A Bayesian Nonparametric Investigation of the Relationship between Commodity Prices and Exchange Rates*

Xin Jin
University of Toronto

November, 2011

Abstract

This paper proposes a full Bayesian nonparametric procedure to investigate the predictive power of exchange rates on commodity prices for 3 commodity-exporting countries: Canada, Australia and New Zealand. We examine the predictive effect of exchange rates on the entire distribution of commodity prices and how this effect changes over time. For this purpose, a time-dependent infinite mixture of normal linear regression model is proposed for the conditional distribution of the commodity price index. The mixing weights of the mixture follow a Probit stick-breaking prior and are hence time-varying. As a result, we allow the conditional distribution of the commodity price index given exchange rates to change over time nonparametrically. It is shown that exchange rates do not have consistent predictive power for commodity prices in all countries considered in this paper. We find that exchange rates do have predictive power in some cases, but their impact tends to be constant over time. On the other hand, the intercept in the regression and the lagged dependent variable show signs of parameter change over time in most cases, which is important in predicting both the mean and the density of the commodity prices one period ahead. The results also suggest that for all countries considered, a significant source of time variation in the conditional distribution of commodity prices comes from the variance.

1 Introduction

This paper proposes a full Bayesian nonparametric procedure to investigate the predictive power of exchange rates over commodity prices for 3 commodity-exporting countries: Canada, Australia and New Zealand. We examine the predictive effect of exchange rates on the entire distribution of commodity prices and how this effect changes over time. For this

*I sincerely thank my supervisor John M. Maheu for his guidance and support throughout the whole project. I am also deeply thankful for excellent comments and generous suggestions from Martin Burda, Christian Gourieroux, Angelo Melino, Xianwen Shi, Yuanyuan Wan, Luhang Wang and seminar attendants at University of Toronto. All errors are mine.

purpose, a time-dependent infinite mixture of normal linear regression model is proposed for the conditional distribution of the commodity price index. The mixing weights of the mixture follow a Probit stick-breaking prior and are hence time-varying, As a result, we allow the conditional distribution of the commodity price index given the exchange rate to change over time nonparametrically. In terms of forecasting, we are able to produce density forecasts as well as point forecasts of future commodity prices, which fully incorporates potential time variation in the underlying relationship.

In a recent paper by Chen, Rogoff, and Rossi (2010), a predictive relationship is uncovered between "commodity currency"[1] exchange rates and global commodity prices. In a classical econometric setting, they show that after controlling for parameter instability, the currencies of a panel of small commodity-exporting countries with floating exchange rate regimes have robust predicting power for the prices of the commodities being exported. They show that the relationship holds both in-sample and out-of-sample, and against a variety of benchmarks.

Their study has important implications from both an economic theory point of view and a policy-making perspective. First, testing the predictive power of exchange rates for commodity prices can be viewed as a test to a present-value model of exchange rate determination.[2] At the same time, predicting commodity prices is very important for policy-making. For example, commodity price fluctuations are closely related to inflation dynamics. For small open economies with substantial commodity exports, price changes have a significant impact on their GDP and current accounts.

Chen, Rogoff, and Rossi (2010) study the predictive power of exchange rates for commodity prices one period ahead in a linear predictive regression framework. They show that their success in finding the relationship depend critically on controlling for parameter instabilities in the regression model they used. In their in-sample study, Chen et al. (2010) incorporated the Exp-W* Test by Rossi (2005) into their model to test the null hypothesis that at no point in time there exists predictive power of exchange rates on commodity prices. The results showed that the null is rejected in almost all cases, indicating Granger causality of exchange rates on commodity prices for at least some points in time during the whole sample period. In their out-of-sample studies, to account for parameter instability they used a rolling window forecasting scheme to generate point forecasts of commodity prices. They showed that including the exchange rates in the mean predictive regression generally produces forecasts of commodity prices with lower mean squared error, compared to alternative benchmark models such as AR1 and random walks.

Their findings are encouraging and their ways of testing are clever within a mean regression framework. However there still remain questions that need to be addressed. First, their mean regression framework only concerns the impact of exchange rates on the conditional mean of commodity prices, and hence can only generate point forecasts of future commodity prices. But for policy-making purposes, an accurate forecast of the future distribution of commodity prices would be of greater relevance. Second, their model has no mechanism to describe the possible time variation in the parameters. In their framework, no quantitative inference can be made on the parameters at any point in time. They only answered the ques-

[1]Commodity currencies are referred to currencies of countries which depend heavily on the export of certain raw materials for income.

[2]Engel and West (2005) also use present-value approach to modeling exchange rates, but they use other fundamentals such as money supplies, etc.

tion of whether there is a predictive relationship from exchange rates to commodity prices, given the possibility of parameter change; but we still do not know if there is any parameter change at all,or how significant the change is, or when the change happens, etc. In other words, they have answered the "yes" or "no" questions, but failed to address "how" and "by how much". Knowing "yes" or "no" is undoubtedly theoretically significant. However to understand "how" and "by how much" will have further implications, particularly for purposes of policy making.

Finally their out-of-sample studies are a disconnected and distinct procedure from their in-sample analysis. The rolling window scheme is an ad hoc way of dealing with time-varying parameters. Without a law of motion of the underlying parameter, it cannot optimally incorporate parameter change to help forecast.here is also the issue of deciding the size of the window. To maintain a fixed size of the rolling window in-sample data points get discarded as out-of-sample data points are added, leading to potential information loss.

The contribution of this paper is motivated by the above issues. This paper revisits the topic on the relationship between exchange rates and commodity prices and addresses all the issues in a unified Bayesian nonparametric setting. It proposes a flexible time-dependent infinite mixture of normal linear regressions to model the conditional distribution of commodity prices given exchange rates. Thus it focuses on the predictive effect of exchange rates on the entire distribution of commodity prices, rather than just the conditional mean. It allows the relationship to change by letting the mixture weights vary over time, thus effectively allowing for time-varying parameters. This is achieved nonparametrically by drawing the mixture weights from a set of Probit stick breaking processes with the underlying latent processes formulated as independent AR(1). Under this framework direct inference on the impact of exchange rates on commodity prices at any point in time in-sample is straightforward using standard Bayesian inference procedure. Meanwhile the latent processes in the PSBP priors provide a law of motion for the time-varying conditional distributions over time, making out-of-sample density forecast and point forecast an integrated and coherent part of whole the procedure that fully incorporates parameter instability.

We examine the predictive power of exchange rates on commodity prices for three commodity-exporting countries: Canada, Australia, New Zealand. Our results suggest that nominal exchange rates do not have a consistent predictive power for commodity prices in all three countries. We find that exchange rates do have predictive power in some cases, but their impact tends to be constant over time. On the other hand, the intercept in the regression and the lagged dependent variable show signs of time-varying parameter change in most cases, which is important in predicting both the mean and the density of the commodity prices one period ahead. The results also suggest that for all countries considered, a significant source of time variation in the conditional distribution of commodity prices comes from the variance.

The rest of the paper is organized as follows. In Section 2, we review the theory on Bayesian density estimation using Dirichlet Process Mixture (DPM). In Section 3, the Time-Dependent PSBP Mixture of Normal Linear Regression Model is proposed to nonparametrically model the time-varying conditional distribution of a scalar response y_t on a set of predictors $\boldsymbol{x}_t$. Section 4 briefly explains the posterior sampling procedure and outlines model comparison. Section 5 provides simulation study for illustration. Section 6 contains the data description. A detailed report of the results is in Section 7 and Section 8 concludes. The

Appendix contains details on posterior simulation.

2 Density Estimation using DPM

Ferguson (1973) opened the door for modern Bayesian nonparametrics by developing theories and properties of the Dirichlet process (DP), which can be used as a nonparametric prior for discrete random distributions. In the application of Bayesian density estimation, the Dirichlet Process Mixture (DPM) model was used by Escobar and West (1995) and West, Müller, and Escobar (1994) to model a continuous density by an infinite mixture of normal densities, with the mixture distribution drawing from a DP prior. More specifically, a typical DPM model for an unknown continuous distribution $f(y)$ can be formalized as the following:

$$\begin{aligned} f(y) &= \int f(y|\phi)G(\mathrm{d}\phi) && (1) \\ G &\sim \mathrm{DP}(\alpha G_0) && (2) \end{aligned}$$

where $\phi = (\mu, \sigma^2)$, $f(y|\phi) = \mathrm{N}(.|\mu, \sigma^2)$, and $\mathrm{DP}(\alpha G_0)$ is a DP prior with scalar precision parameter α and base measure G_0. It is a well known fact that DPM can approximate any continuous density with arbitrary accuracy (Ghosal, Ghosh, and Ramamoorthi 1999).

In situations where the conditional distribution of the response y_i depends on some covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ as in this paper, we can easily extend equation (1) to

$$f(y_i|\boldsymbol{x}_i) = \int f(y_i|\boldsymbol{x}_i, \phi)G(\mathrm{d}\phi).$$

In particular, if we assume $f(y_i|\boldsymbol{x}_i, \phi) = \mathrm{N}(.|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$, where $\phi = (\boldsymbol{\beta}, \sigma^2)$, then $f(y_i|\boldsymbol{x}_i)$ is modeled as an infinite mixture of normal linear regressions using a DPM (West, Müller, and Escobar (1994)). In the special case of $\boldsymbol{x}_i \equiv 1$, we are back to the unconditional distribution.

According to Sethuraman (1994), a random distribution G follows a DP prior if and only if it has a stick-breaking representation of the form:

$$\begin{aligned} G &= \sum_{j=1}^{\infty} w_j \delta_{\phi_j}(.) \\ w_j &= v_j \prod_{k<j}(1 - v_k), \quad v_j \overset{iid}{\sim} \mathrm{Beta}(1, \alpha) \\ \phi_j &\overset{iid}{\sim} G_0 \end{aligned}$$

w_j are called the stick-breaking weights, ϕ_j are the associated atoms drawn iid from G_0. $\delta_{\phi_j}(.)$ denotes a discrete measure concentrated at ϕ_j. $\mathrm{Beta}(1, \alpha)$ stands for a Beta distribution with parameters 1 and α. ϕ_j and v_i are generated independently. Thus, the above DP mixture of normal linear regression model (DPM) of $f(y_i|\boldsymbol{x}_i)$ can be re-written as:

$$
\begin{aligned}
f(y_i|\boldsymbol{x}_i) &= \sum_{j=1}^{\infty} w_j \mathrm{N}(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}_j, \sigma_j^2) && (3)\\
w_j &= v_j \prod_{k<j}(1-v_k) && (4)\\
v_j &\overset{iid}{\sim} \mathrm{Beta}(1,\alpha) && (5)\\
(\boldsymbol{\beta}_j, \sigma_j^2) &\overset{iid}{\sim} G_0 && (6)
\end{aligned}
$$

Compared to a simple normal linear regression $f(y_i|\boldsymbol{x}_i) = \mathrm{N}(y_i|\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$, the DPM model offers more flexibility in modeling the conditional distribution of $f(y_i|\boldsymbol{x}_i)$, since it relaxes the normal error assumption and allows much more general density form in a nonparametric way. At the same time the linear mean regression structure is preserved. Indeed, assuming that the weights $\{w_j\}$ and the regression coefficients $\{\boldsymbol{\beta}_j\}$ are known,

$$
\begin{aligned}
\mathbb{E}(y_i|\boldsymbol{x}_i) &= \sum_{j=1}^{\infty} w_j \boldsymbol{x}_i'\boldsymbol{\beta}_j\\
&= \sum_{j=1}^{\infty} w_j \sum_{l=1}^{p} x_{il}\beta_{jl}\\
&= \sum_{l=1}^{p} x_{il} \sum_{j=1}^{\infty} w_j\beta_{jl}\\
&= \boldsymbol{x}_i'\bar{\boldsymbol{\beta}}
\end{aligned}
$$

where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)$ and $\bar{\beta}_l = \sum_j w_j\beta_{jl}$. $\bar{\boldsymbol{\beta}}$ represent the linear effect of the covariates on the mean of the dependent variable. This linear structure is often desirable because of its simplicity and ease for interpretation. The limitation of this DPM model is, however, that this linear relationship along with the shape of the conditional density are homogeneous across all observations. This may be too restrictive in certain situations. For example, when y_t and $\boldsymbol{x}_t$ are time series data and we want to examine the relationship between y_t and $\boldsymbol{x}_t$, and how this relationship evolves (or if it evolves at all) over time, we may then need a more flexible model that, on one hand, still maintains the linear conditional mean structure and flexible density form assumptions for the conditional distribution, but on the other hand allows both of the features to change over time. And more than likely, we would prefer the change to occur in a "smooth" fashion. That is, two observations that are close in time should have similar conditional distribution structure.

3 Time-Dependent Probit Stick-Breaking Process Mixture of Normal Linear Regression Model

Rodríguez and Dunson (2011) introduced a new class of Bayesian nonparametric priors called the Probit Stick-Breaking Processes (PSBPs), where probit transformations of normal random variables are utilized in constructing the stick-breaking weights.[3] They show that compared to conventional DPs, PSBPs can easily extend from priors on single distributions to priors on collections of dependent distributions, while preserving computation tractability. Under their framework, we propose the following time-dependent PSBP mixture of normal linear regression model for the conditional distribution of y_t on $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$, $t = 1, \ldots, T$:

$$
\begin{aligned}
f(y_t|\boldsymbol{x}_t) &= \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\boldsymbol{x}_t'\boldsymbol{\beta}_j, \sigma_j^2) && (7)\\
w_{jt} &= \Phi(\alpha_{jt}) \prod_{k<j} (1 - \Phi(\alpha_{kt})) && (8)\\
\alpha_{jt} &\overset{\perp}{\sim} \mathrm{N}(.|\gamma_0 + \gamma_1 \alpha_{jt-1}, \sigma_\alpha^2) && (9)\\
(\boldsymbol{\beta}_j, \sigma_j^2) &\overset{iid}{\sim} G_0(\lambda) && (10)\\
\lambda &\sim \mathcal{G}_1 && (11)\\
(\gamma_0, \gamma_1, \sigma_\alpha^2) &\sim \mathcal{G}_2 && (12)
\end{aligned}
$$

where $\Phi(.)$ denotes the cumulative distribution function for the standard normal variable. In the rest of the paper we refer to the above model as the “time-dependent model”. The time-dependent model resembles the DP mixture of normal linear regression (DPM) in many ways. As is the case with DPM, $f(y_t|\boldsymbol{x}_t)$ is modeled as an infinite mixture of normal linear regressions, so the conditional mean of the response y_t is still linear in the covariates. And in both models, the mixing weights are drawn from some stick-breaking processes (SBPs). There are, however, some important differences between the two models. First of all, the stick-breaking processes that produce the mixing weights are different for the two models. In DPM, the SBP uses i.i.d. beta random variables, which corresponds to a DP prior. While in the time-dependent model, the SBPs make use of Probit transformations of marginally normal random variables, so they belong to the class of PSBP priors. This probit structure in the stick breaking mechanism has important implications in posterior sampling and will be discussed in more details in Section 9. Secondly, instead of having a constant set of fixed weights $\{w_j\}_{j=1}^{\infty}$ for all t, which is the case in the DPM, the time-dependent model allows the weights w_{jt} to change over time, hence allowing $f(y_t|\boldsymbol{x}_t)$ to change over time as well. More specifically, at *each point t in time*, a set of $\{w_{jt}\}_{j=1}^{\infty}$ is produced from a Probit stick-breaking process associated with time t (equation (8)). That is, the single SBP for individual sets of random weights in DPM is replaced by a collection, or a “time series” of dependent Probit SBPs (indexed by t) that produce a “time series” of sets of random weights.

In the meantime, this time series of PSBPs are linked together through the underlying infinitely many latent processes $\{\boldsymbol{\alpha}_j\}_{j=1}^{\infty}$. For each j, $\boldsymbol{\alpha}_j = \{\alpha_{jt}\}_{t=1}^{T}$ is an independent process assumed to have a Gaussian marginal distribution. The Gaussian marginal distribution together with the probit structure serves for the purpose of computation tractability and again will be discussed in more detail in Section 9.

Compared to the DPM, the time-dependent model maintains the linear conditional mean and flexible density form assumptions for $f(y_t|\boldsymbol{x}_t)$, but it includes the possibility that both

[3] PSBP priors are also used in Chung and Dunson (2009).

the features are time varying. Indeed, assuming that $\{w_{jt}\}_{j=1}^{\infty}$ and $\{\boldsymbol{\beta}_j\}_{j=1}^{\infty}$ are known, the conditional mean at time t is $\mathbb{E}(y_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\bar{\boldsymbol{\beta}}_t$, where $\bar{\boldsymbol{\beta}}_{\boldsymbol{t}} = (\bar{\beta}_{1t}, \ldots, \bar{\beta}_{pt})$ and $\bar{\beta}_{lt} = \sum_j w_{jt}\beta_{jl}$. By allowing the mixing weights w_{jt} to change over time, the conditional mean of y_t may respond to $\boldsymbol{x}_t$ differently at different points in time; or $\boldsymbol{x}_t$ may be more powerful in explaining the mean of y_t at one point than at other points.

By the same token, the variance of y_t is also allowed to change. Define $\bar{\sigma}_t^2 = \sum_j w_{jt}\sigma_j^2$. $\bar{\sigma}_t^2$ is the average variance of the regression errors in the mixture at time t, and it changes over time as the weights w_{jt} change. In this way, heteroscedasticity is easily accounted for and changes in the scale of $f(y_t|\boldsymbol{x}_t)$ can be detected. Time variation in the higher moments can be produced in similar fashions, thus allowing the whole distribution structure to vary over time.

The property of time-varying for $f(y_t|\boldsymbol{x}_t)$ is desirable, but we would also like to avoid the problem of overfitting, so that $f(y_t|\boldsymbol{x}_t)$ does not over react to individual observations. That is, we prefer that $f(y_t|\boldsymbol{x}_t)$ has some sort of time dependence or persistence, and in the case where a fundamental change does occur over time, the model has the capability to capture it in a smooth fashion. The smoothness of the PSBP model (in terms of total variation distance between distributions) is guaranteed by Rodríguez and Dunson (2011, Theorem 5) under some general conditions. The specific time-dependent structure is however generated by the specific dynamics of the set of latent series $\{\boldsymbol{\alpha}_j\}_{j=1}^{\infty}$. Different choices for the dynamics of $\{\boldsymbol{\alpha}_j\}_{j=1}^{\infty}$ impose different implications on the inter-temporal relationship among the set of PSBPs, and ultimately induce various kinds of dependence and levels of persistence among $f(y_t|\boldsymbol{x}_t)$ across t. In equation (9), the common AR1 specification for all the $\{\boldsymbol{\alpha}_j\}_{j=1}^{\infty}$ series offers a parsimonious way to introduce dependence. It also fully takes advantage of the computational tractability offered by the probit structure in posterior sampling, see Section 9.[4] The AR1 coefficient γ_1 and the error variance σ_α^2 control the degree of persistence for $f(y_t|\boldsymbol{x}_t)$ over time. Intuitively, if γ_1 is close to 1 and σ_α^2 is small, we would expect that for adjacent points in time, $f(y_t|\boldsymbol{x}_t)$ would have similar structure. In a special case, let $\gamma_1 = 0$, then each $\boldsymbol{\alpha}_j$ becomes a series of iid normal random variables, so $\{w_{jt}\}_{j=1}^{\infty}$ are independent sets of weights across t. As a result, $f(y_t|\boldsymbol{x}_t)$ are a priori independent. In another extreme case where $\gamma_0 = 0$, $\gamma_1 = 1$ and $\alpha_{j1} \overset{\perp}{\sim} \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$ for some μ_α and σ_α^2, there is no time variation in $\boldsymbol{\alpha}_j$ for all j and the same set of $\{w_j\}_{j=1}^{\infty}$ is produced for each t. Hence $f(y_t|\boldsymbol{x}_t)$ is the same for all t, a case that is parallel with the DPM, with the only difference in the stick-breaking structure.[5]

The time-dependent model in equation (7) through equation (12) is very general in the sense that it allows all the predictors to have a time-varying effect on the dependent variable. Indeed, in equation (7), each predictor x_{lt} may have a different value for its coefficient β_{jl} across all regression components in the mixture. Depending on the specific questions of interest, some simplifications can be made to the regression mixture structure in the model, resulting in different nested model versions. For example if we are confident that some of the covariates $\boldsymbol{x}_t = (\boldsymbol{x}_{1t}', \boldsymbol{x}_{2t}')'$, say $\boldsymbol{x}_{1t}$ have a constant effect on the conditional mean of y_t over time, but suspect that $\boldsymbol{x}_{2t}$ may have a time-varying effect. We can replace equation (7)

[4] In one example, Rodríguez and Dunson (2011) used a simple random walk dynamic for $\{\boldsymbol{\alpha}_j\}_{j=1}^{\infty}$.

[5] This process belongs to the PSBP prior for single distribution. Details are provided in Rodríguez and Dunson (2011).

by

$$f(y_t|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt}\mathrm{N}(y_t|\boldsymbol{x}'_{1t}\tilde{\boldsymbol{\beta}} + \boldsymbol{x}'_{2t}\boldsymbol{\beta}_j, \sigma_j^2)$$

and add a prior for $\tilde{\boldsymbol{\beta}}$:

$$\tilde{\boldsymbol{\beta}} \sim \mathcal{G}_3$$

This nested model forces the coefficient on $\boldsymbol{x}_{1t}$ to be constant, but allows the effect of $\boldsymbol{x}_{2t}$ and the shape of $f(y_t|\boldsymbol{x}_t)$ to change. Or if we suspect the linear conditional mean structure in a whole is stable over time, and only the shape of the conditional distribution is subject to change, we can make adjustments accordingly and arrive at another nested model version:

$$\begin{aligned} f(y_t|\boldsymbol{x}_t) &= \sum_{j=1}^{\infty} w_{jt}\mathrm{N}(y_t|\boldsymbol{x}'_t\tilde{\boldsymbol{\beta}}, \sigma_j^2) \\ \sigma_j^2 &\overset{iid}{\sim} G_0(\lambda) \\ \tilde{\boldsymbol{\beta}} &\sim \mathcal{G}_3 \end{aligned}$$

In this case, the location-scale mixture structure is reduced to a scale mixture.

4 Posterior sampling and model comparison

4.1 Posterior sampling

To carry out Bayesian inference, we first apply a finite truncation on the infinite number of components in equation (7):

$$f(y_t|\boldsymbol{x}_t) = \sum_{j=1}^{L} w_{jt}\mathrm{N}(y_t|\boldsymbol{x}'_t\boldsymbol{\beta}_j, \sigma_j^2)$$

where L is a large number (for example 40). This truncation is justified in Rodríguez and Dunson (2011), where they showed that the posterior distribution based on a L-finite PSBP converges in distribution to the one based on the infinite PSBP as L goes to infinity. Accordingly, since the stick-breaking weights have to add up to one for all t, we have:

$$\begin{aligned} w_{jt} &= \Phi(\alpha_{jt})\prod_{k<j}(1-\Phi(\alpha_{kt})), \quad \alpha_{jt} \overset{\perp}{\sim} \mathrm{N}(.|\gamma_0 + \gamma_1\alpha_{jt-1}, \sigma_\alpha^2), \quad j = 1, \ldots, L-1 \\ w_{Lt} &= 1 - \sum_{j=1}^{L-1} w_{jt} \end{aligned}$$

As a second step to facilitate posterior sampling, we introduce a sequence of group indicators $S = \{s_t\}_{t=1}^T$ and rewrite the time-dependent model in an equivalent form as below:

$$\begin{aligned} y_t|S, \Theta, \boldsymbol{x}_t &\sim \mathrm{N}(y_t|\boldsymbol{x}_t'\boldsymbol{\beta}_{s_t}, \sigma_{s_t}^2) & (13)\\ s_t &\sim \sum_{j=1}^{L} w_{jt}\delta_j(.) & (14)\\ w_{jt} &= \Phi(\alpha_{jt})\prod_{k<j}(1-\Phi(\alpha_{kt})) & (15)\\ \alpha_{jt} &\overset{\perp}{\sim} \mathrm{N}(.|\gamma_0 + \gamma_1\alpha_{jt-1}, \sigma_\alpha^2) & (16)\\ (\boldsymbol{\beta}_j, \sigma_j^2) &\overset{iid}{\sim} G_0(\lambda_1) & (17)\\ \lambda_1 &\sim \mathcal{G}_1 & (18)\\ (\gamma_0, \gamma_1, \sigma_\alpha^2) &\sim \mathcal{G}_2 & (19) \end{aligned}$$

where $\Theta = \{(\boldsymbol{\beta}_j, \sigma_j^2)\}_{j=1}^L$. So, s_t indicates which of the L normal regression components y_t is drawn from. i.e. $s_t = j \Rightarrow y_t \sim \mathrm{N}(y_t|\boldsymbol{x}_t'\boldsymbol{\beta}_j, \sigma_j^2)$. Rewriting model this way with S included, data augmentation can be employed in the sampling, making the procedure straightforward. The complete joint likelihood function of $\{y_t\}_{t=1}^T$ and S is $f(\{y_t\}_{t=1}^T, S|\{\boldsymbol{x}_t\}_{t=1}^T, \Theta, \boldsymbol{\alpha}) = \prod_{t=1}^T w_{s_t t}\mathrm{N}(y_t|\boldsymbol{x}_t'\boldsymbol{\beta}_{s_t}, \sigma_{s_t}^2)$.

The distribution G_0 in equation (17) assumes a normal-gamma distribution commonly used in the Bayesian literature as a standard conjugate prior for linear models:

$$\sigma_j^{-2} \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{d}{2}), \quad \boldsymbol{\beta}_j|\sigma_j^2 \sim \mathrm{N}(m, \sigma_j^2 H^{-1}) \tag{20}$$

ν and d are positive scalars, m is a $q \times 1$ vector, H is a $q \times q$ positive definte matrix. The mean and the variance of the Gamma distribution is ν/d and $2\nu/d^2$, respectively. In equation (18), the hyperprior $\mathcal{G}_1$ on the hierachical parameters $\lambda_1 = \{m, H, d\}$ is the following:[6]

$$\begin{aligned} H &\sim \mathrm{Wishart}_q(a_0, H_0), \quad m|H \sim \mathrm{N}(m_0, \tau_0 H^{-1}) & (21)\\ d &\sim \mathrm{Gamma}(\frac{c_0}{2}, \frac{d_0}{2}) & (22) \end{aligned}$$

$\mathrm{Wishart}_q(a_0, H_0)$ denotes a Wishart distribtution with a_0 degrees of freedom and scale matrix H_0. a_0 is a positive scalar satisfying $a_0 > q$ and H_0 is a $q \times q$ positive definite matrix. m_0 is a q-dimensional vector. τ_0, c_0, d_0 are each a positive scalar. In equation (19), the prior distribution $\mathcal{G}_2$ on $\{\gamma_0, \gamma_1, \sigma_\alpha^2\}$ admits another normal-gamma distrition:

$$\sigma_\alpha^{-2} \sim \mathrm{Gamma}(\frac{\nu_\alpha}{2}, \frac{d_\alpha}{2}), \quad (\gamma_0, \gamma_1)|\sigma_\alpha^2 \sim \mathrm{N}(m_\alpha, \sigma_\alpha^2 H_\alpha^{-1}) \tag{23}$$

ν_α and d_α are positive scalars, m_α is a 2×1 vector, H_α is a 2×2 positive definte matrix.

We carry out posterior sampling based on Markov chain Monte Carlo (MCMC) techniques. A modified blocked Gibbs sampler (Ishwavan and James, 2001) is proposed. We

[6] In the paper we fix ν at 5, but it can also be estimated by placing a prior (e.g. Exponential distribution) and using a Metropolis-Hastings step to sample from its distribution, see for example Song (2011).

divide the parameter set Υ into 5 blocks, S, Θ, $\boldsymbol{\alpha}$, λ_1 and λ_2, where $\lambda_2 = \{\gamma_0, \gamma_1, \sigma_\alpha^2\}$. We iteratively sample through the conditional posterior distribution of each block

$$\begin{array}{l}
S|\Theta, \boldsymbol{\alpha}, \{\boldsymbol{x}_t\}_{t=1}^T, \{y_t\}_{t=1}^T \\
\Theta|S, \lambda_1, \{\boldsymbol{x}_t\}_{t=1}^T, \{y_t\}_{t=1}^T \\
\boldsymbol{\alpha}|S, \lambda_2 \\
\lambda_1|\Theta \\
\lambda_2|\boldsymbol{\alpha}
\end{array}$$

See Appendix for details of the sampling algorithm. Taking a draw from all of the conditional distributions constitutes one sweep of the sampler. After dropping an initial set of draws as burn-in we collect M draws $\{S^{(i)}, \Theta^{(i)}, \boldsymbol{\alpha}^{(i)}, \lambda_1^{(i)}, \lambda_2^{(i)}\}_{i=1}^M$ for posterior inference. Simulation consistent estimates of posterior moments can be obtained as sample averages of the draws. For instance, the posterior mean of λ_2 can be estimated as $M^{-1} \sum_{i=1}^M \lambda_2^{(i)}$.

4.2 Model comparison

Model comparison is carried out using predictive likelihoods. In a Bayesian framework, the predictive likelihood is a key input into model comparison through predictive Bayes factors (Geweke 2005). Given a whole sample $(y_1, \ldots, y_T)$, the cumulative log-predictive likelihood of the subsample $(y_{T_0}, \ldots, y_T)$ for a candidate model $\mathcal{A}$ is calculated as

$$\hat{p}^{\mathcal{A}} = \sum_{t=T_0-1}^{T-1} \log(p(y_{t+1}|I_t, \mathcal{A})), \tag{24}$$

where $T_0 < T$, and $I_t = \{y_1, \ldots, y_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t+1}\}$[7]. $\hat{p}^{\mathcal{A}}$ measures the forecast performance based on the out-of-sample data points: $y_{T_0}, ..., y_T$. Better models, in terms of more accurate predictive densities, will have larger (24). The one-period predictive likelihood $p(y_{t+1}|I_t, \mathcal{A})$ is approximated as:

$$p(y_{t+1}|I_t, \mathcal{A}) \approx \frac{1}{M} \sum_{i=1}^M p(y_{t+1}|\Upsilon^{(i)}, I_t, \mathcal{A}) \tag{25}$$

where Υ denotes the set of model parameters and $\{\Upsilon^{(i)}\}_{i=1}^M$ are the MCMC draws from the posterior distribution $p(\Upsilon|I_t, \mathcal{A})$ of the model parameters given the information I_t. In particular, for the time-dependent model,

$$p(y_{t+1}|I_t) \approx \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^L w_{jt+1}^{(i)} \mathrm{N}(y_{t+1}|\boldsymbol{x}_{t+1}' \boldsymbol{\beta}_j^{(i)}, \sigma_j^{2(i)}). \tag{26}$$

After calculating the one-period period likelihood of y_{t+1}, the in-sample data is updated by adding the observation y_{t+1}. we then *re-estimate* the model to obtain a new set of draws

[7] x_{t+1} is included in I_t because it is pre-determined at time $t+1$

from the posterior to compute (25). In other words the model is recursively estimated for $t = T_0 - 1, \ldots, T - 1$.

Given a model $\mathcal{A}$ with log-predictive likelihood $\hat{p}^{\mathcal{A}}$, and model $\mathcal{B}$ with log-predictive likelihood $\hat{p}^{\mathcal{B}}$, based on the common data $\{y_{T_0}, \ldots, y_T\}$, the predictive Bayes factor in favor of model $\mathcal{A}$ versus model $\mathcal{B}$ is $BF_{\mathcal{AB}} = \frac{\hat{p}^{\mathcal{A}}}{\hat{p}^{\mathcal{B}}}$. The Bayes factor is a relative ranking of the ability of the models to account for the data. A value greater than 1 means that model $\mathcal{A}$ is better able to account for the data compared to model $\mathcal{B}$. Kass and Raftery (1995) suggest interpreting the evidence for $\mathcal{A}$ as: not worth more than a bare mention if $0 \leq \log(BF_{\mathcal{AB}}) < 1$; positive if $1 \leq \log(BF_{\mathcal{AB}}) < 3$; strong if $3 \leq \log(BF_{\mathcal{AB}}) < 5$; and very strong if $\log(BF_{\mathcal{AB}}) \geq 5$.

4.3 Predictive mean

For the time dependent model, the predictive mean of y_{t+1} conditional on the information set $I_t = \{y_1, \ldots, y_t, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t+1}\}$ is

$$\begin{aligned} \hat{y}_{t+1} &= \mathbb{E}(y_{t+1}|I_t) \\ &= \mathbb{E}(\mathbb{E}(y_{t+1}|\boldsymbol{x}_{t+1}, \Upsilon)|I_t) \\ &= \mathbb{E}(\boldsymbol{x}'_{t+1}\bar{\boldsymbol{\beta}}_{t+1}|I_t) \end{aligned} \tag{27}$$

Equation (27) is approximated as

$$\hat{y}_{t+1} \approx \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{x}'_{t+1}\bar{\boldsymbol{\beta}}^{(i)}_{t+1}$$

where $\bar{\boldsymbol{\beta}}^{(i)}_{t+1}$ is generated from the i^{th} MCMC draw of the posterior distribution $p(\Upsilon|I_t)$ of the model parameters, given the information I_t. Over the whole out-of-sample period $T_0, \ldots, T$, the mean squared predictive error based on predictive mean is calculated as

$$MSE = \frac{1}{T - T_0 + 1}\sum_{t=T_0-1}^{T-1} (y_{t+1} - \hat{y}_{t+1})^2$$

A smaller number in MSE indicates that the model does a better job in predicting the mean value of a future response.

5 Simulation study

In this section, we conduct a simulation study to illustrate the time-dependent models and evaluate their performance in comparison with some benchmark models. A univariate time series of predictor x_t is generated as $x_t \overset{iid}{\sim} \text{Uniform}(0, 5)$ for $t = 1, \ldots, T = 500$. The response

y_t is generated as:

$$\begin{aligned} y_t &= b_0 + b_{1t} x_t + \epsilon_t \\ b_{1t} &= \begin{cases} 0.3 & \text{if } t < \tau_1 \\ 0.3 + 0.004 \times (t - \tau_1) & \text{if } \tau_1 \le t \le \tau_2 \\ 0.3 + 0.004 \times (\tau_2 - \tau_1) & \text{if } t > \tau_2 \end{cases} \\ \epsilon_t &\sim \mathrm{N}(.|0, \sigma_t^2) \\ \sigma_t^2 &= 0.1 + 0.03 \times \epsilon_{t-1}^2 + 0.95 \times \sigma_{t-1}^2 \end{aligned}$$

where $b_0 = 0.2$, $\tau_1 = 200$, $\tau_2 = 300$, and $\sigma_1^2 = 5$ is set as the initial condition. The intercept coefficient b_0 is constant over time, but the slope coefficient b_{1t} has a time pattern: it is constant at 0.3 for the first 200 observations, after that, it increases linearly with t, until after $t = 300$, where it remains constant at 0.7 hereafter. The error term ϵ_t follows a stationary GARCH specification, with a long run variance equal to 5.

We want to uncover the predicting power of x_t on y_t over time. Two time-dependent models (M1,M2) are considered here, each corresponds to a specific regression mixture structure:

$$\begin{aligned} M1: \quad f(y_t|x_t) &= \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_{0j} + \beta_{1j} x_t, \sigma_j^2) \\ M2: \quad f(y_t|x_t) &= \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_0 + \beta_{1j} x_t, \sigma_j^2) \end{aligned}$$

The implications of the different regression mixture structures of the two models should be clear. $M1$ allows that both the intercept and the slope coefficients to change over time; $M2$ is a nested version of $M1$ and only allows the slope coefficient to change while assuming the intercept to be constant. We estimate the model using the entire sample. The posterior summary statistics for some of the parameters are in Table 1. The posterior means of γ_1 are 0.991 and 0.977, respectively, indicating strong time persistence of the conditional distribution of $f(y_t|x_t)$ in both models. To directly investigate the effects of x_t on y_t through the entire sample, we calculate the smoothed $\bar{\beta}_{1t} = \sum_j w_{jt} \beta_{1j}$ at each point t in time for both models, and plot them against t, together with b_{1t} from DGP, see Figure 1. We see that for both models, $\bar{\beta}_{1t}$ track the true b_{1t} reasonably well through the whole sample, demonstrating their ability to detect change in the predicting power of x_t over time. Similarly, from Figure 2, we can see that both models do a good job picking up the time variation in the variance of the error term. Meanwhile, note that for M2, the posterior mean of β_0 is 0.231, with a 95% density interval of $(-0.139, 0.601)$, which is a good estimate of the true value of b_0.

Now we compare the out-of sample performance of the time-dependent models with the benchmark models based on the predictive likelihoods, as illustrated in Section 4.2. We add another two time-dependent models ($M3$, $M4$) to the comparison, their regression mixture structures are:

$$M3: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_{0j}, \sigma_j^2)$$

$$M4: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_0 + \beta_1 x_t, \sigma_j^2)$$

Obviously, $M3$ allows the intercept to change, but assuming that x_t has no predicting power on y_t at all. $M4$ assume constant β_0 and β_1 over time. Compared to the DGP, $M2$ is "closest" to the truth, while $M3$ and $M4$ are both "mis-specified" in some way. For comparison, a benchmark model is also estimated for the data, namely, normal linear regression (Lin).

$$Lin: \quad f(y_t|x_t) = \mathrm{N}(y_t|\beta_0 + \beta_1 x_t, \sigma^2) \tag{28}$$

The model assumes a constant effect of x_t on y_t, and also assumes that the error term is homoskedastic.

The last 80 observations are preserved as out-of-sample data and are used to calculate the the predictive likelihoods. All models are recursively estimated over the whole out-of-sample period with a burn-in of 10000 iterations after which M=10000 draws are collected to compute the predictive likelihoods. The results are reported in Table 2. Both $M1$ and $M2$ are favorable than the benchmark model Lin. Indeed, the log predictive Bayes factor in favor of $M1$ versus Lin is $-186.466 - (-188.108) = 1.642 > 1$, which is positive evidence that $M1$ accounts for the data better than Lin. For $M2$, the evidence of better accountability over Lin is strong, with a difference of $-184.607 - (-188.108) = 3.501 > 3$ in the log predictive likelihoods. This is not surprising, since both $M1$ and $M2$ accommodate time variation in the slope coefficient and heteroscedasticity, while Lin does not. Between $M1$ and $M2$, however, there is a difference of $-184.607 - (-186.466) = 1.859 > 1$ in the log predictive likelihoods, in favor of $M2$.

Next, we investigate the out-of-sample performance among the models by comparing the MSE of predicting the mean of y_t, see Table 2. The ranking of all candidate models in MSE is perfectly consistent with the ranking in log predictive likelihoods. Model with higher log predictive likelihood also produce lower MSE.

The above results render some interesting implications. First, the general time-dependent model $M1$, which is the most flexible version, does a good job picking up the time variation from both the regression coefficient and the variance, hence can help forecast compared to a simple linear model. Second, if we know a priori that some predictor has a constant effect, we should take advantage and incorporate the information into our model (e.g., $M2$), which would improve forecast. Third, if on the other hand, we make the wrong restrictions (as in $M3$ and $M4$), the model can do worse in terms of forecasting. Last, $M4$ and Lin both assume constant effects from the intercept and the predictor, the difference is that Lin also assumes homoscedasticity while $M4$ does not. The fact that $M4$ does worse than Lin suggests that when there is both time variation in the regression coefficient and variance as in the simulated data, simply accounting for the latter while ignoring the former might not necessarily improve forecast compared to the simple linear model that ignores both.

6 Data

We are interested in the relationship between exchange rates and commodity prices among each of the following countries: Canada, Australia and New Zealand. We use monthly data on exchange rates and commodity prices. The exchange rates data are obtained from Statistics Canada, the Reserved Bank of Australia, and the Reserve Bank of New Zealand, and the end-of-period U.S. dollar rates are used. For commodity prices, we use the country-specific commodity price indices issued by the Bank of Canada, the Reserved Bank of Australia, ANZ National Bank Limited, respectively. Each commodity price index is a weighted average of the prices for a range of commodities the corresponding country exports. All three indices are in terms of US dollars. For all countries, the sample periods end at May 2011, but have different starting points: The Canadian data start at January 1972, the Australian data start at January 1984 and the New Zealand data start at January 1986.[8]

7 Results

We study the predictive power of exchange rates (ER_t) on commodity prices (CP_{t+1}) one period ahead and how it changes through time. Let $y_{t+1} = 100\times(\log(CP_{t+1})-\log(CP_t))$, $x_t = 100\times(\log(ER_t)-\log(ER_{t-1}))$, and $\boldsymbol{x}_t = (1, x_t, y_t)'$. The dependent variable is the percentage change of commodity prices for time $t+1$, the independent variable is the percentage change of exchange rates for time t. The lagged dependent variable is included as a common practice in the literature. We model $f(y_{t+1}|\boldsymbol{x}_t)$ using the time-dependent models in Section 3. This specification has the same linear conditional mean structure as the regression model in Chen, Rogoff, and Rossi (2010). In all cases, we let the truncation $L = 40$. Posterior sampling is carried out using the block sampler in Section 4.1. The first 10000 draws of the MCMC chain are discarded as burn-in and the next 10000 draws are used for inference.

7.1 In-Sample Results

We apply a general model $P1$ on the data for in-sample analysis. $P1$ has the the following regression mixture structure

$$P1: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt}\mathrm{N}(\beta_{0j} + \beta_{1j}x_t + \beta_{2j}y_t, \sigma_j^2) \tag{29}$$

Model $P1$ assumes that the exchange rates x_t as well as the intercept and the lagged dependent variable y_t all have a time-varying effect on future commodity prices y_{t+1}. The priors

[8]Canada, Australia and New Zealand started their floating exchange rate regimes respectively in 1970, 1983 and 1985.

of $P1$ are the following:

$$\begin{aligned}
H &\sim \text{Wishart}_3(4, 0.2Id) \\
m|H &\sim \text{N}(\mathbf{0}, H^{-1}) \\
d &\sim \text{Gamma}(2.5, 0.5) \\
\sigma_\alpha^{-2} &\sim \text{Gamma}(5, 0.5), \quad (\gamma_0, \gamma_1)|\sigma_\alpha^2 \sim \text{N}\left((0, 0.95), \sigma_\alpha^2 \begin{pmatrix} 1 & 0 \\ 0 & 0.32 \end{pmatrix}\right)
\end{aligned}$$

We estimate the model based on the entire sample for all three countries respectively. The posterior summary statistics for some of the parameters are in Table 3. The posterior mean of γ_1 is 0.982, 0.972 and 0.996 for the 3 countries, respectively, indicating strong time persistence of the conditional distribution of $f(y_t|x_t)$ in all cases.

One way to look at the effects of the predictors on the dependent variable is through the effects on the mean. For each country, we calculate the posterior mean and 95% density interval of $\bar{\beta}_{lt}$ for $l = 0, 1, 2$ at each point t in time, where $\bar{\beta}_{lt} = \sum_{j=1}^{L} w_{jt}\beta_{lj}$ represent the linear effect at time t on the mean of the dependent variable from the intercept, the exchange rates and the lagged dependent variable, repectively. We also calculate the posterior mean of $\bar{\sigma}_t^2 = \sum_{j=1}^{L} w_{jt}\sigma_j$, which measures the average variance of the regression errors at time t. We plot the set of quantities against t for each country (Figure 3 for Canada, Figure 4 for Australia, and Figure 5 for New Zealand) and examine their time patterns over the whole sample period.

For Canada, there seems to be some weak evidence of time variation in $\bar{\beta}_{0t}$, $\bar{\beta}_{1t}$ and $\bar{\beta}_{2t}$, suggesting all predictors may have time-varying linear mean effects on commodity prices. For instance, the posterior mean of $\bar{\beta}_{1t}$ starts around 0 and ends around 0.4. But there is substantial uncertainty. The significance of $\bar{\beta}_{1t}$ seems increasing over time, while that of $\bar{\beta}_{2t}$ is decreasing. The overall significance of the coefficients is weak. For the majority of the sample period 0 is within the 95% density boundaries for all three coefficients. For $\bar{\sigma}_t^2$, the evidence of time variation is very strong. The posterior mean of $\bar{\sigma}_t^2$ fluctuates around 4 from the start of the sample in 1972 until 1996, when it starts rising abruptly, peaking at 19 by 2002 and maintaining the high level thereafter. The pattern strongly suggests a structural change in the variance of the dependent variable.

In the Australian case, the pattern is different. The linear effect of exchange rates on the mean of commodity prices represented by $\bar{\beta}_{1t}$ is in general constant through the whole sample. It is clearly identified and significant for the mid one third of the sample from 1992 to 2003. Its 95% density interval is relatively narrow with an almost constant width of about 0.2. However for the first 6 years and the last 8 years of the sample period, $\bar{\beta}_{1t}$ is less stable and significant, especially towards the end of the sample. Compared to exchange rates, the intercept and the lagged commodity prices both have a much more volatile linear effect on the mean of the dependent variable. Both effects are weak but stable from 1992 to 2002. For the rest of the sample, especially after 2003, both effects experience drastic changes in significance as well as in magnitudes. $\bar{\sigma}_t^2$ also show ups and downs over time, not even though as dramatic as in the Canadian case.

For New Zealand, $\bar{\beta}_{1t}$ is insignificant through out the whole sample, the posterior mean lies within 0.02 of 0 for more than half of the time. It suggests that the mean effect of exchange rates on commodity prices is very weak for the whole time considered. The intercept also has a weak mean effect on commodity prices, since the 95% density interval of $\bar{\beta}_{0t}$ always contains 0. Mean while, the significance and the magnitude of $\bar{\beta}_{2t}$ increase dramatically from the beginning to the end, which indicates that the linear mean effect on commodity prices from the lagged dependent variable has greatly strengthened over the course.

In summary, the above in-sample analyses for the three countries show that there is no strong evidence for a time-varying linear effect of exchange rates on future commodity prices. The effect, if there is any, is weak in most cases. On the other hand, for Australia and New Zealand, the intercept in the regression and the lagged dependent variable show strong signs of parameter change over time, suggesting both have a time-varying relationship with commodity prices. The results also suggest that for all countries considered, a significant source of time variation in the conditional distribution of commodity prices comes from the variance.

7.2 Out-of-Sample Evaluation

In this section, we evaluate models based on their out-of-sample performance on both density forecasts and point forecasts. Besides the general model $P1$, five nested versions of the general model are also considered, each with a different regression mixture structure. A complete list of all six time-dependent models ($P1, P2, P3, P4, P5, P6$) characterized by their regression mixture structures are presented below

$$P1: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{1j} x_t + \beta_{2j} y_t, \sigma_j^2) \tag{30}$$

$$P2: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{2j} y_t, \sigma_j^2) \tag{31}$$

$$P3: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_2 y_t + \beta_{1j} x_t, \sigma_j^2) \tag{32}$$

$$P4: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_1 x_t + \beta_2 y_t, \sigma_j^2) \tag{33}$$

$$P5: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_2 y_t, \sigma_j^2) \tag{34}$$

$$P6: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{2j} y_t + \beta_1 x_t, \sigma_j^2) \tag{35}$$

$P2$ assumes that exchange rates (x_t) has no predicting power for commodity prices, while allows the coefficient on the lagged dependent variable and the intercept coefficient to change over time. $P3$ assumes that the coefficients on the intercept and the lagged dependent variable to be constant (β_0 and β_2) and only allow exchange rates to have a time-varying effect on commodity prices. $P4$ assumes that the coefficient on exchange rates is constant, and so are the other 2 coefficients. $P5$ also assumes that exchange rates has no predicting power, and it assumes that the other 2 coefficients are constant. $P6$ assumes that exchange rates does a constant effect on future commodity prices, but allows the coefficient on the y_t and the intercept coefficient to change over time. Note that all the above time-dependent models can accommodate time variation in the variance. Finally, a Normal linear regression model (Lin) is included as a bench mark:

$$Lin: \quad f(y_{t+1}|\boldsymbol{x}_t) = \mathrm{N}(\beta_0 + \beta_1 x_t + \beta_2 y_t, \sigma^2) \tag{36}$$

It assumes that all predictors have a constant effect on the dependent variable and the error term is homoskedastic.

By comparing all models' forecasting abilities, we hope to identify the most important factors in terms of forecasting. We try to answer the question of whether it is important to account for time variation when doing forecast, and if yes, where does the variation come from.

The last 80 observations are reserved for out-of-sample data, representing the time period from October 2004 to May 2011. All models are recursively estimated over the whole out-of-sample period with a burn-in of 10000 iterations after which M=10000 draws are collected to compute the predictive quantities (predictive likelihoods and predictive mean).

7.2.1 Density Forecast

In this sub-section, we compare all models on density forecasts through predictive likelihoods, as illustrated in Section 4.2. Forecasting (conditional) density is directly relevant here since the main subject of interest of this paper is the conditional distribution of commodity prices on exchange rates. Predictive likelihood is commonly used in the literature for density forecast evaluation(reference). We compare the abilities to forecast the conditional density of $f(y_{t+1}|\boldsymbol{x}_t)$ one period ahead among different models over the same out-of-sample data. This is done by comparing the cumulative log-predictive likelihoods calculated from each model. Models that produce more accurate density forecasts for the common out-of-sample data will have larger cumulative log-predictive likelihoods. The Results are reported in Table 4.

For Canadian data, model $P4$ produces the highest log predictive likelihoods, suggesting that all the predictors including the exchange rates are important for density forecast of future commodity prices, but their effects are stable over time. Compared to the normal linear model Lin which also assumes constant coefficients, the log predictive Bayes factor in favor of $P4$ is $-259.492 - (-307.503) \approx 48$. This is overwhelming evidence against the homoscedasticity assumption, a fact also suggested in the previous in-sample analysis (Figure 3). Among the group of time-dependent models, in general, models that include x_t in the conditional mean of y_{t+1} beat their counterparts that leave out x_t. For instance, the difference between $P4$ and $P5$ is $-259.492 - (-262.545) = 3.053$ in favor of $P4$, and the difference between $P3$ and $P5$ is $-259.530 - (-262.545) = 3.015$ in favor of $P3$, both provide strong evidence that when assuming that the intercept coefficient and the coefficient on y_t are constant, exchange rates have a predictive effect on commodity prices. However, no evidence suggests any time variation in this predictive effect, since letting it change over time as in $P3$ does not improve forecasting.

For the Australian case, there are some different results. The highest log predictive likelihoods come from $P2$ and $P6$, both of which allow $\bar{\beta}_{0t}$ and $\bar{\beta}_{2t}$ to change over time. With the exception of $P1$, $P2$ and $P6$ dominate all the other models with a remarkable margin of at least 5. These facts show strong evidence that both the intercept and the lagged dependent variable have a time-varying effect on commodity prices. At the same time, the difference between $P2$ and $P6$ is negligible, and they both improve on $P1$ by more than 2, suggesting that exchange rates have no predictive power on commodity prices. The simple linear model Lin still has the lowest log predictive likelihoods, and again is most likely due to its homoskedasticity assumption.

For New Zealand, $P6$ has the highest log predictive likelihood, suggesting that the intercept in the regression and the lagged dependent variable have a time-varying effect on y_{t+1}. Furthermore, $P6$ beats $P5$ by more than 2, which is positive evidence that exchange rates can help forecast the density of the dependent variable. In addition, $P6$ beats $P1$ by more than 6, strongly suggesting that the predictive effect from exchange rates is constant over time. Lin still has the lowest log predictive likelihoods, but the difference between Lin and other models are not as overwhelming as it is for Canadian data and Australian data. This means the heteroscedastic effects for New Zealand data is not as eminent as in the other countries, which can also be seen in the in-sample plots in the previous sub-section.

To summarize, there is no single model that is best for all countries in terms of density forecasts. But for Australia and New Zealand, the data seems to suggest that the intercept and the lagged dependent variable have a time-varying effect on commodity prices, while the effect from the exchange rates is either weak or constant. On the contrary, the Canadian data suggest all predictors matter and their effects are all constant.

7.2.2 Predictive mean evaluation

In this sub-section we calculate each model's MSE of the predictive mean of y_{t+1} over the out-of-sample period for all three countries, respectively. The results are the second row of each sub-panel of Table 4.

For Canadian data, $P6$ has the lowest MSE with a value of 32.271. In particular, it indicates exchange rates have a constant but nonzero effect on the mean of the dependent variable. Lin produces the third smallest value in MSE. It suggests that despite its poor density forecasts, a simple normal linear regression has relative advantage in predicting the mean of future commodity prices, where accounting for heteroscedasticity in the error term is not as important. Among the time-dependent models, leaving out exchange rates in the mean regression increases the MSE. Indeed, $P2$ has a greater MSE than $P1$, and $P5$ has a greater MSE than $P4$.

For Australian data, $P6$ again produces the lowest $MSE = 10.375$ followed by $P2$, implying time variation in the intercept coefficient and the coefficient on the lagged dependent variable. It also implies that the exchange rates help predict the mean as well. Mean while, the fact that $P6$ has a much lower MSE than $P1$ suggests the predictive effect from exchange rates is constant rather than time-varying.

Turning to New Zealand data. $P6$ still has the lowest MSE, which again indicates that the intercept and the lagged dependent variable have a time-varying effect on the mean of future commodity prices, while the effect from exchange rates is constant. It is worth noting that *Lin* produces the highest *MSE* among all models. This is interesting since for both Canadian data and Australian data, *Lin* is rather competing in predicting the mean of y_{t+1}, despite its difficulty in density forecast.

The above results show that for point forecasts, $P6$ is the best for all countries, which emphasizes time-varying effects from the intercept and the lagged dependent variable, but a constant effect from the exchange rates.

8 Conclusion

This paper investigates the predictive relationship from exchange rates to commodity prices for 3 commodity-exporting countries: Canada, Australia and New Zealand. A full Bayesian nonparametric procedure which allows for time-varying parameters is proposed. We find that in contrast to the results in Chen et al (2010), nominal exchange rates don't have a consistent predictive power for commodity prices in all countries considered in this paper. We find that exchange rates do have predictive power in some cases, but their impact tends to be constant over time. On the other hand, the intercept in the regression and the lagged dependent variable show signs of time-varying parameter change in most cases, which is important in predicting both the mean and the density of the commodity prices one period ahead. The results also suggest that for all countries considered, a significant source of time variation in the conditional distribution of commodity prices comes from the variance.

9 Appendix: Blocked Gibbs sampler

Let $X = \{\boldsymbol{x}_t\}_{t=1}^T$, $Y = \{y_t\}_{t=1}^T$

9.1 Sampling $S|\Theta, \boldsymbol{\alpha}, X, Y$

The joint likelihood function of Y and S is

$$f(Y, S|X, \Theta, \boldsymbol{\alpha}) = \prod_{t=1}^{T} w_{s_t t} \mathrm{N}(y_t|\boldsymbol{x}_t' \boldsymbol{\beta}_{s_t}, \sigma_{s_t}^2). \tag{37}$$

For each t, the conditional posterior distribution of s_t satisfying following:

$$\begin{aligned} p(s_t = j|X, Y, \Theta, \boldsymbol{\alpha}) &= p(s_t = j|\boldsymbol{x}_t, y_t, \Theta, \{\alpha_{lt}\}_{l=1}^{L-1}) \\ &\propto w_{jt} \mathrm{N}(y_t|\boldsymbol{x}_t' \boldsymbol{\beta}_j, \sigma_j^2) \end{aligned}$$

$j = 1, \ldots, L$.

9.2 Sampling $\Theta|S, \lambda_1, X, Y$

$\Theta = \{\boldsymbol{\beta}_j, \sigma_j^2\}_{j=1}^L$. We use the standard conjugacy results for linear models to sample the posterior distribution of $(\boldsymbol{\beta}_j, \sigma_j^2)$. The prior is

$$(\boldsymbol{\beta}_j, \sigma_j^2) \sim \mathrm{N} - \mathrm{G}(m, H, \nu, d)$$

$\mathrm{N-G}(m, H, \nu, d)$ denotes a normal-gamma distribution with parameters $\lambda_1 = m, H, \nu, d$, see equation (20). By conjugacy, the posterior is

$$(\boldsymbol{\beta}_j, \sigma_j^2)|S, \lambda_1, X, Y \sim \mathrm{N-G}(\overline{m}_j, \overline{H}_j, \overline{\nu}_j, \overline{d}_j)$$

where

$$\begin{aligned}
\overline{H}_j &= H + X_j'X_j \\
\overline{m}_j &= (Hm + X_j'Y_j)\overline{H}_j^{-1} \\
\overline{\nu}_j &= \nu + n_j \\
\overline{d}_j &= d + Y_j'Y_j + m'Hm - \overline{m}_j'\overline{H}_j\overline{m}_j
\end{aligned}$$

$Y_j = \{y_t : s_t = j\}$, $X_j = \{\boldsymbol{x}_t : s_t = j\}$. n_j is the number of elements in Y_j.

9.3 Sampling $\lambda_1|\Theta$

$\lambda_1 = \{m, H, \nu, d\}$. The conditional posterior of (m, H) is

$$(m, H)|\{\boldsymbol{\beta}_j, \sigma_j^2\}_{j=1}^L \sim \mathrm{N-W}(m_1, \tau_1, a_1, H_1) \tag{38}$$

$\mathrm{N-W}(m_1, \tau_1, a_1, H_1)$ denotes a normal-Wishart distribution with parameters m_1, τ_1, a_1, H_1, where

$$\begin{aligned}
m_1 &= \frac{\tau_0^{-1}m_0 + \sum_{j=1}^L \sigma_j^{-2}\boldsymbol{\beta}_j}{\tau_0^{-1} + \sum_{j=1}^L \sigma_j^{-2}} \\
\tau_1 &= \frac{1}{\tau_0^{-1} + \sum_{j=1}^L \sigma_j^{-2}} \\
a_1 &= a_0 + L \\
H_1 &= \left(H_0^{-1} + \sum_{j=1}^L \sigma_j^{-2}\boldsymbol{\beta}_j\boldsymbol{\beta}_j' + \tau_0^{-1}m_0m_0' - \tau_1^{-1}m_1m_1'\right)^{-1}
\end{aligned}$$

The conditional posterior distribution of d is

$$d|\nu, \{\sigma_j^2\} \sim \mathrm{Gamma}(c_1/2, d_1/2)$$

where $c_1 = c_0 + L$ and $d_1 = d_0 + \sum_{j=1}^L \sigma_j^{-2}$.
As mentioned earlier, we fix ν at 5, but it can also be estimated by placing a prior (e.g. Exponential distribution) and using a Metropolis-Hastings step to sample from its distribution, see for example Song (2011).

9.4 Sampling $\lambda_2|\boldsymbol{\alpha}$

$\lambda_2 = \{\gamma_0, \gamma_1, \sigma_\alpha^2\}$. The conditional posterior is

$$(\gamma_0, \gamma_1, \sigma_\alpha^2)|\boldsymbol{\alpha} \sim \mathrm{N-Gamma}(\overline{m}_\alpha, \overline{H}_\alpha, \overline{\nu}_\alpha, \overline{d}_\alpha)$$

with

$$
\begin{aligned}
\overline{H}_\alpha &= H_\alpha + X'_\alpha X_\alpha \\
\overline{m}_\alpha &= (H_\alpha m_\alpha + X'_\alpha Y_\alpha)\overline{H}_\alpha^{-1} \\
\overline{\nu}_\alpha &= \nu_\alpha + (L-1)(T-1) \\
\overline{d}_\alpha &= d_\alpha + Y'_\alpha Y_\alpha + m'_\alpha H_\alpha m_\alpha - \overline{m}'_\alpha \overline{H}_\alpha \overline{m}_\alpha
\end{aligned}
$$

$Y_\alpha = (\alpha_{1,2}, \alpha_{1,3}, \ldots, \alpha_{1,T}, \alpha_{2,2}, \alpha_{2,3}, \ldots, \alpha_{2,T}, \ldots, \alpha_{L-1,2}, \alpha_{L-1,3}, \ldots, \alpha_{L-1,T})'$. That is, Y_α is a $(L-1)(T-1) \times 1$ vector.
$X_\alpha = ((1, \alpha_{1,1})', \ldots, (1, \alpha_{1,T-1})', (1, \alpha_{2,1})', \ldots, (1, \alpha_{2,T-1})', \ldots, (1, \alpha_{L-1,1})', \ldots, (1, \alpha_{L-1,T-1})')'$. So X_α is a $(L-1)(T-1) \times 2$ matrix, the first column are all 1.

9.5 Sampling $\boldsymbol{\alpha}|S, \lambda_2$

To facilitate the sampling of $\{\boldsymbol{\alpha}_j\}_{j=1}^{L-1}$, we introduce a set of indicator variables $\{Z_{jt}\}_{j=1}^{L-1}{}_{t=1}^{T}$. Given S, let $Z_{jt} = 0$ if $j < s_t$, and $Z_{jt} = 1$ if $j = s_t$, and if $j > s_t$, then Z_{jt} is undefined. So the information on S is mapped into the information on $\{Z_{jt}\}$. Indeed, any particular sequence of of values for S correspondes to a unique set of values for $\{Z_{jt}\}$. We then once again apply data augmentation and define latent variables $\{Z^*_{jt}\}_{j=1}^{L-1}{}_{t=1}^{T}$ such that:

$$
\begin{aligned}
Z_{jt} &= \mathbb{I}(Z^*_{jt} > 0), \\
Z^*_{jt} &\sim \mathrm{N}(.|\alpha_{jt}, 1)
\end{aligned}
$$

Thus, we set ourselves up for another block of Gibbs sampler: conditional on $\{\boldsymbol{\alpha}_j\}$ and $\{Z_{jt}\}$ (or S), sample $\{Z^*_{jt}\}$ from its full conditional distribution

$$
Z^*_{jt}|\alpha_{jt}, s_t \sim \begin{cases} \mathrm{N}(.|\alpha_{jt}, 1)\mathbb{I}_{\mathbb{R}^-} & \text{if } j < s_t \ (Z_{jt} = 0) \\ \mathrm{N}(.|\alpha_{jt}, 1)\mathbb{I}_{\mathbb{R}^+} & \text{if } j = s_t \ (Z_{jt} = 1) \end{cases}
$$

This makes use of the standard Bayesian inference for the Probit model. For derivation, see for example Koop (2003). If $j > s_t$ (Z_{jt} undefined), record Z^*_{jt} as missing value.

Second, conditional on $\{Z^*_{jt}\}$ and $(\gamma_0, \gamma_1, \sigma^2_\alpha)$, the $L-1$ latent processes $\{\boldsymbol{\alpha}_j\}_{j=1}^{L-1}$ are then sampled one process at a time using an efficient block sampler. The block sampler used here is constructed in the same spirit as the sampler proposed by Jensen and Maheu (2010) to sample a latent volatility process in a stochastic volatility model.

References

Chen, Y., K. Rogoff, and B. Rossi (2010): "Can Exchange Rates Forecast Commodity Prices?," *The Quarterly Journal of Economics*, 125(3), 1145.

Chung, Y., and D. Dunson (2009): "Nonparametric Bayes conditional distribution modeling with variable selection," *Journal of the American Statistical Association*, 104(488), 1646–1660.

Engel, C., and K. D. West (2005): "Exchange Rates and Fundamentals," *Journal of Political Economy*, 113(3), 485–517.

Escobar, M. D., and M. West (1995): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90(430), 577–588.

Ferguson, T. (1973): "A Bayesian analysis of some nonparametric problems," *The annals of statistics*, 1(2), 209–230.

Geweke, J. (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Ghosal, S., J. Ghosh, and R. Ramamoorthi (1999): "Posterior consistency of dirichlet mixtures in density estimation," *The Annals of Statistics*, 27(1), 143–158.

Jensen, M., and J. Maheu (2010): "Bayesian semiparametric stochastic volatility modeling," *Journal of Econometrics*, 157(2), 306–316.

Kass, R. E., and A. E. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90(420), 773–795.

Koop, G. (2003): *Bayesian econometrics*. Wiley-Interscience.

Rodríguez, A., and D. Dunson (2011): "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6(1), 145–178.

Rossi, B. (2005): "Optimal tests for nested model selection with underlying parameter instability," *Econometric Theory*, 21(5), 962.

Sethuraman, J. (1994): "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Song, Y. (2011): "Modelling Regime Switching and Structural Breaks with an Infinite Dimension Markov Switching Model," *Working Papers*.

West, M., P. Müller, and M. D. Escobar (1994): "Hierarchical priors and mixture models, with application in regression and density estimation," *Aspects of uncertainty: A Tribute to DV Lindley*, pp. 363–386.

Table 1: Posterior summary of common parameters from $M1$ and $M2$ in the simulation study

	$M1$			$M2$		
Par	Mean	Std	0.95 DI	Mean	Std	0.95 DI
γ_0	0.001	0.004	$(-0.005, 0.008)$	0.014	0.013	$(-0.009, 0.039)$
γ_1	0.991	0.003	$(0.985, 0.996)$	0.977	0.006	$(0.964, 0.981)$
σ_α^2	0.109	0.036	$(0.056, 0.191)$	0.141	0.052	$(0.073, 0.274)$

Table 2: Model comparison based on simulated data

	M1	M2	M3	M4	LR
LPL	−186.466	**− 184.607**	−193.043	−190.445	−188.108
MSE	5.909	**5.865**	6.929	6.637	6.330

$$M1: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_{0j} + \beta_{1j}x_t, \sigma_j^2)$$

$$M2: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_0 + \beta_{1j}x_t, \sigma_j^2)$$

$$M3: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_{0j}, \sigma_j^2)$$

$$M4: \quad f(y_t|x_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(y_t|\beta_0 + \beta_1 x_t, \sigma_j^2)$$

Table 3: Posterior summary of seleted parameters of $P1$

	CAN			AUS			NZL		
Par	Mean	Std	0.95 DI	Mean	Std	0.95 DI	Mean	Std	0.95 DI
γ_0	−0.015	0.008	$(-0.034, -0.002)$	−0.018	0.010	$(-0.039, 0.002)$	−0.004	0.003	$(-0.011, 0.002)$
γ_1	0.982	0.004	$(0.972, 0.989)$	0.972	0.014	$(0.938, 0.989)$	0.996	0.001	$(0.992, 0.998)$
σ_α^2	0.112	0.046	$(0.057, 0.216)$	0.099	0.033	$(0.049, 0.181)$	0.085	0.022	$(0.051, 0.151)$

Table 4: Out-of-sample forecast results in terms of Log predictive likelihood(LPL)(top) and mean square error(MSE)(bottom)

		$P1$	$P2$	$P3$	$P4$	$P5$	$P6$	Lin
CAN	LPL	−262.840	−264.366	−259.530	**−259.492**	−262.545	−264.587	−307.503
	MSE	32.582	34.268	33.316	33.199	35.067	**32.271**	33.048
AUS	LPL	−217.646	**−215.241**	−224.792	−220.956	−220.759	−215.460	−248.017
	MSE	11.296	10.796	11.565	11.280	11.397	**10.375**	11.161
NZL	LPL	−181.485	−177.196	−179.811	−180.543	−178.354	**−175.201**	−189.689
	MSE	5.546	5.720	6.108	6.346	6.405	**5.486**	6.431

$$P1: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{1j}x_t + \beta_{2j}y_t, \sigma_j^2)$$

$$P2: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{2j}y_t, \sigma_j^2)$$

$$P3: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_2 y_t + \beta_{1j}x_t, \sigma_j^2)$$

$$P4: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_1 x_t + \beta_2 y_t, \sigma_j^2)$$

$$P5: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_0 + \beta_2 y_t, \sigma_j^2)$$

$$P6: \quad f(y_{t+1}|\boldsymbol{x}_t) = \sum_{j=1}^{\infty} w_{jt} \mathrm{N}(\beta_{0j} + \beta_{2j}y_t + \beta_1 x_t, \sigma_j^2)$$

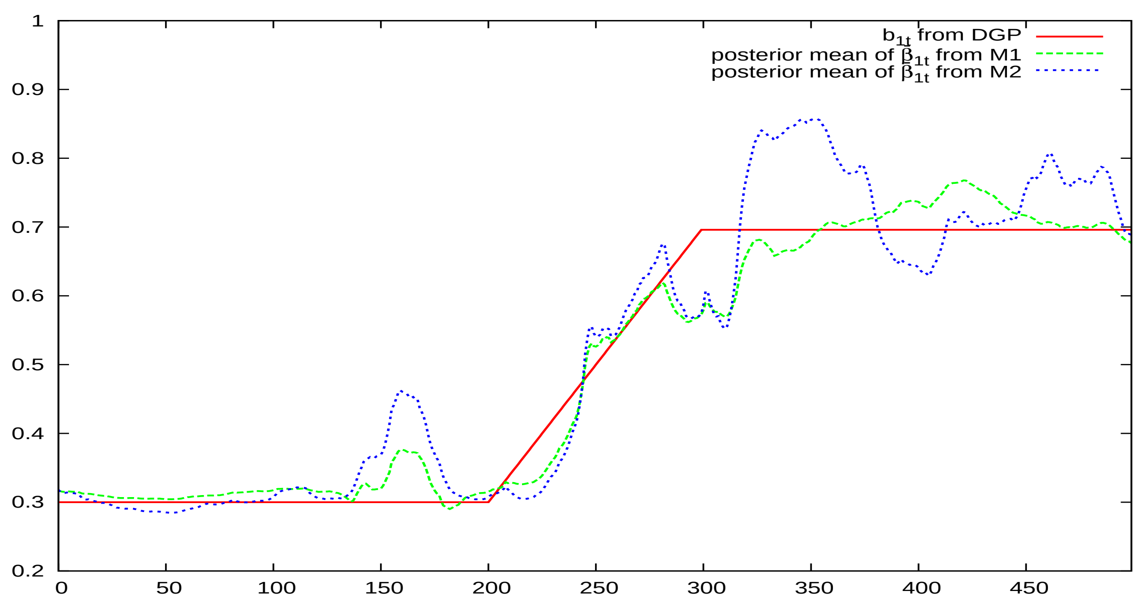

Figure 1: Posterior mean of $\bar{\beta}_{1t}$ from $M1$ and $M2$, respectively, and b_{1t} from DGP.

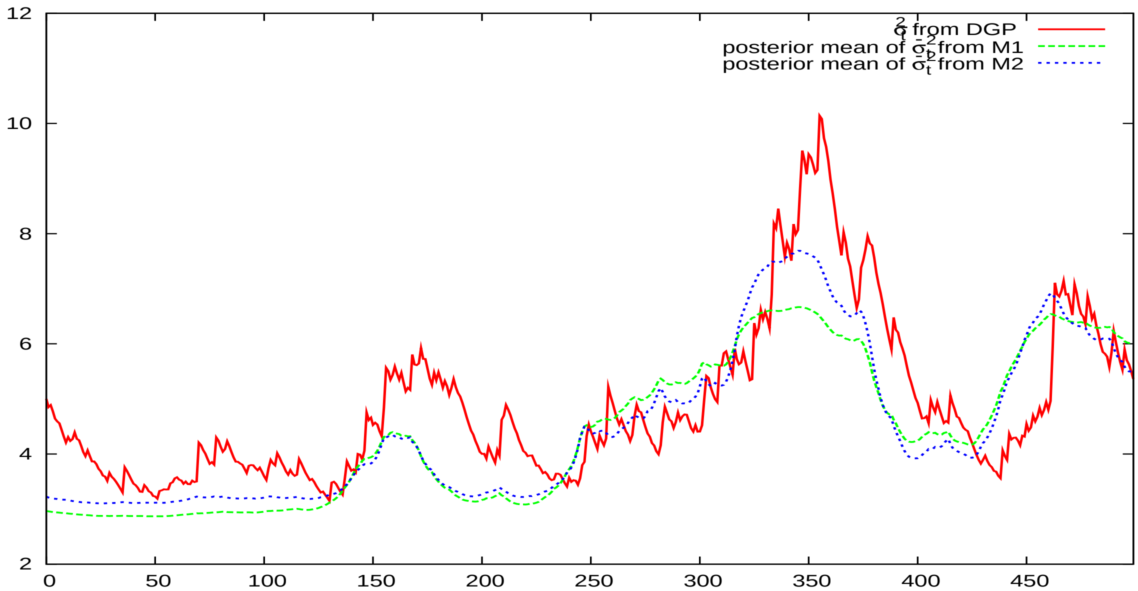

Figure 2: Posterior mean of $\bar{\sigma}_t^2$ from $M1$ and $M2$, respectively, and σ_t^2 from DGP.

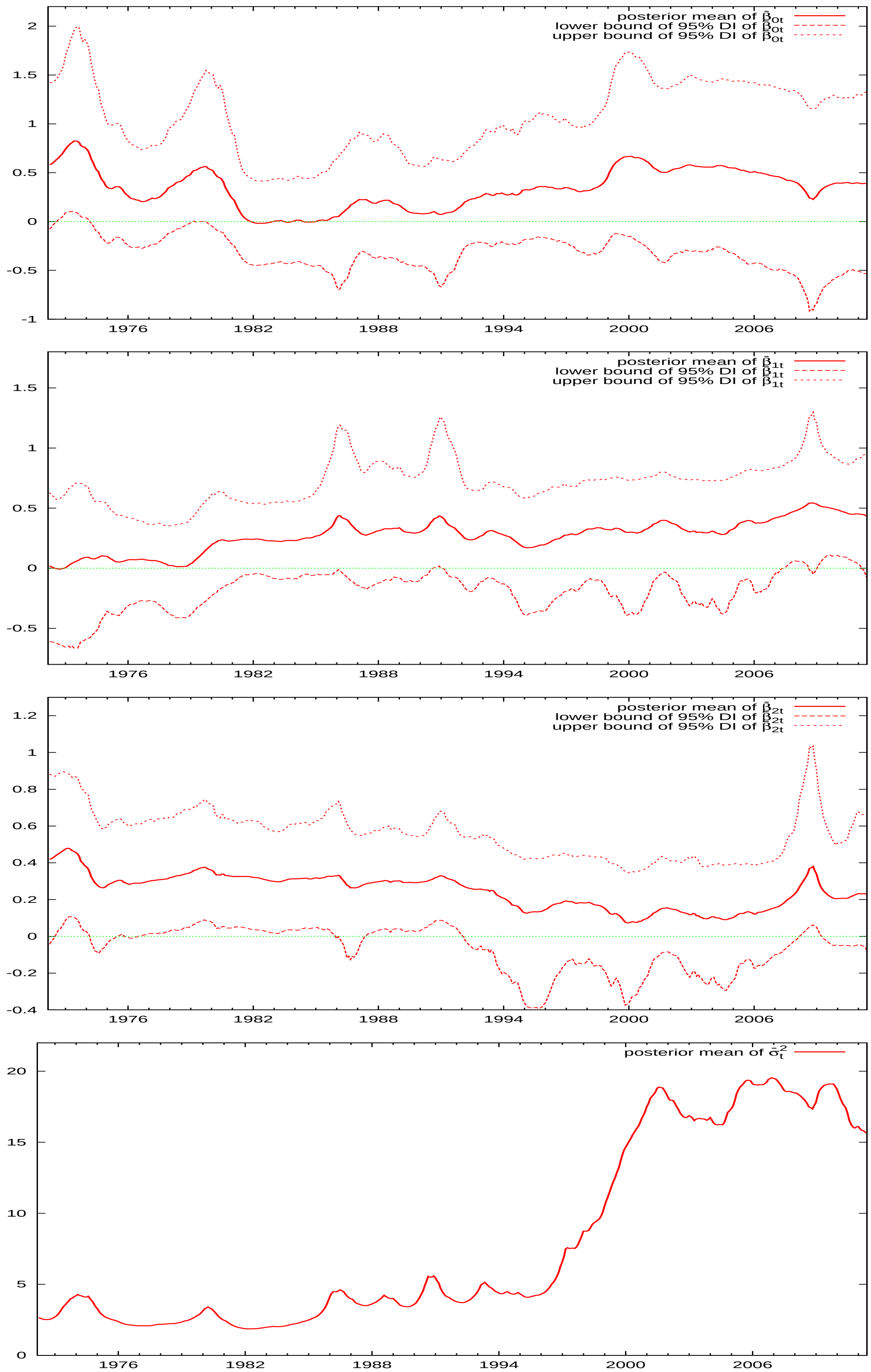

Figure 3: Posterior mean and 95% density interval of $\bar{\beta}_{lt}$ for $l = 0, 1, 2$ and posterior mean for $\bar{\sigma}_t^2$ in model $P1$ over the full sample for Canadian data

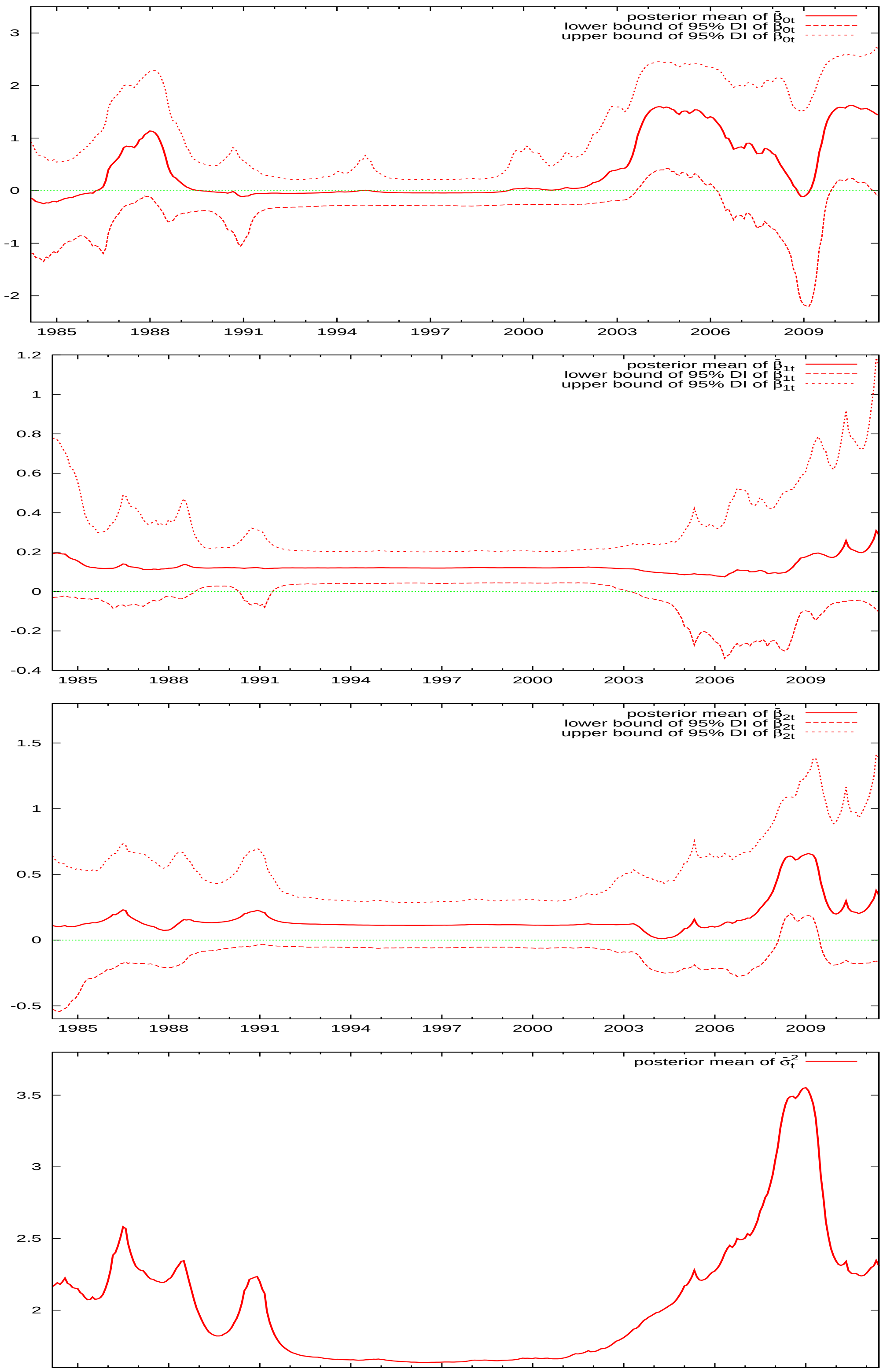

Figure 4: Posterior mean and 95% density interval of $\bar{\beta}_{lt}$ for $l = 0, 1, 2$ and posterior mean for $\bar{\sigma}_t^2$ in model $P1$ over the full sample for Australian data

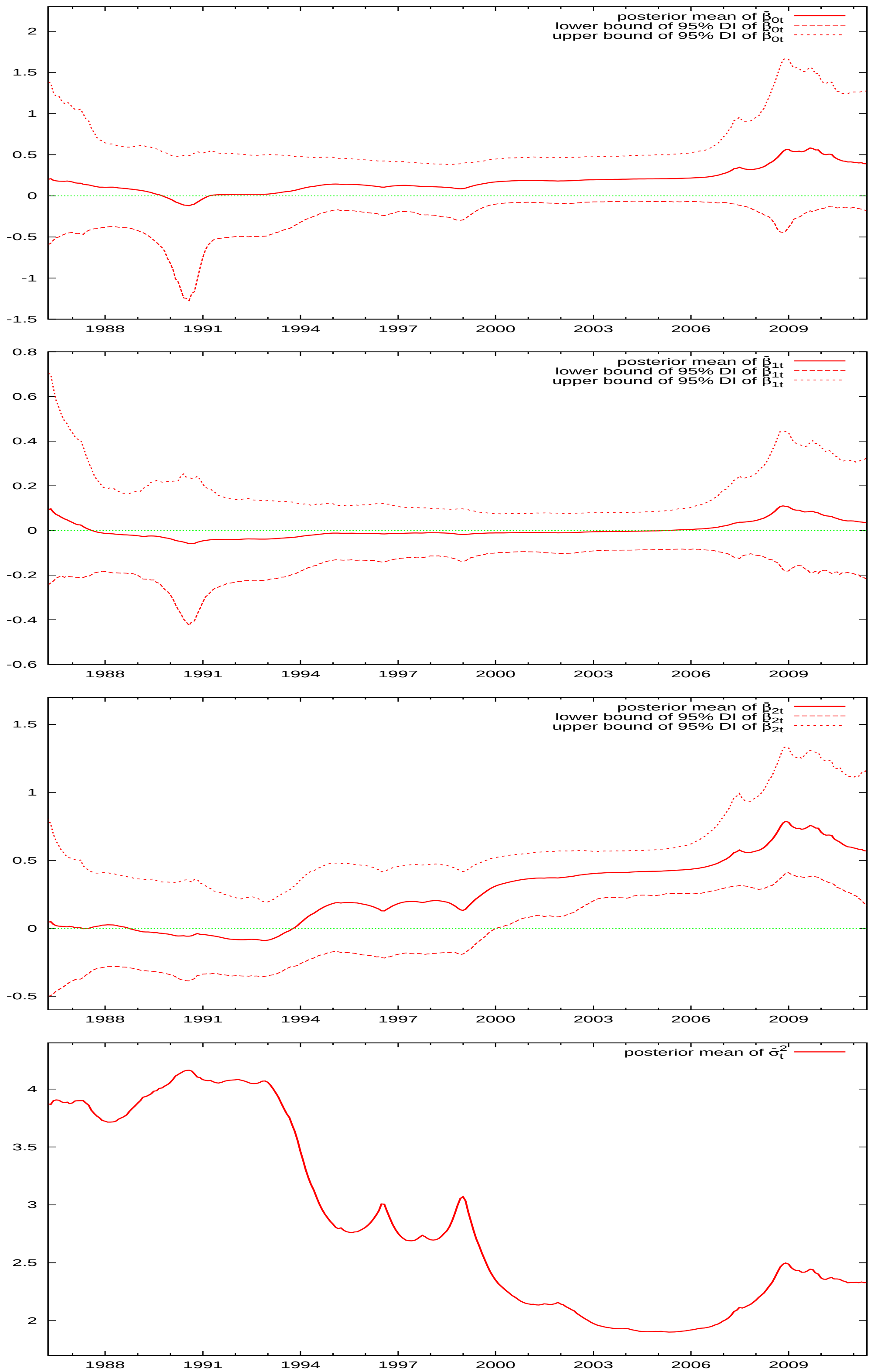

Figure 5: Posterior mean and 95% density interval of $\bar{\beta}_{lt}$ for $l = 0, 1, 2$ and posterior mean for $\bar{\sigma}_t^2$ in model $P1$ over the full sample for New Zealand data