EUROPEAN CENTRAL BANK
EUROSYSTEM

WORKING PAPER SERIES
NO 969 / NOVEMBER 2008

COMPARING AND EVALUATING BAYESIAN PREDICTIVE DISTRIBUTIONS OF ASSET RETURNS

by John Geweke
and Gianni Amisano

EUROPEAN CENTRAL BANK

EUROSYSTEM

WORKING PAPER SERIES

NO 969 / NOVEMBER 2008

COMPARING AND EVALUATING BAYESIAN PREDICTIVE DISTRIBUTIONS OF ASSET RETURNS[1]

by John Geweke[2] and Gianni Amisano[3]

In 2008 all ECB publications feature a motif taken from the €10 banknote.

This paper can be downloaded without charge from http://www.ecb.europa.eu or from the Social Science Research Network electronic library at http://ssrn.com/abstract_id=1299538.

1 Acknowledgment: We wish to acknowledge constructive comments from a co-editor and three referees that have substantially contributed to this paper. Responsibility for errors and omissions remains with the authors. The first author acknowledges partial financial support from NSF grant SBR-0720547. The views expressed in this paper do not necessarily reflect the views of the European Central Bank and the European System of Central Banks.

2 Departments of Statistics and Economics, University of Iowa, 430 N. Clinton St., Iowa City IA 52242-2020, USA; e-mail: john-geweke@uiowa.edu

3 European Central Bank; Kaiserstrasse 29, D-60311 Frankfurt am Main, Germany; e-mail: gianni.amisano@ecb.europa.eu; Università degli Studi di Brescia, Piazza del Mercato, 15 - 25121 Brescia, IT; e-mail: amisano@eco.unibs.it

Address
Kaiserstrasse 29
60311 Frankfurt am Main, Germany

Postal address
Postfach 16 03 19
60066 Frankfurt am Main, Germany

Telephone
+49 69 1344 0

Website
http://www.ecb.europa.eu

Fax
+49 69 1344 6000

ISSN 1561-0810 (print)
ISSN 1725-2806 (online)

CONTENTS

Abstract

Bayesian inference in a time series model provides exact, out-of-sample predictive distributions that fully and coherently incorporate parameter uncertainty. This study compares and evaluates Bayesian predictive distributions from alternative models, using as an illustration five alternative models of asset returns applied to daily S&P 500 returns from 1976 through 2005. The comparison exercise uses predictive likelihoods and is inherently Bayesian. The evaluation exercise uses the probability integral transform and is inherently frequentist. The illustration shows that the two approaches can be complementary, each identifying strengths and weaknesses in models that are not evident using the other.

Keywords: forecasting; GARCH; inverse probability transform; Markov mixture; predictive likelihood; S&P 500 returns; stochastic volatility.

JEL Classifications: C11, C53.

Non-technical summary

Probability distributions for magnitudes that are unknown at a time a decision must be made, but will become known afterward, are required for the formal solutions of most decision problems in economics. Increasing awareness of this context, combined with advances in modeling and computing, is leading to a sustained emphasis on these distributions in econometric research.

For important decisions there are typically competing models and methods that produce predictive distributions. The question of how these predictive distributions should be compared and evaluated then becomes relevant.

This study compares and evaluates the quality of predictive distributions over multiple horizons for asset returns using five different models. We use the daily returns of the Standard and Poors 500 index over the period 1972-2005, a series that is widely employed in academic work and is also one of the most important indexes in the finance industry. The models compared are two from the ARCH family, a stochastic volatility model, the Markov normal mixture model, and an extension of the last model that we have described in detail elsewhere (Geweke and Amisano (2007)).

The basis of comparison used in this study is the predictive likelihood function, i.e. the model's probability density for the return at the relevant horizon before it is observed, evaluated at the actual value of the return after it is observed. This function reflects the logical positivism of the Bayesian approach: a model is as good as its predictions.

Each model produces a predictive distribution for each return ex ante, and therefore a predictive likelihood ex post. Comparison of these predictive likelihoods across models decomposes posterior odds one observation at a time. One of the objectives of this study is to illustrate how this decomposition provides insight into conventional Bayesian model comparison. The basis of evaluation used in this study is the probability integral transform (PIT), which is the inverse of the sequence of ex ante predictive cumulative distribution function (c.d.f.) evaluated at the sequence of actual returns ex post. If returns are in fact generated from this c.d.f. sequence then the ex ante distribution of the PIT is i.i.d. uniform. As a practical matter this condition will not be met precisely even in ideal circumstances: while observed values might come from the model under consideration, uncertainty about parameter values implies that the predictive distributions will not be exactly the same as in the data generating process. Nevertheless the PIT provides a well-recognized and useful paradigm against which any sequence of predictive distributions can be

evaluated. A second objective of this study is to illustrate how the PIT also provides insight into the deficiencies of models.

Model comparison using predictive likelihoods and model evaluation using the PIT are quite distinct methodologically. The predictive likelihood function is inherently Bayesian, while the PIT is inherently frequentist. Taken together the two methods provide insight into the strengths and weaknesses of alternative prediction models.

This study details these comparisons and evaluations using daily S&P 500 returns, and shows how the HMNM model predictive likelihood is comparable to that of the t-GARCH model (and superior to its competitors). At the same time the HMNM model is well calibrated to observed returns as indicated by the PIT.

The final objective of this study is to compare the quality of the predictive distributions of the five models for daily S&P 500 returns, and to identify deficiencies in these models that might be addressed by future research. Briefly, we find that the predictive distributions of the HMNM and t-GARCH models prove superior to those of the other three models considered.

In particular, the predictive likelihood analysis narrowly favors Bayesian t-GARCH over HMNM, but for MLE t-GARCH the predictive likelihoods are nearly identical. By contrast the PIT analysis narrowly favours HMNM predictive distributions over t-GARCH. However the latter analysis also shows that the normalized PIT for the HMNM model is not ideal. In particular, PIT's for successive one-day predictions are not independent, and the performance of HMNM in this dimension is no better than those of the other four models.

A new predictive density can always be formed as a weighted average of predictive densities from different models, the best known example being Bayesian model averaging. The analysis in our paper indicates that for most combinations of models and substantial sub-periods of the sample considered Bayesian model averaging is for all practical purposes equivalent to model selection, with one model receiving a weight very close to 1. This is often the outcome for Bayesian model averaging when the sample is large, as it is here. The notable exception arises when the models averaged include both t-GARCH and HMNM: in that case these two models can have substantial weight in Bayesian model averaging, depending on the days included in the sample. Geweke and Amisano (2008) shows that a weighted average of the HMNM and t-GARCH models compares quite favorably with both models, using predictive likelihood. That paper also shows that, in general, optimization of the predictive likelihood leads to non-trivial weights on several models,

weights that are quite different from those that result from conventional Bayesian model averaging. Ultimately, analysis of this kind provides the elements from which better models may be constructed, as illustrated in the introduction by our experience in developing the HMNM model. The predictive density evaluations presented here, as well as in Geweke and Amisano (2008), show that there is scope for substantial further improvement.

1 Introduction and motivation

Probability distributions for magnitudes that are unknown at a time a decision must be made, but will become known afterward, are required for the formal solutions of most decision problems in economics – in the private and public sectors as well as academic contexts. Increasing awareness of this context, combined with advances in modeling and computing, is leading to a sustained emphasis on these distributions in econometric research (Diebold et al. (1998); Christoffersen (1998)); Corradi and Swanson (2006) provides a survey. For important decisions there are typically competing models and methods that produce predictive distributions. The question of how these predictive distributions should be compared and evaluated then becomes relevant.

This study compares and evaluates the quality of predictive distributions over multiple horizons for asset returns using five different models. We use the daily returns of the Standard and Poors 500 index over the period 1972-2005, a series that is widely employed in academic work and is also one of the most important indexes in the finance industry. The models compared are two from the ARCH family, a stochastic volatility model, the Markov normal mixture model, and an extension of the last model that we have described in detail elsewhere (Geweke and Amisano (2007)).

The basis of comparison used in this study is the predictive likelihood function – the model's probability density for the return at the relevant horizon before it is observed, evaluated at the actual value of the return after it is observed. This function lies at the heart of the Bayesian calculus for posterior model probabilities, reflecting the logical positivism of the Bayesian approach: a model is as good as its predictions. Each model produces a predictive distribution for each return *ex ante*, and therefore a predictive likelihood *ex post*. Comparison of these predictive likelihoods across models decomposes posterior odds one observation at a time. One of the objectives of this study is to illustrate how this decomposition provides insight into conventional Bayesian model comparison. The study does this in Sections 3 and 4.

The basis of evaluation used in this study is the probability integral transform (PIT), which is the inverse of the sequence of *ex ante* predictive cumulative distribution function (c.d.f.) evaluated at the sequence of actual returns *ex post*. If returns are in fact generated from this c.d.f. sequence then the *ex ante* distribution of the PIT is i.i.d. uniform. As a practical matter this condition will not be met precisely even in ideal circumstances: while observed values might come from the model under consideration, uncertainty about parameter values implies that the predictive distributions will not be exactly the same as in the data generating process. Nevertheless the PIT provides a well-recognized and useful paradigm against which any sequence of predictive distributions can be evaluated. A second objective of this study is to illustrate how the PIT also provides insight into the deficiencies of models. The study does this in Sections 5 and 6.

Model comparison using predictive likelihoods and model evaluation using the PIT are quite distinct methodologically. The predictive likelihood function is inherently Bayesian: it is a component of the likelihood function, integrated over the posterior distribution of the unobservables (parameters and latent variables) at the time the prediction is made. The product of predictive likelihood functions over all observations in the sample is the marginal likelihood of the model over the same observations. By contrast the PIT is inherently frequentist, comparing a function of the data with the *ex ante* distribution that function would have if the data were generated by a process coinciding with the model used by the analyst. Both methods can be applied to predictive distributions arising from Bayesian inference, which we do in this study.

Taken together the two methods provide insight into the strengths and weaknesses of alternative prediction models. The choice of models used here reflects our own experience in developing the hierarchical Markov normal mixture (HMNM) model. When we began the research leading to that model, roughly the year 2000, we were aware that using predictive likelihoods t-GARCH models compared favorably with most alternative models for daily financial returns. This finding is driven strongly by days with extreme returns that are much more reasonable in t-GARCH than in competitors like the GARCH and stochastic volatility (SV) models. Yet evaluations of t-GARCH using PIT are poor because it tends to ascribe too much probability to extreme returns relative to what is observed. At the time we began our research we were also aware that normal mixture models were well calibrated relative to competing prediction models but suffered in comparisons of predictive likelihoods. This led us to extend the conventional Markov normal mixture (MNM) model using the hierarchical structure summarized in the next section and described in detail in Geweke and Amisano (2007). Going beyond that paper, this study details these comparisons and evaluations using daily S&P 500 returns, and shows how the HMNM model predictive likelihood is comparable to that of the t-GARCH model (and superior to its competitors). At the same time the HMNM model is well calibrated to observed returns as indicated by the PIT. The findings reported here suggest that there is still scope for improvement in predicting daily S&P 500 returns, a conclusion that is echoed using an alternative approach in Geweke and Amisano (2008).

The final, and overriding, objective of this study is to compare the quality of the predictive distributions of the five models for daily S&P 500 returns, and to identify deficiencies in these models that might be addressed by future research. Briefly, we find that the predictive distributions of the HMNM and t-GARCH models prove superior to those of the other three models considered. This is not an unqualified conclusion; Sections 4 and 6 provide more detail, summarized in the final section.

Amisano and Giacomini (2007) use frequentist tests based on weighted log predictive distributions to compare alternative models. Their method can be applied either to Bayesian or frequentist predictive distributions, but they do not use the PIT approach. Other studies have employed both the predictive likelihood and the PIT to compare and evaluate predictive densities, some with large samples of daily

returns like the one used in this study. Hong et al. (2004) is perhaps closest in these dimensions; see also Bauwens et al. (2004). However none of these studies incorporate parameter uncertainty in their predictive distributions. As discussed in the next section, the coherent combination of intrinsic and parameter uncertainty is the hallmark of Bayesian predictive distributions.

2 Data and Bayesian predictive distributions

This study compares and evaluates the predictive performance of five alternative predictive distributions of asset returns using daily percent log returns of the Standard & Poors (S&P) 500 index. The daily index p_t for 1972-2005 was collected from three different electronic sources: the Wharton WRDS data base;[1] Thompson/Data Stream;[2] and Yahoo Finance.[3] For days on which all three sources did not agree we consulted the periodical publication *Security Price Index Record* of Standard & Poor's Statistical Service. From the price series $\{p_t\}$ assembled in this way the daily percent log returns $y_t = 100 \log (p_t/p_{t-1})$ were constructed. The total number of returns in the sample is 8574.

Each of the five alternative predictive distributions arises from a model A for the time series of S&P 500 asset returns $\mathbf{y}_T = (y_1, \ldots, y_T)$. Each model A for a time series $\mathbf{y}_T = (y_1, \ldots, y_T)$ specifies a density $p(\mathbf{y}_T \mid \boldsymbol{\theta}_A, A)$ for the observables $\mathbf{y}_T$ conditional on a vector of unobservables $\boldsymbol{\theta}_A \in \Theta_A$ that may include latent variables as well as parameters. It also specifies a prior density $p(\boldsymbol{\theta}_A \mid A)$, and through the usual Bayesian calculus the posterior distribution of $\boldsymbol{\theta}_A$ from a sample of t observations is

$$p(\boldsymbol{\theta}_A \mid \mathbf{y}_t^o, A) \propto p(\boldsymbol{\theta}_A \mid A)\, p(\mathbf{y}_t^o \mid \boldsymbol{\theta}_A, A). \tag{1}$$

The superscript o in (1) denotes the *ex post*, observed, value of $\mathbf{y}_t$; that is, *ex post* $\mathbf{y}_t = \mathbf{y}_t^o$ is known and fixed whereas *ex ante* it is random. The posterior distribution represented by (1) is usually accessed using a posterior simulator that produces an ergodic sequence $\left\{\boldsymbol{\theta}_{A,t}^{(m)}\right\}$ $(m = 1, \ldots, M)$.

Conditional on the data $\mathbf{y}_{t-1}^o$ and the model A the predictive density for y_t is

$$p\left(y_t \mid \mathbf{y}_{t-1}^o, A\right) = \int_{\Theta_A} p\left(y_t \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A, A\right) p\left(\boldsymbol{\theta}_A \mid \mathbf{y}_{t-1}^o, A\right) d\boldsymbol{\theta}_A. \tag{2}$$

This distribution can be accessed by the simulating one value $y_t^{(m)}$ from each of the distributions represented by the density $p\left(y_t \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_{A,t-1}^{(m)}, A\right)$ $(m = 1, \ldots, M)$. This simulation is usually straightforward and less demanding than the simulation of

[1] http://wrds.wharton.upenn.edu
[2] http://www.datastream.com/default.htm
[3] http://finance.yahoo.com/

$\boldsymbol{\theta}_{A,t}^{(m)}$ from (1). The predictive density (2) integrates uncertainty about the vector of parameters $\boldsymbol{\theta}_A$ and intrinsic uncertainty about the future value y_t, both conditional on the history of returns $\mathbf{y}_{t-1}^o$ and the assumptions of the model A.

This integration is a hallmark of Bayesian predictive distributions. The use of simulation methods to produce $\left\{\boldsymbol{\theta}_{A,t}^{(m)}\right\}$ and then $\left\{y_t^{(m)}\right\}$ makes these predictive distributions applicable in real time. A key advantage of Bayesian predictive distributions is the combination of the two sources of uncertainty in a logically coherent framework. To consider two alternatives suppose, first, that one were to use the predictive density

$$p\left(y_t \mid \mathbf{y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{(t-1)}, A\right) \tag{3}$$

where the estimate $\widehat{\boldsymbol{\theta}}_A^{(t-1)}$, a function of $\mathbf{y}_{t-1}^o$, replaces the unknown $\boldsymbol{\theta}_A$. This does not account for parameter uncertainty at all. In a second alternative one could work with

$$\int_{\Theta_A} p\left(y_t \mid \mathbf{y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{(t-1)}, A\right) \widehat{p}\left(\widehat{\boldsymbol{\theta}}_A^{(t-1)} \mid A\right) d\widehat{\boldsymbol{\theta}}_A^{(t-1)} \tag{4}$$

where $\widehat{p}\left(\widehat{\boldsymbol{\theta}}_A^{(t-1)} \mid A\right)$ is an asymptotic approximation of the sampling distribution of the estimator $\widehat{\boldsymbol{\theta}}_A^{(t-1)}$. This alternative conditions on the actual history $\mathbf{y}_t^o$ in the first component of the integration, while treating the history as a random variable in the second component. The resulting distribution for y_t thus has no clear interpretation. For further discussion of these issues, see Geweke and Whiteman (2006), Section 2.4.2.

The first model A considered in this study is the generalized autoregressive conditional heteroscedasticity model with parameters $p = q = 1$ in which the distribution of the innovations is Gaussian ("GARCH"). The second model is the same as the first, except that the distribution of the innovations is Student-t ("t-GARCH"). The third model is the stochastic volatility model of Jacquier et al. (1994) ("SV").

The fourth model is a Markov normal mixture model ("MNM"), which dates at least to Lindgren (1978) and has since been applied in statistics and econometrics (Tyssedal and Tjøstheim (1988); Chib (1996); Ryden et al. (1998); Weigend and Shi (2001)). In the MNM model a latent state variable s_t takes on discrete values $s_t = 1, \ldots, m$ and obeys a first-order discrete Markov process $P\left(s_t = j \mid s_{t-1} = i\right) = p_{ij}$. Then

$$y_t \mid (s_t = j) \sim N\left(\mu_j, \sigma_j^2\right). \tag{5}$$

The model is used here with $m = 4$ components, which is the choice made by Weigend and Shi (2001) using S&P 500 return data.

The final model is a generalization of the MNM model proposed in Geweke and Amisano (2007). This generalization replaces the normal conditional normal distribution (5) with a conventional finite mixture of normal distributions. The latent variable s_t becomes the first component s_{1t} of a bivariate latent state vector

$\mathbf{s}_t = (s_{t1}, s_{t2})$; thus $P(s_{t1} = j \mid s_{t-1,1} = i) = p_{ij}$ $(j = 1, \ldots, m_1)$. For the second component $P(s_{t2} = j \mid s_{t1} = i) = r_{ij}$ $(j = 1, \ldots, m_2)$. Then

$$y_t \mid (s_{t1} = i, s_{t2} = j) \sim N\left(\mu_{ij}, \sigma_{ij}^2\right).$$

This generalization is termed the hierarchical normal mixture model ("HMNM") in Geweke and Amisano (2007). The model is used here with $m_1 = m_2 = 5$, the choice being made based on predictive likelihoods as explained in the next section. The HMNM model can also be regarded as a first-order Markov normal mixture with m^2 states and with substantial structure imposed on the Markov transition matrix. Yet a third interpretation is that of an artificial neural network with two hidden layers. Geweke and Amisano (2007) provides further detail about the model, prior distributions, and the posterior simulation algorithm.

These five models are illustrative examples. Bayesian predictive distributions arise naturally in any complete model for time series that specifies a conditional distribution of the form $p(y_t \mid \mathbf{y}_{t-1}, \boldsymbol{\theta}_A, A)$ and a prior distribution of the form $p(\boldsymbol{\theta}_A, A)$.

3 Model comparison with predictive likelihood functions

The one-step-ahead predictive likelihood, which can be evaluated only at time t or later, is the real number

$$PL_A(t) = p\left(y_t^o \mid \mathbf{y}_{t-1}^o, A\right) = \int_{\Theta_A} p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A, A\right) p\left(\boldsymbol{\theta}_A \mid \mathbf{y}_{t-1}^o, A\right) d\boldsymbol{\theta}_A. \quad (6)$$

In most time series models evaluation of $p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A, A\right)$ is straightforward, leading to the approximation of (6),

$$M^{-1} \sum_{m=1}^{M} p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_{A,t-1}^{(m)}, A\right), \quad (7)$$

using an ergodic sequence $\left\{\boldsymbol{\theta}_{A,t}^{(m)}\right\}$ from a posterior simulator.

For the data set $\mathbf{y}_T^o$ the marginal likelihood of the model A is

$$p\left(\mathbf{y}_T^o \mid A\right) = \prod_{t=1}^{T} p\left(y_t^o \mid \mathbf{y}_{t-1}^o, A\right)$$

implying the additive decomposition

$$\log p\left(\mathbf{y}_T^o \mid A\right) = \sum_{t=1}^{T} \log PL_A(t). \quad (8)$$

Given two competing models A_1 and A_2, the log Bayes factor may be decomposed

$$\log\left[\frac{p\left(\mathbf{y}_T^o \mid A_1\right)}{p\left(\mathbf{y}_T^o \mid A_2\right)}\right] = \sum_{t=1}^{T} \log\left[\frac{PL_{A_1}\left(t\right)}{PL_{A_2}\left(t\right)}\right] \tag{9}$$

where $PL_{A_1}\left(t\right)/PL_{A_2}\left(t\right)$ is the predictive Bayes factor in favor of A_1 over A_2 for observation t. Predictive Bayes factors may be approximated using the output of a posterior simulator by means of (7). These approximations are usually quite accurate; the cost is that the posterior simulator must executed for each time period t.

The decomposition (8) shows the intimate relationship between the evaluation of the predictive performance of a model by means of the predictive likelihood, on the one hand, and the evidence in favor of a model in the conventional Bayesian comparison of models by means of Bayes factors, on the other. The corresponding decomposition (9) shows how individual observations contribute to the evidence in favor of one model versus a second. See Geweke (2001) or Geweke (2005), Section 2.6.2, for further details and elaboration.

A generalization of (8) is

$$\log p\left(\mathbf{y}_T^o \mid \mathbf{y}_S^o, A\right) = \sum_{t=S+1}^{T} \log PL_A\left(t\right) \tag{10}$$

for $S < T$, and the corresponding generalization of (9) is

$$\log\left[\frac{p\left(\mathbf{y}_T^o \mid \mathbf{y}_S^o, A_1\right)}{p\left(\mathbf{y}_T^o \mid \mathbf{y}_S^o, A_2\right)}\right] = \sum_{t=S+1}^{T} \log\left[\frac{PL_{A_1}\left(t\right)}{PL_{A_2}\left(t\right)}\right]. \tag{11}$$

In (10) and (11) the cumulation of evidence begins at time $t = S + 1$ rather than at time $t = 1$. If one were to regard $p\left(\boldsymbol{\theta}_A \mid \mathbf{y}_S^o, A\right)$ as the prior distribution for $\boldsymbol{\theta}_A$ – that is, $\mathbf{y}_S^o$ were interpreted as a training sample – then (10) would have the same interpretation as (8) and (11) would have the same interpretation as (9). The analysis in the next section uses (10) and (11) with $S = 1250$ (about five years of data) and $T = 8574$, so that there are 7324 terms in the sums in these two expressions. The same sample is used for the analysis in Section 6. For small values of t $PL_A\left(t\right)$ is sensitive to the prior distribution, whereas for $t \geq 1250$ the results reported here are for all practical purposes invariant with respect to substantial changes in the prior distribution. This result is unsurprising if one interprets $\mathbf{y}_S^o$ as a training sample: the information in these 1250 observations dominates the information in the original prior distribution.

The decomposition (11) shows how individual observations contribute to the evidence in favor of one model versus a second. For example, it may show that a few observations are pivotal in evidence $\mathbf{y}_T^o$ strongly favoring one model over another. Comparison of the predictive Bayes factors $PL_{A_1}\left(t\right)/PL_{A_2}\left(t\right)$ with characteristics of

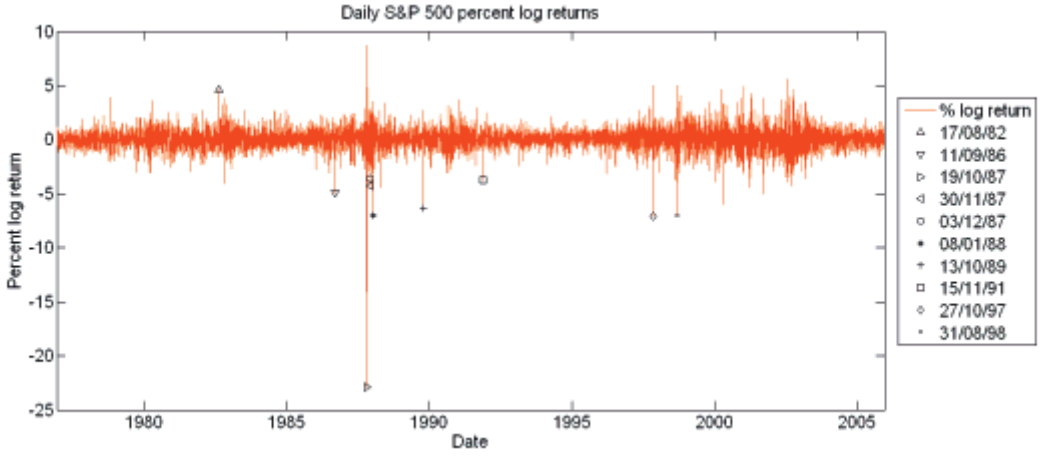

Figure 1: S%P 500 percent log return observations for which predictive likelihood was evaluated. The symbols identify nine specific observations.

the sample y_s^o for $s = t$ and observations s leading up to t can provide insight into *why* the evidence favors one model over the other. The comparison can be carried out using predictions over horizons greater than one period, but the decomposition for multiple-period horizons is exactly the same as that for single-period horizons as explained in Geweke (2001) and Geweke (2005), Section 2.6.2.

The generalization (10) of the marginal likelihood (8) amounts to the evaluation of the predictive densities $p\left(y_t \mid \mathbf{y}_{t-1}^o\right)$ $(t = S+1, \ldots, T)$ using a log scoring rule; see Gneiting and Raftery (2007), Section 7. Non-Bayesian predictive densities, like (3) and (4), may also be evaluated using a log scoring rule. In the case of (3), for example, the score

$$\sum_{t=S+1}^{T} \log p\left(y_t \mid \mathbf{y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{(t-1)}, A\right)$$

is directly comparable with (10). It is therefore possible to compare Bayesian and non-Bayesian methods directly by means of their difference in log scores

$$\sum_{t=S+1}^{T} \log \left[\frac{p\left(y_t^o \mid \mathbf{y}_{t-1}^o, A\right)}{p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{(t-1)}, A\right)} \right]. \tag{12}$$

4 Comparison of five models of S&P 500 returns

Figure 1 shows the familiar S&P 500 percent log return series for the period beginning with December 15, 1976, corresponding to $S + 1 = 1251$ and ending with December

16, 2005, corresponding to $T = 8574$. (Since the data set goes through the end of 2005, and because the analysis in Section 6 utilizes prediction horizons of up to 10 trading days, this exercise ends short of the last trading day of 2005.) Symbols in this figure identify particular dates for reference in the analysis of the predictive likelihood functions and Bayes factors that follows.

Corresponding to (11), the cumulative log predictive Bayes factor through period r, in favor of model A_1 over model A_2, is

$$\log\left[\frac{p\left(\mathbf{y}_r^o \mid \mathbf{y}_S^o, A_1\right)}{p\left(\mathbf{y}_r^o \mid \mathbf{y}_S^o, A_2\right)}\right] = \sum_{t=S+1}^{r} \log\left[\frac{PL_{A_1}(t)}{PL_{A_2}(t)}\right]. \tag{13}$$

Figure 2 shows these cumulative log predictive Bayes factors for $r = S+1, \ldots, T$. For each prediction model posterior inference was carried out by Markov chain Monte Carlo in each of 7324 samples, applying (7) to approximate $p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A, A\right)$. In each panel the comparison model A_2 is GARCH, and the other model is the one indicated. All of these results are out-of-sample: that is, $PL_A(t)$ reflects inference for the parameter vector $\boldsymbol{\theta}_A$ using the sample consisting of observations $1, \ldots, t-1$.

The right endpoint of the plotted points in each panel of Figure 2 provides (13) with $r = T$. For SV versus GARCH the value is 144.36, for t-GARCH 208.44, for MNM 151.65, and for HMNM 199.31. The evidence strongly favors the t-GARCH and HMNM models, with SV, EGARCH and GARCH rounding out the rankings. More than one-third the log predictive likelihood in favor of the other four models over GARCH is due to returns on just two days: the record log return of -22.9% on October 19, 1987, and the log return of -6.3% on October 13, 1989. The returns of -3.7% on November 15, 1991, -7.1% on October 27, 1997, and -7.0% on August 31, 1998 also lead to predictive likelihoods for those days that strongly favor the other four models over GARCH.

Figure 3 provides further comparison of the predictive performance of the t-GARCH and HMNM models as measured by predictive likelihoods. The sequence of cumulative log predictive Bayes factors, panel (a), is not dominated by any single date. Until May 23, 1984, predictive Bayes factors on average favor t-GARCH. From then until November 27, 1987, they favor HMNM on average. From July 20, 1993, through the end of 2005 predictive Bayes factors again favor t-GARCH on average. Log predictive Bayes factors for all ten dates marked by symbols in Figure 1 can be read from panels (c) and (d) of Figure 3.

Panel (b) shows all the log predictive likelihoods for the two models. Combinations above the 45^o line favor the HMNM model and those below it favor t-GARCH. The symbols specifically designate all combinations for which the t-GARCH log predictive likelihood was less than -8 or the log predictive Bayes factor in favor of HMNM was less than -1.5. (That is how the dates indicated in Figure 1 were selected.) The record return of October 19, 1987, has by far the lowest log predictive likelihood in the t-GARCH model, whereas October 13, 1989, has the lowest predictive likelihood in the HMNM model. Panel (d) of Figure 3 shows that there is no simple relationship

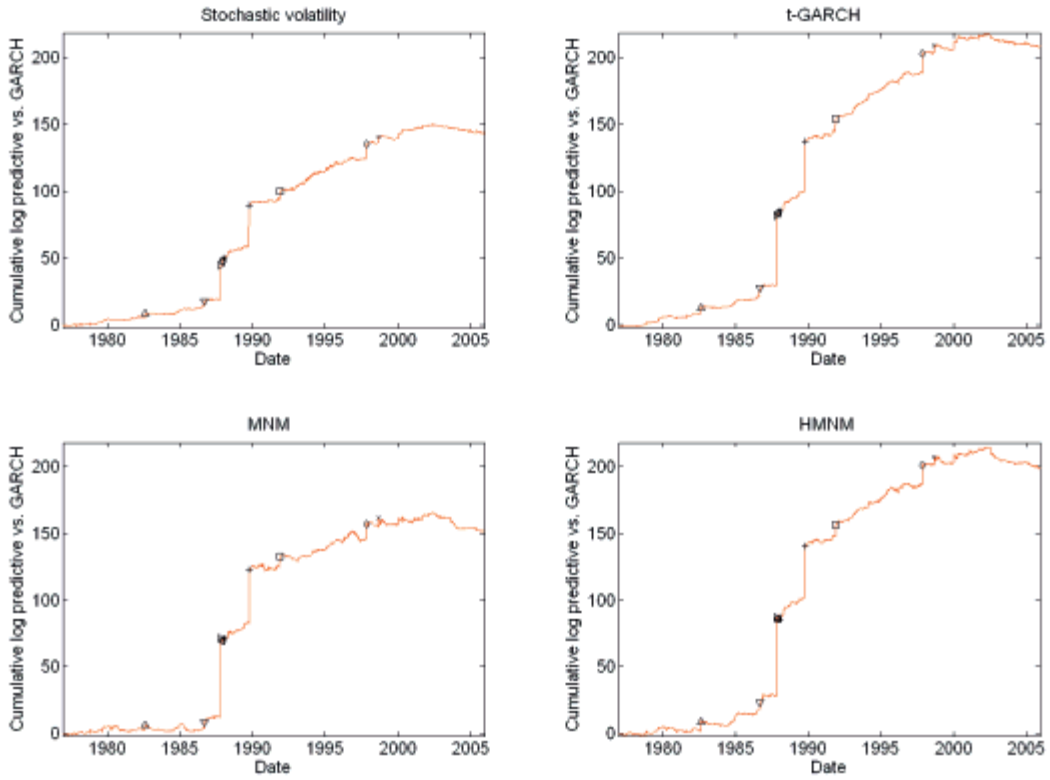

Figure 2: Cumulative predictive log predictive Bayes factors in favor of each of four models over GARCH. Symbols identify dates as indicated in Figure 1.

between returns that are large in magnitude and log predictive Bayes factors, and comparison of panels (b) and (d) shows that there is no simple relationship between these returns and log predictive likelihoods. Panels (b) through (d) show that for most days the predictive Bayes factor in favor of one model or the other is small.

Panel (c) shows a weak but systematic relationship between absolute returns and log predictive Bayes factors: the HMNM model tends to be favored by log Bayes factors when returns are less than 0.5% in magnitude, whereas t-GARCH tends to be favored when return magnitude is between 0.5% and 1%. As return magnitude rises above 1% the range of log predictive Bayes factors tends to increase, with no systematic tendency for one model or the other to be favored. The important exception to this pattern is October 19, 1987.

The exploratory analysis illustrated in Figure 3 can be used to compare the relative predictive performance of the two models (as captured by log predictive Bayes factors) and any function of returns over the preceding days. In all five models more

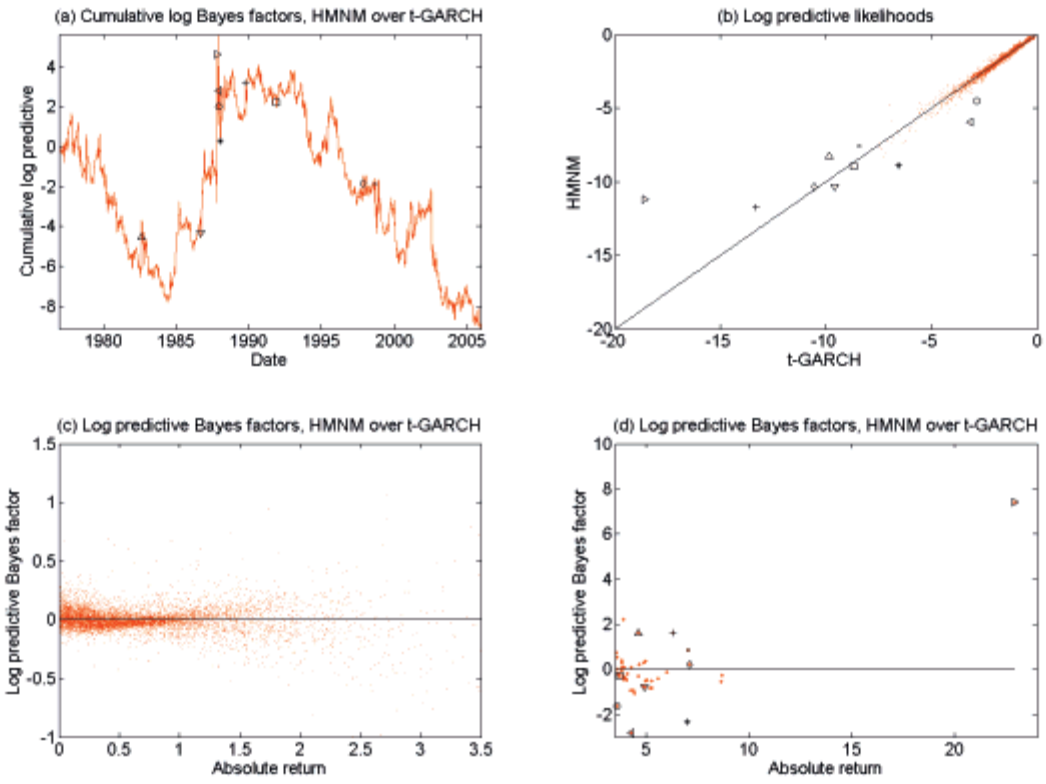

Figure 3: Some comparisons of the t-GARCH and HMNM models using the log predictive Bayes factor in favor of HMNM. Symbols identify dates as indicated in Figure 1.

volatile recent returns lead to greater dispersion in predictive distributions, but the mechanisms are distinct – especially in the HMNM model as opposed to models in the ARCH family. This characteristic of the models suggests that the magnitude of the return relative to recent magnitudes might be systematically related to log predictive Bayes factors. Figure 4 pursues this analysis, capturing return relative to recent magnitudes as the ratio of $|y_t^o|$ to the standard deviation in $\{y_s^o\}$ $(s = t - 80, \ldots, t - 1)$. Call this ratio q.

For the ten dates initially identified in Figure 3(b), the correlation between q and the corresponding log predictive Bayes factors exceeds 0.9: the near-linear relationship is evident in panel (a) of Figure 4. Panel (b) plots log predictive Bayes factors against q for all days on which q is less than 4. (The rescaling of the vertical axis in panel (b) excludes no days.) The pattern in Figure 4(b) is similar to that in Figure 3(c).

As measured by log predictive Bayes factors, the predictive performance of the

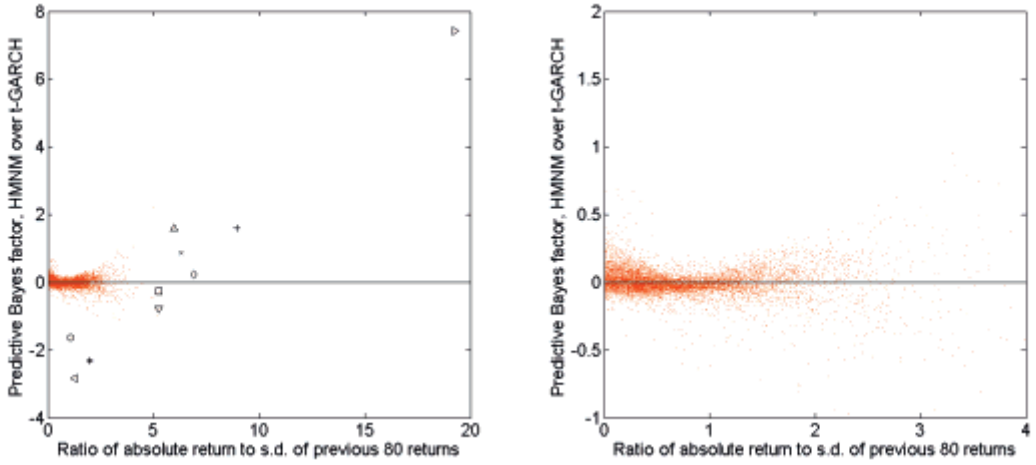

Figure 4: Comparison of the ratio of absolute return to the standard deviation of returns in the past 80 days, with the predictive Bayes factor for HMNM over t-GARCH . Symbols identify dates as indicated in Figure 1.

HMNM model dominates that of the t-GARCH model when returns are very large in magnitude relative to recent volatility – that is, for returns whose absolute value exceeds five standard deviations of returns over the past 80 days. Equivalently, extreme returns that occur roughly once or twice per decade are assigned substantially more probability in the HMNM model than in the t-GARCH model. Overall, the log predictive likelihoods of the two models are nearly identical. Elsewhere (Geweke and Amisano (2008)) we have shown that this implies that neither model corresponds to a true data generating process D, and there must exist models with higher log predictive likelihoods. Figure 4 suggests that in such models the predictive density function might resemble more the HMNM predictive density for returns that are small or quite large relative to recent volatility, and for the remainder might resemble more the t-GARCH model.

For the GARCH and t-GARCH models we prepared an alternative set of predictive densities for each of the 7324 days, using (3) and the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}_A^{(t-1)}$. We then compared the Bayesian and MLE predictive densities using the difference in log scores (12). For the GARCH model the outcome is 18.09 and for the t-GARCH model it is 9.76: both comparisons favor the Bayesian predictive densities over the MLE predictive densities.

This outcome is not surprising. Bayesian predictive densities account for parameter uncertainty, whereas the MLE predictive density (3) does not. Figure 5 provides the daily decomposition of (12). One would expect the advantage of Bayesian predictive densities to be more pronounced in smaller samples, corresponding to earlier data in Figure 5. The daily decomposition for t-GARCH is at least roughly consistent

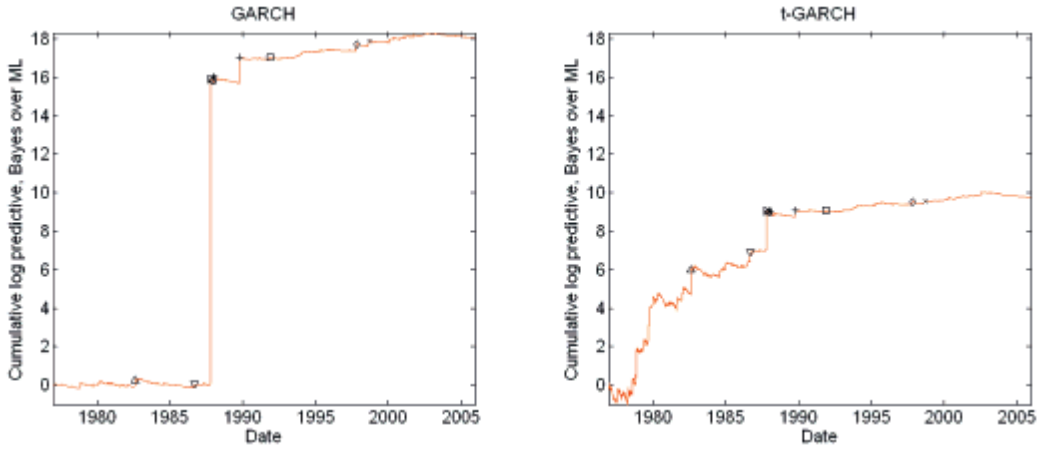

Figure 5: $\log\left[p\left(y_t^o \mid \mathbf{y}_{t-1}^o, A\right)/p\left(y_t^o \mid \mathbf{y}_{t-1}^o, \widehat{\boldsymbol{\theta}}_A^{(t-1)}, A\right)\right]$ for the ARCH and t-GARCH models.

with this understanding.

The decomposition for GARCH is dominated by the inferior performance of the MLE predictive density, relative to the Bayesian predictive density, on October 19, 1987. The GARCH MLE log predictive likelihood for that date was -84.7, whereas the GARCH Bayesian log predictive density was -69.2; for t-GARCH the figures are -20.4 and -17.3, respectively, whereas the HMNM model log predictive density is -11.2. The superior performance of Bayesian GARCH relative to MLE GARCH on this date can be traced to uncertainty about the mean parameter μ. The maximum likelihood estimate and posterior mean of μ are quite close and the posterior distribution of μ is nearly symmetric about the MLE. However the predictive likelihood is a strongly convex function of μ because the observed value is so far in the left tail of the predictive density. Jensen's inequality accounts for the predictive likelihood being higher in the Bayesian predictive likelihood than in the MLE predictive likelihood. This effect is muted in the t-GARCH model because the observed value is not nearly as far in the left tail and the predictive likelihood is not as strongly a convex function of μ.

5 Model evaluation with probability integral transforms

Predictive likelihoods are local measures of the predictive performance of models: that is, they depend only on the predictive probability density evaluated at the realized return. Moreover predictive densities measure only the relative performance of models – indeed, as discussed in Section 3 they are components of Bayes factors that are critical in Bayesian model comparison and model averaging.

The probability integral transform (PIT) provides an alternative assessment of the predictive performance of a model that is based on non-local assignment of predictive probability and is therefore complementary to the assessment based on predictive likelihoods discussed in the previous two sections. Unlike predictive likelihoods, however, the comparison is non-Bayesian.

Suppose that a model A assigns one-step-ahead predictive densities $p(y_t \mid \mathbf{y}_{t-1}, A)$. Denote the corresponding sequence of cumulative distribution functions

$$F(c \mid \mathbf{y}_{t-1}, A) = P(y_t \leq c \mid \mathbf{y}_{t-1}, A).$$

The PIT corresponding to the model A and the sequence $\{y_t\}$ is $p_1(t; \mathbf{y}_T, A) = F(y_t \mid \mathbf{y}_{t-1}, A)$. If $A = D$, the true data generating process for $\{y_t\}$, then $\{p_1(t; \mathbf{y}_T, A)\}$ is i.i.d. with distribution uniform on $(0, 1)$. This result dates at least to Rosenblatt (1952), and was brought to the attention of the econometrics community by Diebold et al. (1998). Following Smith (1985) and Berkowitz (2001), if $A = D$ then

$$f_1(t; \mathbf{y}_T, A) = \Phi^{-1}[p_1(t; \mathbf{y}_T, A)] \overset{iid}{\sim} N(0, 1),$$

and for analytical purposes it is often more convenient to work with $\{f_1(t; \mathbf{y}_T, A)\}$ than with $\{p(t; \mathbf{y}_T, A)\}$. For h-step ahead predictive distributions, let $F_h(c \mid \mathbf{y}_{t-h}, A) = P(y_t \leq c \mid \mathbf{y}_{t-h}, A)$ and $p_h(t; \mathbf{y}_T, A) = F(y_t \mid \mathbf{y}_{t-h}, A)$. If $A = D$ then the distribution of $p_h(t; \mathbf{y}_T, A)$ is uniform on $(0, 1)$, but $p_h(t; \mathbf{y}_T, A)$ and $p_h(s; \mathbf{y}_T, A)$ are independent if and only if $|t - s| \geq h$.

These characteristics of $\{p_h(t; \mathbf{y}_T, D)\}$ would only be approximately true of $\{p_h(t; \mathbf{y}_T, A)\}$ even if $p(\mathbf{y}_T \mid D) = p(\mathbf{y}_T \mid \boldsymbol{\theta}_A, A)$ for some value of $\boldsymbol{\theta}_A$, because $\boldsymbol{\theta}_A$ is unobservable. (The approximation would improve as T increased and $p(\mathbf{y}_T \mid A)$ incorporated an increasingly tight posterior distribution for $\boldsymbol{\theta}_A$.) More important, we know that $A \neq D$. The departure of the sequence $\{p_h(t; \mathbf{y}_T, A)\}$ from these ideal characteristics provides an informal evaluation of A against an absolute standard.

In many models, including all five in this study, analytical evaluation of

$$P\left(y_t \leq c \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A, A\right)$$

is possible, and this is all that is required for a posterior simulation approximation of $F\left(y_t^o \mid \mathbf{y}_{t-1}^o, A\right)$, because

$$M^{-1} \sum_{m=1}^{M} P\left(y_t \leq y_t^o \mid \mathbf{y}_{t-1}^o, \boldsymbol{\theta}_A^{(m)}, A\right) \overset{a.s.}{\to} p_1(t; \mathbf{y}_T^o, A)$$

so long as $\left\{\boldsymbol{\theta}_A^{(m)}\right\}$ is an ergodic sequence whose invariant distribution is the posterior. For $h > 1$ analytical evaluation of $P\left(y_t \leq c \mid \mathbf{y}_{t-h}^o, \boldsymbol{\theta}_A, A\right)$ is very awkward or impossible in all of these models, and that is true of econometric prediction models generally. Instead we employ a simulation approximation using the recursion

$$y_s^{(m)} \sim p\left(y_s \mid \mathbf{y}_{t-h}^o, y_{t-h+1}^{(m)}, \ldots, y_{s-1}^{(m)}, \boldsymbol{\theta}_A^{(m)}, A\right) (s = t - h + 1, \ldots, t). \qquad (14)$$

Since the posterior MCMC sample is large, only one such recursion need be carried out for each $\boldsymbol{\theta}_A^{(m)}$ in the posterior sample, and

$$M^{-1}\sum_{m=1}^{M} I_{(-\infty,y_t^o)}\left(y_t^{(m)}\right) \overset{a.s.}{\to} p_h\left(t;\mathbf{y}_T^o,A\right). \tag{15}$$

6 Evaluation of five models of S&P 500 returns

For the S&P 500 return series $\mathbf{y}_T^o$ we evaluated $p_h\left(t;\mathbf{y}_T^o,A\right)$ for each of the five models A described in section 2, for $h = 1,\ldots,10$, and for $t = 1,\ldots,T$. The computations are based on the 7324 Markov chain Monte Carlo posterior samples in each of the five models, one for each sample, using (7) followed by followed by (14) and (15). We then computed the corresponding transformation to the standard normal distribution, $f_h\left(t;\mathbf{y}_T^o,A\right) = \Phi^{-1}\left[p_h\left(t;\mathbf{y}_T^o,A\right)\right]$.

To describe the departure from the PIT paradigm for each model, we determined the values of

$$(T/h)^{-1}\sum_{s=1}^{[T/h]} I_{(a_{j-1},a_j)}\left[p_h\left((s-1)\,h+k;\mathbf{y}_T^o,A\right)\right] \quad (j=1,\ldots,n;h=1,\ldots,10) \tag{16}$$

with $n = 10$, $a_j = j/10$ $(j = 0,\ldots,n)$ and $k = 1$; that is, we determined the fraction of $p_h\left(t;\mathbf{y}_T^o,A\right)$ in each decile using non-overlapping prediction horizons h. Figure 6 presents the values of (16) for a different model A in each row of panels, and for $h = 1$, 5 and 10 in each of the three columns of panels. Values for the deciles $j = 1,\ldots,10$ are shown in each panel. The paradigm value 0.10 is indicated by the solid horizontal line, and the dotted horizontal lines provide a conventional 95% confidence interval for these values under the condition that $\{p_h\left(sh+1;\mathbf{y}_T^o,A\right)\}$ $(s = 1,\ldots,[T/h])$ is i.i.d. Bernoulli $(p = 0.1)$. Plotted values above 0.1 indicate deciles in which more than 10% of the realized returns occurred; equivalently, the model underpredicts probability in this range of the predictive density. Values below 0.1 correspond to overprediction in the relevant range.

The performance of the GARCH model is markedly inferior to the other four, and the performance of SV is not quite as good as the t-GARCH, MNM and HMNM models. The tendency of the latter three models to over- or under-predict different deciles is roughly the same for all three horizons studied in Figure 6. At horizon $h = 1$ they assign too little probability in the interquarile range and too much in the lower and upper quartiles. At $h = 5$ and $h = 10$ they assign too much probability below the median (with an exception for the lowest decile in some cases) and too little above the median of their predictive distributions. These characteristics are also evident in the GARCH model, where the performance is markedly poorer especially for $h = 10$.

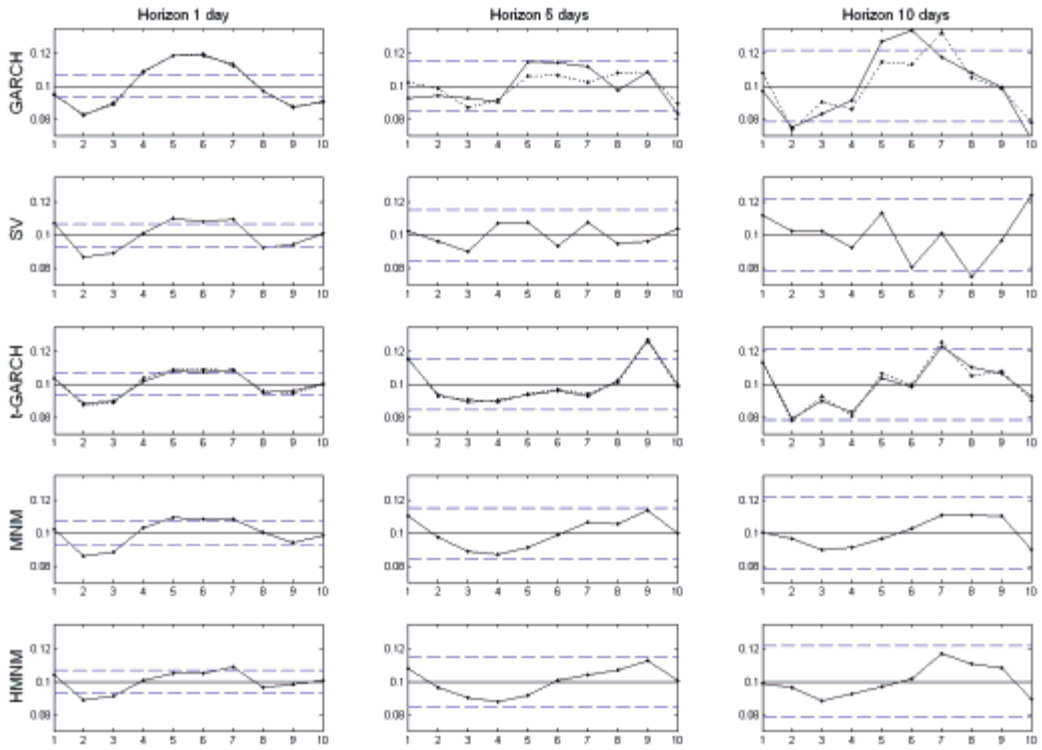

Figure 6: Each panel shows the frequency of observations occuring in each decile of the predictive distribution (16), as a function of the prediction horizon h indicated on the horizontal axis. Solid lines indicate results using Bayesian inference, dotted lines results using maximum likelihood in the GARCH and t-GARCH models. Dashed lines provide a centered 95% confdence interval under the PIT paradigm.

Table 1: Formal PIT goodness of fit tests

	Deciles			Left tail		
Horizon→	1	5	10	1	5	10
Model↓						
GARCH	0	.0088	.0018	1.0×10^{-4}	.042	6.2×10^{-4}
SV	5.4×10^{-8}	.0019	.052	8.2×10^{-5}	.032	.026
t-GARCH	2.6×10^{-5}	.066	>1	.29	.028	.14
MNM	7.9×10^{-7}	.11	.58	.063	5.2×10^{-5}	.19
HMNM	.0014	.13	.18	.15	.0035	.34
GARCH(ML)	0	.39	.73	5.9×10^{-5}	1.3×10^{-6}	1.6×10^{-5}
t-GARCH(ML)	3.7×10^{-6}	.076	.65	.076	.031	.17

Table 1 provides two sets of chi-square goodness of fit test results for the PIT. The entries in the table correspond to p-values for the tests. The results for horizon $h = 1$ are conventional. The results for horizons $h = 5$ and $h = 10$ are based on h separate tests, using (16) for $k = 1, \ldots, h$. These tests are not independent across k because the horizons overlap; the entries in Table 1 are Bonferroni p-values for the h separate tests in each case.

The first set of tests, under the "Deciles" heading, corresponds to the results shown in Figure 6. The results reinforce the conclusion that the PIT fit of the GARCH and SV models is inferior to those of the other three models. At horizon $h = 1$, only the HMNM model comes close to passing a PIT goodness of fit test at conventional significance levels. At longer horizons power is substantially less.

Evaluation of the predictive distributions over particular regions may be of concern in specific applications, particularly for negative returns; see the discussion in Diks et al. (2008). Our second set of tests explores PIT goodness of fit in the lower tail of the predictive distribution, based on (16) with $n = 5$, $a_j = j/200$ $(j = 0, \ldots, n)$, and $h = 1$, 5 and 10. As in the first set of tests, conventional p-values are given for $h = 1$ and Bonferroni test p-values are given for $h = 5$ and $h = 10$. The power of these tests is, of course, much lower than the decile tests. The GARCH and SV models fail at horizon $h = 1$, MNM fails at $h = 5$, and GARCH again at $h = 10$; HMNM has some difficulty at horizon $h = 5$. There is little in these results to suggest that the evaluation of left-tail performance is very different from overall performance using PIT.

Figures 7 and 8 provide some additional evidence on the relationship of the predictive distributions to the PIT paradigm. Each panel plots a different transformation of $p_h\left(t; \mathbf{y}_T^o, A\right)$ as a function of the prediction horizon h shown on the horizontal axis.

Panel (a) displays the interdecile range for each model: for each model A and each horizon h, it is the difference between the maximum and minimum values of (16) taken over $j = 1, \ldots, 10$. For each combination of A and h, larger values of the interdecile range constitute greater evidence against the PIT paradigm. The MNM and HMNM interdecile ranges display greater consistency with PIT than do the ranges for the other three models.

The remaining panels of Figures 7 and 8 pertain to $f_h\left(t; \mathbf{y}_T^o, A\right) = \Phi^{-1}\left[p_h\left(t; \mathbf{y}_T^o, A\right)\right]$. Under the PIT paradigm values of

$$f_h\left(t + sh; \mathbf{y}_T^o, A\right) \quad (s = 1, 2, \ldots) \tag{17}$$

are realizations of a standard i.i.d. normal process, implying that there are many functions of $f_h\left(t + sh; \mathbf{y}_T^o, A\right)$ with well-established distributions. Panels (b) through (e) pertain to the first four moments of $f_h\left(t; \mathbf{y}_T^o, A\right)$, and therefore provide indications of the discrepancy between the predictive and observed distributions. Panel (f) pertains to the first-order autocorrelation coefficient of (17) and therefore provides an indication of the departure from independence in successive quantiles implied by PIT. Unlike the interdecile range, both large and small values of the statistics constitute

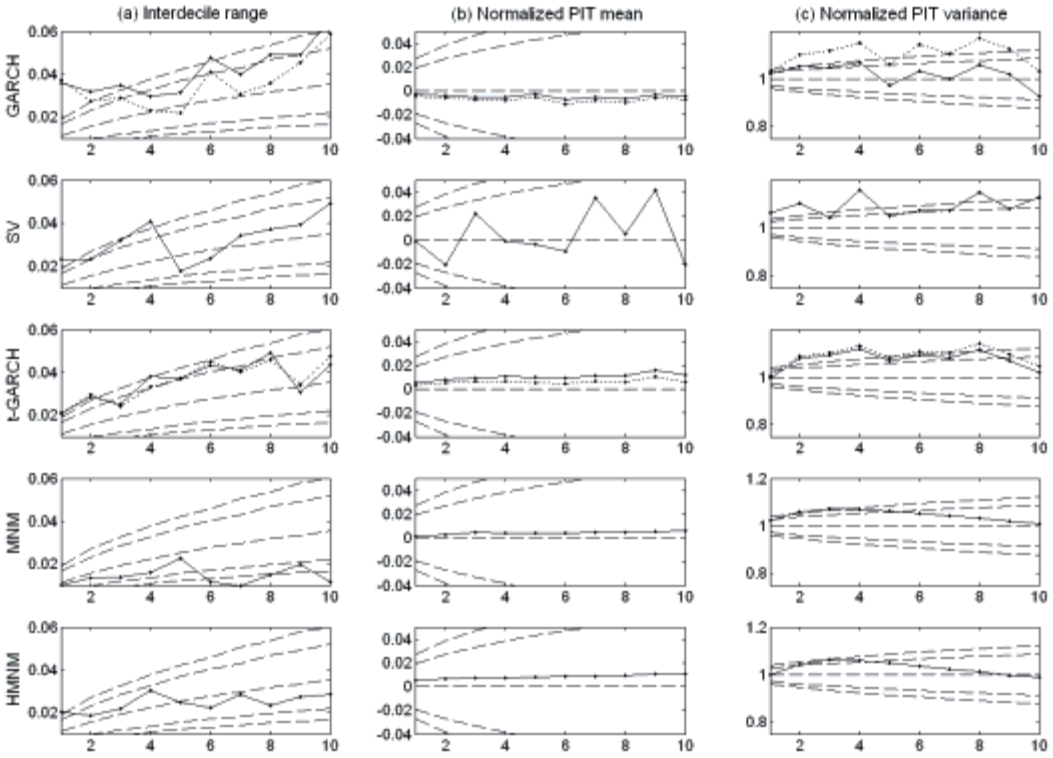

Figure 7: Each panel plots a transformation of $p_h(\mathbf{Y}_T, A)$ as a function of the prediction horizon h shown on the horizontal axis. Solid lines indicate results using Bayesian inference; dotted lines indicate results using maximum likelihood in the GARCH and t-GARCH models. The dashed lines provide the.01, .05, .50, .95 and .99 quantiles of the transformation under the PIT paradigm.

evidence that the PIT paradigm is not appropriate.

The evaluations of the t-GARCH, MNM and HMNM model fits in columns (b) and (c) of these figures are all similar. Means are slightly and insignificantly higher than zero at all horizons. Variances are greater than one, with significant departures for t-GARCH and marginal significant departures for MNM. The distribution for t-GARCH is significantly and negatively skewed (column (d)) whereas for MNM and HMNM it is insignificantly positively skewed. The evaluation of all three models using the kurtosis (column (e)) of the normalized PIT is satisfactory. By contrast the GARCH and stochastic volatility models all have severe departures from the paradigm skewness and kurtosis of the normalized PIT.

The PIT evaluations of the ML and Bayesian predictive distributions do not pro-

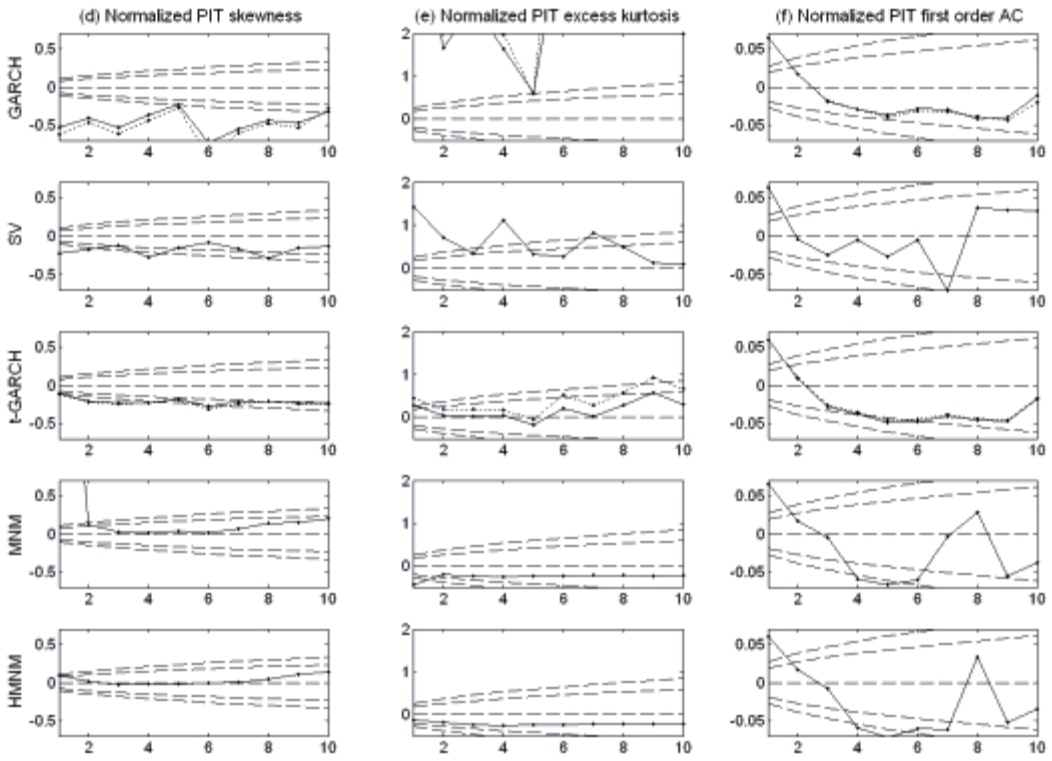

Figure 8: Each panel plots a transformation of $p_h(\mathbf{Y}_T, A)$ as a function of the prediction horizon h shown on the horizontal axis. Solid lines indicate results using Bayesian inference; dotted lines indicate results using maximum likelihood in the GARCH and t-GARCH models. The dashed lines provide the.01, .05, .50, .95 and .99 quantiles of the transformation under the PIT paradigm.

vide any striking comparisons, either in the GARCH or t-GARCH model. Posterior distributions are concentrated about the maximum likelihood estimates and consequently deciles of predictive distributions are much the same in the Bayesian and ML predictive distributions. The extreme convexity of predictive likelihoods in the far tails, responsible for the dominance of Bayesian predictive densities in the analysis in Section 4, is not a factor here.

For one-step-ahead predictive distributions ($h = 1$) the first order autocorrelation coefficient for the series (17) is about 0.06 in all five models, well outside the range of values plausible under PIT. For larger values of h the autocorrelation coefficient is smaller, with a notable tendency to be negative, in most cases, for horizons 3 or greater. Moreover, the pattern of evaluations across horizons is roughly the same in

all cases. These results suggest that there is a persistence in day-to-day returns that is not adequately captured by any of the five models.

7 Summary and conclusions

This study compares and evaluates Bayesian predictive distributions from alternative models, using as an illustration five alternative models of asset returns applied to a time series of 7324 daily S&P 500 returns. The comparison exercise uses predictive likelihoods and is inherently Bayesian. The evaluation exercise uses the probability integral transform (PIT) and is inherently frequentist. The illustration shows that the two approaches can be complementary, each identifying strengths and weaknesses in models that are not evident using the other.

Both the predictive likelihood and PIT analyses lead to the conclusion that the GARCH Bayesian predictive distributions are inferior to the other four, and the same is true of GARCH predictive distributions constructed from maximum likelihood estimates. Each analysis also leads to the conclusion that the Bayesian stochastic volatility (SV) and Markov normal mixture (MNM) predictive distributions are dominated by the t-GARCH and hierarchical Markov normal mixture (HMNM) predictive distributions. These comparisons are readily apparent in Figures 2, 6, 7 and 8, and in Table 1.

The predictive likelihood analysis narrowly favors Bayesian t-GARCH over HMNM, but for MLE t-GARCH the predictive likelihoods are nearly identical; see Figures 2 and 5. By contrast the PIT analysis narrowly favors HMNM predictive distributions over t-GARCH. This is evident in Table 1 and Figures 6, 7 and 8. However the latter analysis also shows that the normalized PIT for the HMNM model is not ideal. In particular, PIT's for successive one-day predictions are not independent, and the performance of HMNM in this dimension is no better than those of the other four models.

A new predictive density can always be formed as a weighted average of predictive densities from different models, the best known example being Bayesian model averaging (Geweke (2005), Section 2.6). The analysis in Section 4 indicates that for most combinations of models and substantial subperiods of the sample considered Bayesian model averaging is for all practical purposes equivalent to model selection, with on model receiving a weight very close to 1. This is often the outcome for Bayesian model averaging when the sample is large, as it is here. The notable exception arises when the models averaged include both t-GARCH and HMNM: in that case these two models can have substantial weight in Bayesian model averaging, depending on the days included in the sample. Geweke and Amisano (2008) shows that a weighted average of the HMNM and t-GARCH models compares quite favorably with both models, using predictive likelihood. That paper also shows that, in general, optimization of the predictive likelihood leads to non-trivial weights on several models, weights that are quite different from those that result from conventional Bayesian

model averaging.

Ultimately, analysis of this kind provides the elements from which better models may be constructed, as illustrated in the introduction by our experience in developing the HMNM model. The predictive density evaluations presented here, as well as in Geweke and Amisano (2008), show that there is scope for substantial further improvement. Awaiting such new research, real-world demands for predictive distributions will continue. In this context analyses of predictive distributions of the kind conducted here provide indications of model limitations with which actual decisions based on these models must always be tempered.

References

Amisano G, Giacomini R. 2007. Comparing density forecasts via weighted likelihood ratio tests, *Journal of Business and Economic Statistics*, 2007 25: 177-190

Bauwens L, Giot, P, Grammig J, Veredas D. 2004. A comparison of financial duration models via density forecasts. *International Journal of Forecasting* 20: 589-609.

Berkowitz J. 2001. The accuracy of density forecasts in risk management. *Journal of Business and Economic Statistics* 19: 465-474.

Chib, S. 1996. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics* 75: 79-97.

Christoffersen PF. 1998. Evaluating interval forecasts. *International Economic Review* 39: 841-862.

Corradi V, Swanson NR 2006. Predictive density evaluation. Elliott G, Granger CWJ, Timmermann A (eds.), *Handbook of Economic Forecasting.* Amsterdam: North-Holland. Chapter 5, 197-284.

Diebold FX, Gunther TA, Tay AS. 1998. Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39: 863-883.

Diks C, Panchenko V, van Dijk D. 2008. Partial likelihood-based scoring rules for evaluating density forecasts in the tails. Tinbergen Institute Discussion paper No. 2008-050/4.

Geweke J. 2001. Bayesian econometrics and forecasting. *Journal of Econometrics* 100: 11-15.

Geweke J. 2005. *Contemporary Bayesian Econometrics and Statistics.* Hoboken: Wiley.

Geweke J, Amisano G. 2007. Hierarchical Markov normal mixture models with applications to financial asset returns. *European Central Bank working paper* 831. http://www.ecb.int/pub/pdf/scpwps/ecbwp831.pdf

Geweke J, Amisano G. 2008. Optimal Prediction Pools.
http://www.biz.uiowa.edu/faculty/jgeweke/papers/paperA/paper.pdf

Geweke J, Whiteman C. 2006. Bayesian forecasting. Granger CWJ, Timmermann A (eds.) *Handbook of Economic Forecasting.* Elsevier: Amsterdam, The Netherlands.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102: 359-378.

Hong YM, Li HT, Zhao F. 2004. Out of sample performance of discrete time spot interest rate models. Journal of Business and Economic Statistics 22: 457-473.

Lindgren, G. 1978. Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics* **5**: 81-91.

Jacquier E, Polson NG, Rossi PE. 1994. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics* 12: 371-389.

Rosenblatt M. 1952. Remarks on a multivariate transformation. *Annals of Mathematical Statistics* 23: 470-472.

Rydén T, Teräsvirta T, Åsbrink S. 1998. Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics* 13: 217-244.

Smith JQ. 1985. Diagnostic checks of non-standard time series models. *Journal of Forecasting* 4: 283-291.

Tyssedal, JS, Tjøstheim, D. 1988. An autoregressive model with suddenly changing parameters and an application to stock market prices. *Applied Statistics* 37: 353-369.

Weigend AS, Shi S. 2000. Predicting daily probability distributions of S&P 500 returns. *Journal of Forecasting* 19: 375-392.

European Central Bank Working Paper Series

For a complete list of Working Papers published by the ECB, please visit the ECB's website (http://www.ecb.europa.eu).

923 "Resuscitating the wage channel in models with unemployment fluctuations" by K. Christoffel and K. Kuester, August 2008.

924 "Government spending volatility and the size of nations" by D. Furceri and M. Poplawski Ribeiro, August 2008.

925 "Flow on conjunctural information and forecast of euro area economic activity" by K. Drechsel and L. Maurin, August 2008.

926 "Euro area money demand and international portfolio allocation: a contribution to assessing risks to price stability" by R. A. De Santis, C. A. Favero and B. Roffia, August 2008.

927 "Monetary stabilisation in a currency union of small open economies" by M. Sánchez, August 2008.

928 "Corporate tax competition and the decline of public investment" by P. Gomes and F. Pouget, August 2008.

929 "Real convergence in Central and Eastern European EU Member States: which role for exchange rate volatility?" by O. Arratibel, D. Furceri and R. Martin, September 2008.

930 "Sticky information Phillips curves: European evidence" by J. Döpke, J. Dovern, U. Fritsche and J. Slacalek, September 2008.

931 "International stock return comovements" by G. Bekaert, R. J. Hodrick and X. Zhang, September 2008.

932 "How does competition affect efficiency and soundness in banking? New empirical evidence" by K. Schaeck and M. Čihák, September 2008.

933 "Import price dynamics in major advanced economies and heterogeneity in exchange rate pass-through" by S. Dées, M. Burgert and N. Parent, September 2008.

934 "Bank mergers and lending relationships" by J. Montoriol-Garriga, September 2008.

935 "Fiscal policies, the current account and Ricardian equivalence" by C. Nickel and I. Vansteenkiste, September 2008.

936 "Sparse and stable Markowitz portfolios" by J. Brodie, I. Daubechies, C. De Mol, D. Giannone and I. Loris, September 2008.

937 "Should quarterly government finance statistics be used for fiscal surveillance in Europe?" by D. J. Pedrega and J. J. Pérez, September 2008.

938 "Channels of international risk-sharing: capital gains versus income flows" by T. Bracke and M. Schmitz, September 2008.

939 "An application of index numbers theory to interest rates" by J. Huerga and L. Steklacova, September 2008.

940 "The effect of durable goods and ICT on euro area productivity growth?" by J. Jalava and I. K. Kavonius, September 2008.

941 "The euro's influence upon trade: Rose effect versus border effect" by G. Cafiso, September 2008.

942 "Towards a monetary policy evaluation framework" by S. Adjemian, M. Darracq Pariès and S. Moyen, September 2008.

943 "The impact of financial position on investment: an analysis for non-financial corporations in the euro area" by C. Martinez-Carrascal and A. Ferrando, September 2008.

944 "The New Area-Wide Model of the euro area: a micro-founded open-economy model for forecasting and policy analysis" by K. Christoffel, G. Coenen and A. Warne, October 2008.

945 "Wage and price dynamics in Portugal" by C. Robalo Marques, October 2008.

946 "Macroeconomic adjustment to monetary union" by G. Fagan and V. Gaspar, October 2008.

947 "Foreign-currency bonds: currency choice and the role of uncovered and covered interest parity" by M. M. Habib and M. Joy, October 2008.

948 "Clustering techniques applied to outlier detection of financial market series using a moving window filtering algorithm" by J. M. Puigvert Gutiérrez and J. Fortiana Gregori, October 2008.

949 "Short-term forecasts of euro area GDP growth" by E. Angelini, G. Camba-Méndez, D. Giannone, L. Reichlin and G. Rünstler, October 2008.

950 "Is forecasting with large models informative? Assessing the role of judgement in macroeconomic forecasts" by R. Mestre and P. McAdam, October 2008.

951 "Exchange rate pass-through in the global economy: the role of emerging market economies" by M. Bussière and T. Peltonen, October 2008.

952 "How successful is the G7 in managing exchange rates?" by M. Fratzscher, October 2008.

953 "Estimating and forecasting the euro area monthly national accounts from a dynamic factor model" by E. Angelini, M. Bańbura and G. Rünstler, October 2008.

954 "Fiscal policy responsiveness, persistence and discretion" by A. Afonso, L. Agnello and D. Furceri, October 2008.

955 "Monetary policy and stock market boom-bust cycles" by L. Christiano, C. Ilut, R. Motto and M. Rostagno, October 2008.

956 "The political economy under monetary union: has the euro made a difference?" by M. Fratzscher and L. Stracca, November 2008.

957 "Modeling autoregressive conditional skewness and kurtosis with multi-quantile CAViaR" by H. White, T.-H. Kim, and S. Manganelli, November 2008.

958 "Oil exporters: in search of an external anchor" by M. M. Habib and J. Stráský, November 2008.

959 "What drives U.S. current account fluctuations?" by A. Barnett and R. Straub, November 2008.

960 "On implications of micro price data for macro models" by B. Maćkowiak and F. Smets, November 2008.

961 "Budgetary and external imbalances relationship: a panel data diagnostic" by A. Afonso and C. Rault, November 2008.

962 "Optimal monetary policy and the transmission of oil-supply shocks to the euro area under rational expectations" by S. Adjemian and M. Darracq Pariès, November 2008.

963 "Public and private sector wages: co-movement and causality" by A. Lamo, J. J. Pérez and L. Schuknecht, November 2008.

964 "Do firms provide wage insurance against shocks? Evidence from Hungary" by G. Kátay, November 2008.

965 "IMF lending and geopolitics" by J. Reynaud and J. Vauday, November 2008.

966 "Large Bayesian VARs" by M. Bańbura, D. Giannone and L. Reichlin, November 2008.

967 "Central bank misperceptions and the role of money in interest rate rules" by V. Wieland and G. W. Beck, November 2008.

968 "A value at risk analysis of credit default swaps" by B. Raunig and M. Scheicher, November 2008.

969 "Comparing and evaluating Bayesian predictive distributions of asset returns" by J. Geweke and G. Amisano, November 2008.

ISSN 1561-0810