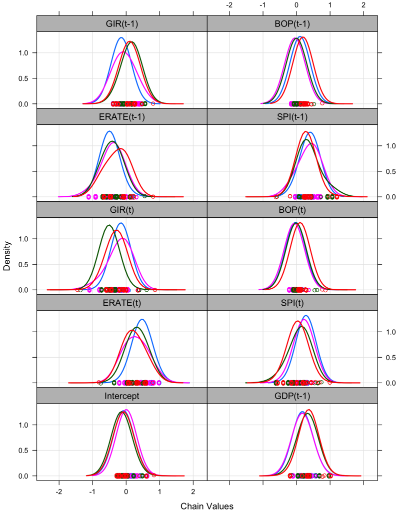
KDE of the Unfiltered Chains using MH Case 3



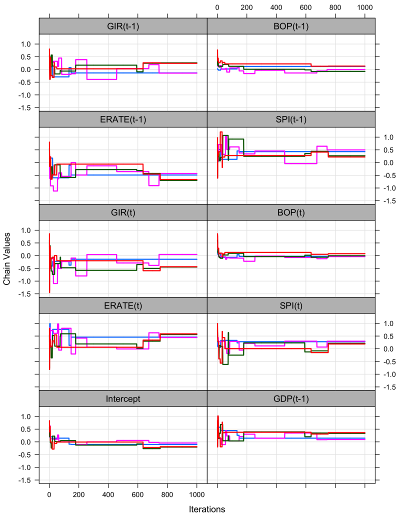
Trace of the Unfiltered Chains using MH Case 3



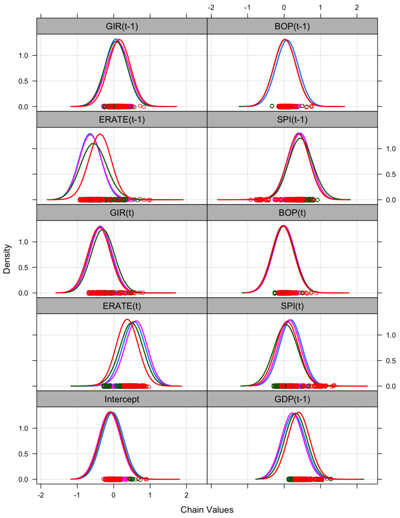
KDE of the Unfiltered Chains using HMC Case 1



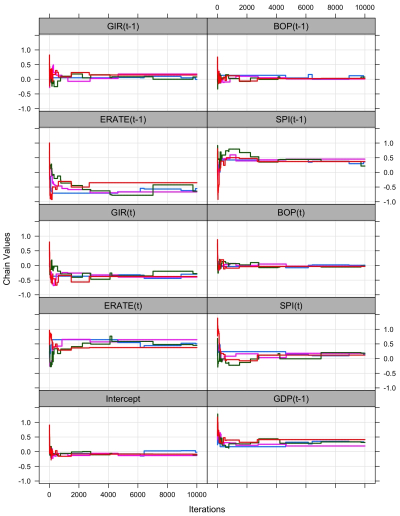
Trace of the Unfiltered Chains using HMC Case 1



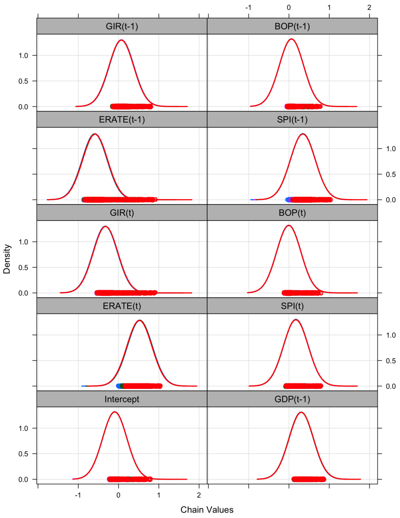
KDE of the Unfiltered Chains using HMC Case 2



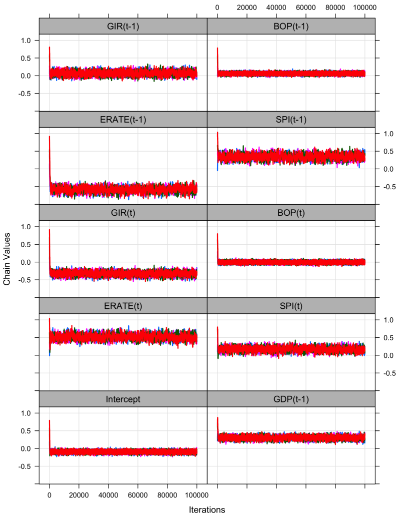
Trace of the Unfiltered Chains using HMC Case 2



KDE of the Unfiltered Chains using HMC Case 3



Trace of the Unfiltered Chains using HMC Case 3



Heidelberger-Welch's Stationarity Test on SGHMC

	Variables	Stationarity Test			Halfwidth Test		
		Status	Start	p-value	Status	Mean	Halfwidth
1st Case	w_0	✓	101	0.457	✗	-0.072	0.0081
	GDP$(t-1)$	✓	1	0.963	✓	0.434	0.0168
	ERATE(t)	✓	1	0.363	✗	0.072	0.0326
	SPI(t)	✓	1	0.530	✗	0.097	0.0256
	GIR(t)	✓	1	0.336	✓	-0.317	0.0254
	BOP(t)	✓	1	0.130	✗	0.039	0.0065
	ERATE$(t-1)$	✓	1	0.224	✗	-0.204	0.0337
	SPI$(t-1)$	✓	1	0.946	✗	0.225	0.0305
	GIR$(t-1)$	✓	1	0.056	✗	0.064	0.0234
	BOP$(t-1)$	✓	1	0.516	✗	0.040	0.0074
2nd Case	w_0	✓	1	0.18	✗	-0.066	0.0095
	GDP$(t-1)$	✓	1	0.66	✓	0.430	0.0110
	ERATE(t)	✓	1	0.20	✗	0.080	0.0249
	SPI(t)	✓	1	0.41	✗	0.110	0.0267
	GIR(t)	✓	1001	0.67	✓	-0.304	0.0132
	BOP(t)	✓	1001	0.76	✗	0.036	0.0063
	ERATE$(t-1)$	✓	1	0.13	✗	-0.206	0.0251
	SPI$(t-1)$	✓	1	0.22	✗	0.220	0.0378
	GIR$(t-1)$	✓	1	0.64	✗	0.049	0.0165
	BOP$(t-1)$	✓	1001	0.41	✗	0.041	0.0072
3rd Case	w_0	✓	1	0.063	✗	-0.042	0.0619
	GDP$(t-1)$	✓	20001	0.146	✓	0.429	0.0057
	ERATE(t)	✓	10001	0.105	✗	0.082	0.0232
	SPI(t)	✓	1	0.176	✗	0.148	0.0893
	GIR(t)	✓	40001	0.052	✗	-0.285	0.0385
	BOP(t)	✓	1	0.178	✗	0.066	0.0526
	ERATE$(t-1)$	✓	40001	0.079	✗	-0.210	0.0300
	SPI$(t-1)$	✓	30001	0.052	✓	0.221	0.0137
	GIR$(t-1)$	✗	NA	0.047	NA	NA	NA
	BOP$(t-1)$	✓	20001	0.122	✗	0.050	0.0124

Table 7.2: *Heidelberger-Welch's Stationarity and Halfwidth Tests of the Stochastic Gradient Hamiltonian Monte Carlos's Mean Chain Across Cases.*

Heidelberger-Welch's Stationarity Test on MH

	Parameters	Stationarity Test			Halfwidth Test		
		Status	Start	p-value	Status	Mean	Halfwidth
1st Case	w_0	✓	1	0.370	✗	-0.1266	0.123
	GDP($t-1$)	✓	201	0.354	✗	0.3851	0.085
	ERATE(t)	✓	1	0.832	✗	-0.1016	0.051
	SPI(t)	✓	301	0.106	✗	0.0081	0.249
	GIR(t)	✓	301	0.082	✗	-0.2375	0.180
	BOP(t)	✓	301	0.052	✗	0.0445	0.102
	ERATE($t-1$)	✓	201	0.248	✗	0.2284	0.101
	SPI($t-1$)	✓	401	0.114	✗	0.3931	0.178
	GIR($t-1$)	✓	1	0.433	✗	0.1330	0.127
	BOP($t-1$)	✓	1	0.102	✗	0.1622	0.056
2nd Case	w_0	NA	NA	NA	NA	NA	NA
	GDP($t-1$)	NA	NA	NA	NA	NA	NA
	ERATE(t)	NA	NA	NA	NA	NA	NA
	SPI(t)	NA	NA	NA	NA	NA	NA
	GIR(t)	NA	NA	NA	NA	NA	NA
	BOP(t)	NA	NA	NA	NA	NA	NA
	ERATE($t-1$)	NA	NA	NA	NA	NA	NA
	SPI($t-1$)	NA	NA	NA	NA	NA	NA
	GIR($t-1$)	NA	NA	NA	NA	NA	NA
	BOP($t-1$)	NA	NA	NA	NA	NA	NA
3rd Case	w_0	✓	1	0.475	✗	-0.095	0.043
	GDP($t-1$)	✓	1	0.100	✗	0.273	0.069
	ERATE(t)	✓	1	0.299	✗	0.693	0.138
	SPI(t)	✓	20001	0.051	✗	0.377	0.070
	GIR(t)	✓	40001	0.425	✗	-0.211	0.042
	BOP(t)	✓	1	0.156	✗	0.018	0.019
	ERATE($t-1$)	✓	1	0.390	✓	-0.822	0.061
	SPI($t-1$)	✓	10001	0.117	✗	0.158	0.086
	GIR($t-1$)	✓	30001	0.051	✗	-0.042	0.122
	BOP($t-1$)	✓	1	0.133	✗	0.081	0.018

Table B.1: *Heidelberger-Welch's Stationarity and Halfwidth Tests of the Metropolis-Hasting's Mean Chain Across Cases.*

Heidelberger-Welch's Stationarity Test on HMC

	Parameters	Stationarity Test			Halfwidth Test		
		Status	Start	p-value	Status	Mean	Halfwidth
1st Case	w_0	✓	101	0.10	✗	-0.08290	0.0521
	GDP$(t-1)$	✓	1	0.65	✓	0.26134	0.0247
	ERATE(t)	✓	1	0.67	✗	0.33301	0.1406
	SPI(t)	✓	1	0.39	✗	0.15495	0.0626
	GIR(t)	✓	1	0.90	✗	-0.28754	0.0297
	BOP(t)	✓	1	0.55	✗	0.00048	0.0098
	ERATE$(t-1)$	✓	1	0.18	✗	-0.39938	0.1304
	SPI$(t-1)$	✓	1	0.62	✗	0.35920	0.0912
	GIR$(t-1)$	✓	401	0.17	✗	0.04093	0.0550
	BOP$(t-1)$	✓	401	0.21	✗	0.04424	0.0155
2nd Case	w_0	✓	1	0.863	✓	-0.076	0.0047
	GDP$(t-1)$	✓	3001	0.335	✓	0.312	0.0088
	ERATE(t)	✓	1001	0.289	✓	0.513	0.0183
	SPI(t)	✓	1	0.083	✗	0.114	0.0302
	GIR(t)	✓	1	0.612	✓	-0.355	0.0100
	BOP(t)	✓	1001	0.225	✗	-0.022	0.0089
	ERATE$(t-1)$	✓	1	0.414	✓	-0.548	0.0383
	SPI$(t-1)$	✓	2001	0.113	✓	0.406	0.0228
	GIR$(t-1)$	✓	2001	0.054	✗	0.102	0.0137
	BOP$(t-1)$	✓	2001	0.059	✗	0.035	0.0149
3rd Case	w_0	✓	30001	0.084	✓	-0.095	0.00082
	GDP$(t-1)$	✓	1	0.110	✓	0.304	0.00226
	ERATE(t)	✓	1	0.189	✓	0.523	0.00480
	SPI(t)	✓	1	0.420	✓	0.172	0.00325
	GIR(t)	✓	30001	0.661	✓	-0.329	0.00265
	BOP(t)	✓	10001	0.437	✗	-0.005	0.00058
	ERATE$(t-1)$	✓	10001	0.652	✓	-0.591	0.00420
	SPI$(t-1)$	✓	1	0.520	✓	0.345	0.00421
	GIR$(t-1)$	✓	30001	0.636	✓	0.073	0.00293
	BOP$(t-1)$	✓	10001	0.126	✓	0.062	0.00059

Table B.2: *Heidelberger-Welch's Stationarity and Halfwidth Tests of the Hamiltonian Monte Carlo's Mean Chain Across Cases.*

Gelman-Rubin's Convergence Test

	Parameters	BADL(1,1)-HMC		BADL(1,1)-SGHMC	
		Point Estimate	95% UCI	Point Estimate	95% UCI
2nd Case	w_0	NA	NA	0.9999	0.9999
	GDP($t-1$)	NA	NA	0.9999	0.9999
	ERATE(t)	NA	NA	0.9999	0.9999
	SPI(t)	NA	NA	0.9999	0.9999
	GIR(t)	NA	NA	0.9999	0.9999
	BOP(t)	NA	NA	0.9999	0.9999
	ERATE($t-1$)	NA	NA	0.9999	0.9999
	SPI($t-1$)	NA	NA	0.9999	0.9999
	GIR($t-1$)	NA	NA	0.9999	0.9999
	BOP($t-1$)	NA	NA	0.9999	0.9999
3rd Case	w_0	1.0006	1.0015	1.0129	1.0393
	GDP($t-1$)	1.0040	1.0115	1.5102	2.1765
	ERATE(t)	1.0080	1.0203	1.2089	1.5239
	SPI(t)	1.0024	1.0058	1.9008	3.0432
	GIR(t)	1.0036	1.0103	1.0201	1.0605
	BOP(t)	1.0002	1.0006	1.0039	1.0119
	ERATE($t-1$)	1.0085	1.0208	1.0172	1.0515
	SPI($t-1$)	1.0011	1.0028	1.9789	3.1677
	BOP($t-1$)	1.0003	1.0008	1.0078	1.0239

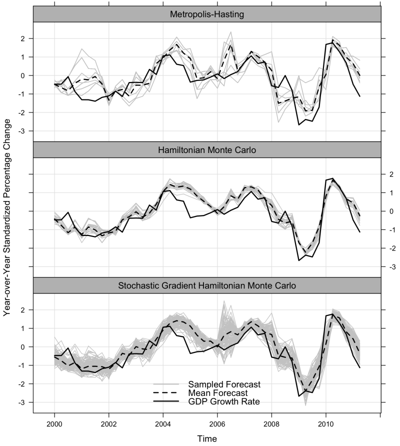
Table 7.3: *Gelman-Rubin's Estimated Potential Scale Reduction or Shrink Factor of the Four Markov Chains.*

Empirical Posterior Means of Model Parameters

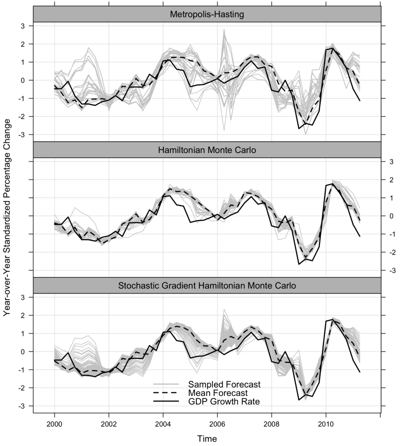
Case	Variables	BADL(1,1)-MH Coefficient	BADL(1,1)-MH Std. Error	BADL(1,1)-HMC Coefficient	BADL(1,1)-HMC Std. Error	BADL(1,1)-SGHMC Coefficient	BADL(1,1)-SGHMC Std. Error
1st Case	w_0	-0.127	0.123	-0.066	0.079	-0.068	0.103
	GDP$(t-1)$	0.439	0.144	0.261	0.052	0.434	0.14
	ERATE(t)	-0.102	0.087	0.333	0.154	0.072	0.199
	SPI(t)	0.159	0.264	0.155	0.086	0.097	0.165
	GIR(t)	-0.129	0.198	-0.288	0.081	-0.317	0.17
	BOP(t)	0.094	0.108	0	0.031	0.039	0.093
	ERATE$(t-1)$	0.323	0.223	-0.399	0.144	-0.204	0.186
	SPI$(t-1)$	0.247	0.235	0.359	0.092	0.225	0.192
	GIR$(t-1)$	0.133	0.131	0.013	0.084	0.064	0.166
	BOP$(t-1)$	0.162	0.085	0.062	0.033	0.04	0.095
2nd Case	w_0	-0.082	0.13	-0.076	0.022	-0.066	0.045
	GDP$(t-1)$	0.292	0.105	0.304	0.046	0.43	0.039
	ERATE(t)	0.099	0.091	0.5	0.055	0.08	0.068
	SPI(t)	-0.049	0.122	0.114	0.055	0.11	0.065
	GIR(t)	-0.26	0.102	-0.355	0.039	-0.294	0.071
	BOP(t)	0.03	0.115	-0.019	0.021	0.04	0.042
	ERATE$(t-1)$	-0.268	0.166	-0.548	0.077	-0.206	0.074
	SPI$(t-1)$	0.459	0.154	0.419	0.07	0.22	0.074
	GIR$(t-1)$	-0.033	0.101	0.098	0.023	0.049	0.052
	BOP$(t-1)$	-0.006	0.101	0.038	0.023	0.045	0.038
3rd Case	w_0	-0.095	0.094	-0.094	0.022	-0.042	0.098
	GDP$(t-1)$	0.273	0.142	0.304	0.027	0.446	0.039
	ERATE(t)	0.693	0.181	0.523	0.043	0.108	0.089
	SPI(t)	0.345	0.09	0.172	0.035	0.148	0.093
	GIR(t)	-0.343	0.196	-0.325	0.038	-0.202	0.144
	BOP(t)	0.018	0.049	-0.004	0.02	0.066	0.078
	ERATE$(t-1)$	-0.822	0.169	-0.586	0.056	-0.124	0.158
	SPI$(t-1)$	0.173	0.113	0.345	0.041	0.266	0.093
	GIR$(t-1)$	0.048	0.17	0.071	0.033	0.045	0.061
	BOP$(t-1)$	0.081	0.053	0.062	0.017	0.084	0.091

Table 7.1: *Estimated Coefficients of the BADL(1, 1) using Different MCMCs Across Cases.*

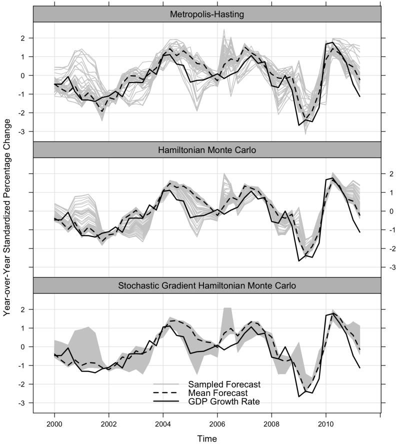
Forecasts on Training Data: Case 1



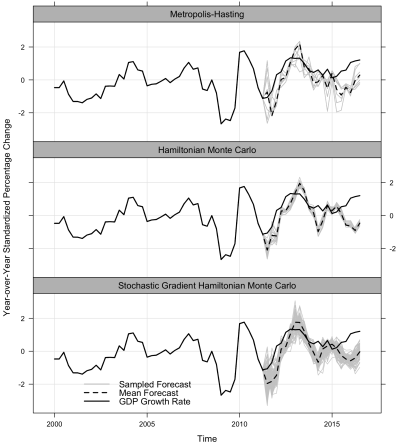
Forecasts on Training Data: Case 2



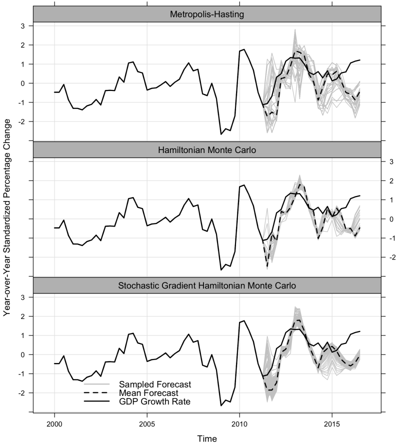
Forecasts on Training Data: Case 3



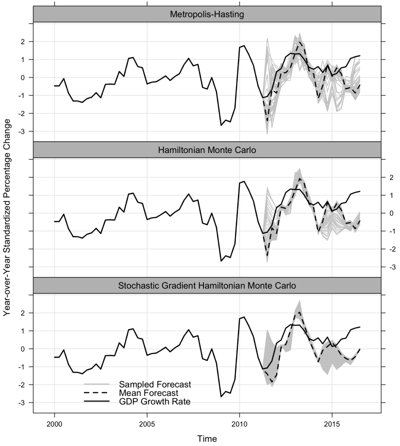
Forecasts on Testing Data: Case 1



Forecasts on Testing Data: Case 2



Forecasts on Testing Data: Case 3



Forecasts' Root Mean Squared Error

Models	Root Mean Squared Error			
	Error Type	1st Case	2nd Case	3rd Case
BADL(1, 1)-**MH**	In-Sample	0.774	0.612	0.567
	Out-Sample	0.952	0.982	0.985
BADL(1, 1)-**HMC**	In-Sample	0.585	0.574	0.538
	Out-Sample	0.979	0.984	0.981
BADL(1, 1)-**SGHMC**	In-Sample	0.606	0.605	0.557
	Out-Sample	0.906	0.911	0.935

Table 7.4: *Root Mean Square Error of the Mean Forecasts of Bayesian ADL.*

Summary

The integration of the Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) in estimating the Bayesian Autoregressive Distributed Lag (BADL) model were presented in two propositions:

- Proposition 6.0.1
  the posterior distribution of the parameters of the BADL;
- Proposition 6.0.2
  the potential energy or the negative log *a posteriori*.

Summary

The theoretical results were applied to forecasting the economic growth of the Philippines. The study considers a simple model of the BADL with order $p = q = 1$, and hence the model of interests are the following:

- BADL$(1,1)$-MH
- BADL$(1,1)$-HMC
- BADL$(1,1)$-SGHMC.

Summary

The theoretical results were applied to forecasting the economic growth of the Philippines. The study considers a simple model of the BADL with order $p = q = 1$, and hence the model of interests are the following:

- BADL$(1, 1)$-MH
- BADL$(1, 1)$-HMC
- BADL$(1, 1)$-SGHMC.

Summary

- The diagnostics of the sampled chains were assessed through different statistical tests.
    - Heidelberger-Welch's Stationary Test;
    - Gelman-Rubin's Convergence Test.
- The result suggests that the BADL$(1,1)$-SGHMC performs well compared to the contenders.

Summary

- The diagnostics of the sampled chains were assessed through different statistical tests.
  - Heidelberger-Welch's Stationary Test;
  - Gelman-Rubin's Convergence Test.
- The result suggests that the BADL$(1,1)$-SGHMC performs well compared to the contenders.

Summary

- The diagnostics of the sampled chains were assessed through different statistical tests.
  - Heidelberger-Welch's Stationary Test;
  - Gelman-Rubin's Convergence Test.
- The result suggests that the BADL$(1,1)$-SGHMC performs well compared to the contenders.

Usage of Programming Languages in the Study

- The use of Julia did save a lot of time in simulation compared to R. The 1.5 hour MH MCMC simulation in R is only 10.7 seconds in Julia.
- And since Julia has the advantage of using multiple processor, that 10.7 seconds for one processor (by default in Julia) is expected to further decrease if multiple processors were initialized.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- The use of Julia did save a lot of time in simulation compared to R. The 1.5 hour MH MCMC simulation in R is only 10.7 seconds in Julia.
- And since Julia has the advantage of using multiple processor, that 10.7 seconds for one processor (by default in Julia) is expected to further decrease if multiple processors were initialized.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- The use of Julia did save a lot of time in simulation compared to R. The 1.5 hour MH MCMC simulation in R is only 10.7 seconds in Julia.
- And since Julia has the advantage of using multiple processor, that 10.7 seconds for one processor (by default in Julia) is expected to further decrease if multiple processors were initialized.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- Finally, in line with the above discussions, this thesis developed the first R and Julia packages for SGHMC.
- These tools will help the readers explore the proposed model for their own datasets, and use SGHMC for other interesting models.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- Finally, in line with the above discussions, this thesis developed the first R and Julia packages for SGHMC.
- These tools will help the readers explore the proposed model for their own datasets, and use SGHMC for other interesting models.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- Finally, in line with the above discussions, this thesis developed the first R and Julia packages for SGHMC.
- These tools will help the readers explore the proposed model for their own datasets, and use SGHMC for other interesting models.
- And one more thing, Julia's syntax is very close to Matlab which is not difficult to learn.

Usage of Programming Languages in the Study

- Finally, in line with the above discussions, this thesis developed the first R and Julia packages for SGHMC.
- These tools will help the readers explore the proposed model for their own datasets, and use SGHMC for other interesting models.

Usage of Programming Languages in the Study

- Finally, in line with the above discussions, this thesis developed the first R and Julia packages for SGHMC.
- These tools will help the readers explore the proposed model for their own datasets, and use SGHMC for other interesting models.

Usage of Programming Languages in the Study

Show the Documentation of the Packages

Conclusion

- One can conclude that the proposed BADL-SGHMC is a good alternative for BADL-MH and BADL-HMC models.
- In particular, lower leapfrog step size is good for BADL-SGHMC even for 1,000 iterations only.
- Further, the training dataset used in modeling the economic growth of the Philippines uses 70% of the entire dataset.
- The unpartitioned dataset contains 67 total data points. The 70% accounts to 46 observations and hence the remaining 21 observations are reserved for the testing dataset.

Conclusion

- One can conclude that the proposed BADL-SGHMC is a good alternative for BADL-MH and BADL-HMC models.
- In particular, lower leapfrog step size is good for BADL-SGHMC even for 1,000 iterations only.
- Further, the training dataset used in modeling the economic growth of the Philippines uses 70% of the entire dataset.
- The unpartitioned dataset contains 67 total data points. The 70% accounts to 46 observations and hence the remaining 21 observations are reserved for the testing dataset.

Conclusion

- One can conclude that the proposed BADL-SGHMC is a good alternative for BADL-MH and BADL-HMC models.
- In particular, lower leapfrog step size is good for BADL-SGHMC even for 1,000 iterations only.
- Further, the training dataset used in modeling the economic growth of the Philippines uses 70% of the entire dataset.
- The unpartitioned dataset contains 67 total data points. The 70% accounts to 46 observations and hence the remaining 21 observations are reserved for the testing dataset.

Conclusion

- One can conclude that the proposed BADL-SGHMC is a good alternative for BADL-MH and BADL-HMC models.
- In particular, lower leapfrog step size is good for BADL-SGHMC even for 1,000 iterations only.
- Further, the training dataset used in modeling the economic growth of the Philippines uses 70% of the entire dataset.
- The unpartitioned dataset contains 67 total data points. The 70% accounts to 46 observations and hence the remaining 21 observations are reserved for the testing dataset.

Conclusion

- According to Chen et al. (2014), the central limit theorem (CLT) for the stochastic gradient noise of the potential energy is satisfied for at least 100 observations, which is far from 46 data points considered in this study.
- Interestingly, the SGHMC algorithm works well even for 46 observations only, this is evident on the performance of the BADL-SGHMC discussed in the precedings chapter compared to BADL-MH and BADL-HMC.

Conclusion

- According to Chen et al. (2014), the central limit theorem (CLT) for the stochastic gradient noise of the potential energy is satisfied for at least 100 observations, which is far from 46 data points considered in this study.
- Interestingly, the SGHMC algorithm works well even for 46 observations only, this is evident on the performance of the BADL-SGHMC discussed in the precedings chapter compared to BADL-MH and BADL-HMC.

Conclusion

- The limitation of the data points to 67 observations only depends heavily on the historical data of the indicators used in this study.
- Therefore, if at least 100 observations were used in the training dataset, the proposed BADL-SGHMC might perform even better compared to the contender. And that would be part of the recommendation.

Conclusion

- The limitation of the data points to 67 observations only depends heavily on the historical data of the indicators used in this study.
- Therefore, if at least 100 observations were used in the training dataset, the proposed BADL-SGHMC might perform even better compared to the contender. And that would be part of the recommendation.

Recommendation

The following are possible extensions for future work:

- Explore higher orders of the BADL-SGHMC using at least 100 observations;
- Set priors on the precision parameters of the model likelihood, which in this study is assumed to be known, fixed to constant α^{-1};

Recommendation

The following are possible extensions for future work:

- Explore higher orders of the BADL-SGHMC using at least 100 observations;
- Set priors on the precision parameters of the model likelihood, which in this study is assumed to be known, fixed to constant α^{-1};

Recommendation

The following are possible extensions for future work:

- Explore complicated BADL-SGHMC by setting priors on all hyperparameters;
- Explore different values on the parameters of the SGHMC, especially for $\mathfrak{C}$ and $\mathfrak{A}$ which were considered as identity matrix for this study;

Recommendation

The following are possible extensions for future work:

- Explore complicated BADL-SGHMC by setting priors on all hyperparameters;
- Explore different values on the parameters of the SGHMC, especially for $\mathfrak{C}$ and $\mathfrak{A}$ which were considered as identity matrix for this study;

Recommendation

The following are possible extensions for future work:

- Consider other Stochastic Gradient MCMC algorithms, like Stochastic Gradient Langevin Monte Carlo (SGLMC) or Langevin Dynamics (SGLD) (*see* Welling and Teh 2011); and Stochastic Gradient Fisher Scoring (*see* Ahn et al. 2012); and,
- Obtain the prediction interval of the forecast by computing the predictive distribution for each items mentioned above.

Recommendation

The following are possible extensions for future work:

- Consider other Stochastic Gradient MCMC algorithms, like Stochastic Gradient Langevin Monte Carlo (SGLMC) or Langevin Dynamics (SGLD) (*see* Welling and Teh 2011); and Stochastic Gradient Fisher Scoring (*see* Ahn et al. 2012); and,
- Obtain the prediction interval of the forecast by computing the predictive distribution for each items mentioned above.

Thank you!