Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Econometrics 135 (2006) 255–284

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Validating forecasts of the joint probability density of bond yields: Can affine models beat random walk?

Alexei V. Egorov[a], Yongmiao Hong[b,c,*], Haitao Li[d]

[a]*College of Business and Economics, West Virginia University, Morgantown, WV 26506, USA*
[b]*Department of Economics and Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA*
[c]*Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian, China*
[d]*Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109, USA*

Available online 31 August 2005

Abstract

Most existing empirical studies on affine term structure models (ATSMs) have mainly focused on in-sample goodness-of-fit of historical bond yields and ignored out-of-sample forecast of future bond yields. Using an omnibus nonparametric procedure for density forecast evaluation in a continuous-time framework, we provide probably the first comprehensive empirical analysis of the out-of-sample performance of ATSMs in forecasting the joint conditional probability density of bond yields. We find that although the random walk models tend to have better forecasts for the conditional mean dynamics of bond yields, some ATSMs provide better forecasts for the joint probability density of bond yields. However, all ATSMs considered are still overwhelmingly rejected by our tests and fail to provide satisfactory density forecasts. There exists room for further improving density forecasts for bond yields by extending ATSMs.

JEL classification: C4; C5; G1

Keywords: Density forecast; Affine term structure models; Probability integral transform; Financial risk management; Value at risk; Fixed-income portfolio management

*Corresponding author. Department of Economics and Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA. Tel.: +1 607 255 5130; fax: +1 607 255 2818.

E-mail address: yh20@cornell.edu (Y. Hong).

0304-4076/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2005.07.018

1. Introduction

The term structure of interest rates, which concerns the relationship among the yields of default-free bonds with different maturities, is one of the most widely studied topics in economics and finance. Following the pioneering works of Vasicek (1977) and Cox et al. (1985), a large number of multifactor dynamic term structure models (DTSMs) have been developed over the last two decades.[1] These models, by imposing cross-sectional and time series restrictions on bond yields in an internally consistent manner, provide important insights for our understanding of term structure dynamics. They have been widely used in financial industry for pricing fixed-income securities and managing interest rate risk.

Affine term structure models (ATSMs), first introduced in Duffie and Pan (1996), have become the leading DTSMs in the literature due to their rich model specification and tractability. In ATSMs, the short-term interest rate is an affine function of the underlying state variables which follow affine diffusions (the instantaneous drift and variance are affine functions of the state variables) under both the risk-neutral and physical measures. These assumptions allow closed-form solutions for a wide variety of fixed-income securities (see, e.g., Duffie et al., 2000; Chacko and Das, 2002), which greatly simplify empirical implementations of ATSMs. As a result, ATSMs have become probably the most widely studied DTSMs in the academic literature.

Despite the numerous empirical studies on DTSMs, the existing literature has mainly focused on in-sample fit of historical bond yields and ignored out-of-sample forecast of future bond yields. In-sample diagnostic analysis is important and can reveal useful information on possible sources of model misspecifications. However, it is the evolution of the yield curve in the future, not the past, that is most relevant in many financial applications, such as pricing and hedging fixed-income securities and managing interest rate risk. As widely recognized in the literature, the current yield curve contains information about the future yield curve and the state of the economy. Therefore, accurate forecasts of bond yields are also important for savings and investments decisions of households and firms, and for macroeconomic policy decisions of monetary authorities. Furthermore, as pointed out by Duffee (2002, p. 465), “a model that is consistent with finance theory and produces accurate forecasts can make a deeper contribution to finance,” especially for explaining time varying expected bond returns and the failure of the expectation hypothesis.

Unfortunately, there is no guarantee that a model that fits historical data well will also perform well in out-of-sample forecast, due to at least three important reasons. First, the extensive search for more complicated models using the same (or similar) data set(s) may suffer from the so-called data snooping bias, as pointed out by Lo and MacKinlay (1989) and White (2000). In the present context, most studies on

[1] The theoretical and empirical literature on multi-factor dynamic term structure models is too huge and diversed to be summarized here. For excellent reviews of the current literature, see Dai and Singleton (2003) and Piazzesi (2004).

ATSMs have used either U.S. Treasury yields in the past 50 years or U.S. dollar swap rates in the past 15 years. While a more complicated model can always fit a given data set better than simpler models, it may overfit some idiosyncratic features of the data without capturing the true data-generating process. Out-of-sample forecasting evaluation will alleviate, if not eliminate completely, such data snooping bias. Second, an overparameterized model contains a large number of estimated parameters and inevitably exhibits excessive sampling variation in parameter estimation. Such excessive parameter estimation uncertainty may adversely affect the out-of-sample forecast performance. Third, a model that fits in-sample data well may not forecast the future well because of unforeseen structural changes or regime shifts in the data-generating process. Therefore, in-sample analysis is not adequate and it is important to examine the out-of-sample predictive ability of existing term structure models, especially when comparing competing models.

A few studies that do consider the out-of-sample performance of ATSMs have shown that they fail miserably in forecasting the conditional mean of future bond yields. For example, Duffee (2002) shows that the completely ATSMs of Dai and Singleton (2000) have worse forecasts of the conditional mean of bond yields than a simple random walk model in which expected future yields are equal to current yields. Consequently, Duffee (2002, p. 434) concludes that "for the purposes of forecasting, completely ATSMs are largely useless".[2] In fact, it has been shown that the simple random walk model outperforms most sophisticated models in forecasting the conditional mean of many other economic and financial time series. One well-known example is the forecasts of foreign exchange rates: the classic paper of Meese and Rogoff (1983) and many important subsequent studies have shown that the random walk model outperforms most structural and time series models in forecasting the conditional mean of major exchange rates.

However, the full dynamics of an intertemporal model is completely characterized by the conditional density of the state variables, which includes not only the conditional mean, but also higher-order conditional moments. A model that has better forecasts of the conditional mean does not necessarily have better forecasts of higher-order conditional moments as well. For example, it is widely known that GARCH and stochastic volatility models provide better forecasts of the conditional variance of many financial time series than simple random walk models. There is a vast literature on volatility forecasting for the purpose of option pricing and risk management (see, e.g., Andersen et al., 2004). As shown by Dai and Singleton (2000), the ATSMs that have the best (in-sample) empirical performance are those that are flexible in modeling the time varying volatilities and correlations of the state variables. In fact, it is well known that changes of most financial time series have weak or little dependent structure in conditional mean, but persistent dependence in conditional variance and higher-order conditional moments. Therefore, ATSMs

[2] Diebold and Li (2002) provide some encouraging results on the forecast of the level of future bond yields using variants of the Nelson–Siegel exponential components framework to model the entire yield curve. However, as the authors point out, their model is mainly statistical and belongs to neither the no-arbitrage nor the equilibrium approach to term structure modeling.

might have good forecasts for the higher-order moments, or even the whole conditional density of bond yields, although they have poor forecasts of the conditional mean dynamics.

In this paper, we study whether ATSMs can provide accurate forecasts of the joint conditional probability density of bond yields. We focus on forecasting the conditional density because interest rates, like most other financial data, are highly non-Gaussian. One needs to go beyond the conditional mean and variance to get a complete picture of term structure dynamics. The conditional probability density of the state variables characterizes the full dynamics of a term structure model and essentially checks all conditional moments simultaneously (if the moments exist). In fact, all continuous-time models in finance, including ATSMs, are essentially models for the transition density of the underlying economic process. Because of this, density forecast evaluation is a very natural and suitable way to evaluate these financial models.

Density forecasts are important not only for constructing statistical tests, but also for many economic and financial applications. For example, the booming industry of financial risk management is essentially dedicated to provide density forecasts for portfolio returns, and then to track certain aspects of the distribution such as Value at Risk (VaR) to quantify the risk exposure of a portfolio (e.g., Duffie and Pan, 1997; Morgan, 1996; Jorion, 2000).[3] Density forecast in ATSMs is especially important for financial risk management in the huge fixed-income markets. In ATSMs, a finite number of state variables drive the evolution of the whole yield curve. Thus, accurate forecasts of the joint density of the state variables would allow us to forecast the distribution of the whole yield curve. If ATSMs can provide accurate density forecasts, they would be very useful for managing the large fixed-income holdings of many banks given their closed-form solutions for most existing fixed-income securities. For other financial applications of ATSMs, such as derivatives pricing and hedging, density forecasts rather than forecasts of a specific feature of the density will be required. Therefore, when evaluating ATSMs, out-of-sample density forecast is an important dimension of model diagnostics that should not be ignored.

Evaluating density forecasts is not trivial, given that the probability density function is not observable even ex post. Unlike point forecast evaluation, there are relatively few statistical tools for density forecast evaluation.[4] In a pioneering contribution, Diebold et al. (1998) evaluate density forecasts by examining the dynamic probability integral transforms of the data with respect to a model forecast

[3] It is important to point out that many applications, such as VaR forecasts, require only certain specific features of the density, but not the entire density. We choose to focus on the entire density because in any application and for any loss function, it is always preferable to use a model that can capture the whole density rather than other models that can only capture some specific features of the density (see Diebold et al., 1998; Granger and Pesaran, 2000). While it could be difficult to identify a correctly specified density model in practice, our procedures can help reveal potential sources of model misspecifications, which could be useful for improving the forecast model.

[4] See Corradi and Swanson (2005) for an excellent survey on existing approaches to density forecast evaluation.

density. Such a transformed series, often referred to as the "generalized residuals" of the density forecast model, should be i.i.d. U[0, 1] if the density forecast model correctly captures the full dynamics of the underlying process. Any departure from i.i.d. U[0, 1] is evidence of suboptimal forecasts and model misspecification. We extend an omnibus nonparametric in-sample test for i.i.d. U[0, 1] developed in Hong and Li (2005) to out-of-sample density forecast for continuous-time models of a multivariate process, some of whose components may be latent variables. The evaluation statistics, which measure the departure of the "generalized residuals" from i.i.d. U[0, 1], can be viewed as a metric of the distance between the forecast model and the true data-generating process. While researchers have to choose a lag order when implementing Hong and Li's (2005) test, we introduce a portmanteau statistic that combines Hong and Li's (2005) test statistics at different lag orders.[5] As a result, the power of the portmanteau statistic becomes much less dependent on which lag order we use in practice. The portmanteau statistic has the advantage of detecting a wide range of suboptimal density forecasts and is convenient for comparing the performance of different models.[6]

Using the density forecast evaluation procedure just described, we provide probably the first comprehensive empirical analysis of the out-of-sample performance of ATSMs in forecasting the joint conditional probability density of bond yields. While we consider similar models as Hong and Li (2005), i.e., the three-factor completely and essentially ATSMs, the focus of our analysis is mainly on the out-of-sample forecasting performance of these models. We find that although the random walk models tend to have better forecasts of the conditional mean of bond yields, some ATSMs provide better density forecasts of the joint probability density of bond yields. However, all affine models are still overwhelmingly rejected and none of them provides satisfactory density forecasts. This suggests that time series models with more flexible specifications might be able to provide better density forecasts than the affine models.

The rest of this paper is organized as follows. In Section 2, we introduce the nonparametric procedure for density forecast evaluation tailored to a continuous-time framework. In Section 3, we discuss density forecast for multifactor ATSMs. Section 4 investigates the in-sample and out-of-sample performance of ATSMs. In Section 5, we conclude the paper. Appendix provides the asymptotic theory.

[5] Hong and Li (2005) show that their statistics at different lag orders yield similar results in all the applications they consider.

[6] Density forecasts can be used for many different purposes, such as pricing, risk management, and portfolio selections. Consequently, the quality of the forecasts can also be evaluated based on the objective of the application for which they are made. Indeed many studies have evaluated forecasting performance using economic rather than statistical criteria. These include Christofferson and Diebold (1997), Diebold (2001), Elliott and Timmermann (2002), Granger and Pesaran (2000), and many others. We choose to evaluate density forecasts based on our nonparametric statistic for the i.i.d. U[0, 1] hypothesis, again based on the fact that a model that can capture the full dynamics of the data generating process should be preferable to any other model, regardless of the objective function (or loss function) of the application.

2. Nonparametric density forecast evaluation

2.1. Dynamic probability integral transform

Probability distribution or density function is a widely accepted approach to modeling uncertainty in economics and finance (e.g., Rothschild and Stiglitz, 1970). The importance of density forecast has been recognized in the recent literature due to the works of Diebold et al. (1998), Granger (1999), and Granger and Pesaran (2000), among many others. These authors show that accurate density forecasts are essential for decision-making under uncertainty when forecasters' objective functions are asymmetric and the underlying processes are non-Gaussian.

In many areas of economics and finance, density forecasts have become a standard practice. For example, modern risk control techniques, such as VaR, typically involve some form of density forecasts.[7] In macroeconomics, monetary authorities in the U.S. and U.K. (the Federal Reserve Bank of Philadelphia and the Bank of England) have been conducting quarterly surveys on density forecasts for inflation and output growth to help set their policy instruments (e.g., inflation target). There is also a growing literature on extracting density forecasts from options prices to obtain useful information on otherwise unobservable market expectations (e.g., Fackler and King, 1990; Jackwerth and Rubinstein, 1996; Soderlind and Svensson, 1997; Ait-Sahalia and Lo, 1998).

One of the most important issues in density forecast is the evaluation of the quality of a forecast (Granger, 1999), since suboptimal density forecasts could have severe consequences in many applications. For example, an excessive forecast of VaR could force risk managers and financial institutions to hold too much capital, imposing an additional cost. Suboptimal density forecasts for important macroeconomic variables may lead to inappropriate decisions (e.g., inappropriate level and timing in interest rate setting), which could have serious consequences on the economy. In a decision-theoretic context, Diebold et al. (1998) and Granger and Pesaran (2000) show that if a density forecast model coincides with the true conditional density of the data-generating process, it would be preferred by all forecast users regardless of their objective functions. Thus, testing the optimality of a forecast boils down to testing whether the forecast model captures the true data-generating process. This is a challenging job simply because we never observe an ex post density. So far there have been not many statistical evaluation procedures for density forecasts.

In an important paper, Diebold et al. (1998) use a probability integral transform of the data with respect to the density forecast model to assess the optimality of density forecasts. They show that if the conditional density model is correctly specified, then the probability integral transformed series should be i.i.d. U[0, 1]. This fact is first established by Rosenblatt (1952) in a simpler context, but Diebold et al. (1998) is one of the first to use it for density forecast evaluation in econometrics.

[7] Not all VaR forecasts involve density forecasts. One example is the CAViaR model of Engle and Manganelli (2004).

Suppose we have a random sample of interest rates $\{r_{\tau\Delta}\}_{\tau=1}^{L}$ of size L, where Δ is the time interval at which the data are observed or recorded. For a given continuous-time interest rate model, there is a model-implied transition density

$$\frac{\partial}{\partial r}\mathrm{P}(r_{\tau\Delta}\leqslant r|I_{(\tau-1)\Delta},\theta)=\mathrm{p}(r,\tau\Delta|I_{(\tau-1)\Delta},\theta),\quad 0<r<\infty,$$

where θ is an unknown finite-dimensional parameter vector, $I_{(\tau-1)\Delta}=\{r_{(\tau-1)\Delta}, r_{(\tau-2)\Delta},\ldots,r_{\Delta}\}$ is the information set available at time $(\tau-1)\Delta$. We divide the whole sample into two subsamples: an estimation sample $\{r_{\tau\Delta}\}_{\tau=1}^{R}$ of size R, which is used to estimate model parameters and a forecast sample $\{r_{\tau\Delta}\}_{\tau=R+1}^{L}$ of size $n=L-R$, which is used to evaluate density forecast.[8] We can then define the probability integral transform of the data in the forecast sample with respect to the model-implied transition density

$$Z_{\tau}(\theta)\equiv\int_{-\infty}^{r_{\tau\Delta}}\mathrm{p}(r,\tau\Delta|I_{(\tau-1)\Delta},\theta)\,\mathrm{d}r,\quad \tau=R+1,\ldots,L. \tag{1}$$

If the continuous-time model is correctly specified in the sense that there exists some θ_0 such that the model-implied transition density $\mathrm{p}(r,\tau\Delta|I_{(\tau-1)\Delta},\theta_0)$ coincides with the true transition density of interest rates, then the transformed sequence $\{Z_{\tau}(\theta_0)\}$ is i.i.d. U[0, 1]. Intuitively, the U[0, 1] distribution indicates proper specification of the stationary distribution of $r_{\tau\Delta}$, and the i.i.d. property characterizes correct specification of its dynamic structure. If $\{Z_{\tau}(\theta)\}$ is not i.i.d. U[0, 1] for all $\theta\in\Theta$, then $\mathrm{p}(r,\tau\Delta|I_{(\tau-1)\Delta},\theta)$ is not optimal and there exists room for further improvement. Thus, density forecast evaluation boils down to testing whether $\{Z_{\tau}(\theta)\}$, which is often referred to as the "generalized residuals" of the model-implied transition density $\mathrm{p}(r,\tau\Delta|I_{(\tau-1)\Delta},\theta)$, is i.i.d. U[0, 1].

It is nontrivial to test the joint hypothesis of i.i.d. U[0, 1] for $\{Z_{\tau}\}_{\tau=1}^{n}$, where $Z_{\tau}\equiv Z_{\tau}(\theta_0)$. One may suggest the well-known Kolmogorov–Smirnov test, which unfortunately checks U[0, 1] under the i.i.d. assumption rather than tests i.i.d. and U[0, 1] jointly. It would easily miss the non-i.i.d. alternatives with uniform marginal distribution. Moreover, the Kolmogorov–Smirnov test cannot be used directly because it does not take into account the impact of parameter estimation uncertainty on the asymptotic distribution of the test statistic.

Diebold et al. (1998) use the autocorrelograms of the generalized residuals and their powers to check the i.i.d. property, and use the histograms to check the U[0, 1] property. While this approach is simple and informative about possible sources of suboptimal density forecasts, it is preferable to use a single omnibus evaluation criteria that takes into account deviations from both i.i.d. and U[0, 1] when comparing the performances of different models. Otherwise, it would be difficult to decide which model is better in capturing the full dynamics of the data if the generalized residuals of one interest rate model have less serial dependence, but displays more departures from U[0, 1] than the other model.

[8] One can also use rolling estimation or recursive estimation. We expect that our test procedures are applicable to these different estimation methods under suitable regularity conditions. However, we do not provide a formal justification for these estimation procedures in this paper.

Hong and Li (2005) recently proposed a class of nonparametric tests of the i.i.d. U[0, 1] hypothesis for in-sample performance of continuous-time models, using the transition density.[9] To apply them to evaluate the out-of-sample performance of ATSMs, we first extend these tests to the out-of-sample setting tailored to multivariate continuous-time model. We explicitly consider the impact of parameter estimation uncertainty and the choice of relative sample sizes between the estimation and prediction samples on the evaluation procedure, two issues that have been ignored by most of the existing procedures. The main idea of our procedure is to use the i.i.d. U[0, 1] property for optimal density forecasts to develop a metric that measures how far a continuous-time model is away from the true data generating process of the underlying process.

2.2. Nonparametric omnibus evaluation procedure

Following Hong and Li (2005), we measure the distance between a forecast density model and the true transition density by comparing a kernel estimator $\hat{g}_j(z_1, z_2)$ for the joint density of $\{Z_\tau, Z_{\tau-j}\}$ with unity, the product of two U[0, 1] densities, where j is a lag order. One advantage of this approach is that since there is no serial dependence in $\{Z_\tau\}$ under correct model specification, nonparametric joint density estimators and related test statistics are expected to perform well in finite samples. This is appealing because there exists persistent dependence in interest rate time series data. Another advantage is that there is no asymptotic bias for nonparametric density estimators under the null hypothesis of correct model specification either, because the conditional density of Z_τ given $\{Z_{\tau-1}, Z_{\tau-2}, \ldots\}$ is uniform (i.e., a constant). Moreover, our test can also be applied to time-inhomogeneous continuous-time processes, because $\{Z_\tau\}$ is always i.i.d. U[0, 1] under correct model specification.[10] Simulation studies in Hong and Li (2005) show that the tests perform well in small samples even for highly persistent financial data.

Our kernel estimator of the joint density is, for any integer $j > 0$,

$$\hat{g}_j(z_1, z_2) \equiv (n-j)^{-1} \sum_{\tau=R+j+1}^{L} K_h(z_1, \hat{Z}_\tau) K_h(z_2, \hat{Z}_{\tau-j}), \quad 0 \leqslant z_1, z_2 \leqslant 1, \tag{2}$$

where $\hat{Z}_\tau = Z_\tau(\hat{\theta}_R)$, $\hat{\theta}_R$ is *any* $\sqrt{R}$-consistent estimator for θ_0, and $K_h(z_1, z_2)$ is a boundary-modified kernel function defined as follows. For $x \in [0, 1]$, we define

$$K_h(x, y) \equiv \begin{cases} h^{-1} k(\frac{x-y}{h}) / \int_{-(x/h)}^{1} k(u)\,\mathrm{d}u & \text{if } x \in [0, h), \\ h^{-1} k(\frac{x-y}{h}) & \text{if } x \in [h, 1-h], \\ h^{-1} k(\frac{x-y}{h}) / \int_{-1}^{(1-x)/h} k(u)\,\mathrm{d}u & \text{if } x \in (1-h, 1], \end{cases} \tag{3}$$

[9] While the transition densities for most continuous-time models have no closed form, many methods exist in the literature to provide accurate approximations of the transition density (e.g., Ait-Sahalia, 2002; Ait-Sahalia and Kimmel, 2002; Duffie et al., 2003). This simplifies the computation of the generalized residuals for continuous-time models.

[10] Egorov et al. (2003) extend Ait-Sahalia's (2002) Hermite expansion approach to obtain accurate closed-form approximation for the transition density of time-inhomogeneous diffusion models.

where $k(\cdot)$ is a prespecified symmetric probability density and $h \equiv h(n)$ is a bandwidth such that $h \to 0, nh \to \infty$ as $n \to \infty$. Throughout our empirical analysis, we use the quartic kernel

$$k(u) = \tfrac{15}{16}(1-u^2)^2\mathbf{1}(|u| \leqslant 1),$$

where $\mathbf{1}(\cdot)$ is the indicator function. In practice, the choice of bandwidth h is more important than the choice of the kernel $k(u)$. Like Scott (1992), we choose $h = \hat{S}_Z n^{-1/6}$, where $\hat{S}_Z$ is the sample standard deviation of $\{\hat{Z}_\tau\}_{\tau=R+1}^{L}$. This simple bandwidth rule attains the optimal rate for bivariate kernel density estimation.

The modified kernel in (3) can automatically deal with the boundary bias problem associated with standard kernel estimation. As is well known (e.g., Härdle, 1990, pp. 130–133), a standard kernel density estimator gives biased estimates near the boundaries of data, because a standard kernel provides an asymmetric coverage of the data in the boundary regions. In contrast, the weighting functions in the denominators of $K_h(x,y)$ for $x \in [0,h) \cup (1-h,1]$ account for the asymmetric coverage and ensure that estimator (2) is asymptotically unbiased uniformly over the entire support $[0,1]$ for the generalized residuals. The modified kernel in (3) has several advantages over some existing alternative solutions to the boundary bias problem in the literature. One alternative is to simply ignore the data in the boundary regions and only use the data in the interior region. Such a trimming procedure is simple, but in the present context, it would lead to the loss of significant amount of information. For a nearly uniformly distributed transformed sequence $\{Z_\tau\}$, the data in the boundary region is still about 10% when the sample size $n = 5000$ and the bandwidth $h = \hat{S}_Z n^{-1/6}$, where $\hat{S}_Z$ is the sample standard deviation of $\{\hat{Z}_\tau\}_{t=R+1}^{L}$. For financial time series such as interest rates, one may be particularly interested in the tail distribution of the underlying process, which is exactly contained in (and only in) the boundary regions! Alternatively, we can also use the so-called jackknife kernel to eliminate the boundary bias, as in Chapman and Pearson (2000) and Diebold et al. (1999). In the present context, the jackknife kernel, however, has the undesired property that it may generate negative density estimates in the boundary regions. It also induces a relatively large variance for the kernel estimates in the boundary regions, adversely affecting the power of the test in finite samples. In contrast, our modified kernel always produces nonnegative density estimates with a smaller variance in the boundary regions.

Hong and Li (2005) propose an in-sample specification test that uses a quadratic form between $\hat{g}_j(z_1,z_2)$ and 1, the product of two U[0, 1] densities. This test, when extended to the out-of-sample context, is given as

$$\hat{Q}(j) \equiv \left[(n-j)h\int_0^1\int_0^1 [\hat{g}_j(z_1,z_2)-1]^2\,\mathrm{d}z_1\,\mathrm{d}z_2 - hA_h^0\right]\Big/ V_0^{1/2}, \quad j = 1,2,\ldots, \tag{4}$$

where j is a prespecified lag order, the nonstochastic centering and scaling factors

$$A_h^0 \equiv \left[(h^{-1}-2)\int_{-1}^{1} k^2(u)\,\mathrm{d}u + 2\int_0^1\int_{-1}^{b} k_b^2(u)\,\mathrm{d}u\,\mathrm{d}b\right]^2 - 1,$$

$$V_0 \equiv 2\left[\int_{-1}^{1}\left[\int_{-1}^{1} k(u+v)k(v)\,\mathrm{d}v\right]^2 \mathrm{d}u\right]^2 \tag{5}$$

and $k_b(\cdot) \equiv k(\cdot)/\int_{-1}^{b} k(v)\,\mathrm{d}v$. Note that the modification of the kernel $k(\cdot)$ in the boundary regions affects the centering constant A_h^0, although not the asymptotic variance V_0.[11]

We first extend Hong and Li's (2005) in-sample specification test to an out-of-sample evaluation procedure in a possibly multivariate continuous-time framework. Under suitable regularity conditions stated in Appendix, we can show that $\hat{Q}(j) \to$ N(0, 1) in distribution when the continuous-time model is correctly specified (see Theorem 1 in Appendix). In a simulation experiment mimicking the dynamics of the U.S. interest rates via the Vasicek model, Hong and Li (2005) find that the in-sample version of $\hat{Q}(j)$ has good sizes for $n \geqslant 250$ (i.e., about 1 year of daily data). This is a substantial improvement over other nonparametric tests (see Ait-Sahalia, 1996; Pritsker, 1998).

With various choices of lag order j, $\hat{Q}(j)$ can reveal useful information of at which lag order significant departures from i.i.d. U[0, 1] occur. This is analogous to the use of the sample autocorrelation function in the linear time series context. If a large set of $\{\hat{Q}(j)\}$ is considered, then some of them will probably be significant even if the null is true, due to statistical sampling variation. In fact, on average 1 out of 20 will be significant at the 5% level under the null. On the other hand, the choice of lag order j is expected to have significant impact on the power of $\hat{Q}(j)$. Moreover, when comparing two different models, it is desirable to use a single portmanteau test statistic. For this purpose, we consider the following portmanteau evaluation statistic

$$\hat{W}(p) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p} \hat{Q}(j). \tag{6}$$

Like many time series test statistics, we still have to choose the lag truncation order p. The power of $\hat{W}(p)$ is still affected by the choice of p, but not as much as the power of $\hat{Q}(j)$ is affected by the choice of individual lag order j. We can show that for any p, $\hat{W}(p) \to$ N(0, 1) in distribution when the continuous-time model is correctly specified. Intuitively, when the forecast model is correctly specified, we have cov$[\hat{Q}(i), \hat{Q}(j)] \to 0$ in probability for $i \neq j$ as $n \to \infty$. That is, $\hat{Q}(i)$ and $\hat{Q}(j)$ are asymptotically independent, whenever $i \neq j$. Thus, the portmanteau test statistic $\hat{W}(p)$ is a normalized sum of approximately i.i.d. N(0,1) random variables, and so is

[11] In Eq. (9) of Hong and Li (2005), a factor of the bandwidth h that should have been multiplied with A_h^0 is missing. However, their GAUSS code has incorporated the h factor correctly and therefore all simulation and empirical results reported in Hong and Li (2005) are not affected.

asymptotically N(0,1). This test may be viewed as a generalization of the popular Box–Pierce–Ljung type autocorrelation test from a linear time series context to a continuous-time context with an out-of-sample setting.

Under model misspecification, we can show that as $n \to \infty, \hat{Q}(j) \to \infty$ in probability whenever $\{Z_\tau, Z_{\tau-j}\}$ are not independent or U[0, 1]. As long as model misspecification occurs such that there exists some lag order $j \in \{1, \ldots, p\}$ at which $\hat{Q}(j) \to \infty$, we have $\hat{W}(p) \to \infty$ in probability (see Theorem 3 in Appendix). Therefore, the portmanteau test statistic $\hat{W}(p)$ can be used as an omnibus procedure to evaluate the out-of-sample density forecast performance of a continuous-time model.[12]

As an important feature of the test, it is only required that the parameter estimator $\hat{\theta}_R$ be $\sqrt{R}$-consistent. One needs not use asymptotically most efficient estimator. The sampling variation in estimator $\hat{\theta}_R$ has no impact on the asymptotic distribution of $\hat{Q}(j)$ or $\hat{W}(p)$. This delivers a convenient procedure in practice, because the asymptotically most efficient estimators such as maximum likelihood estimation (MLE) or approximated MLE may be difficult to obtain or implement in practice. One could choose a suboptimal but convenient estimator in implementing the procedure.

After a continuous-time model is rejected using either $\hat{Q}(j)$ or $\hat{W}(p)$, it would be interesting to explore the possible reasons of the rejection. Hong and Li (2005) develop a class of rigorous separate inference procedures that can utilize the rich information contained in the generalized residuals $\{Z_\tau(\theta)\}$ of a continuous-time model. We also extend this result to the out-of-sample setting. Specifically, we consider the following test statistics:

$$\mathrm{M}(m,l) \equiv \left[\sum_{j=1}^{n-1} w^2(j/p)(n-j)\hat{\rho}_{ml}^2(j) - \sum_{j=1}^{n-1} w^2(j/p)\right] \bigg/ \left[2\sum_{j=1}^{n-2} w^4(j/p)\right]^{1/2}, \tag{7}$$

where $\hat{\rho}_{ml}(j)$ is the sample cross-correlation between $\hat{Z}_\tau^m$ and $\hat{Z}_{\tau-|j|}^l$, and $w(\cdot)$ is a weighting function for the lag orders $\{j\}$.[13] The tests M(m,l) are an extension of Hong's (1996) spectral density tests for the adequacy of discrete-time linear dynamic models. Extending the proof of Hong (1996), we can show that for each given pair of

[12] We note that one could also construct a χ^2 test, such as $\hat{C}(p) = \sum_{j=1}^{p} \hat{Q}^2(j)$. Under the same conditions as for $\hat{W}(p)$, the statistic $\hat{C}(p)$ is asymptotically χ^2 with p degrees of freedom when the forecast model is optimal. However, we expect that it is less powerful than $\hat{W}(p)$, because the latter exploits the one-sided nature of the $\hat{Q}(j)$ statistic under the alternative hypothesis (i.e., $\hat{Q}(j) \to \infty$ in probability under model misspecification).

[13] We assume that $w(\cdot)$ is symmetric around 0 and continuous on the real line except for a finite number of points. An example is the Bartlett kernel $w(z) = (1 - |z|)\mathbf{1}(|z| \leqslant 1)$. If $w(\cdot)$ has bounded support, p is a lag truncation order; if $w(\cdot)$ has unbounded support, all $n-1$ lags in the sample are used. Usually $w(\cdot)$ discounts higher-order lags. This will give better power than equal weighting when $|\rho_{ml}(j)|$ decays to 0 as lag order j increases. This is typically the case for most financial markets, where more recent events tend to have bigger impact than the remote past events.

positive integers (m, l)

$$\mathrm{M}(m, l) \to \mathrm{N}(0, 1) \quad \text{in distribution}$$

under correct model specification, provided the lag truncation order $p \equiv p(n) \to \infty, p/n \to 0$. Moreover, parameter estimation uncertainty in $\hat{\theta}_R$ has no impact on the asymptotic distribution of $\mathrm{M}(m, l)$. Although the moments of the generalized residuals $\{Z_\tau\}$ are not exactly the same as that of the original data $\{r_{\tau\Delta}\}$, they are highly correlated. In particular, the choice of $(m, l) = (1, 1), (2, 2), (3, 3), (4, 4)$ is very sensitive to autocorrelations in level, volatility, skewness, and kurtosis of $\{r_{\tau\Delta}\}$, respectively (see, e.g., Diebold et al., 1998). Furthermore, the choice of $(m, l) = (1, 2)$ and $(2, 1)$ is sensitive to ARCH-in-mean and leverage effects, respectively. Different choices of orders (m, l) can thus examine various dynamic aspects of the underlying process. Like $\hat{Q}(j)$, upper-tailed N(0,1) critical values are suitable for $\mathrm{M}(m, l)$.

3. Density forecast in affine term structure models

We now apply the evaluation procedure described in Section 2 to the ATSMs, given their important roles in the academic literature and industry practice.

In ATSMs, a finite number of state variables drive the evolution of the whole yield curve. By assuming that the state variables follow affine diffusions, ATSMs can generate rich term structure dynamics while still allowing closed-form pricing for a wide variety of fixed-income securities (e.g., Duffie et al., 2000; Chacko and Das, 2002). Therefore, if these models can provide accurate forecasts of the joint probability density of the state variables, they can forecast the evolution of the whole yield curve and will be useful for managing the large fixed-income holdings of many banks. While ATSMs have been widely studied in the literature, there is little work on testing their out-of-sample forecast performance. Our empirical work will help fill this gap in the literature.

3.1. Affine term structure models

In ATSMs, it is assumed that the spot rate $r(t)$ is an affine function of N latent state variables $X(t) = [X_1(t), X_2(t), \ldots, X_N(t)]'$:

$$r(t) = \delta_0 + \delta' X(t), \tag{8}$$

where δ_0 is a scalar and δ is an $N \times 1$ vector. In the absence of arbitrage opportunities, the time t-price of a zero-coupon bond that matures at $t + \tau_m$ $(\tau_m > 0)$ equals

$$P(t, \tau_m) = \mathrm{E}_t^Q \left[\exp\left(-\int_t^{t+\tau_m} r(s)\, \mathrm{d}s \right) \right],$$

where the expectation $\mathrm{E}_t^Q(\cdot)$ is taken under the risk-neutral measure Q. Thus, the whole yield curve is determined by $X(t)$, which follows an affine diffusion under the risk-neutral measure:

$$\mathrm{d}X(t) = \tilde{\kappa}\left[\tilde{\theta} - X(t)\right]\mathrm{d}t + \Sigma S_t\,\mathrm{d}\tilde{W}(t), \tag{9}$$

where $\tilde{W}_t$ is an $N \times 1$ independent standard Brownian motion under measure Q, $\tilde{\kappa}$, and Σ are $N \times N$ matrices, and $\tilde{\theta}$ is an $N \times 1$ vector. The matrix S_t is diagonal with (i, i)th elements

$$S_{t(ii)} \equiv \sqrt{\alpha_i + \beta_i' X(t)}, \quad i = 1, \ldots, N, \tag{10}$$

where α_i is a scalar and β_i is an $N \times 1$ vector.

Under assumptions (8)–(10),, the yields of zero-coupon bonds, $Y(X_t, \tau_m) \equiv -(1/\tau_m)\log P(X_t, \tau_m)$, are an affine function of the state variables:

$$Y(X_t, \tau_m) = \frac{1}{\tau_m}[-A(\tau_m) + B(\tau_m)' X(t)],$$

where the scalar function $A(\cdot)$ and the $N \times 1$ vector-valued function $B(\cdot)$ either have a closed form or can be easily solved via numerical methods.

Completely ATSMs of Dai and Singleton (2000) assume that the market prices of risk

$$\Lambda_t = S_t \lambda_1, \tag{11}$$

where λ_1 is an $N \times 1$ vector. This implies that the compensation for risk is a fixed multiple of the variance of the state vector and the market prices of risk cannot change signs over time. These restrictions make it difficult to replicate some stylized facts of historical excess bond returns. Duffee (2002) shows that completely ATSMs provide poor forecasts of future bond yields and forecast errors are large when the slope of the term structure is steep. Duffee (2002) extends completely ATSMs to essentially ATSMs by assuming

$$\Lambda_t = S_t \lambda_1 + S_t^- \lambda_2 X(t), \tag{12}$$

where S_t^- is an $N \times N$ matrix with (i, i)th elements

$$S_{t(ii)}^- = \begin{cases} (\alpha_i + \beta_i' X(t))^{-1/2} & \text{if } \inf(\alpha_i + \beta_i' X(t)) > 0, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, N$$

and λ_2 is an $N \times N$ matrix. The essentially ATSMs break down the tight link between the market prices of risk and the variances of the state variables. In particular, they allow the market prices of risk to change signs over time.

Under the specifications of Λ_t in (11) and (12), $X(t)$ is also affine under the physical measure

$$\mathrm{d}X(t) = \tilde{\kappa}\left[\tilde{\theta} - X(t)\right]\mathrm{d}t + \Sigma S_t \Lambda_t\,\mathrm{d}t + \Sigma S_t\,\mathrm{d}W(t),$$

where $W(t)$ is an $N \times 1$ standard Brownian motion under the physical measure.

Dai and Singleton (2000) greatly simplify the econometric analysis of ATSMs by systematically classifying all admissible N-factor ATSMs into $N + 1$ subfamilies,[14]

[14] Admissibility means that $\alpha_i + \beta_i' X(t) \geqslant 0$, for all i and all possible values of $X(t)$.

denoted as $\mathbf{A}_m(N)$, where $m \in \{0, 1, \ldots, N\}$ is the number of the state variables that affect the instantaneous variance of $X(t)$. They also introduce a canonical representation for $\mathbf{A}_m(N)$, which has the most flexible specification within each subfamily, as it either nests or is equivalent (via an invariant transformation) to all the models in $\mathbf{A}_m(N)$.

We follow Dai and Singleton (2000) and Duffee (2002) to consider the canonical forms of the three-factor completely ATSMs $\mathbf{A}_m(3)$, $m = 0, 1, 2, 3$, and essentially ATSMs $\mathbf{E}_m(3)$, $m = 0, 1, 2$. In the canonical representation, Σ is normalized to the identity matrix and the state vector $X(t)$ is ordered so that the first m elements of $X(t)$ affect the instantaneous variance of $X(t)$. Setting $\alpha_i = 0$ for $i = 1, \ldots, m$ and $\alpha_i = 1$ for $i = m+1, \ldots, N$, we have $S_{t(ii)} = X_i(t)^{1/2}$ and $S^{-}_{t(ii)} = 0$ for $i = 1, \ldots, m$, and $S_{t(ii)} = [1 + \beta_i' X(t)]^{1/2}$ and $S^{-}_{t(ii)} = [1 + \beta_i' X(t)]^{-1/2}$ for $i = m+1, \ldots, N$, where $\beta_i = (\beta_{i1}, \ldots, \beta_{im}, 0, \ldots, 0)'$.

As the transition density of an affine model generally has no closed form, MLE is infeasible. Following Duffee (2002), we estimate model parameters via quasi-MLE, which is rather convenient for ATSMs because the conditional mean and variance of $X(t)$ are known in closed form (see Duffee, 2002, for details).[15]

3.2. Dynamic probability integral transform for ATSMs

The key to evaluate multifactor ATSMs is to compute their generalized residuals. Suppose we have a time series observations of the yields of N zero-coupon bonds with different maturities, $\{Y_{\tau\Delta,k}\}_{\tau=1}^{L}$, $k = 1, \ldots, N$. Assuming that the yields are observed without error, given a parameter estimator $\hat{\theta}$ using the estimation sample $\{Y_{\tau\Delta,k}\}_{\tau=1}^{R}$, $k = 1, \ldots, N$, we can solve for the underlying state variables $\{X_{\tau\Delta,k}\}_{\tau=1}^{L}$, $k = 1, \ldots, N$. To examine whether the model transition density $\mathrm{p}(X_{\tau\Delta}|I_{(\tau-1)\Delta}, \theta)$ of $X_{\tau\Delta}$ given $I_{(\tau-1)\Delta} \equiv \{X_{(\tau-1)\Delta}, \ldots, X_{\Delta}\}$ under the physical measure provides accurate forecasts of the joint density of the process $X(t)$, we can test whether the probability integral transforms of $\{Y_{\tau\Delta,k}\}_{\tau=R+1}^{L}$, $k = 1, \ldots, N$, with respect to the model-implied transition density is i.i.d. U[0, 1].

There are different ways to conduct the probability integral transform for ATSMs. Following Diebold et al. (1999), we partition the joint density of the N different yields $(Y_{\tau\Delta,1}, \ldots, Y_{\tau\Delta,N})$ at time $\tau\Delta$ under the physical measure into the products of N conditional densities,

$$\mathrm{p}\Big(Y_{\tau\Delta,1}, Y_{\tau\Delta,2}, \ldots, Y_{\tau\Delta,N}|I_{(\tau-1)\Delta}, \hat{\theta}\Big) = \prod_{k=1}^{N} \mathrm{p}\Big(Y_{\tau\Delta,k}|\,Y_{\tau\Delta,(k-1)}, \ldots, Y_{\tau\Delta,1}, I_{(\tau-1)\Delta}, \hat{\theta}\Big),$$

where the conditional density $\mathrm{p}(Y_{\tau\Delta,k}|\,Y_{\tau\Delta,(k-1)}, \ldots, Y_{\tau\Delta,1}, I_{(\tau-1)\Delta}, \hat{\theta})$ of $Y_{\tau\Delta,k}$ depends on not only the past information $I_{(\tau-1)\Delta}$ but also $\{Y_{\tau\Delta,l}\}_{l=1}^{k-1}$, the yields at $\tau\Delta$ with

[15] We could also use other estimation methods, such as the EMM method of Gallant and Tauchen (1996), the approximated MLE of Ait-Sahalia and Kimmel (2002) and Duffie et al. (2003), the simulated MLE of Pedersen (1995) and Brandt and Santa-Clara (2002), and the empirical characteristic function method of Singleton (2001) and Jiang and Knight (2002).

shorter maturities.[16] We then transform the yield $Y_{\tau\Delta,k}$ via its corresponding model-implied transition density

$$Z_{\tau,k}^{(1)}(\hat{\theta}) = \int_0^{Y_{\tau\Delta,k}} \mathrm{p}\left(y|\, Y_{\tau\Delta,(k-1)}, \ldots, Y_{\tau\Delta,1}, I_{(\tau-1)\Delta}, \hat{\theta}\right) \mathrm{d}y, \quad k = 1, \ldots, N. \tag{13}$$

This approach produces N generalized residual samples, $\{Z_{\tau,k}^{(1)}(\hat{\theta})\}_{\tau=1}^{L}$, $k = 1, \ldots, N$. We can use $\{Z_{\tau,k}^{(1)}(\hat{\theta})\}_{\tau=1}^{R}$ and $\{Z_{\tau,k}^{(1)}(\hat{\theta})\}_{\tau=R+1}^{L}$ to evaluate the in-sample and out-of-sample performances of ATSMs in capturing the dynamics of the kth yield, respectively. For each k, both series should be approximately i.i.d. U[0, 1] under correct model specification.

We can also combine the N generalized residuals $\{Z_{\tau,k}^{(1)}(\hat{\theta})\}_{\tau=1}^{L}$ in (13) in a suitable manner to generate a long sequence, which we may call the combined generalized residuals of an ATSM. Define

$$U = (Y_{\Delta,1}, Y_{\Delta,2}, \ldots, Y_{\Delta,N}, Y_{2\Delta,1}, Y_{2\Delta,2}, \ldots, Y_{2\Delta,N}, \ldots, Y_{L\Delta,1}, Y_{L\Delta,2}, \ldots, Y_{L\Delta,N}).$$

We can then conduct the probability integral transforms of U_τ with respect to the model-implied transition density that depends on all the past yields and contemporaneous yields with shorter maturities:

$$Z_{\tau}^{(2)}(\hat{\theta}) = \int_0^{U_\tau} \mathrm{p}(y|\, U_{\tau-1}, \ldots, U_1, \hat{\theta})\, \mathrm{d}y, \quad \tau = 1, \ldots, LN. \tag{14}$$

We could use $\{Z_{\tau}^{(2)}(\hat{\theta})\}_{\tau=1}^{RN}$ and $\{Z_{\tau}^{(2)}(\hat{\theta})\}_{\tau=RN+1}^{LN}$ to measure the in-sample and out-of-sample performances of ATSMs, respectively. Both series should also be approximately i.i.d. U[0, 1] under correct model specification and can be used to check the overall performance of an ATSM. In contrast, each individual sequence of generalized residuals $\{Z_{\tau,k}^{(1)}\}_{\tau=R+1}^{L}$ in (13) can be used to check the performance of an ATSM in forecasting the probability density of each individual yield.

Because the transition density has no closed form for most ATSMs, we use the simulation methods of Pedersen (1995) and Brandt and Santa-Clara (2002) to obtain an approximation for the transition density. This method is applicable to not only affine diffusion models, but also to other general multivariate diffusion models. We could use other approximation methods mentioned earlier (see footnote 15).

In the empirical analysis below, we will focus on the performance of ATSMs in forecasting the joint conditional density of the state variables under the physical measure. While the conditional density under the risk-neutral measure is more

[16] In general, there are $N!$ ways of factoring the joint transition density of yields with different maturities. In our application, the transition density of the yields of long-term bonds depend on the contemporaneous yields of shorter maturity bonds, because the short end of the yield curve is generally more sensitive to various economic shocks and is more volatile. In fact, one is often interested in knowing how short-term interest rate movements, which may be initiated or changed by the central banks, can be transmitted into long-term interest rate movements.

relevant for the pricing purpose, the density under the physical measure is more important for financial risk management. For example, to calculate the VaR of a large fixed-income portfolio over a certain horizon, one needs to forecast the probability distribution of the value of the portfolio under the physical measure. In ATSMs, a finite number of state variables determine the evolution of the whole yield curve and thus the prices of most fixed-income securities. Consequently, under the physical measure, if we can accurately forecast the joint conditional density of the state variables, we would also be able to forecast the conditional distribution of the prices of most fixed-income securities, which is particularly important for VaR calculation.

4. Empirical results

Our empirical analysis focuses on monthly yields of zero-coupon bonds with 6-month, 2 and 10 year maturities from January 1952 to December 1998, the same data as used in Duffee (2002). The zero-coupon bond yields are interpolated from coupon bond prices using the method of McCulloch and Kwon (1993), whose sample has been extended by Bliss (1997) beyond February 1991. We choose the first half of the sample (from January 1952 to June 1975) as the estimation sample and the second half (from July 1975 to December 1998) as the forecast sample. Fig. 1 displays the time series plots of the level, change, and squared change series of the three yields. It is clear that in the second half of the sample, both the level and change series of the three yields exhibit higher mean and volatility, stronger volatility clustering, and more extreme positive and negative moves.

We estimate the seven ATSMs using the three yields from the first half of the sample via QMLE.[17] In addition, we also consider two simple random walk models, denoted as RW1 and RW2, in which yield changes follow multivariate random walks with correlated increments with and without drift, respectively.[18] Based on the estimated parameters, we calculate the in-sample generalized residuals $\{Z^{(1)}_{\tau,k}(\hat{\theta})\}^{R}_{\tau=1}$ in (13), for the 6-month ($k=1$), 2-year ($k=2$), and 10-year ($k=3$) yields, and the combined generalized residuals $\{Z^{(2)}_{\tau}(\hat{\theta})\}^{RN}_{\tau=1}$ in (14). The in-sample performance of the nine models are measured by $\hat{W}(p)$ for $p=5,10$, and 20 in columns 3–5 of Table 1.

We first examine the overall model performance measured by $\hat{W}(p)$ for the combined generalized residuals. One of the most important results is that all models are overwhelmingly rejected by our tests, suggesting that none of them can adequately capture the full dynamics of the three yields. Among the seven ATSMs, $\mathbf{A}_0(3)$ has the best overall performance with a $\hat{W}(5)$ statistic around 44. The models

[17] We assume that the three yields are observed without error and use them to infer the state variables $X(t)$. Duffee (2002) also includes three other yields which are observed with measurement errors in his estimation.

[18] We also consider random walk models with uncorrelated increments. But they generally have worse performance than the models shown here.

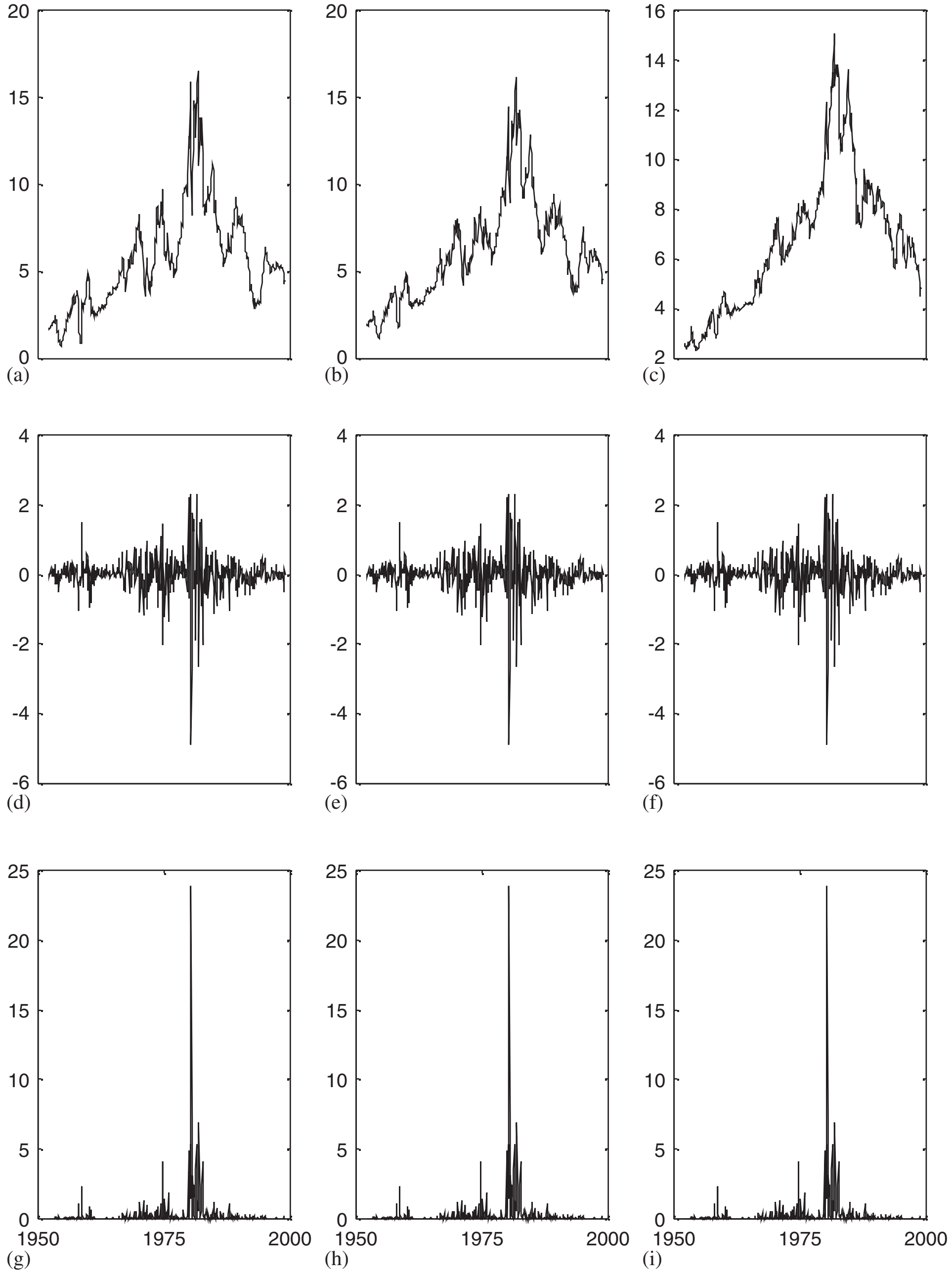

Fig. 1. Time series plots of level, change, and squared change of monthly 6-month, 2- and 10-year yields: (a) 6-month yields; (b) 2-year yields; (c) 10-year yields; (d) change in 6-month yields; (e) change in 2-year yields; (f) change in 10-year yields; (g) squared change (6-month); (h) squared change (2-year); (i) squared change (10-year).

Table 1
Nonparametric portmanteau statistics for in-sample and out-of-sample performance of affine and random walk models

Model	Maturity	In-sample			Out-of-sample		
		$W(5)$	$W(10)$	$W(20)$	$W(5)$	$W(10)$	$W(20)$
$A_0(3)$	Combined	44.12	47.92	57.59	80.05	100.90	129.86
	6-month	19.39	24.55	33.21	64.57	88.87	119.80
	2-year	4.69	5.76	8.51	10.95	14.89	16.50
	10-year	8.80	11.01	14.04	16.07	22.53	28.89
$A_1(3)$	Combined	51.19	56.82	64.37	73.81	91.75	115.47
	6-month	20.15	26.34	35.19	16.01	21.17	29.10
	2-year	6.94	8.08	8.12	12.50	16.73	21.96
	10-year	3.07	4.01	4.74	33.03	43.49	55.90
$A_2(3)$	Combined	69.49	83.35	102.35	149.34	196.82	265.65
	6-month	36.30	49.46	67.67	44.06	61.36	82.35
	2-year	13.96	17.82	22.24	44.04	61.23	83.31
	10-year	6.51	9.35	11.05	57.84	78.41	102.56
$A_3(3)$	Combined	105.73	139.59	188.10	112.81	150.63	198.71
	6-month	69.84	95.23	127.08	44.22	58.93	78.43
	2-year	24.26	32.34	41.00	21.52	29.08	39.24
	10-year	19.02	24.70	33.34	56.86	76.95	100.41
$E_0(3)$	Combined	63.09	70.90	82.77	71.65	83.65	99.77
	6-month	14.78	19.16	27.27	23.69	31.96	42.63
	2-year	11.48	15.21	22.84	12.76	16.23	16.56
	10-year	10.25	13.67	18.73	13.34	18.19	21.83
$E_1(3)$	Combined	51.95	57.43	64.85	57.44	69.39	86.32
	6-month	18.20	23.65	31.42	12.50	16.17	22.47
	2-year	6.28	8.00	8.07	7.38	9.38	13.03
	10-year	2.20	3.39	3.60	27.64	35.90	46.21
$E_2(3)$	Combined	56.27	66.38	80.82	157.06	208.92	283.89
	6-month	28.99	40.27	55.74	40.94	57.04	75.68
	2-year	14.67	18.62	23.84	56.69	78.87	107.06
	10-year	2.08	2.69	2.13	64.44	86.54	112.12
RW1	Combined	50.06	69.12	91.45	118.35	161.87	218.20
	6-month	19.31	25.89	35.53	17.27	22.90	30.94
	2-year	16.37	19.42	22.58	30.99	44.72	63.56
	10-year	26.54	32.84	36.68	82.88	118.55	170.51
RW2	Combined	48.96	80.31	97.79	139.43	181.73	240.66
	6-month	18.31	24.41	33.60	16.03	21.25	28.38
	2-year	15.34	12.60	14.86	34.54	49.70	70.03
	10-year	23.65	15.79	20.57	91.99	130.84	187.31

This table reports the nonparametric portmanteau statistics $W(p)$ defined in Eq. (7), $p = 5, 10$, and 20, for in-sample and out-of-sample performance of completely and essentially affine models and the random walk models. The estimation sample is from January 1952 to June 1975, and the forecast sample is from July 1975 to December 1998. RW1 is a random walk model with correlated increments but no drift. RW2 is a random walk model with correlated increments and drift. The $W(p)$ statistics have a standard normal distribution.

$\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$ have slightly worse performance, with $\hat{W}(5)$ statistics around 51.[19] The models $\mathbf{A}_2(3)$, $\mathbf{E}_2(3)$, and $\mathbf{E}_0(3)$ have $\hat{W}(5)$ statistics range from 56 to 69, and $\mathbf{A}_3(3)$ has the worst performance with a $\hat{W}(5)$ statistic around 105. While the $\hat{W}(5)$ statistics of the two RW models are comparable to those of $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$, their $\hat{W}(10)$ and $\hat{W}(20)$ statistics are much higher. The$\hat{W}(p)$ statistics for the individual generalized residuals in Table 1 show that the models with the best overall performance, such as $\mathbf{A}_0(3)$, $\mathbf{A}_1(3)$, and $\mathbf{E}_1(3)$, also capture each individual yield better than other models. Moreover, most of the above models capture the 2- and 10-year yields much better than the 6-month yields, which could be due to the fact that the 6-month yields are more sensitive to Federal Reserve policies and thus are more volatile than the 2- and 10-year yields.

The relative good in-sample performance of the Gaussian models ($\mathbf{A}_0(3)$ and RWs) are consistent with the well-known trade-offs in ATSMs between modeling the conditional volatilities and correlations of bond yields.[20] The trade-offs suggest that $\mathbf{A}_0(3)$ has the greatest flexibility (difficulty) in modeling the conditional correlation (volatility) of the bond yields due to its Gaussian state variables. In contrast, $\mathbf{A}_3(3)$ is most flexible in modeling the time varying volatility of the bond yields because each of its state variables follows a square-root process. However, to ensure that $X(t)$ is positive, the correlations among the state variables must be positive in $\mathbf{A}_3(3)$. Given that the bond yields in the first half of the sample do not exhibit high-volatility and strong-volatility clustering, the disadvantages (advantages) of $\mathbf{A}_0(3)$ ($\mathbf{A}_1(3)$) become much less important for model performance. The in-sample results also show that the sophisticated market prices of risk in the essentially ATSMs do not necessarily improve the modeling of the conditional density of the bond yields. While certain essentially affine models outperform their completely affine counterparts, the opposite happens for other essentially affine models.

In addition to $\hat{W}(p)$, we also separately examine the $U[0,1]$ and i.i.d. properties of the generalized residuals. Fig. 2 displays the kernel estimators of the marginal densities of the individual and combined generalized residuals of $\mathbf{A}_0(3)$, $\mathbf{A}_1(3)$, $\mathbf{A}_2(3)$, and RW1.[21] The Gaussian models (RW1 and $\mathbf{A}_0(3)$) fail to capture the heavy tails of all three yields: the marginal densities of all the generalized residuals exhibit high peaks at both ends (especially the left end) of the distribution. On the other hand, the marginal densities of the non-Gaussian models ($\mathbf{A}_1(3)$ and $\mathbf{A}_2(3)$) exhibit high peaks in the center of the distribution, particularly for the 6-month and 2-year yields,

[19] For both in-sample and out-of-sample empirical studies, we compare the relative performance between any two models based on their distances to the true data generating process. Strictly speaking, to assess the statistical significance of the relative performance between two potentially misspecified models, we should develop a Diebold and Mariano's (1995) type test. The derivation of the asymptotic distribution for such a test statistic is not trivial in the present context, because nonparametric estimation is involved. The approach by Corradi and Swanson (2004) is expected to be useful here. We leave this to future research.

[20] For more detailed discussions on the trade-offs in ATSMs, see Dai and Singleton (2000).

[21] The marginal densities of $\mathbf{E}_2(3)$ and $\mathbf{A}_3(3)$ are very similar to that of $\mathbf{A}_2(3)$, while the marginal densities of $\mathbf{E}_0(3)$ and $\mathbf{E}_1(3)$ are similar to those of $\mathbf{A}_0(3)$ and $\mathbf{A}_1(3)$, respectively. The two RW models also have very similar marginal densities. Similar results also hold for the second half of the sample.

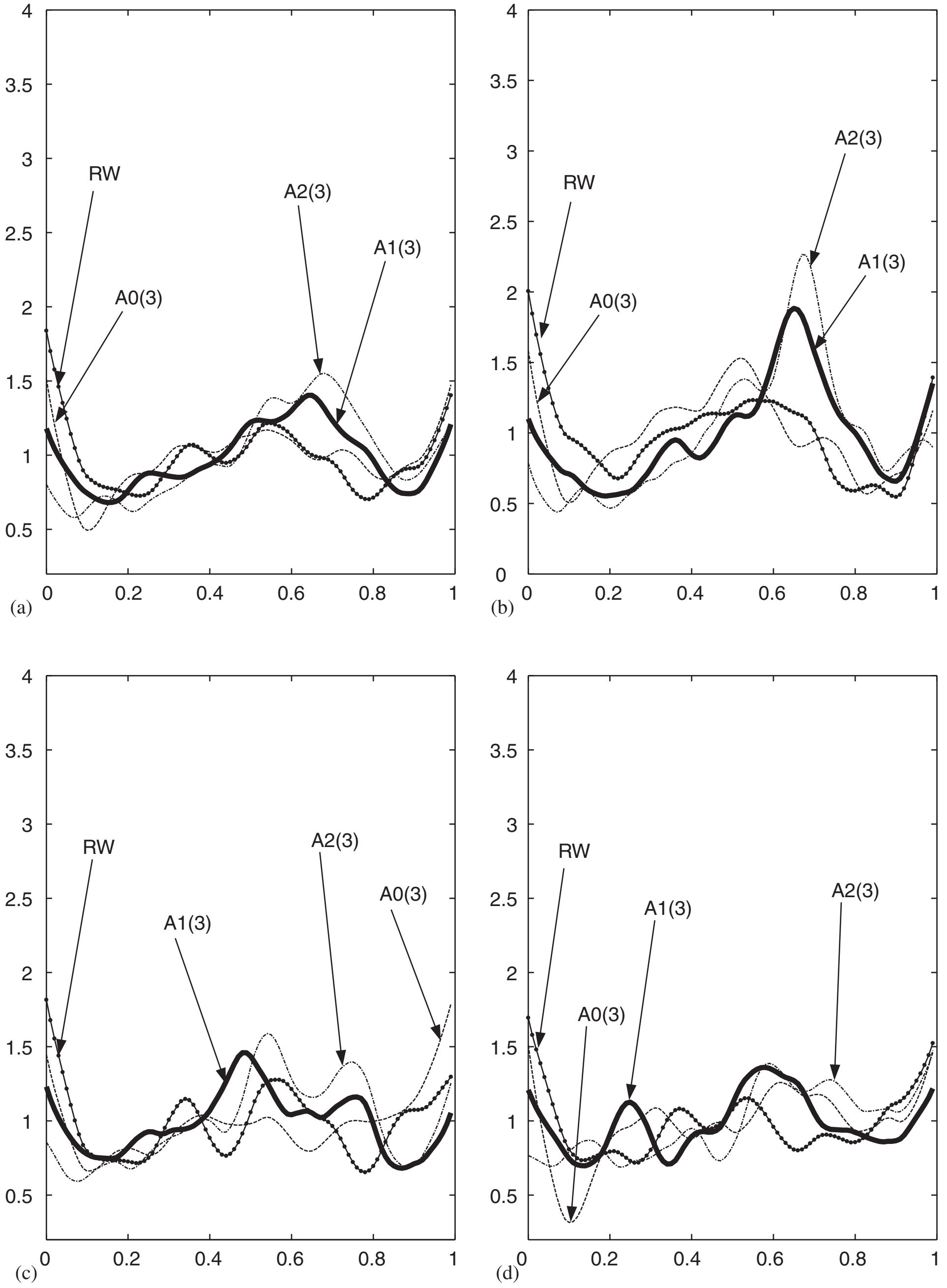

Fig. 2. Nonparametric marginal density of combined and individual generalized residuals (in-sample): (a) marginal density (combined); (b) marginal density (6-months); (c) marginal density (2-years); (d) marginal density (10-years).

suggesting that $\mathbf{A}_1(3)$ and $\mathbf{A}_2(3)$ underpredict the likelihood of small changes. Consistent with the $\hat{W}(p)$ statistics in Table 1, the generalized residuals of $\mathbf{A}_0(3)$ and $\mathbf{A}_1(3)$ are closer to U[0, 1] than those of other models, and in all models the marginal densities of the generalized residuals of the 2- and 10-year yields are much closer to U[0, 1] than that of the 6-month yields.

The $\mathrm{M}(m, l)$ statistics in Panel A of Table 2 summarize the performance of each model in capturing the dynamic dependence of the generalized residuals of all yields. It is obvious that all models, especially the Gaussian models ($\mathbf{A}_0(3)$, $\mathbf{E}_0(3)$, and the RW models), fail to capture the dependence in the conditional variance and kurtosis of the generalized residuals: the M(2, 2) and M(4, 4) statistics are large and overwhelmingly significant for all yields. All models have better performance for the ARCH-in-mean effect (M(1, 2)) and the "leverage" effect (M(2, 1)) for all yields. While the RW models underperform the ATSMs in modeling the conditional variance and kurtosis of the generalized residuals, they capture the dependence in the conditional mean and skewness of the generalized residuals better than the ATSMs, especially for the 6-month yields.

Based on the parameter estimates from the first half of the sample, we calculate the individual and combined generalized residuals using the second half of the sample for each model. These generalized residuals allow us to examine the performance of all models in forecasting the joint probability density of future bond yields. The out-of-sample $\hat{W}(p)$ statistics in columns 6–8 of Table 1 are generally higher that the in-sample counterparts for all models, which should not be surprising given that the models are estimated using only the first half of the sample. The differences between in-sample and out-of-sample performances are quite dramatic for the Gaussian models. While $\mathbf{A}_0(3)$ has the best in-sample performance, it underperforms $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$ in density forecasts. While the RW models have comparable in-sample performance as $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$, they have much worse out-of-sample performance, with $\hat{W}(5)$ statistics between 120 and 140. In contrast, $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$, which are among the best in-sample models, also have the best out-of-sample performance. In fact, $\mathbf{E}_1(3)$ has the smallest out-of-sample $\hat{W}(5)$, which is around 57. As the bond yields become much more volatile and exhibit stronger volatility clustering in the second half of the sample, models that are flexible in capturing both time varying volatilities and correlations of the bond yields, such as $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$, should have better performance. The models $\mathbf{A}_2(3), \mathbf{E}_2(3)$, and $\mathbf{A}_3(3)$ are among the worst out-of-sample models, with $\hat{W}(5)$ statistics range from 120 to 150, which are similar to those of the RW models. Consistent with their overall performance, $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$ also have the best density forecasts for each individual yield. The $\mathbf{E}_1(3)$ model captures the 2-year yields particularly well with $\hat{W}(p)$ statistics in single digits. But it is difficult to capture the 6-month and 10-year yields in all models.

To better understand the possible reasons of suboptimal density forecasts for ATSMs, we separately examine the U[0, 1] and i.i.d. properties of the generalized residuals. Fig. 3 displays the kernel estimators of the marginal densities for the generalized residuals of $\mathbf{A}_0(3), \mathbf{A}_1(3)$, $\mathbf{A}_2(3)$ and RW1. Consistent with the out-of-sample $\hat{W}(p)$ statistics, we find that the generalized residuals become much more nonuniform in the second half of the sample. For example, the high peaks at both

Table 2
Separate inference statistics for affine and random walk models

Model	Maturity	M(1,1)	M(1,2)	M(2,1)	M(2,2)	M(3,3)	M(4,4)
Panel A: *In-sample performance* (*January 1952 to June 1975*)							
$A_0(3)$	6-month	3.39	4.33	0.16	25.73	1.44	20.65
	2-year	0.60	1.17	4.65	29.06	1.55	30.73
	10-year	1.80	−0.74	1.22	23.80	3.44	26.80
$A_1(3)$	6-month	4.91	4.42	−0.33	11.30	2.51	10.92
	2-year	4.70	−0.79	1.70	22.36	−0.004	22.96
	10-year	−0.29	−1.35	−0.24	20.61	−0.21	24.52
$A_2(3)$	6-month	5.07	3.65	1.16	4.97	2.64	4.25
	2-year	0.60	−0.04	0.81	15.37	−0.25	17.42
	10-year	0.22	−0.04	0.66	12.22	−0.31	15.24
$A_3(3)$	6-month	9.03	3.40	0.61	15.03	4.34	11.00
	2-year	0.37	−0.10	2.06	22.01	0.82	15.52
	10-year	−0.06	−1.16	−0.76	15.67	1.35	18.81
$E_0(3)$	6-month	6.69	−1.35	4.90	69.01	7.35	91.57
	2-year	4.19	0.61	−1.06	10.82	2.64	20.73
	10-year	2.64	−0.38	−0.65	30.50	2.41	43.69
$E_1(3)$	6-month	4.16	4.90	−0.01	15.41	2.26	13.34
	2-year	0.72	−0.63	1.38	25.21	0.27	25.25
	10-year	0.45	−1.41	−0.12	23.30	0.02	24.84
$E_2(3)$	6-month	4.37	5.08	1.60	7.06	2.69	5.43
	2-year	0.53	−0.28	1.16	20.26	−0.26	19.08
	10-year	0.51	−0.51	1.14	15.03	−0.25	15.86
RW1	6-month	1.33	3.20	0.17	36.29	0.34	28.89
	2-year	3.20	−0.28	0.34	52.97	6.05	46.44
	10-year	9.09	−0.53	−1.04	31.46	10.73	35.89
RW2	6-month	1.21	3.03	0.058	37.39	0.42	29.99
	2-year	3.20	−0.50	0.43	52.58	6.03	45.81
	10-year	8.80	−0.33	−0.13	28.67	10.79	34.16
Panel B: *Out-of-sample performance* (*July 1975 to December 1998*)							
$A_0(3)$	6-month	8.33	−1.58	8.45	104.59	10.43	131.11
	2-year	3.22	1.17	−0.34	7.10	2.35	16.64
	10-year	0.23	0.58	0.05	31.24	0.04	45.79
$A_1(3)$	6-month	9.77	−0.63	15.51	87.66	12.40	119.07
	2-year	5.77	−0.24	8.12	34.5	5.26	54.23
	10-year	3.55	1.28	14.97	79.26	5.33	92.81
$A_2(3)$	6-month	9.78	0.49	18.78	96.44	9.65	110.07
	2-year	4.21	−0.40	9.70	45.53	5.82	61.57
	10-year	0.96	1.74	12.46	80.88	4.14	84.47
$A_3(3)$	6-month	16.39	−0.94	16.07	81.87	13.81	83.09
	2-year	10.85	−0.41	0.37	1.65	5.09	4.46
	10-year	13.13	0.45	2.68	14.35	5.64	12.46
$E_0(3)$	6-month	3.11	5.02	0.09	28.31	1.10	25.57
	2-year	0.56	1.51	4.25	46.84	0.16	40.86
	10-year	0.52	−0.62	1.08	38.04	1.30	34.14
$E_1(3)$	6-month	11.32	−0.63	15.36	77.48	12.59	106.06
	2-year	3.88	−0.45	8.44	26.11	3.18	45.20
	10-year	2.60	1.96	18.42	73.05	3.46	86.26
E_2 (3)	6-month	11.85	0.07	19.19	86.14	9.56	93.46

Table 2 (*continued*)

Model	Maturity	M(1,1)	M(1,2)	M(2,1)	M(2,2)	M(3,3)	M(4,4)
	2-year	5.27	−0.76	12.65	36.84	5.14	49.46
	10-year	2.40	1.88	13.06	49.77	2.26	49.06
RW1	6-month	4.45	−1.03	−0.11	40.21	3.04	48.29
	2-year	2.27	−1.09	−0.53	0.013	2.23	0.83
	10-year	−0.32	−1.26	0.038	6.13	0.53	8.25
RW2	6-month	4.49	−0.96	0.099	38.51	3.05	48.0
	2-year	2.24	−1.08	−0.49	0.15	2.39	0.96
	10-year	−0.30	−1.24	0.12	4.63	0.55	6.87

This table reports the separate inference statistics M(m,l) in (I.10) for completely and essentially affine models and the random walk models. The statistic M(m,l) can be used to test whether the cross-correlation between the mth and lth moments of $\{Z_{\tau\Delta}\}$ is significantly different from zero. The choice of $(m,l) = (1,1),(2,2),(3,3),(4,4)$ is very sensitive to autocorrelations in mean, variance, skewness, and kurtosis of $\{Y_{\tau\Delta}\}$, respectively. We only show results for lag truncation order $p = 20$, the results for $p = 10$ and 30 are similar. RW1 is a random walk model with correlated increments but no drift. RW2 is a random walk model with correlated increments and drift.

ends of the marginal densities of the RW models become more pronounced, suggesting that the RW models fail to capture the heavy tails of the marginal distribution given that the bond yields exhibit much higher volatility and more extreme positive and negative moves in the second half of the sample. The kernel estimators of the marginal densities of all other ATSMs again exhibit high peaks in the center of the distribution. While all models are far from perfect, $\mathbf{A}_1(3)$ seems to capture the uniform distribution properties of most generalized residuals better than other models.

However, $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$ still have difficulties in modeling the dynamic aspects of the bond yields as measured by the M(m,l) statistics in Panel B of Table 2. For example, all the ATSMs fail to capture all serial dependence in the conditional mean, variance, skewness, and kurtosis of the generalized residuals of all the yields. Interestingly, the two RW models have much better performance in capturing most of the dynamic aspects of the generalized residuals. Except for M$(2,2)$ and M$(4,4)$ for the 6-month yields, the M(m,l) statistics for the RW models are all quite small. This shows that the advantages of the sophisticated ATSMs are in modeling the marginal distribution rather than the serial dependence of the generalized residuals. This is consistent with the important finding of Duffee (2002) that the simple random walk models outperform all the completely ATSMs in predicting the conditional mean of future bond yields.

Our empirical analysis documents some interesting results on the out-of-sample performance of the ATSMs. While the RW models tend to have better forecasts for the conditional mean dynamics of the bond yields, we find that some ATSMs have better forecasts for the probability density of the bond yields. For example, we find that $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$ are not only among the best in-sample models, but also have the best out-of-sample density forecasts. In contrast, although the RW models have

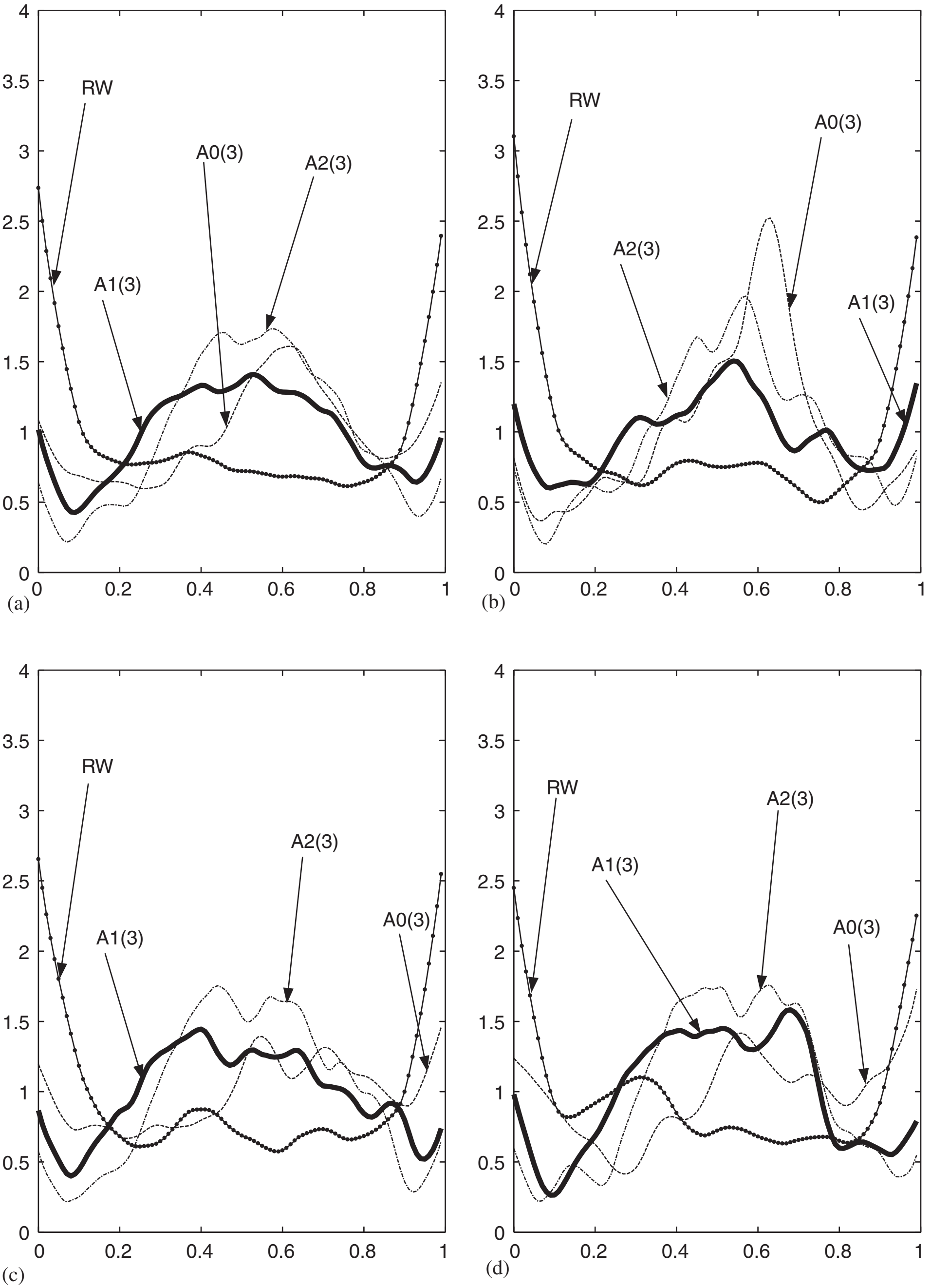

Fig. 3. Nonparametric marginal density of the combined and individual generalized residuals (out-of-sample): (a) marginal density (combined); (b) marginal density (6-month); (c) marginal density (2-years); (d) marginal density (10-years).

comparable in-sample performance as $\mathbf{A}_1(3)$ and $\mathbf{E}_1(3)$, they provide much worse density forecasts. Diagnostic analysis shows that while the best ATSMs capture the marginal densities of the bond yields better, the RW models have advantages in modeling the serial dependence in the bond yields, particularly, the serial dependence of odd-order conditional moments of the generalized residuals. While the answer to the question raised in the title of this paper is a definite "yes," we emphasize that all the ATSMs considered still fail to provide satisfactory density forecasts for the bond yields. Part of the poor performance could be due to the dramatic differences between the two samples, which suggest that regime-switching models might perform better for this data. The reasonably good performance of the RW models in capturing the dynamic dependence in the bond yields suggests that more sophisticated time series models, such as those with more flexible specifications of the error term or dependence structures in various conditional moments, might be able to provide better density forecasts than the affine models.

5. Conclusion

The affine term structure models have become one of the most popular term structure models in the literature due to their rich model specification and tractability. In spite of the numerous empirical studies of the ATSMs, little effort has been devoted to examining their out-of-sample performance in forecasting future bond yields. In this paper, we have contributed to the literature by providing probably the first comprehensive empirical analysis of the performance of three-factor completely and essentially ATSMs in forecasting the joint conditional probability density of bond yields. Density forecasts of bond yields are important for many financial applications, such as pricing and hedging fixed-income securities and managing interest rate risk. Using a new nonparametric omnibus procedure for density forecast evaluation tailored for ATSMs, we find that although the random walk models tend to have better forecasts of the conditional mean of the bond yields, some ATSMs provide better forecasts of the joint probability density of the bond yields. However, all ATSMs we consider are still overwhelmingly rejected and none of them can provide satisfactory density forecasts. There exists room for further improving the density forecasts for future bond yields by extending ATSMs. This is left for future investigation.

Acknowledgments

We would like to thank Torben Andersen, Zongwu Cai, Graham Elliott, Eric Ghysels, David Hendry, Jingzhi Huang, Nick Kiefer, Changjin Kim, Yoon-Jin Lee, Bruce Lehmann, Ernst Schaumburg, George Tauchen, Yoonjae Whang, an anonymous referee, and seminar participants at Cornell University, Korea University, the University of North Carolina at Charlotte, West Virginia University, the 2003 North American Econometric Society Winter Meeting, the 2003 Midwestern

Econometric Society Meeting, the 2003 NBER Time Series Conference, the 2003 European Finance Association Meeting, the 2003 International Symposium on Risk Management at Chinese Academy of Science, the 2004 Mini-Conference on Predictive Methodology and Application in Economics and Finance in honor of Clive Granger, the 2004 Credit Risk and Scoring Conference at Office of Currency Control, U.S. Department of Treasury, Washington D.C., and the 2004 China International Conference in Finance, for helpful comments and discussions. We are responsible for any remaining errors. Hong's research is supported by the National Science Foundation via Grant SES-0111769 and the Changjiang Visiting Scholarship Program by the Chinese Ministry of Education and Xiamen University.

Appendix A. Mathematical appendix

We consider evaluation of density forecasts for a multivariate continuous-time process $Y(t)$ some of whose components are latent variables. Throughout, we use C to denote a generic bounded constant, $|\cdot|$ to denote the usual Euclidean norm. We first provide regularity conditions.

Assumption A.1. Let $(\Omega, \mathscr{F}, P)$ be a complete probability space. (i) $Y(t) \equiv Y(t, \omega)$, where $\omega \in \Omega$ and $t \in [0, T] \subset \mathbb{R}^+$, is a stationary multidimensional continuous-time process with a well-defined transition density that may not have a closed form; (ii) a discrete sample $\{X_{\tau\Delta}\}_{\tau=1}^{L}$ of the observable subvector $X(t)$ of $Y(t)$ is observed, where Δ is a fixed sample interval and L is the sample size.

Assumption A.2. Let $M = \{m(\theta), \theta \in \Theta\}$ be a class of multivariate continuous-time model for $Y(t)$, where Θ is a finite-dimensional parameter space. (i) For each $\theta \in \Theta$, $\mathrm{p}(x, t|I_s, s, \theta)$ is the transition density of the observable subvector $X(t)$ implied by a multivariate continuous-time model $m(\theta)$ for $\{Y(t)\}$, where I_s is the observed information set for $X(t)$ available at time $s<t$; (ii) for each $\theta \in \Theta$, $\mathrm{p}(x, t|I_s, s, \theta)$ is a measurable function of (x, I_s), and there exists $\theta_0 \in \Theta$ such that $\mathrm{p}(x, t|I_s, s, \theta_0)$ coincides with the true transition density of $X(t)$; (iii) with probability 1, $\mathrm{p}(x, t|I_s, s, \cdot)$ is twice-continuously differentiable with respect to θ in a neighborhood Θ_0 of $\theta_0 \in \Theta$; (iv) put $Z_\tau(\theta) = \int_{-\infty}^{X_{\tau\Delta}} \mathrm{p}(x, t|I_{(\Delta-1)\tau}, (\Delta-1)\tau, \theta)\,\mathrm{d}x$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\tau=R+1}^{L} \mathrm{E} \sup_{\theta\in\Theta_0} \left| \frac{\partial}{\partial\theta} Z_\tau(\theta_0) \right|^{2v} \leqslant C$$

for some constant $v>1$ and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\tau=R+1}^{L} \mathrm{E} \sup_{\theta\in\Theta_0} \left| \frac{\partial^2}{\partial\theta\partial\theta'} Z_\tau(\theta_0) \right|^{2} \leqslant C,$$

where $n \equiv L - R$.

Assumption A.3. (i) $G_{\tau-1}(z) \equiv \mathrm{E}[(\partial/\partial\theta)Z_\tau(\theta_0)|Z_\tau(\theta_0)=z, I_{(\tau-1)\Delta}]$ is a stationary measurable function of $(z, I_{(\tau-1)\Delta})$; (ii) with probability 1, $G_{\tau-1}(z)$ is continuously differentiable with respect to z, and $\lim_{n\to\infty} n^{-1}\sum_{\tau=R+1}^{L}\mathrm{E}|G'_{\tau-1}[Z_\tau(\theta_0)]|^2 \leqslant C$.

Assumption A.4. $\{X_{\tau\Delta}, (\partial/\partial\theta)Z_\tau(\theta_0)\}'$ is a stationary strong mixing process with strong mixing coefficient $\alpha(j)$ satisfying $\sum_{j=0}^{\infty}\alpha(j)^{(v-1)/v} \leqslant C$, where $v > 1$ is as in Assumption A.2.

Assumption A.5. $\hat{\theta}_R \equiv \hat{\theta}(\{X_{\tau\Delta}\}_{\tau=1}^{R}) \in \Theta$ is a parameter estimator based on the first subsample $\{X_{\tau\Delta}\}_{t=1}^{R}$ such that $R^{1/2}(\hat{\theta}_R - \theta^*) = \mathrm{O_P}(1)$, where $\theta^* \equiv \mathrm{p}\lim_{R\to\infty}\hat{\theta}_R$ is an interior element in Θ and $\theta^* = \theta_0$ under correct model specification.

Assumption A.6. The kernel function $k: [-1,1] \to \mathbb{R}^+$ is a symmetric, bounded, and twice continuously differentiable probability density such that $\int_{-1}^{1} k(u)\,\mathrm{d}u = 1$, $\int_{-1}^{1} uk(u)\,\mathrm{d}u = 0$, and $\int_{-1}^{1} u^2 k(u)\,\mathrm{d}u < \infty$.

Assumption A.7. (i) The bandwidth $h = cn^{-\delta}$ for $c \in (0,\infty)$ and $\delta \in (0, \frac{1}{5})$, where $n \equiv L - R$; (ii) $n^{\lambda}/R \to 0$, where $\lambda < \max[1-\delta, \frac{1}{2}(1+5\delta), (5-2/v)\delta]$.

Assumption A.8. For each integer $j > 0$, the joint density $g_j(z_1, z_2)$ of the transformed random vector $\{Z_\tau, Z_{\tau-j}\}$, where $Z_\tau \equiv Z_\tau(\theta^*)$ and θ^* is as in Assumption A.5, exists and is continuously differentiable on $[0,1]^2$.

Assumption A.1 is a regularity condition on the data-generating process $\{Y(t)\}$; we allow the stationary multivariate process $Y(t)$ to be time-inhomogeneous with jump components, and some components of $Y(t)$ to be unobservable. Assumptions A.2 and A.3 are regularity conditions on the transition density $\mathrm{p}(x, t|I_s, s, \theta)$ of the observable subvector $X(t)$ implied by a multivariate continuous-time model for $X(t)$. We do not require $\mathrm{p}(x, t|I_s, s, \theta)$ has a closed form. Many methods available in the literature can be used to obtain accurate approximations for the model-implied transition density. Assumption A.4 characterizes temporal dependence in $\{X_{\tau\Delta}, (\partial/\partial\theta)Z_\tau(\theta_0)\}$. The strong mixing condition is often used in nonlinear time series analysis, as is the case here. For the definition of the strong mixing condition, see (e.g.) White (1984, p. 45). We note that although $\{Z_\tau(\theta_0)\}$ is i.i.d. when the multivariate continuous-time model is correctly specified, the sequence of its gradients, $\{(\partial/\partial\theta)Z_\tau(\theta_0)\}$ is no longer i.i.d. In general, the gradient $(\partial/\partial\theta)Z_\tau(\theta_0)$ depends on the past information set $I_{(\tau-1)\Delta}$. Assumption A.5 allows for any in-sample $R^{1/2}$-consistent estimator for θ_0 under correct model specification, which need not be asymptotically most efficient. This provides a convenient procedure to implement our tests, because one can use a suboptimal but computationally simple estimation method. Assumption A.6 is a standard regularity condition on kernel function $k(\cdot)$. Assumption A.7 provides conditions on the bandwidth h and the relative speed between R and n, the sizes of the estimation sample and the prediction sample. We allow the optimal bandwidth rate (e.g., $h \propto n^{-1/6}$) for bivariate kernel estimation. Moreover, we allow the size of the prediction sample, n, to be larger than, or smaller than, or the same as the size of the estimation sample, R. This offers a wide scope of applicability of our procedure, particularly when the whole sample

$\{X_{\tau\Delta}\}_{\tau=1}^{L}$ is relatively small. Finally, Assumption A.8 is a regularity condition under model misspecification.

We now state the asymptotic theory for the $\hat{Q}(j)$ test defined in (4) and the $\hat{W}(p)$ test defined in (6) under the null hypothesis and the alternative hypothesis.

Theorem 1 (*Asymptotic distribution of the out-of-sample evaluation statistic at individual lags*). *Suppose Assumption* A.1–A.7 *in Appendix hold, and let* $R \to \infty$, $n \to \infty$ *as the total sample size* $L \to \infty$. *Then for any given lag* $j > 0$, $\hat{Q}(j) \xrightarrow{d} \mathrm{N}(0,1)$ *when the continuous-time model for* $Y(t)$ *is correctly specified.*

Theorem 2 (*Asymptotic distribution of the out-of-sample portmanteau evaluation statistic*). *Suppose the conditions of Theorem* 1 *hold. Then for any given lag truncation order* p, $\hat{W}(p) \to \mathrm{N}(0,1)$ *in distribution as* $n \to \infty$ *when the continuous-time model for* $Y(t)$ *is correctly specified.*

Theorem 3 (*Asymptotic power of the out-of-sample evaluation statistics*). *Suppose Assumptions* A.1–A.8 *in Appendix hold. Then* (i) $(nh)^{-1}\hat{Q}(j) \to^{\mathrm{p}} V_0^{-1/2} \int_0^1 \int_0^1 [g_j(z_1, z_2) - 1]^2 \, \mathrm{d}z_1 \, \mathrm{d}z_2$ *for any fixed integer* $j > 0$; (ii) *for any sequence of constants* $\{C_n = \mathrm{o}(nh)\}$, *we have* $\mathrm{P}[\hat{W}(p) > C_n] \to 1$, *whenever* Z_τ *and* $Z_{\tau-j}$ *are not independent or* U[0, 1] *at some lag* $j \in \{1, 2, \ldots, p\}$.

Theorems 1 and 2 imply that both $\hat{Q}(j)$ and $\hat{W}(p)$ are asymptotically N(0,1) under correct specification of the continuous-time model for $Y(t)$. Theorem 3 characterizes the asymptotic power properties of $\hat{Q}(j)$ and $\hat{W}(p)$. These tests have asymptotic unit power whenever the generalized residuals Z_τ and $Z_{\tau-j}$ are not independent or not U[0,1] at some lag j. Because $\hat{Q}(j)$ and $\hat{W}(j)$ diverge to positive infinity as $n \to \infty$ under model misspecification, one should use upper-tailed N(0,1) critical values (e.g., 1.65 at the 5% level). Due to space constraint, we omit the details of the proofs for these theorems here. We refer interested readers to Hong and Li (2005) and Hong et al. (2005). The latter consider density forecast evaluations for a univariate time series conditional density model in a discrete-time setting. Although our current multivariate continuous-time framework is more complicated, the proof strategies are similar.[22]

References

Ait-Sahalia, Y., 1996. Testing continuous-time models of the spot interest rate. Review of Financial Studies 9, 385–426.

Ait-Sahalia, Y., 2002. Maximum-likelihood estimation of discretely sampled diffusions: a closed-form approach. Econometrica 70, 223–262.

Ait-Sahalia, Y., Kimmel, R., 2002. Estimating affine multifactor term structure models using closed-form likelihood expansions. Working paper, Princeton University.

Ait-Sahalia, Y., Lo, A., 1998. Nonparametric estimation of state-price densities implicit in financial asset prices. Journal of Finance 53, 499–547.

Andersen, T., Bollerslev, T., Diebold, F.X., 2004. Parametric and nonparametric volatility measurement. Handbook of Financial Econometrics, forthcoming.

[22] Alternatively, proofs of Theorems 1–3 here are available directly from the authors upon request.

Bliss, R., 1997. Testing term structure estimation methods. Advances in Futures and Options Research 9, 197–231.

Brandt, M., Santa-Clara, P., 2002. Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. Journal of Financial Economics 63, 161–210.

Chacko, G., Das, S., 2002. Pricing interest rate derivatives: a general approach. Review of Financial Studies 15, 195–241.

Chapman, D., Pearson, N., 2000. Is the short rate drift actually nonlinear. Journal of Finance 55, 355–388.

Christofferson, P., Diebold, F.X., 1997. Optimal predication under asymmetric loss. Econometric Theory 13, 806–817.

Corradi, V., Swanson, N., 2004. A test for comparing multiple misspecified conditional interval models. Working paper, University of Rutgers.

Corradi, V., Swanson, N., 2005. Predictive density evaluation. Handbook of Economic Forecasting, forthcoming.

Cox, J.C., Ingersoll, J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

Dai, Q., Singleton, K., 2000. Specification analysis of affine term structure models. Journal of Finance 55, 1943–1978.

Dai, Q., Singleton, K., 2003. Term structure dynamics in theory and reality. Review of Financial Studies 16, 631–678.

Diebold, F.X., 2001. Elements of Forecasting, second ed. South-Western Publishing.

Diebold, F.X., Li, C., 2002. Forecasting the term structure of government bond yields. Working paper, University of Pennsylvania.

Diebold, F.X., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business and Economic Statistics 13, 253–265.

Diebold, F.X., Gunther, T., Tay, A., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review 39, 863–883.

Diebold, F.X., Hahn, J., Tay, A.S., 1999. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns of foreign exchange. Review of Economics and Statistics 81, 661–673.

Duffee, G., 2002. Term premia and interest rate forecasts in ATSMs. Journal of Finance 57, 405–443.

Duffie, D., Pan, J., 1997. A yield-factor model of interest Rates. Mathematical Finance 4, 13–32.

Duffie, D., Pan, J., 1997. An overview of value at risk. Journal of Derivatives 4, 13–32.

Duffie, D., Pan, J., Singleton, K., 2000. Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68, 1343–1376.

Duffie, D., Pedersen, L., Singleton, K., 2003. Modeling credit spreads on sovereign debt: a case study of Russian bonds. Journal of Finance 58, 119–160.

Egorov, A., Li, H., Xu, Y., 2003. Maximum-likelihood estimation of time-inhomogeneous diffusions. Journal of Econometrics 114, 107–139.

Elliott, G., Timmermann, A., 2002. Optimal forecast combination under general loss functions and forecast error distributions. Working paper, University of California, San Diego.

Engle, R.F., Manganelli, S., 2004. Caviar: conditional autoregressive value at risk by regression quantiles. Journal of Business and Economic Statistics 22, 367–381.

Fackler, P.L., King, R.P., 1990. Calibration of option-based probability assessments in agricultural commodity markets. American Journal of Agricultural Economics 72, 73–83.

Gallant, A.R., Tauchen, G., 1996. Which moments to match? Econometric Theory 12, 657–681.

Granger, C., 1999. The evaluation of econometric models and of forecasts, Invited Lecture in Far Eastern Meeting of the Econometric Society, Singapore.

Granger, C., Pesaran, M.H., 2000. A decision theoretic approach to forecasting evaluation. In: Chan, W.S., Li, W.K., Tong, H. (Eds.), Statistics and Finance: An Interface. Imperial College Press, London, pp. 261–278.

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, New York.

Hong, Y., 1996. Consistent testing for serial correlation of unknown form. Econometrica 64, 837–864.

Hong, Y., Li, H., 2005. Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. Review of Financial Studies 18, 37–84.

Hong, Y., Li, H., Zhao, F., 2005. Can the random walk model be beaten in out-of-sample density forecasts? Evidence from intraday foreign exchange rates. Working paper, Cornell University.

Jackwerth, J.C., Rubinstein, M., 1996. Recovering probability distributions from option prices. Journal of Finance 51, 1611–1631.

Jiang, G., Knight, J., 2002. Estimation of continuous-time processes via the empirical characteristic function. Journal of Business and Economic Statistics 20, 198–212.

Jorion, P., 2000. Value at Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York.

Lo, A., MacKinlay, A.C., 1989. Data-snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 175–208.

McCulloch, J., Kwon, H., 1993. U.S. term structure data, 1947–1991. Working paper 93–6, Ohio State University.

Meese, R., Rogoff, K., 1983. Empirical exchange rate models of the seventies: do they fit out-of-sample? Journal of International Economics 14, 3–24.

Morgan, J.P., 1996. Risk Metrics—Technical Document, fourth ed. Morgan Guaranty Trust Company, New York.

Pedersen, A.R., 1995. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scandinavian Journal of Statistics 22, 55–71.

Piazzesi, M., 2004. Affine term structure models. Handbook of Financial Econometrics, forthcoming.

Pritsker, M., 1998. Nonparametric density estimation and tests of continuous time interest rate models. Review of Financial Studies 11, 449–487.

Rosenblatt, M., 1952. Remarks on a multivariate transformation. Annals of Mathematical Statistics 23, 470–472.

Rothschild, M., Stiglitz, J., 1970. Increasing risk. I. A definition. Journal of Economic Theory 2, 225–243.

Scott, D.W., 1992. Multivariate Density Estimation, Theory, Practice and Visualization. Wiley, New York.

Singleton, K., 2001. Estimation of affine asset pricing model using the empirical characteristic function. Journal of Econometrics 102, 111–141.

Soderlind, P., Svensson, L., 1997. New techniques to extract market expectations from financial instruments. Journal of Monetary Economics 40, 383–429.

Vasicek, O., 1977. An equilibrium characterization of the term structure. Journal of Financial Economics 5, 177–188.

White, H., 1984. Asymptotic Theory for Econometricians. Academic Press, San Diego.

White, H., 2000. A reality check for data snooping. Econometrica 69, 1097–1127.