KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Selecting a bond-pricing model for trading: benchmarking, pooling, and other issues

Piet Sercu and Tom Vinaimont

DEPARTMENT OF ACCOUNTANCY, FINANCE AND INSURANCE (AFI)

AFI 0612

Selecting a bond-pricing model for trading: benchmarking, pooling, and other issues*

Piet Sercu† and Tom Vinaimont‡

First draft: March 2003; this version: August 2006.
Keywords: Term Structure of Interest Rates, Bonds.
JEL classification: G12; G14

*This is substantially revised version of WP 0337 (KU Leuven FETEW) with the same title. We thank Wolfgang Buehler, Frank de Jong, Marco Lyrio, Isabelle Platten, Linda Van de Gucht, Lambert Vanthienen, and many others at workshops in City University of Hong Kong, Facultés Universitaires NDP, KU Leuven, Universität Mannheim and at the 2004 EFA meeting in Maastricht. All remaining errors are ours.

†K.U.Leuven, Graduate School of Business Studies, Naamsestraat 69, B-3000 Leuven; email: piet.sercu@econ.kuleuven.ac.be.

‡Corresponding Author; Department of Economics and Finance, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong; email: eftomv@cityu.edu.hk

Abstract

Does one make money trading on the deviations between observed bond prices and values proposed by bond-pricing models? We extend Sercu and Wu (1997)'s work to more models and more data, but we especially refine the methodology. In particular, we provide a normal-return benchmark that markedly improves upon the Sercu-Wu ones in terms of noisiness and bias, and we demonstrate that model errors contribute more to the variance of residuals—actual minus fitted prices—than pricing errors made by the market. Trading on the basis of deemed mispricing is profitable indeed no matter what model one uses. But there is remarkably little difference across models, at least when one re-estimates and trades daily; and with pooling and/or longer holding periods the results seem to be all over the place, without any relation with various measures of fit in the estimation stage. We also derive and implement an estimator of how much of the typical deviation consists of mispricing and how much is model mis-estimation or mis-specification. Lastly, we find that pooled time-series and cross-sectional estimation, as applied by e.g. De Munnik and Schotman (1994), does help in stabilizing the parameter, but hardly improves the trader's profits.

Keywords: Term Structure of Interest Rates, Bonds

JEL classification: G12; G14

Selecting a bond-pricing model for trading: benchmarking, pooling, and other issues

I Introduction

Since the late 1970s, term-structure (TS) theory has evolved from qualitative propositions about shapes of interest-rate curves to very specific, non-linear models that price both bonds and derivatives. Following Sercu and Wu (1997), our test of eight such models center on the question how much money can be made by trading on the deviations between observed bond prices and values proposed by bond-pricing models. Sercu and Wu (SW) report that such trading generates abnormal returns. One improvement, in the present paper, is that we extend their work to a longitudinally as well as cross-sectionally larger sample and add more models, especially two-factor models. But the more interesting contributions, we think, concern the methodology and the conclusions. First, we come up with just one benchmark, and it is not biased and is more efficient. SW, in contrast, use three benchmarks of (then) untested validity and efficiency. But when in a trial run we applied the SW trading rule to a-select portfolios (like buying short-term bonds only, or long-term bonds only), we found that some of these naive buy-and-hold strategies seemed to provide abnormal returns too, by SW standards. If a-select portfolios already seem to provide abnormal returns, then the finding that a selective trading rule is profitable sounds less impressive: the cause may just be a flawed benchmark for the normal return. So this prompted us to look for a new benchmark-return strategy that avoids such biases and minimizes noise. A second methodological improvement is that we decompose the deviation between observed prices and a model's fitted prices into (*i*) a pricing error made by the market (and subsequently corrected); and (*ii*) a model specification and estimation error that is rightly ignored by the market, and we are able to show that model errors are more important, in terms of variance, than pricing errors. A third improvement over SW is that we check whether the results from these various trading strategies bear any relation to more conventional statistical criteria one could invoke to compare TS models; we find no such relation. Fourth, we also test whether panel estimation helps. SW estimate the TS models from single-day cross-sections, ignoring the intertemporal constraints on the model's parameters and, thus, arguably demoting these structured TS models to little more than clever curves. If the intertemporal constraints are of any value, mixed (or "panel") models would do better (De Munnik and Schotman, 1994), but we find no such effect. Lastly, like SW, we find that there are moderate abnormal profits to be made from using formal models, of the order of two to four percent per year. Unlike in the earlier SW results, however, no model or group of models seems to do reliably better, and the rankings across models differ a lot depending on the criterion.

In the remainder of this introduction we position our work relative to other empirical TS work, we justify some fundamental choices in the research design, and we outline the paper.

In general, the empirics spawned by (and providing feedback to) theoretical work relate to either the appropriateness of the models' assumptions, or the prices it produces, or its delta's or hedge ratios.

In the first category one strand of studies, illustrated by Chan, Karolyi, Longstaff and Sanders (1992), attempts to accept or reject the stochastic form of the factors put forward in TS theory. Others pragmatically let the data decide on the data-generating process for often-used factors in TS modelling like the short term interest rate, and also try out additional features like non-linearities (Aït-Sahalia, 1996a, 1996b; Stanton, 1997; Chapman and Pearson, 1999) or volatility clustering (Bali, 2003) or regime shifts (Ang and Bekaert, 2002a, 2002b). In a more recent development, some papers attempt to link the factors to macro-economic variables. See for instance Ang and Piazessi (2003), Dewachter and Lyrio (2003), and Rudebusch and Wu (2003).

Work related to the prices produced by these models, rather than to the underlying processes, ranges from analyzing the fit between the term structure to bond prices (as do *e.g.* Chen and Scott, 1993; De Munnik and Schotman, 1994; Brown and Dybvig, 1986; Bliss, 1997; and Eom, Subrahmanyam and Uno, 1998) to determining whether swaps derivative prices calculated from estimated TSs are close to observed market prices. In practice, deriving sound prices for all types of instruments at the same time proves to be a difficult task.

Rather than studying underlying processes or fitted prices, one can also verify the correctness of the deltas or hedge ratios proposed by these models, like *e.g.* Driessens, Klaassen and Melenberg (2003) and Gupta and Subrahmanyam (2001).

Our work fits in the second category—prices—but has linkages to the third strand of empirical work: like studies of hedge ratios, it has a dynamic, intertemporal flavor and adopts the professional user's point of view. So while we do look at the static goodness of fit within cross-sections (and even provide a new measure of flexibility), this is mainly to see whether statistical goodness of fit bears any relation to price-change predictability and practical use.

The raw material we work with is a class of highly liquid Belgian government bonds called *Obligations Linéaires / Lineaire Obligaties* (OLOs). The advantage of a simple instrument is that there is absolute clarity with respect to terms and conditions. True, we could also include derivative products in the analysis. However, the BEF interest-derivatives market was entirely OTC; thus, there is no organized market, no records of transaction prices nor a coherent data set of quotes; and the terms and conditions are not standardized.

The models we select are all closed-form as far as zero-bond prices are concerned. This does limit the range of the work. However, selecting these models makes the estimation procedure in essence straightforward, as there is no need for numerical approximations. A concomitant advantage is that all models can be estimated in essentially the same way, non-linear least squares. While an assessment of whether the estimation procedure influences the performance can be interesting as well, we prefer to keep this outside this particular paper. Within the range of closed-form models we limit our selection to a few one- and two-factor models. Our aspiration is not to cover all possible specifications, but to sample a range of models that differ in terms of complexity and ability to fit the data.

We close our introduction with an outline of the paper. In Section II we present our shortlist of TS models; we describe the data; and we provide some statistical measures on how each of the models fit the bond prices cross-sectionally. In Section III, we determine whether models are able to detect mispricing. This consists of a review and validity check of various measures of normal returns, a regression analysis of abnormal returns, and a decomposition of the residual variance into a pricing error, a model error, and a covariance. Section IV describes the result of various trading strategies and a discussion of the question whether anything is gained by doing the estimation in pooled cross-sections. Section IV connects the results from Sections II and III, and concludes.

II Statistical Fit

A The models

The models we work with are, in order of complexity, (*i*) the cubic spline; (*ii*) two seminal one-factor models, (*iii*) four two-factor models. Most of these are widely known, but to identify the parameter estimates presented below we nevertheless need to agree on a notation. Thus, the key factor processes or equations are presented below.

The Vasicek model. Vasicek (1977) assumes a mean-reverting Gaussian process for the instantaneous interest rate,

$$dr(t) = \alpha\left(\beta - r(t)\right)dt + \sigma dW(t), \tag{1}$$

where $\alpha > 0$ is the mean reversion parameter, β the unconditional mean of r(t), σ the volatility of the spot rate, and $W(t)$ a standard Brownian motion. The price of risk is assumed to be constant.

The Cox-Ingersoll-Ross Model. The second model, by Cox, Ingersoll and Ross (1985), is general-equilibrium in nature. It assumes log-utility investors facing a mean-reverting squareroot process for output, and from these derives a mean-reverting square-root process for the instantaneous rate and an endogenous price of risk. The process for r is

$$dr(t) = \alpha\left(\beta - r(t)\right)dt + \sigma\sqrt{r(t)}dW(t) \quad , \tag{2}$$

where $\alpha > 0$ is the mean reversion parameter, β the unconditional mean of $r(t)$, σ a measure of volatility of the spot rate, and $W(t)$ a standard Brownian motion.

The Richard Model. Starting from the Fisher equation, Richard (1978) assumes that the instantaneous real interest rate (R) and the expected inflation rate (π) each follow a mean-reverting squareroot process:

$$\begin{aligned} dR(t) &= a(R^* - R)dt + \sigma_R\sqrt{R}\, dZ_R(t), \text{ and} & (3)\\ d\pi(t) &= c(\pi^* - \pi)dt + \sigma_\pi\sqrt{\pi}\, dZ_\pi(t). & (4)\end{aligned}$$

The correlation between Z_R and Z_π is assumed to be zero. Actual inflation is expected inflation plus noise, and the nominal rate is the real rate plus expected inflation:

$$\begin{aligned} dP(t)/P(t) &= \pi(t)dt + \sigma_P(\pi, R)dZ_P(t), \text{ and} & (5)\\ r(t) &= R(T) + \pi(t)(1 - \sigma_P^2). & (6)\end{aligned}$$

The Longstaff and Schwartz model. Longstaff and Schwartz (1992) develop a two-factor general equilibrium model of the term structure that builds upon CIR. They take the short-term interest rate and the instantaneous variance of the short-term interest rate as the two driving factors. The mathematical structure is very similar to Richards', though. Initially, Longstaff and Schwartz assume two unobservable state variables, X and Y, which follow squareroot processes,

$$
\begin{aligned}
dX &= (a - bX)dt + c\sqrt{X}dW_2(t), \text{ and} \qquad (7)\\
dY &= (d - eY)dt + f\sqrt{Y}dW_3(t), \qquad (8)
\end{aligned}
$$

and which affect expected returns on investment as follows:

$$\frac{dQ}{Q} = (\mu X + \theta Y)dt + \sigma\sqrt{Y}dW_1(t), \qquad (9)$$

where W_2 is assumed to be uncorrelated with W_1 and W_3. Assuming log utility, expected growth in marginal utility—the instantaneous interest rate—is expected output minus variance of output. Thus,

$$r(t) = \alpha x + \beta y \qquad (10)$$

where $\alpha = \mu c^2$, $\beta = (\theta - \sigma^2)f^2$, $x = X/c^2$, $y = Y/f^2$, $\gamma = a/c$, $\delta = b$, $\eta = d/f^2$ and $\xi = e$. The variance of changes in the short-term interest rate is

$$V(t) = \alpha^2 x + \beta^2 y. \qquad (11)$$

The Balduzzi, Das, Foresi and Sundaram model. Balduzzi, Das, Foresi and Sundaram (2000) develop a two-factor model where the first factor r is the short rate and the second factor, θ, is the mean level of the short rate (in the sense of the long-run level to which the rate is attracted, everything else being the same). The short rate follows the same process as in the Vasicek setting,

$$dr(t) = \kappa(\theta - r)dt + \sigma dW_1(t). \qquad (12)$$

except that it is attracted not to a constant mean but to a moving target, with

$$d\theta(t) = (a + b\theta)dt + \eta dW_2(t), \qquad (13)$$

with a, b and η constants. The two processes can be correlated: $dW_1 dW_2 = \rho dt$. The prices of risk are assumed to be constant.

The Baz and Das model. Baz and Das (1996) extend the Vasicek model by adding a Poisson jump process N(t) with intensity rate λ. The process for the short-term rate in the extended Vasicek jump-diffusion process then becomes:

$$dr(t) = \alpha\left(\beta - r(t)\right)dt + \sigma dW(t) + JdN(t). \tag{14}$$

with α the mean reversion coefficient, β the long-term mean of the short interest rate, and σ the instantaneous volatility. The intensity of the jump is defined by J, which is assumed to be a normal variable with mean θ and a standard deviation of δ. This one-factor model jump-diffusion model can be easily extended when one assumes two orthogonal factors. To that end two similar processes can be defined:

$$\begin{aligned} dy_1(t) &= \alpha_1\left(\beta_1 - y_1(t)\right)dt + \sigma_1 dW_1(t) + J_1 dN_1(t) && (15)\\ dy_2(t) &= \alpha_2\left(\beta_2 - y_2(t)\right)dt + \sigma_2 dW_2(t) + J_2 dN_2(t), && (16)\\ r(t) &= y_1(t) + y_2(t) && (17)\end{aligned}$$

where $dy_1(t)$ and $dy_2(t)$ are independent.

The Cubic Spline. McCulloch (1975) uses the cubic spline to curve-fit the TS. The price of a discount bond with remaining life τ is then given by

$$P(\tau) = a_1\tau + a_2\tau^2 + a_3\tau^3 + \sum_{j=1}^{K} d_i\left[max(\tau - k_j, 0)\right]^3 \tag{18}$$

where k_i are the K knot points or knots. These divide the maturity range into $K+1$ distinct sections, within each of which the TS follows a cubic and where the cubics smoothly join at the knots. The choice of the number of knots and their values is rather arbitrary. For comparability with Sercu and Wu, we set two knots, at 2 and 7 years. The parameters a_1, a_2, a_3, d_1 and d_2 can be estimated by an ordinary linear regression.

This finishes our presentation of the models and their notation; the data to which these models are taken come next.

B Data

The test ground for our selection of term structure models are a class of Belgian government bonds called *Obligations Lineaires / Lineaire Obligaties* (OLOs). OLOs have many advantages relative to ordinary Belgian government bonds. OLO maturities nowadays run up to thirty years and contain no embedded option features. Being registered bonds rather than bearer securities, OLOs are mainly held by corporations, making tax clientele effects less likely. Furthermore, Belgian OLOs are actively traded each working day: there are about twenty market makers obliged to quote on request, with a legal bound on the spread. Transaction costs are therefore low and similar across bonds. When reaching the maturity date individual OLO-lines have on average an amount of BEF 200 billion (EUR 5 billion) outstanding. We obtain the OLO mid-prices from the Central Bank of Belgium. Like the Fed in the US market, the Central Bank of Belgium solicits quotes from all market makers every day, at 3 p.m., and publishes the average mid-quote. The averages are not transaction prices but are based on binding quotes.[1] Our sample contains 29 OLOs in total; the number of available issues on any given date varies between 10 and 19. We choose our sample period to include all trading days between June 1, 1992 to December 13, 1998. The decision is based on minimum cross-sectional sample size and an even maturity range. Before June 1992, too few OLOs were traded to meaningfully fit the different models. Secondly, on December 13, 1998, for the first time a 30-year OLO gets introduced. Before that date, the longest issues were OLOs with 20 years to maturity. The first issue of the 30-year OLO creates a serious gap in time to maturity/duration between the 30-year bond and the bond with the next-longest time to maturity (then 18 years). Limiting the sample to December 1998 also reduces potential influence from the introduction of the common currency in the Euro-zone. Table 1 displays maturity dates and coupon rates of individual bonds in the sample. We also include T-bill data for six maturities (two and four weeks; and 2, 3, 6 and 12 months) to enhance the estimation of the short end of the term structure. The T-bills, however, do not enter the performance tests.

C Estimation of the Term Structure Models

Note that we estimate directly from all available raw coupon-bond data, not from a few zero-coupon interest rates or swap quotes. That is, each coupon-bond price is written as the sum of

[1] Until the early nineties, when official market makers were brought in, Belgian government bonds could be traded through a public limit-order book, but even then virtually all of the volume was off the exchange. SW use these limit-order book prices, but these refer to small transactions and are not time-stamped. The advantage of using midpoint quotes instead of transaction prices is that we need to worry less about bid-ask bounce, non-synchronized data or temporal liquidity shocks creating extra noise in transaction data.

Table 1: Description of the sample

We show time of inclusion in the sample, maturity date and annual coupon. The bonds make up the full set of traded fixed coupon Belgian government OLO bonds in the market for the June 1992 - December 1998 period.

OLO Code	Inclusion in sample	Maturity Date	Coupon
239	June 1, 1992	June 1, 1999	8.25%
245	June 1, 1992	April 5, 1996	10.00%
247	June 1, 1992	August 2, 2000	10.00%
248	June 1, 1992	January 2, 1998	9.25%
249	June 1, 1992	February 28, 1994	9.50%
251	June 1, 1992	March 28, 2003	9.00%
252	June 1, 1992	June 27, 2001	9.00%
254	June 1, 1992	August 29, 1997	9.25%
257	June 1, 1992	October 1, 2007	8.50%
259	June 18, 1992	June 25, 2002	8.75%
260	July 24, 1992	July 30, 1998	9.00%
262	December 17, 1992	December 24, 2012	8.00%
264	April 22, 1993	April 29, 1999	7.00%
265	April 22, 1993	April 29, 2004	7.25%
266	May 11, 1993	May 25, 1997	6.75%
268	July 22, 1993	July 29, 2008	7.50%
270	November 18, 1993	November 25, 1996	6.25%
273	March 24, 1994	March 31, 2005	6.50%
275	June 27, 1994	October 15, 2004	7.75%
278	December 19, 1994	December 22, 2000	7.75%
279	January 23, 1995	December 26, 1997	7.50%
282	June 26, 1995	March 28, 2015	8.00%
283	September 25, 1995	May 15, 2006	7.00%
285	January 29, 1996	March 28, 2001	5.00%
286	September 23, 1996	March 28, 2007	6.25%
287	April 18, 1997	January 22, 2000	4.00%
288	October 24, 1997	March 28, 2008	5.75%

the present values of its pay-outs, each of these present values being specified as the zero-coupon-bond pricing equation of the model that is being considered. In the unpooled estimation, our base case, the procedure is that for each day in the sample we estimate the models cross-sectionally by non-linear least squares, that is, by minimizing the sum of squared errors between observed bond prices and fitted values. The optimization method used is a Marquardt procedure.

Averages and medians of the coefficients and of the implied numbers with a ready economic interpretation, like the long-run asymptotic interest rate, produce acceptable values, as can be seen in Table 2.

As expected from pure cross-sectional regressions, and as documented before by *e.g.* Brown and Dybvig (1985) and De Munnik and Schotman (1994), parameters occasionally turn non-sensical for some subperiods and some models. For instance, estimated implicit variances can be negative. The alternative would be to force specific parameters to behave within theoretical constraints—for instance, ≥ 0 for the variance. However, all too often the outcome then is that the parameter is set equal to the bound. Also, estimation then often turns unstable or the models show absolute inability in fitting the term structure. Nonsensical estimates and unstable solutions tend to mean that the objective function is hardly affected by the parameter. By allowing the parameters to free-range, we are mainly assessing whether the functional form of the model provides a good tool to summarize the term structure. Violations of theoretical constraints do not necessarily mean that this specific model is less useful. Indeed, one of the themes in this paper is to investigate the link between practical usefulness, complexity and fit.

D Goodness of fit, cross-sectional and longitudinal

In this section, we explore the characteristics of the regression residual and proceed by ranking the models according to their ability to fit the coupon bond prices in the market. Summary statistics on the bond-price residuals can be found in Table 3. The tables are not set up per individual bond because of the changing time to maturity as time passes on. Instead, we package the bonds into six simple time-to-maturity portfolios. For every day in the sample, the first portfolio combines residuals from bonds with time-to-maturity not exceeding 1 year at that time. The other groups similarly contain bonds from 1 to 2, 2 to 4, 4 to 8, 8 to 15 and over 15 years time to maturity.

A puzzling feature of the average errors per bracket is that, for all time-to-maturity brackets, all models are unanimous about the direction of mispricing. For example, according to all models, very short term bonds are typically underpriced, 1-to-2 and 2-to-4-year paper underpriced, and so on. Even the *ad hoc* spline, with the fewest restrictions on the TS shape, perfectly agrees with the average errors of the more structured models. Especially CIR has problems with estimating the shortest end of the TS, with an average error of minus 13.3 and and average absolute error of 18 basis points. Our concern, initially, had been that no model would be able to capture the short end of the TS well, characterized as it was by a sharp hump during the 1992 and 1993 turmoil in the Exchange Rate Mechanism. However, the period was, apparently, too short to matter in the average. Instead, across all models, the highest Average Absolute Errors (AAEs) are actually found in the bracket containing bonds with time-to-maturity between eight and fifteen years.

The ranking for overall AAEs and average RMSEs is summarized in table 4, alongside other rankings that will be introduced later. The top three models in terms of fit for both criteria are BDFS, Longstaff-Schwartz and the spline. The worst model with respect to these criteria is CIR's, even though the differences are never staggering. Yet that ranking is totally overturned as soon as we adopt two other measures of goodness of fit that have to do with longitudinal properties. These measures are (*i*) the autocorrelation in the residuals, extracted from the daily cross-sectional regressions and grouped bond by bond into time series; and (*ii*) the average run length, *i.e.* the average number of consecutive days where the residuals of a given bond all have the same sign. Both are measures of persistence of unexplained bond values. Of course, these numbers do not tell us whether any high persistence is due to market inefficiency (the market is slow to realize and correct its mistakes) or model mis-specification (some TS shapes cannot be captured, and since the shapes persist, the apparent pricing errors persist too) or persistent mis-estimation; but the same ambiguity exists with respect to the MSE of a regression. Also noisier estimates would directly lead to less persistent errors, so also from that perspective there is ambiguity..

The autocorrelations are quite high, ranging from 0.74 (Baz-Das) to 0.93-0.94 (Vasicek and the spline), with most other models hovering around 0.85; but this is similar to what others found in quite different data bases (*e.g.* Chen and Scott, 1993, or De Jong, 2000). In the same vein, correcting a mistake requires on average anywhere between 7 days (Baz-Das) and 18 (Vasicek). Also, there is little connection between size and persistence of pricing errors. The spline, which does well in terms of cross-sectional fitting, produces quite persistent errors while CIR, rather bad at fitting across bonds, does relatively better in terms of persistence. The distinct performance of Baz-Das in terms of persistence—the difference with the second best is quite marked—is likewise hard to explain from the MSEs. By and large, apart from the Vasicek outlier, higher residual variance tends to go with lower persistence, consistent with the idea that low persistence may just reflect high noisiness of estimates rather than higher flexibility or better fit.

Still, the size of the autocorrelations and the lengths of runs of same-sign residuals remains disconcerting. It is hard to believe that all of this would be pure market inefficiency; rather, inability to capture twists in the TS seems to be at least as plausible an explanation. This second view could fit in with our earlier observation that all models seem to produce similar errors for bonds in the same time-to-maturity bracket. Tests of what part of the error is market mistakes versus model misspecification are the main subject of the next section.

III Market errors v model errors

In the previous section we established a ranking of the competing models based on the natural belief that smaller errors between model and observed prices translate in better pricing capabilities of that model. Especially for the purpose of pricing options, many potential users of a model would balk if that model misprices the underlying. The fact that there still is a residual would be acceptable if these were random and shortlived deviations caused by, say, transaction costs or stale data and causing, in turn, some random estimation error in the coefficients too. The high autocorrelation of the residuals we observed belies this: the market must have been making fairly persistent pricing errors, or we must have introduced persistent error in specifying and estimating the model. In this section we attempt to quantify the relative importance of these two components.

A Measuring autocorrelation, relevance, and relative variance

In standard microstructure style we let the observed transaction price for bond b deviate from its true equilibrium value by a pricing error, which we allow to be autocorrelated. The fitted price produced by a particular model likewise deviates from the true value by a modeling and estimation error. Thus, the observed residual is a compound of both errors. Conceptually, distinguishing between these components is relevant and possible because in a reasonably efficient market the pricing errors should have a tendency to disappear as overpriced bonds are bid down and underpriced bonds bid up. Thus, pricing errors sooner or later lead to abnormal returns of the opposite sign, which makes them very relevant for the purpose of bond selection. Modeling and estimation errors in our computations, in contrast, were not even known to the market at the time and should not lead any patterns in observed subsequent returns on bonds. They are irrelevant for the purpose of portfolio selection. It is not our ambition to estimate the components separately; still, with a minimum of structure we can rank the variances of the modeling error versus the pricing error; we can also estimate the autocorrelation in the latter. This is achieved using a regression of holding-period returns on beginning-of-period residuals as in SW. The details are as follows.

The basic equations are in logs, with p denoting a log trade price, v the true equilibrium price,

$\hat{p}_m$ the log fitted price for model m, and π and μ the pricing error and model error, respectively:

$$\text{Market price:} \qquad p_{b,t} = v_{b,t} + \pi_{b,t}. \tag{19}$$

$$\text{where } \pi_{b,t} = \rho_b \, \pi_{b,t-1} + \nu_{b,t}, \nu \text{ white noise;} \tag{20}$$

$$\text{Model fitted price:} \qquad \hat{p}_{b,m,t} = v_{b,t} + \mu_{b,m,t}, . \tag{21}$$

Equations (19) and (21) allow us to describe the model residual that acts as the regressor in the SW regression:

$$\begin{aligned} RES_{b,t} &\stackrel{\text{def}}{=} p_{b,t} - \hat{p}_{b,m,t}, \\ &= \pi_{b,t} - \mu_{b,m,t}. \end{aligned} \tag{22}$$

Equations (19) and (20) also provide us with a description of the one-holding-period return:

$$\begin{aligned} p_{b,t+1} - p_{b,t} &= v_{b,t+1} - v_{b,t} + \pi_{b,t+1} - \pi_{b,t}, \\ &= v_{b,t+1} - v_{b,t} + (\rho_b - 1)\pi_{b,t} + \nu_{b,t+1}. \end{aligned} \tag{23}$$

Similarly, for the h-period holding return we get

$$p_{b,t+h} - p_{b,t} = v_{b,t+h} - v_{b,t} + (\rho_b^h - 1)\pi_{b,t} + \sum_{l=1}^{h} \rho^{l-1} \nu_{b,t+l}. \tag{24}$$

Lastly, define the abnormal return as the observed return minus the change in the equilibrium prices:

$$\begin{aligned} AR_{b,t,h} &\stackrel{\text{def}}{=} (p_{b,t+h} - p_{b,t}) - (v_{b,t+h} - v_{b,t}), \\ &= (\rho_b^h - 1)\pi_{b,t} + \sum_{l=1}^{h} \rho^{l-1} \nu_{b,t+l}. \end{aligned} \tag{25}$$

In our modified SW regression of AR_t on the beginning-of-period residual RES_t, the slope can be now decomposed into two factors:

$$\begin{aligned} \beta_{b,m,h} &= \frac{\text{cov}(AR_{b,h}, RES_{b,m})}{\text{var}(RES_{b,m})}, \\ &= -(1 - \rho_b^h)\, \gamma_{b,m}, \\ \text{where } \gamma_{b,m} &= \frac{\text{cov}(\pi_b, RES_{b,m})}{\text{var}(RES_{b,m})}, \end{aligned} \tag{26}$$

The term $(1 - \rho_b^h)$ is a speed of adjustment. It is nonnegative, and approaches unity when $h \to \infty$ or $\rho_b \to 0$, that is, when we allow plenty of time for adjustment or when adjustment is complete in one day. In that light, $\gamma_{b,m}$ measures the long-run adjustment, the fraction of $RES_{b,m}$ that, on average, does get reversed in subsequent adjustments. We accordingly dub this ratio the relevance of the residual. It very literally estimates the coefficient we'd get if pricing errors—the relevant part in the residual—had actually been observable and if we had regressed them on the residuals. We expect γ to be a positive number, making the entire β coefficient negative. Also, when ρ_b is positive, β should become more negative when h increases. By running the regressions for two different values of h we can then obtain a point estimate of ρ_b, and back out γ. One can get a better estimate of ρ and γ by simultaneously considering all βs for one particular bond b via a regression of the form

$$\hat{\beta}_{b,m,h} \quad = \quad \left[\alpha_b^h - 1\right]\left[\sum_{i=1}^{M} \zeta_{b,i} I_{i,m}\right] + \epsilon_{b,m,h}. \tag{27}$$

In the above, the regressors are the holding period, h, and a dummy, $I_{i,m}$, which equals unity when $i = m$. The coefficients α_b and $\zeta_{b,m}$ estimate, respectively, the bond's autocorrelation and the relevance (γ) of model m for bond b. Ideally, one should take into account the covariances between the 567 betas, but how to do this is by no means obvious; we content ourselves with consistent estimates from unweighted NLS, equation per equation,[2] without confidence intervals or t-tests.

While a trader may be content with the value of gamma in itself, to the academic this number also tells something about the relative variance of the two error terms. Notably, if γ exceeds 1/2, then irrespective of the sign or size of the covariance, the variance of the pricing error exceeds the variance of the model error:

$$\begin{aligned} \gamma_{b,m} \quad &= \quad \frac{\operatorname{var}(\pi_b) - \operatorname{cov}(\pi_b, \mu_{b,m})}{\operatorname{var}(\pi_b) - 2\operatorname{cov}(\pi_b, \mu_{b,m}) + \operatorname{var}(\mu_{b,m})} \\ &\gtrless \quad 1/2 \text{ iff } \operatorname{var}(\pi_b) \gtrless \operatorname{var}(\mu_{b,m}). \end{aligned} \tag{28}$$

In all of the the above, the residual is the one observed at the beginning of the holding period. SW also work with residuals lagged once (and more, in fact). One reason is that their prices are transaction prices, thus inducing bid-ask bounce correlated with the next-day return.[3] Our data

[2]So we have 27 regressions, one per bond, with each 21 observations: seven models, and three hs per model.

[3]If the last trade at $t-1$ is at the *bid*, then the residual tends to be low, while the subsequent return starting from that low price tends to be high. This biases β negatively.

being midpoint quotes, the problem does not arise here. The second reason for introducing a lag has to do with the practical usefulness for a trader. The trader interested in the information content of a residual might not be able to instantaneously import all bonds' quotes, run a complicated non-linear regression, and still buy or sell the very moment a quote is provided. Even though building-in a full 24-hour delay vastly exaggerates in the other direction, it is the best we can do with daily data.

To make all this operational, however, we still need to handle the unobservable true price change. In stock-market microstructure studies this term is often be postulated to be white noise, which would allow us to move it to the right-hand side, as part of the total regression error. In bond markets, where volatility is much smaller, changes over time in true expected returns are probably more important, making the true return less white-noisy. Our quest for a good proxy of the true price change is discussed in the next section.

B Normal returns and abnormal returns

To approximate the true return on the bond we consider three approaches, two of them already adopted in SW (1997). Anxious to avoid circularity and modeling errors correlated with the regressor, we do not use the bond-pricing models themselves to tell us what the bond prices should have been.[4] Instead, the general idea is to consider the return on a portfolio of bonds that (*i*) has similar characteristics as the one being investigated, and (*ii*) is well diversified, so that pricing errors become minimal:

---

[4]SW do consider the percentage change in the own-model fitted price. We reject this because the fitted price is part of the residual, thus creating correlation between the measurement error and the regressor. For each model we could, instead, have used each of the six competing models to provide fitted values and hence true-return proxies; this would have mitigated the problem but not solved it as we have clear evidence that the model errors are correlated across models (see Section 1.D) .

A rather different approach, suggested by a discussant, is to construct delta-neutral zero-investment portfolios, with the allegedly underpriced bonds held long and the overpriced bonds short, and setting the weights such that the entire position is delta-neutral. This approach is subject to estimation errors in levels and deltas and does not allow a separate analysis of under- and overpriced bonds. The interpretation is ambiguous, too: we are testing, at the same time, the model's ability to get the bond's price as well as its delta(s) right. We tested the model on passive portfolios and found it to be badly biased: passive, well-diversified portfolios seem to generate abnormal returns under this approach.

The duration-based market model. This benchmark return, proposed by Elton and Gruber (1991) and adopted by SW, is based on the "market model" familiar from stock-market studies. In the bond-market version the bond's market sensitivity or beta is not estimated but computed, notably as the ratio of the duration of the target bond to the duration of the market as a whole. Estimation errors in the duration are minimal, and pricing errors are largely diversified away by taking a wide portfolio as the basis. By construction, this benchmark generates a zero abnormal return for the market as a whole, that is, it is correct on average, across all bonds. Long and short positions can be studied separately. The drawback is that it only works under the well-known duration-model assumptions. Non-parallel shifts, like rotation, may (and do) induce serious errors in the estimated normal returns for short or long bonds separately.

The duration-and-convexity matching portfolio. In this benchmark, proposed by SW, one constructs a mimicking portfolio from three equally weighted subportfolios, each consisting of all available short, middle and long bonds, respectively. The mimicking consists of matching price, duration and convexity of the target bond. Because three subportfolios are used and the problem is linear, the weights for each of the subportfolios are uniquely defined. This model has similar pros and cons as the duration market-model. One difference is that, being a quadratic approximation rather than a linear one, this model is better suited to deal with large shifts. Also, since it uses three portfolios, it will price correctly, on average, each of the three subclasses of bonds rather than just the market-wide average bond. However, it may (and does) still misprice the very short or very long bonds. Also, the three benchmark portfolios, consisting of just one third of the (limited) market, are less well diversified than the market portfolio and, therefore, more subject to measurement error.

The minimum-variance duration-and-convexity matching portfolio. In this third approach we form a matching portfolio not from three pre-determined portfolios but from all individual bonds (except, of course, for the bond that is being studied). The weights x_i for each traded bond i are chosen so as to minimize the variance of the portfolio subject to the constraint that the portfolio weights sum to unity and that the portfolio has the same duration and convexity of the bond that is to be matched.[5] To estimate the covariance matrix of the bonds that enter the portfolio, we use

[5] We also stop bonds from taking up more than one quarter of the portfolio, so that the mimicking portfolios are well-diversified; but it turns out that this restriction is never binding.

60 trading days of historical returns. When a bond does not trade over the full 60-day estimation period *(e.g.* when it was issued very recently) it is not included as a possible candidate for the portfolio. Like the other benchmark portfolios, the minimum-variance replica is bond-specific in that the weights for the matching instruments depend on the bond that is being matched. In addition, there now is a constraint that the own-bond weight is zero. Relative to the previous normal-return model, this approach does look for a portfolio that best resembles the bond to be studied. Duration and convexity are just taken as two conveniently familiar characteristics that heavily rely on time to maturity, and also the minvar approach helps guaranteeing that we pick bonds with similar characteristics. Lastly, since the bond itself is excluded from the menu available for the mimicking, all circularity is now excluded; with the other benchmarks this is not *quite* true as the bond that is being matched is also part of the pre-set baskets that are used in the mimicking.

Each of the three approaches provides us with a portfolio that resembles, by pre-specified standards, the bond that is being investigated. Unlike the bond, its match is well-diversified, so the pricing errors should be much smaller. This is why the return on the matching portfolio is taken as the proxy for the normal return.[6]

We have validity-tested the three methods[7] by examining the "abnormal" returns realized by holding static, a-select portfolios (e.g. an equally weighted portfolio of short-lived bonds). Abnormal returns on such test portfolios measured against the benchmark candidates should on average be close to zero. Figure **??** provides plots of the time series of accumulating abnormal returns generated by the candidate benchmarks for four equally weighted portfolios: the total sample, and the bonds in the 4-8, 8-15, and >15 year brackets. Specifically, every day, each bond is matched using a portfolio of all other bonds, and a tracking error is computed for that bond. The tracking error for all bonds, or for all bonds in a maturity class, is then computed as the average tracking error for each of those bonds. Lastly, these average tracking errors are cumulated over

[6]There is a similarity to stock-market event studies, where one uses, for instance, a beta-weighted mixture of market returns and the risk-free rate as the normal return. Our relative duration plays the same role as the stock's beta. Apart from the constraints, our minvar benchmark would be similar to regressing the stock's excess return on each and every other stock's excess return, and building a portfolio with weights equal to the regression coefficients.

[7]The portfolios used to construct Duration-and-Convexity matches are the T-bills, the 1-to-3-year bonds, and the >3-year issues, as in SW.

time. The duration ratio model does well for the all-bond portfolio, by construction, but rather badly fails the test for subportfolios: after 6 years, the cumulative "abnormal" return on this simple investment strategy peaks at 8% for the short bonds and drops to minus 14% for the long bonds. The results for Sercu and Wu's three-portfolio duration-convexity matched investments are only marginally better. The minimum-variance benchmark, by contrast, prices all time-to-maturity-bracket portfolios correctly and performs equally well for the all-sample portfolio, never drifting farther than one percent from the zero line. We therefore use, in what follows, the minimum-variance benchmark to calculate abnormal returns.

C Regression Tests

Using the normal-return model validated in the preceding section we can now compute abnormal returns on each bond. The next step is to regress the abnormal return for bond i between t and $t+h$ on the relative pricing error observed at the beginning of the holding period ($L=0$) or the day before ($L=1$). In the actual computations we used simple percentages rather than log changes. Thus, if M denotes the value of the match portfolio, the regression is finessed as

$$\begin{aligned} AR_{b,m,h,t} &\stackrel{\text{def}}{=} P_{b,t+h}/P_{b,t} - M_{b,t+h}/M_{b,t+h}, \\ &= \alpha_{b,m,h} + \beta_{b,m,h}\frac{P_{b,t-1-L} - \hat{P}_{b,m,t-1-L}}{P_{b,t-1-L}} + \varepsilon_{i,t}. \end{aligned} \tag{29}$$

Recall from (26) that in the regressions with $L=0$, $\beta_{b,m,h}$ equals $-(1-\rho_b^h)\cdot\gamma_{b,m}$, where $(1-\rho_b^h)\geq 0$ is the adjustment speed over de holding period (h days) and $\gamma\in[0,1]$ measures the relevance of the model's fitted price. The version with $L=1$ provides a lower bound for the relevance of the residual from the point of view of a trader who cannot act instantaneously upon observing a quote. For both $L=0$ and $L=1$ we vary the holding period in the abnormal return from 1- to 10- and 20-day periods, meaning about two and four weeks. We run these 2×3 regressions for each individual bond$\times$model combination and test two specific hypotheses H_1: $\beta=0$ (that is, no relevance or no adjustment); and H_2: $\beta=-1$ (perfect relevance—no model errors—and full adjustment within the holding period).

Table 5 summarizes the regression results (average, mean, significance and sign of the estimates) for 1-, 10- and 20-day holding periods, and for $L=0$ (top part) and $L=1$ (bottom part). For virtually all regressions with respect to one-day holding periods we see negative estimates of β for both immediate and one-day-lagged trading. Most of these are also significant; the rare positive estimates, in contrast, are never significant. Thus, statistically there is an information content and the market does react to it. Algebraically, however, the average immediate one-day reaction coefficients are low—between -0.052 (CIR) and -0.083 (LS)—and the next-day reactions are up to one-half lower again.

If these low one-day immediate reaction coefficients reflect sluggishness in the market rather than a low relevance coefficient, then a low β is good news for a trader. In an attempt to extract from this β coefficient the relevance coefficient γ, we increase the holding period for AR to 10 and 20 days. Average slope coefficients for a two-week holding period are now much more seizable, ranging between -0.20 (BDFS) and -0.28 (LS); and adding another 2 weeks further boosts the coefficients to at least -0.28 (Baz and Das) and occasionally even -0.37 (spline). Thus, the news is good from the trader's point of view. First, 30 percent or more of the observed price discrepancy is relevant in the sense that it gets reflected in the price within one month. And second, the adjustment seems to be slow: even a trader that has to wait a full day before reacting loses a mere 3-5 percent of that 30-plus. On the downside, note that the 20-day return is noisier, too: the relative importance of the initial mispricing shrinks because, over a longer horizon, there are so many other influences affecting the price. This noisiness is reflected in the variability of the 20-day β coefficients across bonds. One indicator of this variability is the number of instances with the wrong (positive) sign for the 20-day-AR regressions; another indicator is the difference between median and mean γ's per model, which can be rather large and all over the place.

Recall that we can extract from the betas an estimated autocorrelation for the bond's pricing error. At 0.89, the average across all bonds is quite high. The median is similar: 0.92. This is higher than the autocorrelation for the residuals, which suggests that modeling errors are less autocorrelated than pricing errors. Also shown in the panel for $L = 0$ is, for each model, the average and median implied gamma. These are noisy estimates, being extrapolations for $h \to \infty$ from the three numbers for $h = \{1, 10, 20\}$, so we rely on means and medians. Average gammas are all below 0.50, and only one median exceeds 0.50 (for the spline), implying that the variance of modeling and estimation errors tends to exceed that of pricing errors.

For the trader, the potentially good news is that, even though model errors dominate pricing errors in terms of variance, still a good part of any observed residual eventually gets reversed in later returns. The economic relevance of all this is still unclear as the initial signals are quite small: 30-40 percent of a 15-bp mispricing is not a large gain. Thus, we need to know how often large gains occur, whether it is worthwhile focusing on large gains only, and so on. These issues are addressed in the next section.

IV Measuring the economic relevance: trading-rule tests

A Base-Case Trading Rules: set-up and results

We construct contrarian portfolios by buying underpriced bonds and selling overpriced bonds. Contrarian strategies are based on the deviation of observed asset prices from their fundamental values. The further an observed asset deviates from its fundamental value, the larger should be the correction and, therefore, the higher the weight that should be assigned to the asset in the contrarian portfolio. In implementing this trading strategy, we set up two basic portfolios, a "buy" portfolio, where weights are assigned to undervalued assets, and a "sell" portfolio that contains overpriced assets. When we construct such a time-$(t-1)$ short or long portfolio p (where $p = s$ (sell) or b (buy)) on the basis of the pricing errors observed at $t-1-L$, with $L = 0$ for instant trading and $L = 1$ for delayed trading, then we set the weight for bond i as follows:

$$w_{p,i,t-1-L} = \frac{RES_{i,t-1-L} D_{p,i,t-1-L}}{\sum_{i=1}^{N_{p,t}} RES_{i,t-1-L} D_{p,i,t-1-L}}, p = b, s \quad , \tag{30}$$

where $RES_{i,t-1-L}$ is the residual for bond i as estimated from the time-$(t-1-L)$ cross-section; $D_{b,i,t-1-L} = 1$ if $RES_{i,t-1-L}$ is positive and 0 otherwise; $D_{s,i,t-1-L} = -1$ if $RES_{i,t-1-L}$ is negative and 0 otherwise; and $\mathrm{N}_{p,t}$ the number of assets in portfolio p. Note that $w_{p,i,t-1-L} \geq 0$ and $\sum_{i=1}^{N_t} w_{p,i,t-1-L} = 1$. The abnormal return of a contrarian strategy can then be measured as

$$AR_{p,t,L} = \sum_{i=1}^{N_t} w_{p,i,t-1-L} D_p AR_{i,t}, \quad , \tag{31}$$

where $p = (b, s)$, D_p is equals 1 when $p = b$ and –1 when $p = s$. This is our base-case setup. In variants discussed in the next section we ignore the smaller signals RES and/or trade less frequently than daily.

Table 6 displays percentage profits from contrarian strategies, cumulative over 6.5 years, for $L = 0$ or 1. All the outcomes are statistically very significant, so t-stats are omitted. Our discussion is centered on the combined payoffs from buying and selling ("$b + s$", in the table), which are obtained by adding the accumulated gains from the long and short positions and expressing them as a fraction of the initial notional value. (Since "$b + s$" is a zero-investment strategy, the resulting percentages are not returns in the usual sense.) The table also provides cumulative abnormal returns for buy and sell separately, but there is little to say about these except that they are usually quite similar, and always statistically indistinguishable.[8]

At this stage we are interested in the base-case numbers only, starting with one-day holding periods and instantaneous trading. Although the pricing models seemed rather different in terms of fit, persistence of mispricing, and reaction coefficients, all models produce very similar "$b + s$" CARs, ranging from 21% to 23% over 6.5 years—about 3% *per annum*. The results are not due to one or two freak episodes; rather, they accumulate steadily over time throughout the period, as can be seen from Figure **??** where the evolution of contrarian profits over time is illustrated for immediate trading. Nor are the results due to a few bonds or to one or two maturity classes: when we group the CARs of individual bonds into the six time-to-maturity brackets used before, we find that each of the brackets contributes positively.

When introducing a one-day lag between signal recognition and the actual trading, CARs drop markedly, by about 11% cumulative: a one-day interval between the signal and the execution of trades yields CARs between 11% and 12.5% in total, *i.e.* about 1.5% *p.a.* True, it is unlikely that professional investors need 24 hours to import the data and run a regression, so that the realistically feasible profits are probably closer to the no-lag profits than to the once-lagged result. Still, the "$L = 0$" results are too optimistic. In the next tests, we try and jazz up the base case by being more selective: should we really react to each signal, no matter how small? Also, how much is lost if we trade every 10 or 20 days rather than daily? It turns out that a good dose of selectivity recuperates half of the revenue that would be lost by waiting a full day. We present that evidence below, after a discussion of the trading-cost issue.

The above results are before costs. There are no records of detailed spreads per market maker or best quotes at any moment, but in those days spreads were of the order of magnitude of 6 bp (of the price). Given an annual churn rate of about 25, two-costs would amount to about 1.5% *p.a.* for a buy or a sell strategy and 3% for $b + s$, which would reduce the base-case strategies to mere break-even propositions at $L = 0$, and loss propositions at $L = 1$. However, the selective

---

[8] The buy results do dominate the sell returns in most cases, but in view of the enormous dependencies across the experiment it is probably dangerous to attach much importance to this observation.

applications return far more, as documented below. More fundamentally, many banks trade for liquidity reasons. So their transaction costs are inevitable and, therefore, irrelevant for our purpose. Given that they have to buy or sell, the message is that it is worth pausing two seconds to run a simple spline regression before the trade. For a portfolio manager who faces random in- and outflows every day, a quick look at the residuals would have added about 1.5% to the annual return.

B Filtering out the smaller discrepancies or revising less often

In the preceding section, the bond weights were proportional to the estimated discrepancy; still, we might be able to improve the results by altogether eliminating the bonds with the smallest residuals. Two obvious reasons are that the expected gain is small anyway (a relevant consideration when trading is costly) and that noise is probably important relative to the signal. More subtly perhaps, if mispricing takes time to disappear, mispricing may also take time to build up; if so, it is better for the trader to wait until the discrepancy is peaking before moving in.

When building our selective portfolios, we again construct two groups, one containing bonds with negative residuals and one including bonds with positive residuals. In each group and for each trading day, we now rank the bonds in terms of the size of the absolute residual. We try out two variants of filtering: the first rule keeps only the bonds with the 50% biggest absolute pricing errors in each group, while the second filter is even more selective and considers only bonds in the top quart of absolute pricing errors. Individual bond weights are then again weighted as indicated in equation (30), except that, of course, more of the Ds are set equal to zero.

In Panel A in Table 6, the second and third lines in each cell provide the CARs from the contrarian strategy based upon the 50% and the 25% loudest signals of each day. Introducing the mild filter (called "50% biggest" in the table) has a positive but unspectacular effect for most models; there are even slightly negative effects for the Vasicek model and the spline. The Richard model benefits most (at lag 0), with CARs increasing by 5%, but for $L = 1$ the effect is far smaller. The jump model by Baz and Das still remains the best performing model, with CARs now up to 27% for lag 0 and 17% for lag 1. When introducing the strong filter (called "25% biggest" in the table), in contrast, outcomes do change dramatically, in some instances almost doubling the CARs for the base case. In contrast to the introduction of a weaker filter, now also the Vasicek model and spline function gain from using the stronger filter. The BDFS model and Richard model benefit most: CARs increase with 14% for immediate trading and now attain a level of 36% (almost 5% *p.a.*). Note as well that CARs remain on average very high even for longer lags.

Many private investors would not bother to evaluate and rebalance their portfolios each and every day. Thus, in this section we also investigate to what extent a reduction in the frequency of trading erodes the abnormal returns of the contrarian strategies and filter rules. In a first experiment we consider a holding period of two weeks. After the trade is made based on the contrarian strategy weights, the portfolio holding remains unchanged for two weeks. At the end of the two-week period, we then identify the then prevailing over- and underpricing and adjust the portfolio accordingly. In a second variant, we consider a holding period of one month. As in the previous sections, we investigate, next to immediate trading, the influence of a one-day difference (lag 1) between the mispricing signal and the actual trade.

Earlier, we showed that mispricing tends to gradually disappear, but with the largest adjustments in the days immediately after the detection of the pricing errors. By rebalancing only one every tenth trading day, for instance, we miss nine out of the ten best days; and in a filtered version of the trading rule, we also hold on to positions that would have been liquidated already if rebalancing had been done on a daily basis. Thus, when considering longer holding periods, and therefore less frequent rebalancing, CARs must inevitably erode. The good news, as shown in Panels B and C of Table 6, is that the effects of rebalancing every two weeks and each month are not dramatic: for the base case without filter, CARs remain positive, in the 8-10% range. Predictably, CARs for monthly revisions are lower than for two-week periods. The difference between starting the period immediately (L=0) and leaving one day in-between (L=0) is relatively small. Again, introducing filters seriously enhances the CARs. By and large, the best performing models are the two-factor models. The spline comes out a clear last, this time.

C To pool or not to pool?

A last variant we discuss is about the estimation stage rather than the trading rule itself. Schotman (1996) remarks that day-by-day cross-sectional regressions generate a lot of variability in the parameters and hence in the implied deltas, which would trigger many (probably pointless) trades for the derivatives desk. One recommended solution is to combine several consecutive cross sections. We implement this with 5- and 20-day pooling. In the economic models we constrain the parameters to be equal across cross-sections if they are assumed to be intertemporally constant.

The risk-free rate, an implied number, notably is left to vary from day to day, and so is the other factor in the two-factor models. For the spline, there is no good theoretic reason to fix some parameters; indeed, when we fix all parameters the results are so atrocious that we do not bother to show them. Lastly, the pooled estimations for the Baz-Das model usually failed utterly to converge. In short, we are now down to five competing models.

The results, as summarized in Table 7, are not encouraging. The general rule is that pooling worsens the returns, and pooling 20 days is worse than 5. There are a few exceptions: BDFS tends to improve marginally, and the combination of filtering 50% with pooling 5 days beats the base-case estimation about half of the time. But in the absence of a good reason why these exceptions would be externally valid, the general conclusion seems to be that pooling does not help for current purposes.

V Conclusion

In this paper we fit a set of term structure models to government bonds.[9] One central question was whether a fixed-income-desk trader who faces an in- or outflow can more or less randomly pick a bond in a desirable time-to-maturity bracket, or instead should take a few minutes or seconds to run a cross-sectional regression. We find she should. In contrast, a trader who wants to swap an overpriced bond for an underpriced one should be selective and heed only clear signals, because for these non-liquidity-driven trades transaction costs are not irrelevant. Still, also for this purpose the regression residuals are useful. Another reliable finding is that there is no good case to be made for pooling, at least for our purpose; rather, the indications are mostly against such pooling. A third result is that duration- or duration-and-convexity matched control strategies are not reliable, at least when they work with pre-set portfolios covering a wide time-to-maturity spectrum. What is needed, instead, is a control portfolio with very similar bonds, like the minimum-variance portfolio we adopt here.

Which model to select, if profitability is the criterion? The models are conspicuous in the similarity of their cumulative abnormal returns, at least for the base case of daily rebalancing.

[9]The Belgian data set is not particular in any way: the findings that Sercu and Wu obtained from similar set of Belgian data, have been confirmed by German data.

For filtered applications and less frequent revisions the results are more divergent, but it remains unclear to what extent this is a reliable result or just a reflection of the higher randomness one expects when there are far fewer trades. While applications in other data may shed light on this, we think that, for anyone hoping for a reliable ranking, the omens are not good. Table 8 summarizes some performance measures, both statistical and economic ones, along with the models' rankings for each of the criteria. A comparison of the spline and the Baz-Das model serves to make our case. In terms of MSE the spline looks near-perfect and Baz-Das way below average; yet these rankings switch almost perfectly when we look at another measure of (in)flexibility, the persistence of the deviations between observed and fitted values. True, low persistence may reflect noisier estimates (a bad) rather than measuring flexibility (a good), thus resolving the apparent contradiction. But there seems to be no easy way to explain the contradiction when we consider economic content rather than statistical fit. On the basis of the regressions one would have anticipated a great future for the spline-based trading rule, as the spline's residuals seemed to come out way ahead in terms of predicting subsequent abnormal returns. Yet the spline does bad in the trading experiments. And Baz-Das does very well there, even though its regression coefficients were about the worst among all models. Thus, identifying an unambiguously outstanding model seems to be a bit of search for the Holy Grail.

One possible explanation why we get such blurred results would be that the signals, the residuals, are highly correlated across models. If that would be true, of course the diagnoses would be very similar across models, and so would be the investment results. But this is not, in fact, the case: correlations of residuals per bond across all models range mostly between 0.20 and 0.40. So the picture really is that all models to some extent pick up genuine mispricing (hence the correlations across models) but overlay it with rather similar and substantial amounts of specification and estimation error, with none of the models really sticking out in that respect. The diagnoses can be rather different across bonds, but still are not consistently superior or inferior across models. The relevance parameters (γ) confirm this. Their averages for all bonds are rather similar across models: about 32-42 percent of prima facie errors are ultimately set right, and the differences between these average gamma's per model look even shakier if one takes into account the variability between the bonds' individual estimated gammas for one given model. All models, in the end, fit a rather flexible nonlinear function through the same set of cross-sections. While some shapes must occasionally go better with specific models, it is not necessarily surprising that no functional form systematically beats the other ones in a long series of cross-sections.

References

Aït-Sahalia, Y., 1996a. Testing Continuous-Time Models of the Term Structure of Interest Rates., The Review of Financial Studies 9, 385-426.

Aït-Sahalia, Y., 1996b. "Nonparametric Pricing of Interest Rate Derivative Securities", Econometrica 64, 527-560.

Ang, A., Bekaert G., 2002a. Short Rate Nonlinearities and Regime Switches. Journal of Economic Dynamics and Control 26, 1243–1274.

Ang, A., Bekaert G., 2002b. Regime Switches in Interest Rates. Journal of Business and Economic Statistics 20 (2002b), 163–82.

Ang, A. and M. Piazzesi. (2003). A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables. Journal of Monetary Economics 50, 745-787.

Balduzzi, P., Das, S.R., Foresi, S., Sundaram, R.K., 2000. Stochastic Mean Models of the Term Structure, in: Jegadeesh N., Tuckman, B. (Eds.), Advanced Fixed-Income Valuation Tools, John Wiley and Sons, pp. 162–189.

Bali, T., 2003. Modeling the Stochastic Behavior of Short-Term Interest Rates: Pricing Implications for Discount Bonds. Journal of Banking and Finance 27, 201–228.

Baz, J., Das, S.R., 1996. Analytical approximations of the term structure for jump-diffusion processes: a numerical analysis. *Journal of Fixed Income*, 78–86.

Bekaert, G., Ang, A, 2003. The Term Structure of Real Rates and Expectedc Inflation. Working Paper.

Bliss, R., 1997. Testing term Structure Estimation Methods. Advances in Futures and Options Research, Vol. 9, 197–231.

Brace, A., Gatarek, D., Musiela, M., 1997. The Market Model of Interest Rate Dynamics. Mathematical Finance 7, 127–155.

Brennan, M.J., Schwartz, E.S., 1979. A Continuous Time Approach to the Pricing of Bonds. Journal of Banking and Finance 3, 133–155.

Brown, S., Dybvig, P., 1986. The Empirical Implication of the Cox, Ingersoll, Ross Theory of the Term Structure of Interest Rates. Journal of Finance 41, 617–630.

Buehler, W., Urig-Homburg, M., Walter, U., Weber, T., 1999. An Empirical Comparison of Forward-Rate and Spot-Rate Models for Valuing Interest-Rate Options. Journal of Finance 54, 269–305.

Chapman D., Pearson, N., 2000. Is the Short Rate Drift Actually Nonlinear? Journal of Finance 55, 355–388.

Chan, K.C., Karolyi, G.A., Longstaff, F.A., Sanders, A.B., 1992. An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. Journal of Finance 47, 1209–1227.

Cox John C., Ingersoll, J.E., Ross, S., 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, 1943–78.

De Jong, F., 2000. Time Series and Cross-section Information in Affine term Structure Models. Journal of Business and Economic Studies 18, 300–314.

De Munnik, J. and Schotman, P., 1994. Cross-Sectional versus Time Series Estimation of Term Structure Models: Empirical Results for the Dutch Bond Market. Journal of Banking and Finance 18, 997–1025.

Driessens, J., Klaassen, P., Melenberg, B., 2003. The Performance of Multi-Factor Term Structure Models for Pricing and Hedging Caps and Swaptions. Journal of financial and quantitative analysis 38, .

Duffie, D., Kan, R., 1996. A Yield-Factor Model of Interest Rates. Mathematical Finance 6, 379–406.

Elton, E.J., Gruber, M.J., 1991. Modern Portfolio Theory and Investment Analysis, 4th ed. Wiley, New York.

Eom, Y.H, Subrahmanyam, M.G., Uno, J., 1998, Coupon Effects and the Pricing of Japanese Government Bonds: An Empirical Analysis. Journal of Fixed Income 8. 69–86.

Gupta, A., Subrahmanyam, M., 2001, Pricing and Hedging Interest Rate Options: Evidence from Cap-Floor Markets. NYU Working Paper.

Heath, D., Jarrow, R., Morton, A., 1992. Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claim Valuation. Econometrica 60, 77–105.

Longstaff, F., Schwartz, E., 1992. Interest-rate Volatility and the term Structure: a Two-Factor General-Equilibrium Model. Journal of Finance 47, 1259–1282.

Richard, S., 1978. An Arbitrage Model of the Term Structure of Interest Rates., Journal of Financial Economics 6 , 33–57.

Rogers, L.C.G., 1995. Which Model of the Term Structure Should One Use? Mathematical Finance 65, 93–115.

Rudebusch, G.D., Wu, T., 2003. A Macro-Finance Model of the Term Structure, Monetary Policy, and the Economy. Working Paper, Federal Reserve Bank of San Francisco.

Sercu, P., Wu, X., 1997. The information content in bond model residuals: An empirical study on the Belgian bond market. Journal of Banking and Finance 21, 685–720.

Stanton, R., 1997. A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk. Journal of Finance 52, 1973–2002.

Vasicek, O., 1977. An Equilibrium Characterization of the Term Structure. Journal of Financial Economics 5 , 177–188.

Table 2: Cross-sectional estimation of the competing models: estimated and derived parameters

Key: Averages and medians of the coefficients and of the implied numbers are given along Average Root Mean Square Errors (RMSE). The majority of the variables in the table are defined in equations (1) - (18), while the rest is defined as follows. The one-factor models are estimated using a parameterized form as in Brown and Dybvig (1986). Therefore, we obtain raw estimated parameters for the Vasicek model: $\phi_0 = r/\alpha$, $\phi_1 = (\alpha\beta - q\sigma)/\alpha^2 - 1/2\left(\sigma^2/\alpha^3\right)$ and $\phi_2 = 1/4\left(\sigma^2/\alpha^3\right)$. The estimated parameters for CIR are $\theta_1 = \left[(\alpha+\lambda)^2 + 2\sigma^2\right]^{.5}$, $\theta_2 = (\alpha + \lambda + \theta_1)/2$ and $\theta_3 = 2\alpha\beta/\sigma^2$. μ refers to the model-derived risk adjusted drift of r_t, while R_L represents the derived long yield. Estimation of the Longstaff-Schwartz model produces $\nu = \lambda + \xi$ with λ the market price of risk of changes in Y. The coefficients λ and ξ obtained from the estimation of the BDFS and Baz-Das model respectively refer to the market price of interest rate risk.

	Vasicek																
	ϕ_0	α	ϕ_1	ϕ_2	r	R_L	μ	σ^2									RMSE
average	0.20483	0.37432	0.37575	-0.01756	0.05288	0.08631	-0.00786	-0.00193									0.0018
Median	0.15103	0.31329	0.26753	-0.02941	0.05281	0.08591	-0.00518	0.00436									0.0015
	Cox-Ingersoll-Ross																
	θ_1	θ_2	θ_3	r	R_L	σ^2	μ										RMSE
average	0.31491	0.24019	1.20002	0.04859	0.08255	0.05943	0.0074										0.0021
Median	0.22633	0.10522	1.08310	0.05036	0.08298	0.01885	0.0081										0.0020
	Richard																
	a	c	σ_R	ϕ_R	σ_π	ϕ_π	R^*	π^*	σ_P	R	π	r					RMSE
Average	-0.00027	0.23	0.030	0.54	0.2900	-35.60	573.36	0.245	0.907	0.069	-0.042	0.059					0.0016
Median	0.00000	0.21	0.001	0.54	0.0005	-0.00012	0.00088	0.244	0.919	0.064	-0.000	0.062					0.0015
	Longstaff-Schwartz																
	α	β	δ	γ	ν	η	r	V	R_L								RMSE
average	0.1055	0.2142	0.0272	1.1847	1.1711	-0.4185	0.0547	0.0096	0.0866								0.0015
Median	0.0535	0.0937	0.0728	0.2563	0.2793	0.0510	0.0472	0.0044	0.0870								0.0013
	Balduzzi-Das-Foresi-Sundaram																
	κ	σ	a	b	η	ρ	λ	r	θ								RMSE
Average	0.1673	0.0012	0.0603	-1.3691	0.0943	2.8690	-1.123	0.0529	0.0337								0.0012
Median	0.1302	-0.0001	0.0636	-.5533	0.0744	0.3071	0.4021	0.04507	.03927								0.0010
	Baz-Das																
	α_1	α_2	ξ_1	ξ_2	σ_1	σ_2	β_1	β_2	θ_1	θ_2	δ_1	δ_2	λ_1	λ_2	Y_1	Y_2	RMSE
Average	1.2402	0.2694	0.5199	-3.7014	-0.0063	0.0054	0.0495	0.0410	-0.0461	0.2756	0.0114	0.1741	0.3384	0.0757	0.1188	0.0618	0.0017
Median	1.2634	0.2531	0.0999	-3.3513	0.0000	0.0051	0.0410	0.0138	-0.0117	0.0168	0.0000	0.0001	0.1659	0.0065	0.0364	0.0333	0.0016
	Cubic Spline																
	a_1	a_2	a_3	d_1	d_2												RMSE
average	-0.05086	-0.00061	-0.00002	0.00018	-0.00018												0.0014
Median	-0.04600	-0.00333	0.00030	-0.00009	-0.00020												0.0013

Table 3: Summary numbers on bond-price residuals from pure cross-sectional estimation, grouped by time-to-maturity.

Key: Bond-price residuals for each model are grouped into time-to-maturity brackets. The summary statistics we show are the Average Error (Avg) and the Average Absolute Pricing Error (AAE) per time-to-maturity bracket. All numbers are in basis points and par value for bonds equals 100.

		vasicek	cir	rich	ls	bdfs	b-d	spline
>3m ≤1y	avg	-2.9	-13.3	-4.4	-3.7	-3.8	-1.4	-4.1
	AAE	8.3	18.0	9.3	7.3	7.3	7.8	6.9
>1y ≤2y	avg	3.1	2.6	0.3	-0.8	0.7	2.7	2.0
	AAE	8.6	11.4	7.1	6.4	6.9	8.2	8.2
>2y ≤4y	avg	1.0	2.1	1.6	1.4	1.3	-4.3	1.9
	AAE	10.5	10.9	9.1	7.4	8.6	14.2	8.7
>4y ≤8y	avg	-4.6	-1.7	0.3	-1.2	-2.2	-4.3	-2.5
	AAE	14.2	16.5	14.3	13.5	13.6	14.2	12.5
>8y ≤15y	avg	7.4	5.3	2.7	3.2	4.1	8.1	2.8
	AAE	24.0	24.2	21.6	21.0	21.6	24.3	18.6
>15y	avg	-10.3	-9.0	-5.4	-5.0	-5.6	-12.4	-1.2
	AAE	16.1	13.5	12.6	12.0	11.1	19.8	7.5
overall	avg	0.19	0.30	0.58	0.17	0.18	0.33	0.26
	AAE	15.6	16.9	14.3	13.3	13.7	15.8	12.4

Table 4: Size and persistence of errors, across models

Key: We show two measures of unexplained variability in prices, the Average Absolute Error (AAE) and the Average Root Mean Square, the average standard deviation of the residuals. Both are measured in basis points. Also shown are the autocorrelation, averaged across bonds, of the time series of residuals per bond extracted from each cross section, and the average run length (in days), where a run is defined as a sequence of days where the residuals have the same sign.

	vasicek	cir	rich	ls	bdfs	b-d	spline
	statistics						
AAE	15.6	16.9	14.3	13.3	13.7	15.8	12.4
ARMSE	17.5	20.5	16.0	14.6	12.0	17.1	13.9
autocorr	0.94	0.85	0.74	0.85	0.86	0.73	0.93
avg runl	17.6	12.2	7.7	14.9	13.9	7.4	17.7
	ranking of models						
AAE	5	7	4	2	3	6	1
RMSE	6	7	4	3	1	5	2
autocorr	7	3	2	4	5	1	6
avg runl	6	3	2	5	4	1	7

Table 5: Regression tests on abnormal returns: market v model errors

Key: We regress $AR_{i,t} = a_{i,t} + b_{i,t}[RES_{i,t-1-L}/P_{i,t-1-L}] + \varepsilon_{i,t}$ with $AR_{i,t}$ = abnormal return for bond i between $t-1$ and $t-1+\Delta$, $\Delta = \{1, 10, 20\}$ days; and $RES_{i,t-1-L}/\mathrm{P}_{i,t-1-L}$ = the bond's L-days-lagged relative pricing error, $L = \{1, 2\}$ days. Entries like "pos 27(19)" mean that 27 coefficients were positive, whereof 19 significantly so. The line "meta-γ refers to the second-pass regression analysis where a bond's 21 betas for $L = 1$ are decomposed into a speed of adjustment and a relevance; we show, for each model, the average and median γ across all bonds.

Panel A: Instant Reaction ($L = 0$)

	Vasicek (1F)		CIR (1F)		Richard (2F)		LS (2F)		BDFS (2F)		Baz-Das (2F)		Spline	
1 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.058	5.30E-07	-0.052	1.60E-05	-0.062	2.22E-05	-0.083	6.70E-06	-0.064	1.70E-05	-0.056	9.82E-06	-0.076	1.69E-05
median	-0.037	8.60E-06	-0.040	8.70E-06	-0.041	7.94E-06	-0.039	3.20E-06	-0.041	9.70E-06	-0.038	8.62E-06	-0.054	5.60E-06
# neg	27(19)	12(4)	27(19)	8(1)	27(21)	9(1)	27(19)	11(3)	25(19)	8(1)	26(19)	9(2)	27(20)	12(3)
# pos	0(0)	15(2)	0(0)	19(2)	0(0)	18(2)	0(0)	16(4)	2(0)	19(3)	1(0)	18(2)	0(0)	15(4)
10 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.232	1.60E-05	-0.249	2.77E-05	-0.223	5.07E-05	-0.281	3.74E-05	-0.208	4.15E-05	-0.216	3.24E-05	-0.275	1.21E-05
median	-0.226	5.51E-05	-0.273	7.22E-05	-0.213	6.05E-05	-0.240	5.64E-05	-0.218	8.84E-05	-0.213	6.68E-05	-0.281	3.07E-05
# neg	24(20)	11(5)	27(26)	12(3)	26(21)	11(4)	25(19)	11(3)	23(20)	11(3)	25(20)	9(4)	24(21)	12(5)
# pos	3(0)	16(8)	0(0)	15(10)	1(1)	16(10)	2(0)	16(7)	4(0)	16(10)	2(0)	18(10)	3(0)	15(9)
20 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.311	3.83E-05	-0.362	3.80E-05	-0.311	7.63E-05	-0.353	5.86E-05	-0.298	5.62E-05	-0.279	5.45E-05	-0.373	2.4E-06
median	-0.319	1.47E-04	-0.369	1.50E-04	-0.270	1.38E-04	-0.367	8.73E-05	-0.358	1.64E-04	-0.293	1.47E-04	-0.394	7.3E-05
# neg	23(19)	12(5)	27(21)	12(3)	23(20)	11(6)	22(20)	11(5)	22(20)	10(5)	22(20)	9(4)	23(20)	12(7)
# pos	4(1)	15(11)	0(0)	15(12)	4(1)	16(11)	5(1)	16(10)	5(3)	17(11)	5(1)	18(11)	4(2)	15(10
meta-γ	average	median	average	median	average	median	average	median	average	median	average	median	average	median
	0.32	0.42	0.40	0.47	0.38	0.50	0.31	0.37	0.40	0.42	0.26	0.47	0.42	0.57

Panel B: Delayed Reaction ($L = 1$)

	Vasicek (1F)		CIR (1F)		Richard (2F)		LS (2F)		BDFS (2F)		Baz-Das (2F)		Spline	
1 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.028	3.20E-06	-0.032	7.30E-06	-0.029	8.27E-06	-0.042	8.40E-06	-0.028	8.70E-06	-0.028	2.87E-06	-0.037	4.90E-06
Median	-0.027	8.60E-07	-0.027	5.30E-06	-0.025	7.25E-06	-0.027	4.10E-06	-0.025	6.80E-06	-0.027	5.41E-06	-0.034	6.50E-07
# neg	23(11)	13(1)	27(17)	9(1)	25(8)	10(1)	27(11)	12(2)	25(9)	11(1)	26(11)	9(1)	24(12)	13(2)
# pos	4(0)	14(0)	0(0)	18(1)	2(0)	17(1)	0(0)	15(3)	2(0)	16(3)	1(0)	18(1)	3(0)	14(2)
10 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.181	2.09E-05	-0.209	1.21E-05	-0.170	3.09E-05	-0.206	3.75E-05	-0.154	2.74E-05	-0.163	2.49E-05	-0.209	5.97E-07
Median	-0.210	6.65E-05	-0.226	6.39E-05	-0.184	6.30E-05	-0.211	4.46E-05	-0.145	7.84E-05	-0.188	6.16E-05	-0.240	4.09E-05
# neg	23(18)	12(4)	27(21)	12(3)	24(18)	11(2)	23(16)	11(3)	22(17)	10(3)	23(19)	9(3)	23(18)	12(5)
# pos	4(0)	15(7)	0(0)	15(10)	3(1)	16(7)	4(0)	16(8)	5(1)	17(8)	4(0)	18(9)	4(1)	15(8)
20 day	b	a	b	a	b	a	b	a	b	a	b	a	b	a
average	-0.260	4.27E-05	-0.322	3.09E-05	-0.256	5.98E-05	-0.269	6.15E-05	-0.242	4.51E-05	-0.229	4.79E-05	-0.303	-8.2E-06
Median	-0.301	1.13E-04	-0.324	1.51E-05	-0.273	1.23E-04	-0.315	8.04E-05	-0.297	1.27E-04	-0.254	1.07E-04	-0.338	5.7E-05
# neg	23(18)	11(5)	27(21)	13(3)	23(18)	11(5)	22(18)	10(5)	21(20)	11(6)	22(20)	11(4)	23(19)	12(7)
# pos	4(3)	16(11)	0(0)	14(12)	4(2)	16(10)	5(1)	17(10)	6(4)	16(11)	5(2)	16(11)	4(3)	15(10)

Table 6: Cumulative Abnormal Returns for trading Strategies, in percent

Key: In the base case, all bonds are held (short or long depending on the sign of the initial or lagged mispricing), while in the filtered versions only the top 50% or 25% of the mispricing signals are acted upon, the rest is ignored. The best among the buy strategies and the best among the sell strategies of a given row are indicated by sharps ($^\sharp$); the worst buy and sells are indicated by flats ($^\flat$).

Instant Reaction (*L*=0)

	One-factor models						Two-factor models												(n.a.)		
	Vasicek			CIR			Richard			LS			BDFS			Baz-Das			Spline		
	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell
	Panel A: One-day holding period																				
base case	21.8	11.5	10.3	21.4	11.8	$^\flat$9.5	22.8	$^\sharp$13.0	9.8	21.6	11.6	9.9	21.6	$^\flat$11.3	10.3	23.4	12.7	$^\sharp$10.8	21.6	11.6	9.9
50% biggest	20.0	$^\flat$10.3	9.8	25.0	13.4	11.6	28.0	$^\sharp$15.4	12.6	25.8	13.0	12.8	25.4	12.4	13	27	13.8	$^\sharp$13.2	20.8	11.2	$^\flat$9.6
25% biggest	34.8	$^\flat$16.4	$^\sharp$18.5	33.8	18.4	15.3	34.6	$^\sharp$20.1	$^\flat$14.5	32.8	17.4	15.5	35.6	17.7	18	34.8	17.5	17.3	33.2	17.4	15.7
	Panel B: Two-week Holding Period																				
base case	11.2	6.5	4.6	11.0	6.4	4.7	10.8	$^\sharp$6.8	4.1	10.8	6.4	4.4	10.2	$^\flat$5.6	4.7	12.4	6.6	$^\sharp$5.9	9.8	5.9	$^\flat$4.0
50% biggest	10.0	5.6	4.5	12.6	6.7	5.9	13.0	7.5	5.5	12.8	7.0	5.9	12.0	6.6	5.4	14.6	$^\sharp$8.1	$^\sharp$6.5	9.2	$^\flat$4.9	$^\flat$4.4
25% biggest	17.0	10.6	6.5	15.2	9.1	$^\flat$6.0	17.0	$^\sharp$10.9	6.2	17.8	10.2	7.6	16.6	9.9	6.7	18.4	10.5	$^\sharp$7.8	14.2	$^\flat$7.7	6.6
	Panel C: One-Month Holding Period																				
base case	8.0	4.9	3.1	8.8	5.2	3.6	8.4	$^\sharp$5.4	3.1	7.8	4.7	3.1	7.8	$^\flat$4.2	3.5	9.0	4.9	$^\sharp$4.0	7.4	4.6	$^\flat$2.7
50% biggest	7.2	4.1	$^\flat$3.1	10.2	5.6	$^\sharp$4.6	10.2	5.9	4.3	9.2	4.8	4.4	9.2	5.1	4.0	10.6	$^\sharp$6.1	4.4	6.8	$^\flat$3.6	$^\flat$3.1
25% biggest	11.8	7.2	4.5	11.4	7.2	$^\flat$4.1	13.0	$^\sharp$8.3	4.7	12.4	6.7	$^\sharp$5.7	12.4	7.6	4.8	13.2	7.9	5.3	10.4	$^\flat$5.4	4.9

Delayed Reaction (*L*=1)

	Vasicek			CIR			Richard			LS			BDFS			Baz-Das			Spline		
	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell
	Panel A: One-day holding period																				
base case	13.0	6.9	6.0	13.2	7.6	$^\flat$5.6	13.6	$^\sharp$8.0	$^\flat$5.6	13.4	7.0	6.4	12.8	$^\flat$6.4	6.4	14.6	7.9	$^\sharp$6.8	13.4	7.0	6.4
50% biggest	10.8	5.7	$^\flat$5.2	16.2	$^\sharp$9.3	6.9	15.8	8.8	7.1	15.6	7.5	8.0	15.4	7.8	7.6	17.2	8.8	$^\sharp$8.5	10.8	$^\flat$5.5	5.3
25% biggest	19.0	10.8	8.1	17.8	12.0	$^\flat$5.8	22.0	$^\sharp$12.5	9.5	21.8	10.0	$^\sharp$11.8	17.2	9.4	7.9	17.8	9.4	8.3	14.8	$^\flat$8.8	6.1
	Panel B: Two-week Holding Period																				
base case	9.6	5.8	3.8	9.8	5.7	4.1	9.4	$^\sharp$5.9	3.4	9.4	5.6	3.7	8.8	$^\flat$4.8	4.0	10.6	5.6	$^\sharp$4.9	8.2	5.0	$^\flat$3.3
50% biggest	8.6	4.8	$^\flat$3.7	11.0	5.9	5.2	11.2	6.5	4.7	11.2	6.1	5.1	10.4	5.8	4.6	12.4	$^\sharp$7.0	$^\sharp$5.4	7.8	$^\flat$4.0	3.8
25% biggest	14.4	$^\sharp$9.5	5.0	12.6	7.9	$^\flat$4.7	14.4	9.3	5.1	15.4	8.9	$^\sharp$6.5	14.2	8.6	5.6	15.6	9.2	$^\sharp$6.5	11.8	$^\flat$6.2	5.6
	Panel C: One-Month Holding Period																				
base case	7.2	4.5	2.8	8.0	4.8	3.3	9.4	$^\sharp$5.0	3.4	7.0	4.3	2.7	7.0	$^\flat$3.8	3.1	8.0	4.4	$^\sharp$3.6	6.4	4.1	$^\flat$2.3
50% biggest	6.6	3.8	$^\flat$2.7	9.4	5.2	4.2	9.2	5.4	3.7	8.2	4.3	$^\sharp$3.9	8.2	4.7	3.5	9.6	$^\sharp$5.7	$^\sharp$3.9	6.0	$^\flat$3.2	$^\flat$2.7
25% biggest	10.6	6.7	3.8	10.2	6.7	$^\flat$3.4	11.6	$^\sharp$7.6	4.1	11.0	6.0	$^\sharp$5.1	11.0	7.0	4.0	11.8	7.3	4.5	9.0	$^\flat$4.7	4.3

Table 7: Cumulative Abnormal Returns for trading Strategies, in percent: pooled versus unpooled

Key: In the base case, all bonds are held (short or long depending on the sign of the initial or lagged mispricing), while in the filtered versions only the top 50% or 25% of the mispricing signals are acted upon, the rest is ignored. The best among the buy strategies and the best among the sell strategies of a given row are indicated by sharps ($^\sharp$); the worst buy and sells are indicated by flats ($^\flat$).

filter	pooling	Vasicek			CIR			Richard			Longstaff-Schwartz			BDFS		
		b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell	b+s	buy	sell
Panel A: One-day holding period																
base case	none	21.8	11.5	10.3	21.4	11.8	9.5	22.8	$^\sharp$13.0	9.8	21.2	11.6	9.9	21.6	11.3	10.3
	5 days	18.9	9.1	9.8	19.0	9.4	9.6	21.3	12.0	9.3	21.2	10.6	10.6	20.8	9.9	10.9
	20days	16.8	$^\flat$7.6	9.3	14.4	$^\flat$7.6	$^\flat$6.8	22.9	11.8	$^\sharp$11.1	19.3	10.0	9.4	21.2	10.7	10.5
50% biggest	none	20.0	10.3	9.8	25.0	13.4	11.6	28.0	$^\sharp$15.4	12.6	25.8	13.0	12.8	25.4	12.4	13.0
	5 days	23.9	11.1	12.8	22.9	10.7	12.2	24.9	14.1	10.8	26.8	13.4	13.4	26.1	12.2	$^\sharp$13.9
	20days	20.7	$^\flat$8.4	12.3	18.8	11.1	$^\flat$7.8	26.7	13.6	13.1	24.4	12.5	11.9	25.5	12.5	13.0
25% biggest	none	34.8	16.4	18.5	33.8	18.4	15.3	34.6	$^\sharp$20.1	14.5	32.8	17.4	15.5	35.6	17.7	18.0
	5 days	29.6	13.0	16.6	25.4	13.7	11.7	32.1	18.5	13.6	33.6	16.0	17.6	35.4	16.4	$^\sharp$19.0
	20days	24.8	$^\flat$10.0	14.8	25.4	15.4	$^\flat$10.0	28.6	16.3	12.3	28.5	16.0	12.6	33.6	17.6	15.9
Panel B: two-week holding period																
base case	none	11.2	6.5	4.6	11.0	6.4	4.7	10.9	$^\sharp$6.8	4.1	10.8	6.4	4.4	10.2	5.6	4.7
	20days	9.1	$^\flat$3.9	5.3	7.6	$^\flat$3.9	$^\flat$3.7	11.3	5.7	5.5	9.8	4.7	5.1	12.5	6.2	$^\sharp$6.3
50% biggest	none	10.0	5.6	4.5	12.6	6.7	5.9	13.0	7.5	5.5	12.8	7.0	5.9	12.0	6.6	5.4
	5 days	12.6	6.6	6.1	10.8	5.6	5.3	12.6	6.8	5.8	13.3	6.7	6.6	13.3	6.9	6.5
	20days	11.4	$^\flat$4.8	6.6	9.7	5.5	$^\flat$4.1	13.5	7.0	6.5	12.2	6.0	6.2	15.4	$^\sharp$7.9	$^\sharp$7.6
25% biggest	none	17.0	10.6	6.5	15.2	9.1	6.0	17.0	$^\sharp$10.8	6.2	17.8	10.2	7.6	16.6	9.9	6.7
	5 days	16.0	8.6	7.5	11.6	7.1	4.5	15.6	8.6	7.0	16.9	8.5	8.4	17.2	9.4	7.8
	20days	13.2	$^\flat$4.9	8.2	12.1	7.7	$^\flat$4.4	15.8	8.6	7.3	14.9	7.0	8.0	19.6	10.3	$^\sharp$9.3
Panel C: One-month holding period																
base case	none	8.0	4.9	3.1	8.8	5.2	3.6	8.5	$^\sharp$5.4	3.1	7.8	4.7	3.1	7.8	4.2	3.5
	5 days	7.6	3.9	3.6	7.1	3.6	3.4	8.5	4.7	3.8	8.0	4.1	3.8	8.1	4.2	3.9
	20days	7.1	$^\flat$3.4	3.7	6.1	$^\flat$3.4	$^\flat$2.7	8.4	4.3	4.1	6.8	3.7	3.1	9.1	4.7	$^\sharp$4.4
50% biggest	none	7.2	4.1	$^\flat$3.1	10.2	5.6	4.6	10.3	$^\sharp$5.9	4.3	9.2	4.8	4.4	9.2	5.1	4.0
	5 days	9.2	5.0	4.2	8.3	4.5	3.8	9.7	5.4	4.3	9.5	5.2	4.4	9.9	5.3	4.6
	20days	8.5	$^\flat$4.0	4.5	7.7	4.5	3.2	9.7	5.2	4.5	8.6	4.9	3.7	11.0	5.9	$^\sharp$5.1
25% biggest	none	11.8	7.2	4.5	11.4	7.2	4.1	13.0	$^\sharp$8.3	4.7	12.4	6.7	5.7	12.4	7.6	4.8
	5 days	11.3	6.3	5.0	8.8	5.7	$^\flat$3.0	11.9	6.7	5.2	11.9	6.7	5.2	13.0	7.3	5.7
	20days	10.0	$^\flat$4.4	5.5	9.8	6.3	3.5	11.0	6.2	4.8	10.7	6.2	4.5	14.0	7.7	$^\sharp$6.3

Table 8: Various measures of performance, across models

Key: We show two measures of unexplained variability in prices, the Average Absolute Error (AAE) and the Average Root Mean Square, the average standard deviation of the residuals. Both are measured in basis points. Also shown are the autocorrelation, averaged across bonds, of the time series of residuals per bond extracted from each cross section, and the average run length (in days), where a run is defined as a sequence of days where the residuals have the same sign. Next come the regression coefficients of abnormal returns on initial mispricing, for 1- or 20-day holding periods and with or without lag ($L = \{1, 0\}$). Lastly we show some CARs, for daily and monthly revision frequencies and for trading rules where we act only upon the 50 or 25 percent strongest signals. In the second part of the table we show the ranks of the models rather than the statistics.

	vasicek	cir	rich	ls	bdfs	b-d	spline
	statistics						
AAE	15.6	16.9	14.3	13.3	13.7	15.8	12.4
ARMSE	17.5	20.5	16.0	14.6	12.0	17.1	13.9
autocorr	0.94	0.85	0.74	0.85	0.86	0.73	0.93
avg runl	17.6	12.2	7.7	14.9	13.9	7.4	17.7
β, 1d, $L = 0$	-0.058	-0.052	-0.062	-0.083	-0.064	-0.056	-0.076
β, 20d, $L = 0$	-0.311	-0.362	-0.311	-0.353	-0.298	-0.279	-0.373
β, 20d, $L = 1$	-0.260	-0.322	-0.256	-0.269	-0.242	-0.229	-0.303
CAR, daily, 50%, $L = 0$	20.0	25.0	28.0	25.8	25.4	27.0	20.8
CAR, monthly, 25%, $L = 0$	11.8	11.4	13.0	12.4	12.4	13.2	10.4
CAR, daily, 50%, $L = 1$	10.8	16.2	15.8	15.6	15.4	17.2	10.8
CAR, monthly, 50%, $L = 1$	10.6	10.2	11.6	11.0	11.0	11.8	9.0
	ranking of models						
AAE	5	7	4	2	3	6	1
RMSE	6	7	4	3	1	5	2
autocorr	7	3	2	4	5	1	6
avg runl	6	3	2	5	4	1	7
β, 1d, $L = 0$	5	7	4	1	3	6	2
β, 20d, $L = 0$	3	2	4	3	6	7	1
β, 20d, $L = 1$	4	1	5	3	6	7	2
CAR, daily, 50%, $L = 0$	7	5	1	3	4	2	6
CAR, monthly, 25%, $L = 0$	5	6	2	3	3	1	7
CAR, daily, 50%, $L = 1$	6	2	3	4	5	1	6
CAR, monthly, 50%, $L = 1$	5	6	2	3	4	1	7