Forecasting Asset Returns in State Space Models

D I S S E R T A T I O N
of the University St. Gallen,
Graduate School of Business Administration,
Economics, Law and Social Sciences (HSG)
to obtain the title of
Doctor of Economics

submitted by

Dominik Boos

from

Amden (St. Gallen)

Approved on the application of

Prof. Dr. Alex Keel

and

Prof. Paul Söderlind, PhD

Dissertation Nr. 3812

ADAG Copy AG, St. Gallen 2011

The University of St. Gallen, Graduate School of Business Administration, Economics, Law and Social Sciences (HSG) hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St. Gallen, October 26, 2010

The President:

Prof. Ernst Mohr, PhD

Acknowledgements

First and foremost I am very grateful to the members of my thesis committee, Prof. Dr. Alex Keel and Prof. Paul Söderlind, PhD for their willingness to support my chosen topic and to undertake the tasks of advising me in writing this thesis.

I also owe many thanks to Lukas Wäger for numerous discussions on the Kalman filter and interest rate models as well as for the proofreading of my dissertation. Additional thanks go to Jack Corrigall for his careful proofreading and language quality control.

Finally, I owe very much to Birgit who has been particularly understanding and supportive during all stages of my dissertation.

Winterthur, January 2011 Dominik Boos

Summary

Expectations about the future evolution of the economy are of immense importance for taking the right decisions in a stochastic environment. Econometricians have long been studying forecasting techniques to this end. Most of this work is based on regression techniques such as OLS or GMM.

I propose a different approach: state-space models and their estimation by means of maximum likelihood using the Kalman filter. While the two techniques are often identical in an environment with clean data; state-space models are clearly superior if the observed data is affected by measurement error or displays a seasonal pattern. In this case, state-space models allow the separation of the true underlying signal from the measurement noise. As only the signal is relevant for prediction, this can considerably improve the quality of the forecast.

In particular, I use the state space framework to estimate affine yield curve models and find that the implied return forecasts for long bonds is much more reliable than that implied by a linear regression, although the implied insample R^2 is lower. Moreover, I detect substantial predictability of long/short portfolios not properly revealed by a linear regression.

I then generalize the affine yield curve models such that they can include persistent shocks or state variables not spanned by yields. Firstly, these unspanned factor models are used to further improve the yield-curve forecast by including expected inflation as an additional state variable. In this model, the R^2 of the annual term premium forecast is above 30 percent. Secondly, I build a joint stock-bond model that merges the yield curve model with a stock market model using the price-dividend ratio as an additional variable. This is achieved by linearization using the Campbell-Shiller approximation. Thirdly, the cross-section of assets is enlarged by including size and book-to-market

sorted portfolios. This model provides evidence for substantial variation in the dividend growth rate. Once the model captures this feature, it is able to explain a large fraction of the value premium by a higher exposure of value stocks to the single persistent shock of the system.

Finally, this thesis uses rank-reduction techniques to explore the return predictability pattern. This analysis provides strong evidence for at least two independent predictability factors: the term premium and the equity premium.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The price of an asset reflects the discounted value of the asset's expected cash-flow. Prices are high when the expected cash-flow is high or when the discount rate - and hence expected returns - are low. Conversely, low prices imply either high return or low cash-flow.

Empirically, prices vary greatly. Stock prices, expressed as the aggregate dividend-price ratio, vary from more than 7 percent in 1949 to less than 2 percent in 2001. Consequently, returns, dividend growth or both must be predictable. A first indication of the source of the price variations can be derived from a regression of the dividend yield onto returns and dividend growth. It provides statistical evidence against the null of absence of return predictability, but there is no evidence for predictability of dividend growth (e.g. Fama and French (1988) or Cochrane (2008a)).

However, the inability of the dividend yield to forecast dividend growth does not necessarily imply that dividends are not predictable. When expected returns and expected dividend growth are positively correlated, the dividend yield can completely loose its ability to forecast dividends (see Menzly, Santos, and Veronesi (2004) for a model with this feature). In this situation, which is consistent with the data, there are at least two state variables. Both above-mentioned univariate forecasting regressions thus suffer from an omitted variable bias. More precisely, we observe that the R^2 of the return regressions is reduced and the ability of the price-dividend ratio to forecast dividend growth is eliminated.

Bond markets are remarkably similar. We observe a very rich yield curve structure with three or four factors. These factors do not simply reflect a complex pattern of expected inflation; rather we observe a time-varying real rate, predictable bond returns (Fama and Bliss (1987)), and that the yield curve contains information about economic growth. Moreover, yields do not contain all information about expected inflation. As (expected) inflation is a variable that interacts with yields, this observation says that there is a hidden or unspanned variable in the dynamics of yields. As a consequence, there is an omitted-variable bias in the forecasting regressions as in the case of equities.

By definition, omitted variables justify the inclusion of other variables into the forecasting regression. Unfortunately, there is no consensus as to how this omitted variable should be selected. This lack of theory-based variable selection has - driven by the immense importance of return predictability in the financial industry - produced a huge body of academic papers and practical articles on the subject. They include an almost endless list of instruments that forecast returns for a specific sample or even out-of-sample.

In practice the question arises, which of these models should be chosen. Should that with the highest R^2 be selected? This seemingly plausible answer turns out to be misleading. The paradox occurs because in order to get a high R^2, a forecasting procedure needs to adequately fit returns. But fitting returns is not the same thing as filtering out expected returns! When we search a data set - as the single historic data set we have - extensively for a predictability pattern, the estimator with the best fit found up to that point is necessarily one that fits a lot of noise and does not simply filter the 'true' expected return. Selecting the highest R^2 thus means choosing the model that overestimates predictability the most. This selection bias is not restricted to in-sample analysis. It is also at work out-of-sample and even for funds with a live track record. It is simply slightly more difficult to find high out-of-sample R^2 or to even produce real outperformance.

How can we get out of this dilemma? My solution starts with the above-mentioned condition that prices must be low when expected returns are high or when the cash-flow growth rate is low. Mathematically, we can regard prices as a function of expected returns and expected cash-flows. These two expectations may themselves be a function of some other variables. For instance, the expected market return is the sum of the equity premium and the short interest rate. We can describe this observation more abstractly by

assuming that asset prices are functions of a set of macro or state variables. These functions can, of course, also depend on asset-specific cash-flow expectations. However, for very large asset classes, such as government bonds, the aggregate stock market or aggregate small caps, this idiosyncratic influence will be small.

If we have at least as many assets as state variables, we can invert these price functions and solve for the state variables. Because some prices have an idiosyncratic part, this is rather a statistical inversion than an exact algebraic. Typically, it is a regression across prices. This is a straightforward technique to reveal the state variables, most importantly, expectations, as a function of prices. Since the choice of assets is far less subjective than taking a subset of economic indicators, this process is clearly less affected by selection bias.

Once the state variables are filtered, we can also explore the cross-section of returns. Motivated by Merton's (1973) ICAPM and its extension to state dependent utility in Merton (1995), shocks to state variables are natural candidates for explaining the cross-section of returns, such as value and size anomalies (Fama and French (1992)). For example, Campbell and Vuolteenaho (2004) use a version of the ICAPM to break the CAPM beta of a stock with the market portfolio into two components: one reflecting news about the market's cash-flows and one reflecting news about the market's discount rate. These two components are not, strictly speaking, innovations to state variables but they are closely related to innovations to the expected dividend growth rate and the expected discount rate. Empirically, the paper finds that value and small stocks have considerably higher cash-flow betas than growth and large stocks, and this can explain their higher average return. Unfortunately, the model starts with a VAR that embeds a return-forecasting equation. The authors of the paper are quite open about the fact that their results depend crucially on the variables contained in the VAR. The empirical content of this model thus suffers from a similar selection bias as the forecasting regression.

Another paper that deals with state variables in the equity market is Bansal and Yaron (2004). It models the consumption and dividend growth rate with a small long-run predictable component. Building on this theoretical groundwork, Bernhard (2008) estimates an international model that associates the international value premium with shocks to the predictable component. The calibration of the model starts with a consumption-forecasting regression that determines expected consumption growth. Stock returns can

now be decomposed in a way that is similar to the decomposition in Campbell and Vuolteenaho (2004). The difference is that Bernard starts with forecasting growth rates instead of forecasting returns. The composition of the variables in this regression, however, also suffers from a selection bias.

The empirical implementation of these two models is thus not much different from the very simple and far less theory-based approach of Petkova (2006). By fitting a simple VAR, she finds that unexpected innovations to the credit-spread explain the size effect and shocks to the slope of the yield curve which, to a great extent, explains the value effect. Hahn and Lee (2006) comes to the same conclusion by taking unconditional innovations of the same variables.

Brennan, Wang, and Xia (2004) and Brennan and Xia (2006) are two papers that are also very similar to the present thesis. More precisely, they estimate a yield curve model with the Kalman filter and then use the filtered variables to explain the cross-section of equity returns. The main difference is that they do not use the cross-section of size and book-to-market sorted portfolios to estimate their model.

Many forecasting instruments used in these papers are really prices or price differences and changes to these variables are thus closely related to returns. For instance, if the term spread is used as an indicator for return forecasting, a shock to the expected return is closely related to movements in the yield curve and thus to the return of a specific portfolio. As a consequence, if the model explains the cross-section of returns by different loadings to changes in expected returns, there is a connection between yield curve movements and the rest of the model which must not be ignored. If this link is not included in the model and the term spread is treated simply as an unrelated instrument - more precisely if bond returns are not included in the cross-sectional regression - cross-sectional expectation of the assets explored can be inconsistent with those of bond returns. Therefore, if prices are used as predictive instruments, they should be modeled jointly with other parts of the economy.

Another problem with some of these models is that they explain only a very small fraction of size and value returns. Movements of the state variables, however, are large movements of the economic conditions and should therefore have a large impact on prices. I circumvent these problems by restricting the observables to asset returns and prices. Shocks to the state variables are then necessarily price movements and thus closely related to returns. This forces

the model to explain a large fraction of cross-sectional returns by movements of the state variables.

This defines the approach of this thesis. Firstly, choose a set of assets that should be explored. Secondly, define the basic characteristics of the models, such as the number of yield curve factors and the total number of state variables, and define their dynamics. Thirdly, estimate the model using prices and returns of these assets as the observable quantities. Fourthly, impose sensible and testable restrictions to the model. Fifthly, rotate the model to meaningful state variables, such as expected growth rates and expected returns. Finally, find meaningful economic interpretations of the model. One specific goal is to find an explanation of the variation of risk-compensation across assets.

This procedure will lead us to joint stock-bond models. There are a few other papers that have built such models: Beckaert and Grenadier (2001), Mamaysky (2002), D'Addona and Kind (2006) and Lemke and Werner (2009). The last of these papers also provides estimates based on the Kalman filter.

Chapter 2

State Space Models and the Kalman Filter

This chapter introduces the theory and definitions of linear state space models and their estimation using the Kalman filter. The first two sections focus on what is needed to follow the estimation methods used in this thesis. Specifically, a basic description of the Kalman filter and smoother and their specific applications in this thesis. Finally, the third section introduces the theory required to derive the gradient of the likelihood function and some examples of derivatives. For further reading I refer to Durbin and Koopman (2001) and Harvey (1989), which are both comprehensive treatments of the subject.

2.1 State Space Models

A multivariate linear state space model is defined by two equations. Firstly, a state or transition equation defines the dynamics of the underlying state of the economy. It is, in general, not, or only partially, observable and follows

$$x_{t+1} = c + Tx_t + R\varepsilon_{t+1} \tag{2.1}$$

where x_t $(t = 1 \dots \tau)$ is a $k \times 1$ vector that summarizes the state of the economy at time t. T is a $k \times k$ matrix, c is a $k \times 1$ vector, and R is a $k \times l$ matrix $(l \leq k)$. These three parameters are constant in all models applied

in this thesis, although the model can also be formulated using time-varying parameters. ε_t is a mean zero $l \times 1$ vector of Gaussian shocks with covariance matrix Q. It is assumed to be uncorrelated over time. In summary, the state of the economy is assumed to follow a K-dimensional VAR with a potentially degenerate covariance matrix.

Secondly, the measurement equation constitutes the observable part of the economy. It links the state vector x_t to the observable variables y_t by

$$y_t = d + Zx_t + \eta_t. \tag{2.2}$$

The vector of observable variables y_t has dimension $n \times 1$. Accordingly, Z is $n \times k$ and d is $n \times 1$. The n-dimensional noise term η_t is assumed to be normally distributed with zero mean and covariance matrix H. It is uncorrelated over time and uncorrelated with ε (for any lag). η_t can often be interpreted as a vector of measurement errors. In some models, however, it contains economically meaningful quantities and important aspects of the model are specified in the measurement equation. This is not a problem for the Kalman filter as long as η_t and ε_t remain independent.

The specification of the state space system is completed by assuming that the initial state vector x_1 has mean $x_{1|0}$ and covariance matrix P_1.

I refer to the set of parameters as $\psi = \{T, c, Q, R, Z, d, H\}$. Finally, let $Y_t = [y_1 \dots y_t]$ be the data vector up to time t.

2.2 The Kalman Filter

The Kalman filter is a recursive procedure for computing the conditional mean $(x_{t|t-1})$ and covariance matrix (P_t) of the state vector x_t given Y_{t-1}. It is used to estimate the likelihood function of a state space system or, once the parameters are known, to make inference about the unobservable state of the economy x_t.

After selecting the initial state $x_{1|0}$ and P_1, the Kalman filter is often considered to iterate between two steps: time update or prediction and measurement update. The prediction step uses the current state estimate at time t, i.e $x_{t|t}$, to produce an estimate of the state at $t+1$ using the state equation:

$$x_{t+1|t} = c + Tx_{t|t}.$$

Once the new observation y_{t+1} is available, the measurement update step refines this estimate by the actual measurement equation at time $t+1$

$$\begin{aligned} \nu_{t+1} &= y_{t+1} - d - Z x_{t+1|t} \\ x_{t+1|t+1} &= x_{t+1|t} + K^g_{t+1} \nu_{t+1} \end{aligned}$$

to arrive at a new, more accurate state estimate. The update step is essentially a cross-sectional regression in all applications provided in this thesis. The so-called Kalman gain K^g_{t+1} is defined in step 4 below. Similar equations hold for the covariance matrix P_t.

Algebraically, the two steps of the recursion can be written in one single step going directly from $x_{t|t-1}$ to $x_{t+1|t}$. Throughout the thesis, I run the filter in the following way:

1. The mean of the initial state $x_{1|0}$ is estimated by

$$x_{1|0} = \overline{x} = (1-T)^{-1} c \tag{2.3}$$

and P_1 by solving

$$P_1 = T P_1 T' + RQR'.$$

These are the mean and the covariance of the unconditional distribution of the state vector given the parameters. The equation for the covariance is solved by vectorization:

$$\text{vec}(P) = [1 - T \otimes T]^{-1} \text{vec}(RQR^T). \tag{2.4}$$

2. The algorithm then runs the Kalman filter with varying F, K and L going directly from $x_{t|t-1}$ to $x_{t+1|t}$:

$$\begin{aligned} \nu_t &= y_t - d - Z x_{t|t-1} \\ F_t &= Z P_t Z^T + H \\ K_t &= T P_t Z^T F_t^{-1} \\ L_t &= T - K_t Z \\ x_{t+1|t} &= c + T x_{t|t-1} + K_t \nu_t \\ P_t &= T P_{t-1} L_t + RQR^T \end{aligned}$$

until F, K and L have converged. Numerically, this is the fastest way to run the Kalman filter (see Durbin and Koopman (2001)).

3. If the eigenvalues of T lie within the unit circle, the filter switches to the steady state model as soon as the matrices F, K and L have converged and thus no longer need to be updated (see Harvey (1989) chapter 3.3.4).

4. For forecasting, we use all information available at time t. We take the measurement update step to generate expectation at time t given $x_{t|t-1}$ which has already been estimated in the previous step:

$$x_{t|t} = x_{t|t-1} + K_t^g \nu_t \tag{2.5}$$

where, as already mentioned, $K_{t+1}^g = P_t Z^T F_t^{-1}$ is the Kalman gain. I denote this series as the filtered series. If we intend to compare forecasting results generated by the Kalman filter with some alternative model, we should take the filtered series as the basis as it uses information up to time t.

5. Finally, we can also estimate the states using information up to time τ. I denote these smoothed values with a hat, thus, $x_{t|\tau} = \widehat{x_t}$ and so on. They are given by the following backward recursion:

$$\begin{aligned}
\widehat{\eta}_t &= H(F_t^{-1}\nu_t - K_t^T r_t) \\
\widehat{\varepsilon}_{t+1} &= QR^T r_t \\
\widehat{x}_{t+1} &= c + T\widehat{x}_t + R\widehat{\varepsilon}_{t+1} \\
r_{t-1} &= Z^T F_t^{-1}\nu_t + L_t^T r_t \\
\mathrm{var}(\eta_t|y) &= H - H(F_t^{-1} + K_t^T N_t K_t)H \\
\mathrm{var}(\varepsilon_{t+1}|y) &= Q - QR^T N_t RQ \\
N_{t-1} &= Z^T F_t^{-1} Z + L^T N_t L
\end{aligned}$$

for $t = \tau \ldots 1$ where $r_\tau = 0$ and $N_\tau = 0$. This algorithm is known as the Kalman smoother. Smoothed values can be used to better understand the dynamics of a model. They should not, however, be used when dealing with forecasting because of their forward-looking bias.

Given the values calculated in steps 1 to 3, the log likelihood function $\mathcal{L}$ can

be written as

$$\begin{aligned}\mathcal{L} &= -\frac{n\tau}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{\tau}\log|F_t| - \frac{1}{2}\sum_{t=1}^{\tau}\nu_t^T F_t^{-1}\nu_t \\ &= \text{constant} + \mathcal{L}_0(x_{1|0}, P_1) - \frac{\tau}{2}\log|H| - \frac{\tau-1}{2}\log|Q| \\ &- \frac{1}{2}\sum_{t=1}^{\tau}\eta_t^T H^{-1}\eta_t - \frac{1}{2}\sum_{t=2}^{\tau}\varepsilon_t^T Q^{-1}\varepsilon_t \end{aligned} \tag{2.6}$$

where $\mathcal{L}_0$ captures the influence of the initial distribution on the likelihood. $\mathcal{L}$ as defined in the first line is maximized when we estimate parameters. Note that steps 4 and 5 should be avoided during the likelihood maximization in order to save computer time.

The representation of the log likelihood function in the second and third lines of (2.6) is known as the innovation error form. In the next section, it is used to derive the score vector.

2.3 The Gradient

The likelihood function of large models has many parameters and the quality of the gradient is often poor when estimated numerically. In order to achieve precision, the step size of the finite differencing algorithm must be allowed to vary across variables: a feature that is not common in numerical minimization routines (see appendix B for an example). Moreover, its optimal value can also change when running the optimization algorithm. At some starting point away from the optimum, the gradient is typically not particularly sensitive to the step size. Near the optimum, however, numerical gradients can be very poor if the step size is not controlled for. This problem is most serious when the likelihood function is flat around its maximum, as is the case for some models in this thesis.

To circumvent these convergence problems, it is essential to use analytical rather than numerical gradients or score vectors. This makes optimization faster and more reliable. Empirically, this study has found higher, sometimes, substantially higher, likelihood values when using analytical gradients. This also saves computation time. Furthermore, the results of the numerical optimization become independent of the starting point.

To derive the gradient of the Kalman filter, I follow Durbin and Koopman (2001), section 7.3.3 and Koopman and Shepard (1992). The derivation starts with the observation that the distribution of the initial vector $x_1 \sim N(x_{1|0}, P_1)$ is assumed to be known (see the first step in the previous section). Let $p(x, y|\Psi)$ be the joint density of x and y, let $p(x|y, \Psi)$ be the conditional density of x given y and let $p(y|\Psi)$ be the marginal density of y given the parameter vector Ψ. We now evaluate the score vector

$$\frac{\partial \mathcal{L}}{\partial \Psi} = \frac{\partial \mathcal{L}(y|\Psi)}{\partial \Psi} = \frac{\partial \log p(y|\Psi)}{\partial \Psi}$$

at the value $\widetilde{\Psi}$. We then have

$$\log p(y|\Psi) = \log p(x, y|\Psi) - \log p(x|y, \Psi).$$

Let $\widetilde{\mathrm{E}}$ denote expectation with respect to the density $p(x|y, \Psi)$. Since $p(y|\Psi)$ does not depend on x, taking $\widetilde{\mathrm{E}}$ of both sides gives:

$$\log p(y|\Psi) = \widetilde{\mathrm{E}}[\log p(x, y|\Psi)] - \widetilde{\mathrm{E}}[\log p(x|y, \Psi)].$$

In order to obtain the gradient at $\widetilde{\Psi}$, we differentiate both sides with respect to Ψ and put $\Psi = \widetilde{\Psi}$. Assuming that differentiating under the integral is legitimate:

$$\begin{aligned}
\widetilde{\mathrm{E}}\left[\frac{\partial \log p(x|y, \Psi)}{\partial \Psi}\bigg|_{\Psi=\widetilde{\Psi}}\right] &= \widetilde{\mathrm{E}}\left[\frac{1}{p(x|y, \Psi)}\frac{\partial p(x|y, \Psi)}{\partial \Psi}\bigg|_{\Psi=\widetilde{\Psi}}\right] \\
&= \int \frac{1}{p(x|y, \Psi)}\frac{\partial p(x|y, \Psi)}{\partial \Psi}\bigg|_{\Psi=\widetilde{\Psi}} p(x|y, \Psi)dx \\
&= \frac{\partial}{\partial \Psi}\int p(x|y, \Psi)dx\bigg|_{\Psi=\widetilde{\Psi}} = 0
\end{aligned}$$

since the last integral is equal to 1. Thus,

$$\frac{\partial \log p(y|\Psi)}{\partial \Psi}\bigg|_{\Psi=\widetilde{\Psi}} = \widetilde{\mathrm{E}}\left[\frac{\partial \log p(x, y|\Psi)}{\partial \Psi}\right]\bigg|_{\Psi=\widetilde{\Psi}}.$$

As in equation (2.6) the log likelihood function is written in the innovation error form

$$\begin{aligned}
\log p(x, y|\Psi) = \text{constant} + \mathcal{L}_0(x_{1|0}, P_1) + \frac{\tau}{2}\log|H| + \frac{\tau - 1}{2}\log|Q| \\
- \frac{1}{2}\sum_{t=1}^{\tau} \eta_t^T H^{-1}\eta_t - \frac{1}{2}\sum_{t=2}^{\tau} \varepsilon_t^T Q^{-1}\varepsilon_t.
\end{aligned}$$

On taking the expectation $\widetilde{\mathrm{E}}$ and differentiating with respect to Ψ, this gives the score vector at $\Psi = \widetilde{\Psi}$:

$$\begin{aligned}\left.\frac{\partial \mathcal{L}(y|\Psi)}{\partial \Psi}\right|_{\Psi=\widetilde{\Psi}} &= -\frac{1}{2}\frac{\partial}{\partial \Psi}[\sum_{t=1}^{\tau}(\log|H| + \mathrm{Tr}[H^{-1}\{\widehat{\eta}_t\widehat{\eta}_t^T + \mathrm{var}(\eta_t|y)\}]) \\ &\quad + \sum_{t=2}^{\tau}(\log|Q| + \mathrm{Tr}[Q^{-1}\{\widehat{\varepsilon}_t\widehat{\varepsilon}_t^T + \mathrm{var}(\varepsilon_t|y)\}]) \\ &\quad + \mathcal{L}_0(y)\big|\Psi]\Big|_{\Psi=\widetilde{\Psi}} \end{aligned} \tag{2.7}$$

where $\widehat{\eta}_t$, $\widehat{\varepsilon}_t$, $\mathrm{var}(\eta_t|y)$ and $\mathrm{var}(\varepsilon_t|y)$ are obtained by the Kalman smoother (step 5 of the previous section).

Derivatives with respect to the covariance matrices H and Q are now particularly easy:

$$\begin{aligned}\frac{\partial \mathcal{L}(y|\Psi)}{\partial H} &= -\frac{\tau}{2}\frac{\log|H|}{\partial H} - \frac{1}{2}\frac{\mathrm{Tr}[H^{-1}z_1]}{\partial H} \\ &= -\frac{\tau}{2}H^{-1} + \frac{1}{2}H^{-1}z_1H^{-1}\end{aligned}$$

where z_1 is the expression in the curly brackets in the first line of (2.7). The derivation for Q is obviously identical and z_2 used below can be defined in the same way as z_1 using the second line of the same equation.

The distribution of the measurement error is often defined (by the researcher) as a simple function of the parameter space Ψ. This means that the derivative of $\mathcal{L}$ is often easy to calculate. A particularly easy structure is when H is defined as a scalar multiple h^2 of the identity matrix. We then have

$$\widetilde{\mathrm{E}}[\log p(x, y \mid \Psi_T)] = \text{constant} - \tau n \log(h) - \frac{\mathrm{Tr}[z_1]}{2h^2}$$

where the constant summarizes terms not depending on H. The derivative is thus

$$\frac{\partial L}{\partial h} = \frac{\mathrm{Tr}[z_1]}{h^3} - \frac{\tau n}{h}.$$

Unrestricted covariance matrices are usually parameterized using the Cholesky decomposition to ensure positive definiteness $Q = CC^T$. Then, using the analog definition for the constant,

$$\widetilde{\mathrm{E}}[\log p(x, y \mid \Psi_T)] = \text{constant} - (\tau - 1)\log(\prod c_{ii}) - \frac{1}{2}\mathrm{Tr}[(CC^T)^{-1}z_2]$$

and the derivative is

$$\frac{\partial L}{\partial C} = Q^{-1} z_2 (C^T)^{-1} - (\tau - 1) \begin{bmatrix} \frac{1}{c_{11}} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \frac{1}{c_{ll}} \end{bmatrix}. \tag{2.8}$$

The proof for the derivatives with respect to the other parameters is more involved but the resulting expressions have all a very simple form[1]:

$$\begin{aligned} \frac{\partial \mathcal{L}}{dZ} &= H^{-1}(z_3 - Z z_4) \\ \frac{\partial \mathcal{L}}{dT} &= (RQR^T)^{-1}(z_6 - T z_5) \\ \frac{\partial \mathcal{L}}{dd} &= H^{-1}(\tau d - z_8) \\ \frac{\partial \mathcal{L}}{dc} &= (RQR^T)^{-1}((\tau - 1)c - z_7). \end{aligned} \tag{2.9}$$

$z_3 \dots z_8$ are matrices that can be calculated during the run of the Kalman smoother. R is a predefined matrix in all the models applied in this thesis so its derivative is not of interest here. Rosenbaum and Zetlin-Jones (2006) derive a vectorized version of these equations.

If one of the covariance matrices RQR^T or H is degenerate, the inverse is replaced by the generalized inverse. In this case, the procedure does not deliver all derivatives and I either reparameterize the model or estimate certain elements of the gradient numerically (see appendix B).

Finally, the filter is initialized with the unconditional mean and covariance matrix defined in (2.3) and (2.4). The likelihood function thus also depends on the parameters through $\mathcal{L}_0(x_{1|0}, P_1) = \mathcal{L}_0(c, T, Q)$. This adds several extra terms to the derivatives defined in (2.8) and (2.9).

[1]Proof and details regarding the parameters are available from the author upon request.

Chapter 3

Regression Forecasts of Bond Returns

Cochrane and Piazzesi (2005a) find that a single tent-shaped linear combination of forward rates predicts excess returns on one- to five-year maturity bonds with R^2 up to 0.44. I take up this analysis and extend it to longer maturity bonds. I also compare it with other data sets, using yields that go back to 1946. The broader cross-section offers further insights into predictability. Most importantly, the analysis reveals that the single factor representation proposed by Cochrane and Piazzesi does not fully describe the expected return pattern. It is shown that a second return factor, associated with the slope of the yield curve, is needed to explain the entire cross-section of expected U.S. government bond returns. This contrasts with the findings of Cochrane and Piazzesi (2005a, 2008) who claim that there is only one predictability factor and that this factor predicts compensation for level shocks.

The second factor is lost in Cochrane and Piazzesi's analysis since they only explore single bonds and not portfolios that are long in some maturities and short in others. Over 90 percent of yield curve movement is associated with level shifts, so forecasting single bond returns predominantly means fitting these level shifts. All other movement is so unimportant relative to level shifts that it is completely overlaid in the forecasting analysis. Consequently, Cochrane and Piazzesi (2005a) find that one single factor predicts excess returns for all maturities. Finally, in their second paper, Cochrane and Piazzesi

(2008) find that this factor catches risk compensation for level shifts and that there is no compensation for slope and curvature risk.

In contrast to the analysis of Cochrane and Piazzesi, I also include analysis of duration neutral portfolios. These portfolios are long in some U.S. government bond with shorter maturity and short another bond with longer maturity. Portfolios constructed this way, have an unconditional Sharpe ratio which has been 0.16 over the last 60 years. Although this is at best marginally significant, the Sharpe ratio is higher than that of a simple long only portfolio. In a next step, I also provide strong evidence that these portfolios are predictable. This result is in conflict with the findings of Cochrane and Piazzesi.

3.1 Data and Notation

3.1.1 Notation

I use the same notation as Cochrane and Piazzesi (2005a). The notation for log bond prices is

$$p_t^{(n)} = \log \text{ price of an } n\text{-year discount bond at time } t.$$

Parentheses are used to distinguish maturity from exponential in the superscript. The log yield is

$$y_t^{(n)} = -\frac{1}{n} p_t^{(n)}.$$

The log forward rate at time t for loans between time $t+n-1$ and $t+n$ is denoted by

$$f_t^{(n)} = p_t^{(n-1)} - p_t^{(n)}.$$

The holding period log return from buying an n-year bond at time t and selling it as an $n-1$ year bond at time $t+1$ is then

$$r_{t+1}^{(n)} = p_{t+1}^{(n-1)} - p_t^{(n)}$$

and the corresponding excess log return is

$$rx_{t+1}^{(n)} = r_{t+1}^{(n)} - y_t^{(1)}.$$

3.1.2 Data

The choice of data is mainly a compromise between a long data set and clean data. There are four sources that provide zero yields at fixed maturities. The most current data set is Gürkaynak, Sack, and Wright (2006) which will be denoted as the GSW data set. This data set has been made available by the Federal Reserve[1] and is constantly updated. Unfortunately, this data set only includes data back to 1961 and does not include maturities below one year. Ten year yields are available only since August 1971. However, for the purpose of this paper, this is not the biggest shortcoming of this data set. The data set is constructed using the Svensson (1994) method and this method smoothes away important information necessary for return forecasting. Most importantly, the information contained in the five-year forward rate is smoothed away, as highlighted by Cochrane and Piazzesi (2005a).

The second source is the data set of McCulloch and Kwon (1993), published by J. Huston McCulloch on his web site[2] and denoted as the McK data set. The main advantage of this data set is that it goes back to 1946 and contains more maturities than the GSW data set. It ends in 1993 and is constructed using cubic splines. Although the cubic spline method diverges for longer maturities and hence provides a poor fit for curves that are flat or have a flat long end (Diebold and Li (2006)), the influence of this smoothing technique on the forecasting regression is less severe than that of the Svensson method. In particular, Cochrane and Piazzesi (2005b) show that the same tent-shaped factor that is present in their original data set also occurs with the McK data, but not in the GSW data.

The cleanest data set available is that provided by Diebold and Li (2006), available at Francis X. Diebold's web page[3] and denoted as DL. It contains the fixed-maturity yields from 1970:1 to 2000:12 deduced from unsmoothed Fama and Bliss (1987) forward rates. This data set is the cleanest one, and therefore it serves as a reference point. Cochrane and Piazzesi also use yield curve data constructed by the unsmoothed Fama-Bliss method. This data set is extended until the end of 2006 by a data set contributed by Wäger (2009). The extension uses yields from Datastream and it is also based on the unsmoothed Fama-Bliss method. In order to rule out concerns about quality

[1] available at www.federalreserve.gov/pubs/feds/2006/200628/feds200628.xls
[2] www.econ.ohio-state.edu/jhm/ts/mcckwon/mccull.htm
[3] www.ssc.upenn.edu/∼fdiebold/papers/paper49/

of Datastream bond data, the yields provided by Wäger are compared with DL yields[4]. This is possible as they start before 2000:12 and we can compare the two data sets in the overlapping period. In this overlapping period, the two data sets are nearly identical.

Long data sets are crucial in forecasting; I therefore compose a synthetic one as follows. For yield data, I use the cleanest data available. This implies that I use the McW data from 1946:12 until 1969:12. I then use the DL data until 2000:12 and finally its extension until the end of the sample 2006:12.

Return data is constructed using smoothed data. I therefore use the McW data until 1972:7 and GSW afterwards. Using yields and returns from different sources also has the advantage that estimation becomes slightly more robust against measurement error. See also the next section regarding this point.

3.1.3 Measurement Error

Interest rates are usually expressed in terms of discount rates or equivalently as zero bond prices. Observation of yields are, however, not perfect, rather they are affected by measurement error for at least three reasons. Firstly, if we want to estimate the discount rate for a given maturity, nothing guarantees that we find a zero bond with exactly this maturity. The price of the zero bond must therefore be approximated by other bonds with similar maturities. Secondly, most bonds pay coupon and the value of their coupons must be deduced from other bonds, which themselves are only measured with error. Thirdly, most bonds are not traded very frequently, so we see measurement error due to nonsynchronous trading.

Cochrane and Piazzesi (2005a) show that measurement error leads to spurious predictability with a very specific pattern. A spuriously high price of the n-year bond at time t implies a lower n-year forward rate as well as a lower return $r_{t+1}^{(n)}$. Thus, measurement error implies spuriously high loadings for the n-year forward rate when n-year bond returns are predicted. Similarly, the coefficient of the $(n+1)$-year forward would be spuriously low when measurement error were present.

This effect is reduced when we analyze longer forecasting horizons as the

[4]Clean data is not crucial for the application of the Fama-Bliss method. Indeed, the difficulty of the Fama-Bliss method is to remove illiquid or mispriced bonds and this step also removes data errors. Once clean bond prices are available, the application of the Fama-Bliss bootstrapping algorithm is straightforward.

relative importance of the measurement error gets smaller because its volatility is independent of the horizon, whereas that of the return increases with the horizon. Therefore, Cochrane and Piazzesi (2005a) suggest forecasting one-year excess returns to diminish the influence of measurement. When returns are forecasted over shorter time horizons, measurement error becomes more important relative to 'true returns'. For very short horizons, the signal-to-noise ratio is so low that it completely overlays the predictability pattern.

A second way to reduce the disturbance caused by measurement error is to use yield data and return data based on different smoothing techniques. Measurement error is not the same when the yield curve is fitted in a different way which diminishes the connection between the measurement error in the yields and those in the returns. Of course, yield estimates are always based on the same observation of bond prices, so the measurement error is still correlated from one data set to another.

In this chapter, I adopt the approach of Cochrane and Piazzesi (2005a) and forecast one-year returns. In the next chapter, however, I also identify an alternative way to tackle the problem by explicitly modeling the measurement error. In this framework, the previously observable state variables become unobservable. I therefore apply the Kalman filter for parameter estimation and for filtering the latent yields. The state space approach confirms many findings of the simpler regression approach. In particular, both estimation methods find highly significant predictability of duration-neutral as well as long-only portfolios.

3.2 Bond Risk Premia

3.2.1 Forecasting Long-Only Bond Returns

Following Cochrane and Piazzesi (2005a), I run regressions of bond excess returns at time $t+1$ on forward rates at time t using bonds with annual maturities up to ten years:

$$rx_{t+1}^{(n)} = \beta_0^{(n)} + \beta_1^{(n)} y_t^{(1)} + \beta_2^{(n)} f_t^{(2)} + ... + \beta_{10}^{(n)} f_t^{(10)} + \varepsilon_{t+1}^{(n)}. \qquad (3.1)$$

I use two data sets to estimate this regression. Firstly, the unsmoothed DL data set and then the composed data set. In both cases I use monthly overlapping data because intra-year data still contains valuable information. This in-

creases the sample size and estimation results become more stable. Of course, return data is no longer independent, so we have to adjust test statistics.

Table 3.1 shows t-statistics for the forecasting regression using the extended DL data set. It shows that over all return maturities, the one-, three-, five-, and eight-year forward rates are strongly significant. The four-year rate is also significant but with somewhat lower t-statistics for longer maturities. It is striking how clear the pattern is. The one, three and eight-year rate are extremely significant while the other variables, with the exception of the four- and five-year rate, are not.

The clear pattern is also evidence that the result is not due to measurement error, as it does not resemble spurious predictability pattern implied by measurement error described in the previous section. This is likely to occur. When we have a limited number of factors that describes the yield curve but use more forward rates than the number of factors, we necessarily have variables on the right hand side of the regression that should not be there. Remember that we have ten variables on the right hand side and only 30 independent return observations on the left hand side; so the danger is huge. We do not see this pattern at all and we can conclude that the high R^2 is not due to measurement error. Cochrane and Piazzesi (2005a) give further evidence for this claim. Most importantly, they show that one-month-lagged instruments have similar (often even better) forecasting power for bond returns. Using lagged returns, however, completely rules out spuriously high R^2 due to measurement error.

The results are very similar for the composed data sets. They are given in table 3.2. The main difference is that the four-year rate is now more significant than the five-year rate.

The DL data is cleaner and I use this data to decide which forward rates to use for prediction. I do not want to take the entire forward curve to avoid the danger of overfitting. As outlined in paragraph 3.2.3 below, three factors are not sufficient to explain the entire predictability pattern, so I use four forecasting instruments. By restricting the set of independent variables to the four most significant ones ($y^{(1)}$, $f^{(3)}$, $f^{(5)}$ and $f^{(8)}$) we see only a small reduction of R^2 (see the last two rows of table 3.1 and table 3.2). For the DL data, R^2 lies between $0.29 - 0.38$ instead of $0.34 - 0.41$ for the unrestricted model. In the composed data sets, R^2 lowers to $0.26 - 0.30$ from $0.29 - 0.35$. The reduction is rather small and I will therefore use the restricted set of

Table 3.1: t-Statistics for a Regression of Bond Excess Returns on Forward Rates, 1970-2006

	$rx^{(2)}$	$rx^{(3)}$	$rx^{(4)}$	$rx^{(5)}$	$rx^{(6)}$	$rx^{(7)}$	$rx^{(8)}$	$rx^{(9)}$	$rx^{(10)}$
$y^{(1)}$	-5.54	-5.97	-6.50	-6.89	-6.96	-7.13	-6.44	-6.54	-6.26
$f^{(2)}$	0.88	0.25	0.49	0.72	0.94	1.15	1.06	1.45	1.53
$f^{(3)}$	3.08	4.47	3.93	3.69	3.59	3.55	3.55	3.27	3.34
$f^{(4)}$	2.34	2.10	3.34	3.06	2.66	2.05	1.89	2.05	1.95
$f^{(5)}$	-1.91	-2.36	-2.60	-1.53	-2.01	-2.54	-2.81	-2.99	-3.24
$f^{(6)}$	0.93	0.80	0.62	0.52	1.70	1.61	1.36	1.22	0.91
$f^{(7)}$	-0.74	-0.75	-0.76	-0.98	-1.04	0.49	0.22	-0.03	-0.61
$f^{(8)}$	-4.84	-5.20	-5.20	-5.22	-5.28	-5.20	-3.90	-4.09	-4.29
$f^{(9)}$	-0.86	-0.64	-0.52	-0.45	-0.52	-0.48	-0.30	0.64	0.76
$f^{(10)}$	-0.95	-0.76	-0.67	-0.52	-0.38	-0.29	-0.21	-0.10	1.01
R^2: $f^{(1)} \dots f^{(10)}$	0.34	0.37	0.39	0.38	0.41	0.41	0.39	0.40	0.40
R^2: $f^{(1)}$, $f^{(3)}$, $f^{(5)}$, $f^{(8)}$	0.29	0.34	0.35	0.34	0.36	0.36	0.35	0.35	0.35

Note: Monthly observation of annual returns. t-statistics are estimated using the Newey-West adjustment with 18 lags. R^2 is in the second to the last row and R^2 for the restricted model is in the last row. Returns and forward rates are estimated from the extended DL data set.

Table 3.2: t-Statistics for a Regression of Bond Excess Returns on Forward Rates, 1947-2006

	$rx^{(2)}$	$rx^{(3)}$	$rx^{(4)}$	$rx^{(5)}$	$rx^{(6)}$	$rx^{(7)}$	$rx^{(8)}$	$rx^{(9)}$	$rx^{(10)}$
$y^{(1)}$	-4.07	-4.63	-4.96	-5.15	-5.28	-5.37	-5.44	-5.49	-5.52
$f^{(2)}$	0.12	0.16	0.15	0.19	0.24	0.31	0.40	0.51	0.63
$f^{(3)}$	3.32	3.42	3.34	3.19	3.04	2.89	2.75	2.63	2.51
$f^{(4)}$	2.35	2.54	2.71	2.79	2.81	2.78	2.72	2.64	2.55
$f^{(5)}$	-1.98	-2.08	-2.05	-2.04	-2.04	-2.06	-2.10	-2.14	-2.19
$f^{(6)}$	0.86	0.87	0.97	1.08	1.17	1.25	1.31	1.36	1.39
$f^{(7)}$	-1.13	-0.95	-0.84	-0.73	-0.61	-0.51	-0.43	-0.37	-0.34
$f^{(8)}$	-5.49	-5.65	-5.55	-5.35	-5.14	-4.95	-4.78	-4.63	-4.49
$f^{(9)}$	-0.55	-0.43	-0.36	-0.26	-0.15	-0.02	0.10	0.23	0.36
$f^{(10)}$	-0.94	-0.79	-0.65	-0.49	-0.34	-0.19	-0.05	0.08	0.21
R^2: $f^{(1)} \dots f^{(10)}$	0.29	0.31	0.32	0.33	0.33	0.34	0.34	0.34	0.35
R^2: $f^{(1)}$, $f^{(3)}$, $f^{(5)}$, $f^{(8)}$	0.26	0.28	0.29	0.29	0.30	0.30	0.30	0.30	0.30

Note: Monthly observation of annual returns. t-statistics are estimated using the Newey-West adjustment with 18 lags. R^2 is in the second to the last row and R^2 for the restricted model is in the last row. Forward rates and returns are based on different smoothing techniques as described in paragraph 3.1.2.

independent variables from now on.

We see a somewhat lower R^2 for the longer sample. Cochrane and Piazzesi (2005a) highlight that using unsmoothed data is important for the success of the forecasting regression. If smoothed data are used, R^2 is substantially lower due to the fact that the important information contained in $f^{(5)}$ is smoothed away (this is more severe for the Svensson method than for cubic spline methods). The lower significance of $f^{(5)}$ in the composed data set is further evidence for this claim. Secondly, the two data sets do not cover the same time period. When the fifties and the early sixties are included in the Cochrane Piazzesi regression (not reported) its R^2 is also lower than in the original data used by the authors, which starts in 1964 and ends 2003[5]. It seems that forecasting power is somewhat lower during this period. According to Fama and Bliss (1987), data quality is doubtful in the pre-1964 period. Consequently, the lower R^2 can also be a result of low-quality data.

It is time for a word of caution. We have selected the four most significant instruments out of ten candidates: a dangerous procedure which makes a severe selection bias very likely. I have simply produced a good example which shows how easy it is to construct a parsimonious and plausible forecasting regression with a high R^2. Once we know how the example is constructed, we can expect an upward biased R^2. I return to this point in the next chapter, where the regression results of this chapter are compared with Kalman filter estimates, which do not suffer from this selection bias.

3.2.2 Graphical Analysis of the Regression Coefficients

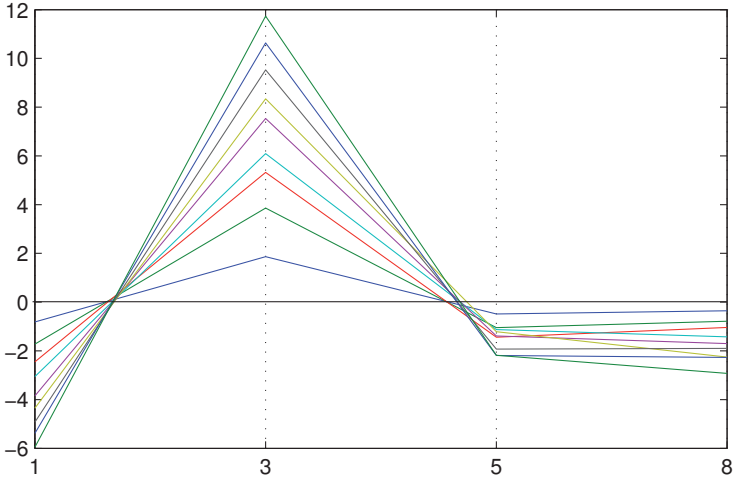
Figure 3.1 plots the coefficient (without the intercept) of the restricted regressions

$$rx_{t+1}^{(n)} = \beta_0^{(n)} + \beta_1^{(n)} y_t^{(1)} + \beta_3^{(n)} f_t^{(3)} + \beta_5^{(n)} f_t^{(5)} + \beta_8^{(n)} f_t^{(8)} + \varepsilon_{t+1}^{(n)}, \qquad (3.2)$$

as a function of maturity. It uses the DL data. Figure 3.2 does the same for the composed data set. There is no legend but the plots are easy to understand. The coefficients for the ten-year bond are those that have the highest absolute values, followed by the nine-year bond and so on until the two-year bond is reached. This has the smallest loadings in absolute terms.

[5]The longer version of the unsmoothed data is available on John Cochranes web page: faculty.chicagogsb.edu/john.cochrane/research/Data_and_Programs/Bond_Risk_Premia/

Figure 3.1: Loadings of the Bond Forecasting Regression: 1970-2006

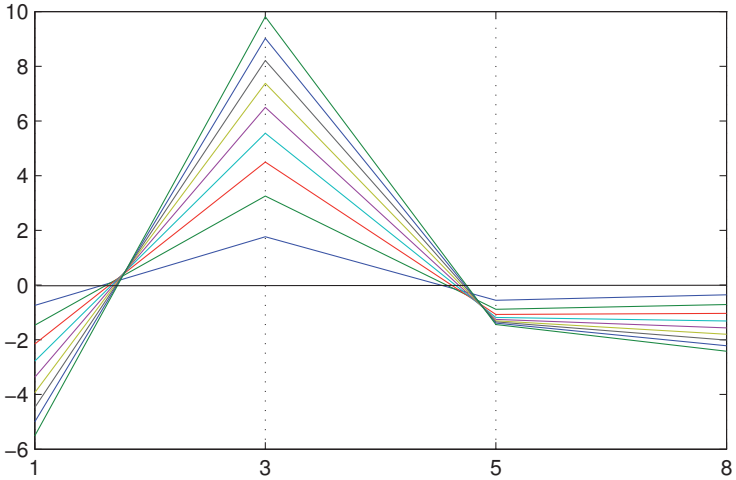


Note: Coefficients of the restricted forecasting regression (3.2) using extended DL data. Maturity for the forward rates is indicated on the x-axis. Monthly observation of annual returns.

These plots look very similar to figure 1 of Cochrane and Piazzesi (2005a). For the one-, three-, and the five-year rates, Cochrane-Piazzesi's tent-shaped forecasting factor is clearly visible. It is slightly more asymmetric than the original picture, which is due to the presence of the eight-year forward rate. Moreover, we see that the coefficients are nearly proportional in both cases (this is because the intersection of the lines is nearly one single point on the graph, and because this point is very close to the abscissa). Following Cochrane and Piazzesi (2005a), we can conclude that these plots make the pattern clear: the same function of the forward rates forecasts holding period returns at all maturities. Longer maturities simply have greater holdings on this function.

While the proportionality of the loadings is still very clear for shorter maturities, it is not so obvious for the eight-year rate. In figure 3.1, the two-year bond line is slightly upward-sloping between the five-year and the eight-year loading while the ten-year line is downward-sloping. This is definitely more pronounced for the composed data set where the ten-year line is

Figure 3.2: Loadings of the Bond Forecasting Regression: 1947-2006



Note: Coefficients of the restricted forecasting regression (3.2) using the composed data set. Maturity for the forward rates is indicated on the x-axis. Monthly observation of annual returns.

clearly downward sloping. Thus the loadings of the eight-year forward rate are not proportional to those of the five-year rates. This, however, is (graphical) evidence against the hypothesis that one single factor predicts all bonds returns as this hypothesis would imply that all coefficients were proportional.

Although the difference is so small as to seem unimportant, we will see in section (3.3) that this is not true.

3.2.3 Is the Forecasting Factor Spanned by Level, Slope and Curvature?

Cochrane and Piazzesi (2005a) find that their forecasting factor is not spanned by level, slope and curvature. However, the forecasting factor estimated here is not exactly the same. Moreover, the first three principal components of one- to five-year rates is not the same thing as those of one to ten-year rates. So their results do not necessarily apply to the present case. Therefore, I redo the analysis.

Table 3.3: Forecasting Level Returns using Level, Slope and Curvature

	$rx^{(2)}$	$rx^{(3)}$	$rx^{(4)}$	$rx^{(5)}$	$rx^{(6)}$	$rx^{(7)}$	$rx^{(8)}$	$rx^{(9)}$	$rx^{(10)}$
Level	−3.63	−3.62	−3.52	−3.36	−3.17	−3.00	−2.83	−2.69	−2.56
Slope	−2.34	−2.73	−3.11	−3.43	−3.71	−3.94	−4.13	−4.28	−4.40
Curvature	2.18	1.96	1.79	1.65	1.54	1.45	1.38	1.33	1.28
R^2	0.21	0.21	0.22	0.23	0.23	0.23	0.23	0.24	0.24
					Adding Forward Rates				
$y^{(1)}$	0.02	−0.13	−0.18	−0.25	−0.34	−0.42	−0.52	−0.61	−0.69
$f^{(3)}$	1.59	1.49	1.35	1.22	1.13	1.06	1.01	0.98	0.95
$f^{(5)}$	−3.52	−3.80	−3.93	−4.02	−4.10	−4.19	−4.29	−4.40	−4.51
$f^{(8)}$	−2.75	−3.26	−3.49	−3.57	−3.59	−3.59	−3.57	−3.54	−3.51

Note: t-statistics and R^2 when the first three principal components are used to forecast bond excess returns using the composed data set. In the second part, t-statistics for additional forecasting variables are displayed. These are from a regression with the three first principal components and the variable indicated on the left-hand side. Monthly forecast of annual returns using the composed data set.

I find that R^2 for the forecasting regression using the first three principal components of yields is about a quarter below that of equation (3.2) as displayed in table 3.3. When the five- or the eight-year forward rate is added, they remain strongly significant (see the second part of table 3.3). We can conclude that the regression-based level forecasting factor is not spanned by level, slope and curvature. This is evidence for the existence of a factor beyond level, slope and curvature which is also important for describing the yield curve. It seems that an affine model should therefore contain (at least) four factors. Cochrane and Piazzesi (2005a) end up with the same conclusion using their forecasting factor.

3.3 A Second Risk Premium for Slope Risk

3.3.1 Duration-Neutral Portfolios

In this section, I analyze portfolios that are nearly orthogonal to single bond returns and thus react predominately to slope shocks. By regarding these slope portfolios instead of single bonds or conventional 'long-only portfolios', I reveal an additional pattern that is hidden in the Cochrane and Piazzesi (2005a, 2008) analysis. I find that not only do level portfolios earn a premium, but so too do slope portfolios. Their premium is comparable to that of long maturity bonds in terms of the Sharpe ratio. However, this premium is tiny if we look at bonds from a conventional long-only view point.

By running forecasting regressions on long-only portfolios as in the previous section, we will only detect the level-related term premium since we predominantly fit the first principal component. Any other movement, including slope movement, is so small that fitting this part of the expected return pattern is of little importance relative to the first principal component and it is not visible in the regressions. To reveal the slope-related predictability pattern, we have to build slope portfolios: portfolios that are not dominated by level movements.

I construct these slope portfolios from bonds with adjacent maturities. First, I run regression of excess returns of the longer maturity bond on the shorter bond's excess return:

$$rx_t^{(n)} + \frac{1}{2}\sigma(rx^{(n)})^2 = a^{(n)} + b^{(n)}(rx_t^{(n+1)} + \frac{1}{2}\sigma(rx^{(n+1)})^2) + \varepsilon_t^{(n)} \qquad (3.3)$$

I also run this regression using ten-year bonds together with two-year bonds. Coefficients of this regression are indicated with a tilde. This portfolio is very close to a linear combination of the portfolios with adjacent maturities that combines these portfolios along the entire yield curve. I then construct the slope portfolios as follows:

$$sr_t^{(n)} = rx_t^{(n)} + \frac{1}{2}\sigma^2(rx^{(n)}) - b^{(n)}(rx_t^{(n+1)} + \frac{1}{2}\sigma^2(rx^{(n+1)})) \tag{3.4}$$

Of course, this is not really a portfolio, since we have summed over two adjusted log returns. However, the goal of this exercise is to show that slope shocks are predictable. Cochrane and Piazzesi (2008) claim that there is only one predictable factor in a log return framework. It is therefore sensible to use log returns to find a counter example to that claim. The same exercise can also be done with simple returns. Results remain very similar in this case (not reported).

The variance adjustments, or Jensens's inequality terms, always appear in a log return factor model. They only affect the unconditional premium and they do not affect predictability. I introduce them to get the Sharpe ratio right. The next chapter rigorously shows how to construct an affine yield curve model that produces exactly the adjustments made in equation (3.4). Section 5.1 treats the general case of log return factor models.

The important property of portfolios constructed as in (3.4) is that they are not sensitive to parallel shifts of the yield curve. They react to slope shifts and possibly to curvature shifts, but certainly not to level shifts. Practitioners would say that they are duration-neutral.

Table 3.4 shows the OLS estimates of equation (3.3). The first row of panel A displays the values of the intercept $a^{(n)}$. Due to the variance adjustment, it can be regarded as the average premium earned by holding the slope portfolio.

The absolute return on these portfolios is so tiny that it seems negligible. However, the volatility of these portfolios is also very low, implying significant risk adjusted returns. (In (3.3) OLS really minimizes the volatility of a portfolio consisting of two bonds with adjacent maturities!). Looking at the Sharpe ratio, defined here as the ratio of the expectation of $sr_t^{(n)}$ divided by its volatility, we find values between 0.14 (for the five- and six-year bond portfolio) and 0.22 (for the portfolio consisting of the two- and ten-year bonds). This Sharpe ratio is lower than that of equity portfolios, yet still significant. Given the 60 years of data, a Sharpe ratio of 0.22 implies a weakly signifi-

Table 3.4: Expected Returns on Slope Portfolios

	Panel A: US, 1947-2006								
	$sr^{(2)}$	$sr^{(3)}$	$sr^{(4)}$	$sr^{(5)}$	$sr^{(6)}$	$sr^{(7)}$	$sr^{(8)}$	$sr^{(9)}$	$\widetilde{sr}$
$a^{(n)}$ (percent p.a.)	0.05	0.06	0.05	0.05	0.05	0.05	0.06	0.06	0.18
t-stat $a^{(n)}$	1.93	1.76	1.47	1.39	1.42	1.47	1.49	1.49	2.02
Sharpe ratio	0.19	0.18	0.15	0.14	0.15	0.15	0.15	0.14	0.22
$b^{(n)}$	0.55	0.72	0.80	0.84	0.87	0.89	0.90	0.91	0.15
Starting point adjustments	0.006	0.007	0.006	0.007	0.004	0.007	0.007	0.008	0.014
	Forecasting Regression								
Intercept ($\times$100)	0.08	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.32
$f^{(10)} - f^{(1)}$	−0.05	−0.07	−0.08	−0.08	−0.07	−0.07	−0.06	−0.06	−0.18
	−5.57	*−6.24*	*−5.65*	*−4.92*	*−4.28*	*−3.78*	*−3.34*	*−2.93*	*−5.22*
intercept (lag,$\times$100)	0.09	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.32
$f^{(10)} - f^{(1)}$ (lag)	−0.05	−0.08	−0.08	−0.08	−0.07	−0.07	−0.06	−0.06	−0.18
	−5.72	*−6.65*	*−6.07*	*−5.14*	*−4.33*	*−3.71*	*−3.24*	*−2.81*	*−5.28*
R^2: $f^{(1)} \dots f^{(10)}$	0.16	0.21	0.21	0.21	0.20	0.18	0.16	0.14	0.21
R^2: $f^{(1)}$, $f^{(10)}$	0.13	0.16	0.16	0.15	0.13	0.10	0.08	0.06	0.15
R^2: $f^{(10)} - f^{(1)}$	0.12	0.16	0.16	0.15	0.13	0.10	0.08	0.06	0.14
	Forecasting with the Level-Forecasting Factor x								
t-stat x: x only	-1.40	-1.29	-1.07	-0.80	-0.48	-0.21	0.00	0.14	0.45
t-stat x: $f^{(10)} - f^{(1)}$ incl.	-0.31	0.00	0.37	0.73	0.95	1.01	1.00	0.97	2.25
R^2: x only	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00

				Panel B: US, 1970-2000					
$a^{(n)}$ (percent p.a.)	0.08	0.03	0.19	−0.20	0.45	−0.18	0.17	0.56	0.32
Sharpe ratio	0.21	0.06	0.34	−0.29	0.41	−0.17	0.15	0.37	0.32
$b^{(n)}$	0.54	0.72	0.81	0.82	0.88	0.85	0.91	0.89	0.14
				Forecasting Regression					
Intercept	0.12	0.09	0.24	−0.13	0.58	−0.21	0.19	0.88	0.51
$f^{10} - f^1$	−0.04	−0.06	−0.03	−0.05	−0.16	0.14	0.05	−0.42	−0.16
	−3.16	*−5.25*	*−1.78*	*−1.64*	*−2.43*	*1.88*	*1.23*	*−4.93*	*−4.87*
R^2: $f^{10} - f^1$	0.06	0.07	0.02	0.03	0.10	0.07	0.01	0.35	0.12
				Panel C: Germany, 1972-2007					
$a^{(n)}$ (percent p.a.)	−0.03	0.01	0.04	0.06	0.07	0.07	0.07	0.06	0.15
Sharpe ratio	−0.12	0.03	0.10	0.13	0.15	0.15	0.13	0.11	0.16
$b^{(n)}$	0.54	0.72	0.80	0.84	0.87	0.89	0.90	0.91	0.14
Starting point adjustments	0.018	0.020	0.017	0.012	0.010	0.008	0.007	0.005	0.043
				Forecasting Regression					
Intercept	−0.01	0.11	0.17	0.21	0.24	0.25	0.25	0.24	0.42
$f^{10} - f^1$	−0.01	−0.04	−0.06	−0.07	−0.08	−0.08	−0.08	−0.08	−0.12
	−0.50	*−1.55*	*−1.95*	*−2.17*	*−2.43*	*−2.44*	*−2.52*	*−2.27*	*−1.59*
R^2: $f^{10} - f^1$	0.00	0.05	0.08	0.09	0.09	0.09	0.08	0.07	0.05

Note: In each panel, the average return, the average Sharpe ratio, and the relative weight of the longer maturity bond $b^{(n)}$ are given. Panels A and C also contain the starting point adjustments (see the appendix to this chapter). The forecasting regression section displays the regression parameter, t-statistics (italics, below) and R^2. t-statistics are estimated using Newey-West adjustments with 18 lags, since monthly observation of annual returns are used. Panel A also includes regression with lagged forward rates and R^2 using more forward rates for forecasting as described in the appendix to this chapter. Panel A uses long data sets; panel B uses DL forward rates and DL returns. The German yield curve in panel C is from the German Federal Bank.

cant t-statistic of 1.73 ($\approx \sqrt{60} \times 0.22$) for the most successful portfolio. The Newey-West adjusted t-statistics for $a^{(n)}$ are, however, higher and equal to 2.02.

Finally, the Sharpe ratio of slope portfolios is similar in magnitude to those of long-only portfolios. For instance, two-year bonds have a Sharpe ratio of 0.22 during the same time span. With higher maturities, the Sharpe ratio decreases monotonically and the Sharpe ratio for ten-year bonds is as low as 0.08.

3.3.2 Forecasting Slope Returns

We have seen that single bond excess returns are predictable with a very high R^2. Can similar predictability be found for slope returns? The answer is yes, but R^2 is certainly much lower than the 30 to 40 percent we see for single bonds.

Panel A of table 3.4 displays results from regressing slope portfolio returns on forward rates using the composed data set. I only use the spread between the ten-year forward rate and the short rate to predict returns. Parameters are estimated by running the following regressions:

$$sr_t^{(n)} = c_0^{(n)} + c_1^{(n)}(f_t^{(10)} - y_t^{(1)}) + \xi_{t+1}^{(n)}. \tag{3.5}$$

In the last row of the forecasting section of panel A, we see that the R^2 of these regressions lies between 0.06 (for $sr^{(9)}$) and 0.16 (for $sr^{(3)}$ and $sr^{(4)}$). t-statistics vary between 6.24 ($sr^{(3)}$) and 2.93 ($sr^{(9)}$) in absolute terms.

Adding more forward rates or relaxing the restriction does not significantly improve the fit of the regression. As we can see in the third last row of the forecasting part, R^2 increases only moderately when all ten forward rates are used. Although nine variables are added on the right hand side, R^2 only increases by an average of 0.05!

The following row shows that using both the one-year rate and the ten-year forward rate separately leaves the regression nearly unchanged. Looking at the coefficients of the unrestricted regression (not reported), it is remarkable how close the negative loadings of the one-year rate are to the loadings of the ten-year forward rate. R^2 for this regression is given in the second to last row of the forecasting section of panel A. It is virtually the same as that of the restricted regression.

The third section of panel A shows t-statistics and R^2 for the forecast using the level forecasting factor x_t. There is no evidence at all that x_t helps predicting slope returns. When only x_t is used on the right-hand side R^2 is close to zero, and the t-statistics insignificant for all maturities. When $f_t^{(10)} - y_t^{(1)}$ is also included, $\widetilde{sr}_t$ is the only portfolio with a significant t-statistics (2.25) for x_t.

From all these regressions we can conclude that the slope of the yield curve predicts returns of slope portfolios and that this predictability is highly significant. Furthermore, the slope of forward rates summarizes the bulk of yield curve information necessary to predict slope returns. Adding other forward rates to the right hand side of equation (3.5) does not significantly improve the forecasting power.

3.4 Curvature and other Risk Premia

So far we have seen that investors are compensated for bearing level and slope risk and that there is evidence that this compensation is time-varying. In order to be rigorous, we should also check if there is a premium for curvature risk. I do this by running regressions similar to (3.3) using two adjacent maturities instead of only one. To be precise, I run a regression of the n-year bonds' excess return on the excess return of the two adjacent bonds with maturity $n-1$ and $n+1$. I build portfolios analogously to (3.4). These portfolios are long in the n-year bond and short in the two neighboring maturities so they are sensitive to changes in the curvature of the term structure. The Sharpe ratio for all these portfolios is negative and never exceeds 0.14 in absolute terms (see table 3.5). Hence, it is not statistically significant different from zero (the t-statistic is $0.14 \times \sqrt{60} = 1.1$). However, the Sharpe ratio is similar in magnitude to the Sharpe ratio of level and slope portfolios.

I also run forecasting regressions using forward rates. There is no sign of a forecasting pattern similar to that of level and slope portfolios. Significant regression parameters occur only infrequently (not reported). R^2 values for the regressions are displayed in table 3.5. They never exceed 0.10.

Finally, I did a similar analysis for other portfolios (not reported). Specifically, I estimated the unconditional Sharpe ratio for all principal components. None of these Sharpe ratios were statistically different from zero. However, the fourth largest principal component has a Sharpe ratio comparable to the

Table 3.5: Forecasting Curvature Returns

	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
Sharpe ratio	−0.08	−0.14	−0.11	−0.08	−0.08	−0.10	−0.11
R^2	0.05	0.04	0.06	0.05	0.07	0.09	0.10

Note: Sharpe ratio of curvature portfolios and R^2 for the forecasting regression using all ten forward rates. Monthly observations of annual returns using the composed data set.

Sharpe ratio of the first three principal components. I also ran forecasting regressions on all principal components and even using all ten forward rates, none of these regressions had a R^2 above 0.10. In short, regression analysis does not provide any evidence of return predictability beyond the first two principal components.

3.5 Summary of the Empirical Findings

In this chapter, the predictability pattern for U.S. government bond returns is explored. This was achieved by running annual forecasting regressions with forward rates as instruments. There are three key findings:

- None of the principal components of bond excess returns has delivered a statistically significant premium over the last sixty years. Some slope portfolios, consisting of a long position in short maturities and a short position in longer bonds, provided positive Sharpe ratios that are weakly significant. The Sharpe ratio on bonds with lower maturities is also significantly positive.

- The first principal component, associated with level shifts, is strongly predictable with an R^2 of about 30 percent. In a simple forecasting regression, we need at least four forward rates to capture all the predictability of level returns.

- The second principal component, associated with slope shifts, is also predictable, though R^2 is lower than for the first principal component. The slope of the yield curve itself summarizes the yield curve information necessary to predict these returns.

In a separate analysis not reported here, these results are confirmed for simple returns. Moreover, the appendix to this chapter adds evidence for predictability of slope returns among German government bonds.

3.A Appendix: Robustness of Slope Return Forecasts

Predictability of slope returns is a new result that is in conflict with previous findings. In order to improve the confidence of this new result, I give a variety of robustness checks for both the existence of a slope premium and its predictability. I show that the results do not depend on the selected data set. I also show that predictability is neither spurious due to measurement error nor are the test statistics highly affected by the small sample size.

3.A.1 Measurement Error and Lagged Instruments

When we run the forecasting regression with lagged forward rates as right hand variables, we can completely rule out spurious predictability due to uncorrelated measurement error. Thus, I also run forecasting regressions using the one-month lagged differences of $f^{10}_{t-1/12} - y^{1}_{t-1/12}$. Regression coefficients and t-statistics can be found in panel A of table 3.4. We can see that using lagged forecasting variables does not change the result. The coefficients remain virtually unchanged and for most of the maturities, the t-statistics are even higher than for the actual value of the forecasting variable.

This result is not surprising. As there is only one variable on the left hand side of the regression, measurement error is only possible for some slope portfolios (e.g. for those including the ten-year bond, since the price of the ten-year bond also enters into the instrument). What is important to see, however, is that the forecasting regression also works with lagged instruments.

3.A.2 Different Data

Panel B of table 3.4 shows the unconditional analysis and the forecasting regression for the DL data. This data is noisier and so the results are not so uniform: two portfolios even have negative Sharpe ratio; others display very positive Sharpe ratios reaching values up to 0.41 ($sr^{(6)}$). On average, there is

still a positive Sharpe ratio of 0.16 and the Sharpe ratio for the portfolio consisting of the two- and ten-year bonds reaches 0.32. This value is marginally significant. The predictability results are in the second part of panel B. They are also noisier and R^2 seems to be lower on average. Notwithstanding, it remains significant for most portfolios.

Of course, the DL data set is not independent from the original one. I also want to test the result with a completely independent data set. I decided to use the default free German term structure because it is easily available from the German Federal Bank[6] and because it is relatively long (it goes back to the year 1972). This data set consists of smoothed yields (Svensson method). Results from the German data are displayed in panel C. The Sharpe ratio is positive for all portfolios with the exception of that with the shortest maturity. The highest value is 0.16 ($sr^{(9)}$) and the Sharpe ratio is quite stable for the longer maturities. Predictability is weaker than for the US data, but remains significant. Overall, the German data provides further evidence for the existence of a small slope premium. The fact that the slope of the yield curve predicts this premium is also confirmed.

3.A.3 Data Range

Historical averages of returns and historical Sharpe ratios are very sensitive to the choice of the time period. The result could be spuriously induced by a lucky starting point. This is not the case here. At the beginning of the data set, the term structure is steeper than at the end. The two-year rate is 0.95 percent and the ten-year rate is 1.83 percent at the beginning (1946:12) versus 4.73 percent and 4.67 percent at the end of the sample (12:2006). We have constructed portfolios that are duration-neutral, so the level effect evens out and we only need to consider changes in the slope. The interest rate increase is lower for the longer rates, so longer maturity returns are relatively higher due to a purely random flattening of the term structure. If we assume that the slope of the yield curve is mean reverting, this effect is a random shift particular to the selected time period. As slope portfolios are short in long maturities, this induces a negative one-off effect for the slope portfolio. It can be approximated as follows:

$$(\text{slope}_{\text{begin}} - \text{slope}_{\text{end}}) \times \text{weight long bond} \times \text{duration}.$$

[6] http://www.bundesbank.de/statistik/statistik.php

For the two-year minus ten-year bond portfolios this is: $\{(1.83-0.95)-(4.73-4,67)\} \times 0.15 \times 10 = 0.87$, or 0.014 percent per year. These adjustments are shown in the fourth row of panel A for the US data. I also report these adjustments for the German data in panel C.

We see that for both data sets, the adjustments are positive. They are rather small and for most portfolios they amount to approximately 10 percent of unconditional average. With regard to these adjustments, it is more likely that the choice of the time period will produce unusually low rather than unusually high returns. The result is thus not the result of some specific starting point, rather the unconditional return of the slope portfolios is underestimated.

Chapter 4

Return Predictability in an Affine Yield Curve Model

Predicting returns is an important task of immense value for practitioners. However, simply running forecasting regressions is not the ultimate goal of academic research. We want to understand why an investor is compensated for holding certain assets, and why this compensation varies over time. A first step in this direction is to incorporate the forecasting evidence into a yield curve model. Such a model must be flexible and allow for independent movement in risk prices. In discrete time, the framework of Ang and Piazzesi (2003) has the desired property.

The first section of this chapter therefore briefly reviews the Ang-Piazzesi framework and points out how the forecasting evidence of the previous chapter can be incorporated into the framework. In particular, it is shown how the empirical finding that the number of factors with time-varying expected returns is smaller than the total number of factors, can be incorporated into the model. This is done by imposing rank restriction to the general version of the model which is derived in the second section.

In the third section, I estimate three and four factor models with all possible number of predictable factors by maximizing the likelihood function using the Kalman filter. I then apply the likelihood ratio test to determine the number of forecastable factors. This analysis favors a four factor model with three forecastable factors.

It appears that estimating affine models by the Kalman filter is straightforward. Unfortunately, this is not the case. The likelihood function is so flat around its maximum value that maximization by standard optimization routines is not reliable at all. Two runs of the same maximization, starting at different initial values, can produce maximum values that vary by as much as 10 or more. Such an inaccurate estimation is, of course, not suitable for an application of the likelihood ratio test.

Therefore, I derive an analytical gradient of the likelihood function. Moreover, in the second section, I derive an economically meaningful parametrization which unequivocally identifies the model. This parametrization is algebraically tractable, so that we can calculate the gradient, and it leads to a likelihood function with one single global optimum. Both these measures are very important for a reliable parameter estimation. Once they are taken, the optimization is much faster and consistently converges to the same maximum. Parameters of two estimates are virtually identical and the variation of the maximum value of the likelihood function using different starting points is below 0.1.

Finally, I discuss some properties of the Ang-Piazzesi model when maturity approaches infinity.

4.1 Constructing Affine Yield Curve Models

4.1.1 Prices and Yields

As already noted, I follow the discrete time homoscedastic essential affine model structure of Ang and Piazzesi (2003) also used in Cochrane and Piazzesi (2008).

Let the vector that describes the state of the economy follow a VAR:

$$x_{t+1} = \alpha + \Phi x_t + \varepsilon_{t+1}, \tag{4.1}$$

where ε_{t+1} is a Gaussian iid error with covariance matrix Σ. The state vector can consist of latent factors, statistical factors such as level, slope and curvature, or economically meaningful factors such as yields. The algebra below is the same in all these cases.

The definition of the model is completed by specifying an exponential

affine pricing kernel:

$$M_{t+1} = \exp(-\delta_0 - \delta_1^T x_t - \frac{1}{2}\lambda_t^T \Sigma \lambda_t - \lambda_t^T \varepsilon_{t+1}) \tag{4.2a}$$

$$\lambda_t = \lambda_0 + \lambda_1 x_t. \tag{4.2b}$$

General asset pricing theory says that the price of an n-period zero bond is equal to the n-period-ahead expectation of the pricing kernel. Based on this proposition, expectations can be calculated recursively following Ang and Piazzesi (2003) or Cochrane and Piazzesi (2005b), which is the web appendix of Cochrane and Piazzesi (2005a):

$$p_t^{(1)} = \log \mathrm{E}_t[M_{t+1}] = -\delta_0 - \delta_1^T x_t \tag{4.3a}$$

$$p_t^{(n)} = \log \mathrm{E}_t[M_{t+1} \exp(p_{t+1}^{(n-1)})]. \tag{4.3b}$$

The result is a relatively simple recursive scheme: log zero bond prices are affine functions of the state variables,

$$p_t^{(n)} = A_n + B_n^T x_t. \tag{4.4}$$

This equation defines how factors are transformed into prices.

Coefficients A_n and B_n can be computed as follows:

$$A_0 = 0 \tag{4.5a}$$

$$B_0 = 0 \tag{4.5b}$$

$$B_{n+1}^T = -\delta_1^T + B_n^T \Phi^* \tag{4.5c}$$

$$A_{n+1} = -\delta_0 + A_n + B_n^T \alpha^* + \frac{1}{2} B_n^T \Sigma B_n \tag{4.5d}$$

where α^* and Φ^* are defined as

$$\Phi^* = \Phi - \Sigma \lambda_1 \tag{4.6a}$$

$$\alpha^* = \alpha - \Sigma \lambda_0. \tag{4.6b}$$

These are also the parameters of the risk-neutral VAR:

$$x_{t+1} = \alpha^* + \Phi^* x_t + \varepsilon_{t+1} \tag{4.7}$$

which defines the dynamics of x_t under the risk-neural measure.

By iterating the definition of B_n, it is easy to see that B_n can be written as a simple geometric series:

$$B_n^T = -\delta_1^T(1 + \Phi^* + \ldots + \Phi^{*n-1}). \tag{4.8}$$

If all the eigenvalues of Φ^* lie in the unit-circle, there is a well known explicit solution for geometric series which in the present case reads:

$$B_n^T = -\delta_1^T(1 - \Phi^*)^{-1}(1 - \Phi^{*n}). \tag{4.9}$$

We are more familiar with yields. From

$$y_t^{(n)} = -\frac{1}{n}p_t^{(n)}$$

we find an affine representation of yields:

$$y_t^{(n)} = A_n^y + B_n^{y,T}x_t$$

with

$$A_n^y = -\frac{1}{n}A_n \tag{4.10a}$$

$$B_n^y = -\frac{1}{n}B_n. \tag{4.10b}$$

Specifically, the risk-free rate reads:

$$y_t^{(1)} = \delta_0 + \delta_1^T x_t. \tag{4.11}$$

4.1.2 Forward Rates

Forward rates are linear functions of prices. So, given prices, we can find forward rate loadings from the definition of the forward rate:

$$\begin{aligned} f_t^{(n)} &= p_t^{(n-1)} - p_t^{(n)} \\ &= (A_{n-1} - A_n) + (B_{n-1}^T - B_n^T)x_t \qquad (4.12) \\ &= A_n^f + B_n^{f,T}x_t. \qquad (4.13) \end{aligned}$$

The resulting coefficients A_n^f and B_n^f are even simpler than those for yields or prices:

$$\begin{aligned} B_n^{f,T} &= B_{n-1}^T - B_n^T \\ &= B_{n-1}^T(1 - \Phi^*) + \delta_1^T \\ &= \delta_1^T\Phi^{*n-1} \end{aligned} \tag{4.14}$$

and

$$\begin{aligned} A_n^f &= A_{n-1} - A_n \\ &= \delta_0 - B_{n-1}^T \alpha^* - \frac{1}{2} B_{n-1}^T \Sigma B_{n-1}. \end{aligned} \tag{4.15}$$

The constant terms also have an explicit representation and need not be estimated recursively as with prices.

4.1.3 Returns

We can also derive similar equations for returns. From the definition of the return we find:

$$\begin{aligned} r_{t+1}^{(n)} &= p_{t+1}^{(n-1)} - p_t^{(n)} \\ &= (A_{n-1} - A_n) + B_{n-1}^T x_{t+1} - B_n^T x_t \\ &= (A_{n-1} - A_n) + B_{n-1}^T \alpha + (B_{n-1}^T \Phi - B_n^T) x_t + B_{n-1}^T \varepsilon_{t+1}. \end{aligned}$$

As returns are stochastic, they also depend on the shocks to the state variables. Thus, there is an extra term multiplying ε_{t+1}. Defining

$$r_{t+1}^{(n)} = A_{n+1}^r + B_{n+1}^{r,T} x_t + C_{n+1}^{r,T} \varepsilon_{t+1}$$

we find the following coefficients:

$$\begin{aligned} B_n^{r,T} &= B_{n-1}^T \Phi - B_n^T \\ &= B_{n-1}^T \Phi + \delta_1^T - B_{n-1}^T \Phi^* \\ &= \delta_1^T + B_{n-1}^T \Sigma \lambda_1 \\ &= \delta_1^T - \delta_1^T (1 - \Phi^*)^{-1} (1 - \Phi^{*n-1}) \Sigma \lambda_1, \end{aligned} \tag{4.16}$$

the constant is

$$\begin{aligned} A_n^r &= A_{n-1} - A_n + B_{n-1}^T \alpha \\ &= \delta_0 + B_{n-1}^T \Sigma \lambda_0 - \frac{1}{2} B_{n-1}^T \Sigma B_{n-1} \\ &= \delta_0 + B_{n-1}^T \Sigma \lambda_0 - \frac{1}{2} \text{var}[rx^{(n)}]. \end{aligned} \tag{4.17}$$

The last equality holds since the coefficients multiplying the shocks are

$$C_n^r = B_{n-1}.$$

Expected returns follow by setting $C_n^r = 0$. Excess returns are also simple, just remove δ_1^T from $B_n^{r,T}$ and δ_0 from A_n^r.

4.1.4 Restricting the Number of Forecastable Factors

The volatility of the pricing kernel, driven by the market price of risk λ_t multiplying the shock ε_{t+1}, generates expected returns in our model. λ_0 captures the permanent part of expected returns and $\lambda_1 x_t$ its time-varying part. There is no other source of variation in expected returns since the model generates homoscedastic returns.

This is most obvious from the affine parameters of returns in (4.16) and (4.17). These equations also show that we can specify the predictability pattern by restricting λ_0 and λ_1. For instance, excess returns are not predictable if we set $\lambda_1 = 0$, because B_n^r is then equal to δ_1 and the expected return thus equal to the risk-free rate plus a constant, as in the multi-factor version of the Vasicek (1977) model. If λ_0 is also zero, the pricing kernel (4.2) is non-stochastic, consequently, the risk neutral VAR (4.7) is equivalent to the state equation (4.1) and there is no risk premium at all.

In the general model with unrestricted λ, each return factor has an associated predictability factor which shows at least some idiosyncratic movement. In this case, it is possible that one factor has a relatively high premium while another factor's premium is relatively low at the same time.

We can restrict the number of independent predictability factors by restricting the rank of λ_1. Then expected returns, but not realized returns, follow a reduced rank factor model. For example, the one factor model of Cochrane and Piazzesi (2005a) implies that λ_1 has rank 1. The regression analysis of the previous chapter indicates a rank of 2. Throughout this chapter, I denote a K-factor model with rank λ_1 equal to L as an $AP(K, L)$ model.

4.2 Parametrization

Kalman filter estimation of the affine model is very demanding for two reasons: firstly, the likelihood function is extremely flat around its maximum, which makes numerical optimization by standard methods impossible. Secondly, many parameterizations are not well identified and the likelihood function can have more than one maximum. I tackle these problems by suggesting an algebraically tractable parametrization with meaningful factors. The relatively simple structure makes it possible to analytically calculate the gradient of the likelihood function (see the appendix to this chapter). Moreover, the unequivocal economic interpretation of the factors identifies the model.

Duffee (2002, 2008b, 2008a) also estimates affine models by maximum likelihood. These papers use the standard rotation of Dai and Singleton (2000) and find that the (quasi-) likelihood function has a large number of local maxima. These papers thus try various starting points and then choose the estimate with the highest likelihood. The approach proposed here obviously differs considerably from this optimization technique. It is much more reliable and far more efficient in terms of computation time.

The first paragraph explains the identification problem and reviews how it has been solved in the literature. The second paragraph describes the parametrization.

4.2.1 Rotating the Factors

If the factors in the VAR (4.1) are latent factors with no particular meaning, there is a fundamental identification problem with affine models. We can perform invariant transformations that leave security prices unchanged but change the factors and the parameter vector. The parameters of the model would remain unidentifiable even if a panel data of all possible fixed income securities were available (Collin-Dufresne, Goldstein, and Jones (2008)).

Typically, the model is identified by rotating the factors such that some parameters are fixed and thus the number of free parameters is reduced. This approach is introduced in Dai and Singleton (2000). Unfortunately, these representations possess the feature that neither the state variables nor the parameters have an economic interpretation. Such latent factor models often lead to models that are locally but not globally identifiable, which is to say that the likelihood function possesses more than one maxima. For instance, Collin-Dufresne, Goldstein, and Jones (2008) show that the so-called central tendency model has this undesirable property. Two estimates of the same model can lead to different parameters and states although the maximized likelihood and the resulting yields are identical.

Alternatively, the model is identified by rotating from the set of latent state variables to a set of observable yields (Duffie and Kan (1996)) or to other economically meaningful or observable variables. States and parameters then are globally identifiable. Unfortunately, this approach is very difficult to implement for many choices of factors because there is not necessarily a closed form solution for yields or prices. For instance, in many models we cannot find a closed-form solution for long-term bond yields (or prices). An

obvious example is the Ang-Piazzesi model just described.

These two observations define requirements for a good parametrization of an affine model: factors should have an economic meaning to achieve global identification and the algebra should be tractable. In the next paragraph, I propose a parametrization that has the desired properties.

4.2.2 An Almost General Parametrization

Inspired by the analysis of Collin-Dufresne, Goldstein, and Jones (2008), I decided to use the risk-free rate and its $1 \dots (K-1)$-period-ahead expectations under the risk-neutral measure as factors[1]. Up to a constant implied by Jensen's inequality, these factors happen to be the first K forward rates. Taking the forward rates as the factors is also possible, though algebraically more cumbersome.

More formally, the factors are defined as follows:

$$x_{t,1} = y_t^{(1)} \tag{4.18a}$$

$$x_{t,k} = \mathrm{E}^*[x_{t+1,k-1}], \tag{4.18b}$$

where the last equation holds for $k = 2 \dots K$ and E^* denotes the expectation under the risk-neutral measure. Assuming that these factors constitute a basis of the state space, the risk-neutral VAR in equation (4.7) implies that Φ^* has the following structure:

$$\Phi^* = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \\ c_1 & c_2 & \cdots & c_K \end{bmatrix}. \tag{4.19}$$

which is known as the companion form. Moreover α^* is

$$\alpha^* = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \alpha_K \end{bmatrix} \tag{4.20}$$

[1]The paper defines a general parametrization for affine yield curve models in continuous time. The state vector comprises infinitesimal maturity yields and their quadratic covariations.

and

$$\delta_0 = 0 \tag{4.21a}$$

$$\delta_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T \tag{4.21b}$$

where $c_1 \dots c_K$ and α_K as well as α, Φ and Σ are free parameters.

The proposed parametrization is very flexible, yet not completely general since the factors defined in (4.18) can be linearly dependent. In this case, the VAR (4.1), together with a subset of prices given by (4.4), is not observable in the sense of control theory (see Harvey (1989) section 3.3 for a definition of observability) even when there are as many or more prices than factors. Non-observability entails that the factors cannot be constructed as affine functions of (observable) bond prices or equivalently so that we cannot take a subset of prices as factors. This links the parametrization (4.18 - 4.21) to the concept of self-consistent models in Cochrane and Piazzesi (2005b). By definition, a model is self-consistent if we can take the first K prices (or equivalently yields or forward rates) as the state variables. Tautologically, the models resulting from the parametrization are thus self-consistent by construction.

Almost equivalently, we could define that a model is self-consistent if any set of K prices can be taken as the factors. The two definitions differ by algebraically interesting yet economically unimportant cases such as $\Phi^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. These cases are thin in the sense of measure theory implying that in empiric work we would detect them with probability zero unless we impose constraints that force Φ^* to have the restricted form. Thus, abusing notation from measure theory, we can say that the two definitions are almost equivalent.

Similarly, for an arbitrary parametrization, the set of parameter vectors implying that the factors defined in (4.18) are linearly dependent is thin. This follows because linear dependence can only be ensured by imposing a restriction on the parameters and thus lowering the dimension of the parameter space. Again, in empiric work we will not detect these cases unless we impose restrictions that force Φ^* to have the required form. In this sense, the parametrization (4.18 - 4.21) is almost general.

Linear dependence amongst the factors in (4.18) has a clear and meaningful economic interpretation. It says that we have more state variables than yield curve factors: a property that is discussed in the next chapter and plays an important role in the rest of this thesis.

A concern about this parametrization is that the factors are highly correlated and we expect some sort of multi-collinearity problem. This, however, is inevitable, or can at least only be transformed to other numerical instabilities, as long as we use factors at the short end of the yield curve, which is necessary to solve for yields. For instance, we could take the consecutive differences of the factors proposed here as the new factors. This means taking $x_{t,1}$ as the first factor, $x_{t,1} - x_{t,2}$ as the second, $(x_{t,1} - x_{t,2}) - (x_{t,2} - x_{t,3})$ as the third and so on. This produces factors that are nearly orthogonal and a matrix Φ^* that also has a treatable structure. The last of these factors, however, has a standard deviation that is some order of magnitude lower than the standard deviation of the first factor, so again, we face a numerical problem.[2]

Empirically, I find that the problem is not serious when parameter estimation is performed in a state space framework. In particular, my experience is that B^y is somewhat unstable, but the resulting yields and the predictability pattern are not affected. This pattern resembles multi-collinearity in a linear regression.

We should also mention that the companion form of a matrix is more flexible than the diagonal form as it allows Φ^* to have not only real but also complex eigenvalues.[3] Note that this generalization is achieved with the same number of parameters.

Finally, there is another interesting property of the companion form of a matrix: the roots of the polynomial

$$P(\lambda) = -c_K - c_{K-1}\lambda - \cdots - c_1\lambda^{K-1} + \lambda^K \tag{4.22}$$

are equivalent to the eigenvalues of Φ^*. This equation is useful to impose parameter bounds.

4.3 Kalman Filter Estimation

In order to perform Kalman filter estimation, we need to put our system into the state space form. This is done by augmenting the model-implied yields in

[2]Formally, the conditioning number of Φ and Φ^*, as a measure of numerical stability, is a similarity invariant and thus, like yields, also an invariant under invariant transformations of affine models.

[3]Duffee (2008a, 2008b) estimate yield curve models with diagonal Φ^*.

equation (4.10) by a measurement error η_t:

$$y_t^{(n)} = A_n^y + B_n^{y,T} x_t + \eta_t^{(n)}. \tag{4.23}$$

As required by the Kalman filter, $\eta_t^{(n)}$ is assumed to be serially and cross-sectionally (not required) uncorrelated, and to be uncorrelated with the innovations in the state equation, which is of course equation (4.1). Furthermore, in order to keep the number of parameters small, it is assumed that the standard deviation h of $\eta_t^{(n)}$ is the same for all maturities. The covariance matrix of η_t is thus a scaled identity matrix. The empirical validity of this measurement error specification is discussed in the last paragraph of this section.

Maximum likelihood estimation by means of the Kalman filter is now straightforward: Take the VAR in (4.1) as the state equation and (4.23) as the measurement equation.

All parameters, including h, are jointly estimated base on quarterly data. The cross-section consists of yields with maturities of 3, 6 and 9 months as well as with 1 - 10 years. In order to compare the estimate with the reminder of this thesis, I start the DL data set only in June 1947.

Unlike the case of the return regression, the Kalman filter approach makes it possible to estimate a quarterly model with implied quarterly return forecasts. Since the measurement error is explicitly modeled, the Kalman filter is able to reveal the predictability pattern for a much lower signal-to-noise ratio as is possible with a linear regression. Moreover, we can use the entire cross-section, which helps smoothing away measurement error of single maturities and thus improves the quality of the forecasting instruments. Consequently, all results in the rest of this thesis are based on quarterly models.

Among others, Kalman filter estimates of affine yield curve models has previously been performed by Duffee (2008a, 2008b) as well as Brennan, Wang, and Xia (2004). Duffee and Stanton (2004) show by numerical simulation that Kalman filter estimates of yield curve model is very efficient.

The first paragraph of this section answers the question under which circumstances forecasts provided by the Kalman filter are more efficient than those of a linear regression. Paragraph 2 deals with model selection. The third and the fourth paragraphs explore the predictability of level and slope portfolios. As already mentioned, the section is completed by an analysis of the measurement error.

4.3.1 Is Kalman Filter Estimation more Efficient?

If a K-factor model is self-consistent, as it is the case here, we can take the first K prices as the state variables. This determines $A_1 \dots A_K$ and $B_1 \dots B_K$ as well as δ_0 and δ_1. Moreover, we can fit the state space VAR in equation (4.1), using OLS, for example, which determines α, Φ and Σ. By the recursive scheme (4.5) we can calculate α^* and Φ^* from the cross-section of prices and finally solve for λ_0 and λ_1. One single linear regression of observable variables determines the full specification of the model. In the absence of measurement error, OLS is equivalent to GLS because all the equations have identical explanatory variables. Thus, the simple OLS estimator is efficient. Consequently, the Kalman filter cannot provide a better forecast. In fact, the two techniques are equivalent because both are maximum likelihood estimates and thus produce exactly the same forecast (see Joslin, Singleton, and Zhu (2010) and Greene (1997), chapter 15).

Real data, however, is noisy and, in this case, the Kalman filter is more efficient. It disentangles 'true yields' from measurement error. In particular, the cross-sectional regression (the measurement update step) involved in the Kalman filter produces a predetermined number of instruments as well as returns for all yields that are more precisely measured than any combination of the original yields. As outlined in the previous paragraph, forecasts based on the Kalman filter and forecasts based on OLS are equivalent once we have clean data. Consequently, it is really the ability to clean the data that makes the Kalman filter a superior estimation technique.

A second advantage of the Kalman filter is that it tackles the selection bias in a self-consistent model. Once the distribution of the measurement noise and the dimension of the state space are defined, the Kalman filter objectively selects the forecasting variable and leaves no subjective judgement to the researcher. There is no possibility of data mining in this set-up. Results from this already unambiguous process can even be improved by the rank reduction technique of paragraph 4.1.4. This technique provides testable restrictions on the parameter space which leave no subjective judgement to the researcher. As a consequence, the forecasting evidence based on the Kalman filter is essentially free of any selection bias.

4.3.2 Model Selection: Testing for the Rank of λ_1

Four Factor Models

I estimate the affine four factor model for all possible ranks of λ_1 $(0-4)$. Results for the DL data are summarized in panel A of table 4.1. Panel B contains the evaluation for the composed data set. The maximum values of the log likelihood function are given in the first line of each panel. I use them to perform a likelihood ratio test for the rank of λ_1 implied by the number of free parameters displayed in the last row of the same table. The last line of the table contains critical values of the likelihood ratio test.

For the extended DL data, we find that the log likelihood of the unrestricted model is nearly identical to the version of the model with rank λ_1 equal to 3. Consequently, the likelihood ratio test is not able to reject this restriction. For the null of the rank 3 model against the alternative of the rank 2 model, the likelihood ratio test statistics (twice the difference between the maximum values of the log likelihood functions of the two models) takes the value $17.9 = 2(6992.3 - 6984.3)$. The 1 percent critical value for 3 degrees of freedom is 11.3, thus the restriction is strongly rejected. Similarly, the likelihood ratio tests indicates that the rank 2 model dominates the rank 1 model, which itself dominates the model with $\lambda_1 = 0$. For both tests, the p-value is far below 0.01. These findings are strongly confirmed in the longer data set.

Overall, the evidence based on the likelihood ratio test is very compelling: among the four factor models, the rank 3 model consistently dominates all other specifications. Therefore, I mainly discuss the properties of the $AP(4,3)$ model in the next section.

Three Factor Model

The results for the three factor models are akin to the four factor case and summarized in table 4.2 (which is organized in the same way as table 4.1). Again, the model with a rank one smaller than the number of factors (the $AP(3,2)$ model) is very similar to the unrestricted specification. This model is also the dominant model and the implied restriction is not rejected by the likelihood ratio test. Further rank restrictions, however, are clearly rejected by the test. p-values are again far below 0.01.

Table 4.1: Kalman Filter Estimates of Four Factor Models

	Panel A: US, 1970-2006				
	$AP(4,0)$	$AP(4,1)$	$AP(4,2)$	$AP(4,3)$	$AP(4,4)$
Log Likelihood	9659.0	9673.2	9684.3	9692.3	9692.4
Std. Dev. of h	9.13	9.11	9.11	9.07	9.07
R^2 of $rx^{(10)}$	-	-0.026	0.234	0.234	0.237
R^2 of $\widetilde{sr}$	-	0.036	0.056	0.161	0.183
Correlation	-	-	0.390	0.007	-0.070
Max. Eigv. of Φ	0.997	$0.97 \pm 0.02i$	$0.96 \pm 0.01i$	0.956	0.954
Max. Eigv. of Φ^*	0.997	0.998	0.998	0.998	0.998
	Panel B: US, 1947-2006				
Log Likelihood	15947.5	15973.8	15987.1	16000.1	16000.1
Std. Dev. of h	7.45	7.43	7.43	7.40	7.40
R^2 of $rx^{(10)}$	-	-0.016	0.230	0.219	0.219
R^2 of $\widetilde{sr}$	-	0.069	0.057	0.181	0.181
Correlation	-	-	0.140	0.035	0.047
Max. Eigv. of Φ	0.994	0.984	0.973	0.979	0.980
Max. Eigv. of Φ^*	0.994	0.994	0.994	0.993	0.993
Free Parameters	20	27	32	35	36
$\chi^2_{0.99}$	18.5	15.1	11.3	6.63	

Note: Panel A contains results for the extended DL data set and panel B for those of the composed data set (June 1947 - June 2006) using quarterly data of the following yields: $y^{(1/4)}, y^{(1/2)}, y^{(3/4)}, y^{(1)}, \ldots, y^{(10)}$. For each panel, the first row gives the maximum value of the likelihood function for the model indicated in the header of panel A; the second row, the estimated value of the standard deviation h. The next two lines give the R^2, as defined by equation (4.24), for the implied return forecast on $rx^{(10)}$ and a $\widetilde{sr}$ (using the filtered series). The fifth line shows the correlation between the two expected returns. The last two lines contain the eigenvalue with the largest absolute value of the two matrices Φ^* and Φ. Finally, the second to last line of the table indicates the number of free parameters and the last line $\chi^2_{0.99}$ for the likelihood ratio test between the model in the corresponding column and that in the next column.

As in the four factor case, these results are strongly confirmed by the composed data set.

Table 4.2: Kalman Filter Estimates of Three Factor Models

	Panel A: US, 1970-2006			
	$AP(3,0)$	$AP(3,1)$	$AP(3,2)$	$AP(3,3)$
Log Likelihood	9444.9	9453.8	9461.7	9461.8
Std. Dev. of h	11.3	11.3	11.3	11.3
R^2 of rx^{10}	0.000	0.218	0.251	0.251
R^2 of $\widetilde{sr}$	0.000	-0.014	0.089	0.096
Correlation	-	-	0.030	-0.05
Max. Eigenvalue of Φ	0.994	0.92+0.03i	0.952	0.949
Max. Eigenvalue of Φ^*	0.994	0.994	0.994	0.994
	Panel B: US, 1947-2006			
Log Likelihood	15553.7	15566.0	15578.6	15578.7
Std. Dev. of h	9.49	9.49	9.49	9.49
R^2 of rx^{10}	-	0.247	0.227	0.228
R^2 of $\widetilde{sr}$	-	-0.027	0.091	0.090
Correlation	-	-	0.158	0.119
Max. Eigenvalue of Φ	0.996	0.970	0.979	0.979
Max. Eigenvalue of Φ^*	0.996	0.996	0.996	0.996
Free Parameters	14	19	22	23
$\chi^2_{0.99}$	15.1	11.3	6.63	

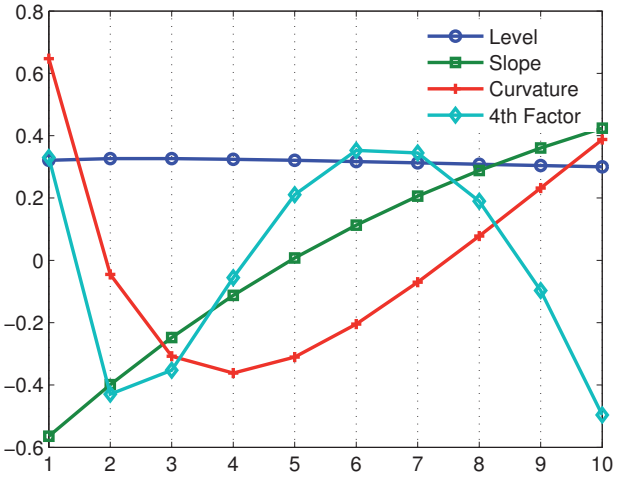
Note: See table 4.1 for explanation.

Three or Four Factors?

Comparing the results of the $AP(3,2)$ model with the favored four factor model ($AP(4,3)$), we see that the difference in the likelihood function is huge. The measurement error is also reduced by about 20 percent. Clearly, $AP(4,3)$ is the dominant model.

In order to better understand the nature of the fourth factor, I also perform a principal component analysis from the yields implied by the $AP(4,3)$ model. The result is plotted in figure 4.1. The new factor looks much like a cosine. It loads positively on the one-year rate, then negatively on the two- and three-year yields, then again positive on the six- and seven-year rate and finally negatively on the ten-year yield. If we include the below-one-year maturities, they all load positively on the fourth factor. It is apparent that we need this

Figure 4.1: Principal Component Analysis of $AP(4,3)$ implied Yields



Note: PCA on one- to ten-year yields implied by the $AP(4,3)$ model based on Kalman filter estimates using quarterly extended DL data (1970-2006).

extra factor for a better fit of the very short end of the yield curve.

This fourth factor is a sensible factor that looks the way we expect it to be. Moreover, the factor is slowly mean-reverting and not highly negatively-autocorrelated as pure measurement error. Note that the principal component analysis based on the raw data does not reveal this factor (not reported). Apparently, this factor is hard to disentangle from measurement noise.

Three factor models such as $AP(3,2)$ and $AP(3,1)$, however, are also interesting as they not only capture the bulk of the cross-sectional variation of yields, but also the forecasting evidence of long only portfolios.

4.3.3 Predicting the Excess Returns on Long Bonds

Four Factor Models

The forecasting ability of the Kalman filter estimates is judged by calculating R^2 as 1 minus the forecasting variance divided by the variance as follows:

$$R^2 = 1 - \frac{\sum(rx_t^{(10)} - \mathrm{E}_{t-s}[rx_t^{(10)}])^2}{\sum(rx_t^{(10)} - \overline{rx}^{(10)})^2} \tag{4.24}$$

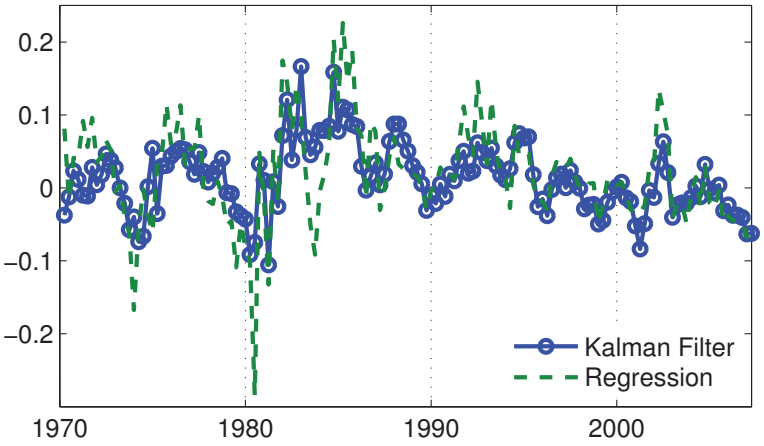
$rx_t^{(10)}$ is the annual GSW return and $\mathrm{E_{t-s}}[rx_t^{(10)}]$ is the forecast from the Kalman filter estimates (filtered series) using data known at $t-4$. The third line of panel A in table 4.1 shows that we find an R^2 of 0.23 for the annual returns implied by $AP(4,3)$. This is considerably below the 0.35 of the regression analysis, but still very high. This decrease, however, is no surprise. I have already discussed this point in section 3.2.1: we expect a selection bias in the regression approach as we have selected the 4 most significant instruments out of 10 candidates. This point is also discussed in the next paragraph.

A similar R^2 is also found for the rank 2 and the rank 4 models. The $AP(4,1)$ model, however, has only very little forecasting power for long bonds. This is further evidence against the one factor specification of Cochrane and Piazzesi (2008).

The forecasts generated by the $AP(4,3)$ model have a correlation of 0.73 with the regression forecast. Moreover, as we can see from figure 4.2, they also follow a very similar pattern. The main difference between the two is that the regression forecast sometimes takes more extreme values. In particular, the most negative expected return implied by the Kalman Filter is -11 percent. In comparison, this value is -28 percent using the regression approach. In our homoscedastic model, this value implies a conditional Sharpe ratio of more than 2.5, a number that is probably hard to explain by any utility-based model.

Finally, note that the Kalman filter estimates quarterly forecasts by fitting the quarterly VAR of the state variables. The one-year forecast is then produced by a 4-step-ahead forecast of this VAR. Nevertheless, these forecasts are, except for some unrealistic outliers, very similar to those of the annual regression. This shows how well the quarterly VAR, fitted by the Kalman filter, captures the one-year dynamic of the yield curve.

Figure 4.2: Expected 1-year Excess Return on a 10-year Bond: 1970-2006



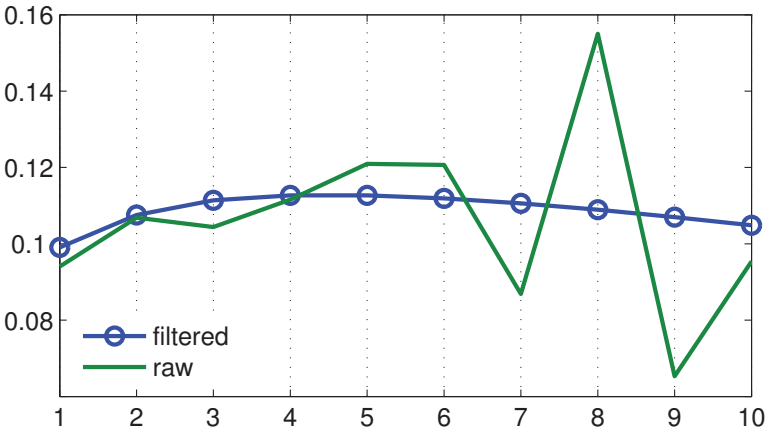
Note: Quarterly Kalman filter estimates are based on the extended DL data set. Parameters of the regression approach are estimated from monthly overlapping data as in table 3.1.

Understanding the Bias in the Forecasting Regression

In a separate analysis that is not reported here, I estimate regression implied excess returns by randomly selecting four yields from the yield curve. These forecast are typically much closer to those of the Kalman filter (correlation is about 0.85) and they have an R^2 that is also close to the 0.23 found for the Kalman filter. Hardly any quadruple produces more extreme outliers than the original regression.

This analysis also reveals that a large fraction of the higher R^2 for the original forecasting regression stems from a much better fit of two data points only: March and June 1983. The difference in the forecast is clearly visible in figure 4.2. Note that most of the yield quadruples produce a forecast that is much closer to that of the Kalman filter. Furthermore, the Kalman smoother detects a huge 5 percent measurement error for the eight-year forward rate at these two data points. Figure 4.3 plots DL forward rates (remember they are calculated by the unsmoothed Fama-Bliss method) together with the forward rate implied by the Kalman smoother. The unsmoothed yield curve with its

Figure 4.3: Smoothed and Raw Forward Curve: June 1983



Note: The smooth forward curve is based on the Kalman filter. The raw data is from the DL data set.

sawtooth long end, is so unrealistic that it is hard to interpret as the 'true yield curve'.

Apparently, a huge measurement error has produced a boost in the fit of the forecasting regression. This is simply a very instructive example of how easy it is to overfit a forecasting regression. Among the $\binom{10}{4}$ quadruples we have selected one such that a huge measurement error at the right time produces a very accurate (insample) forecast for a few extreme return observations.

Note that this kind of bias is different from the type of measurement error discussed in the previous chapter (and in Cochrane and Piazzesi (2005a)). Here a few outliers followed by extreme returns boost the R^2. In the previous section we ruled out the possibility that many small measurement errors are responsible for the very good fit of the regression.

This should have aroused our suspicion earlier. In table 3.1, we see that some forward rates are highly significant while others are not. This looks meaningful in a pure regression framework, as it helps us to select the forecasting instruments. If we, however, assume that yields are generated by an

affine model, we expect another pattern. In this case, none of the forward rates should be significant! In a four factor model, four yields already contain all information. Adding an extra yield on the left hand side of the forecasting regression does not add any new information and thus its t-statistics should be zero. Consequently, when we use the entire yield curve for forecasting, we should observe a significant overall regression with no single t-statistics different from zero.

What about the tent-shaped forecasting factor?

From the smoothed yields, we can construct forward rates and the annual expected excess return on the 10-year bond. This expectation is, by construction, an affine combination of any quartuple of forward rates. Taking the one-, three-, five-, and eight-year rates as in the annual regression, we can try to reconstruct the tent-shaped factor of figure 3.1. Figure 4.4 presents these annual excess return forecasting coefficients implied by the $AP(4,3)$ model estimated by the Kalman filter.

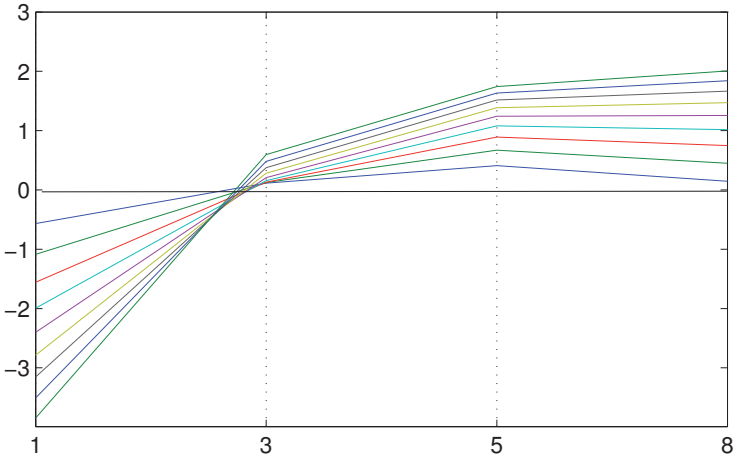
The coefficients are nothing at all like the tent-shaped factor we find in direct annual regression. Rather, they look like a combination of the conventional slope and curvature factors. In fact, the three factors explain more than 0.99 percent of expected excess returns on the ten-year bond. Moreover, the slope factor alone explains as much as 73 percent of this return. We are very close to the original bond forecasting regression of Fama and Bliss (1987). Further evidence on this claim is given in the next paragraph where the results of the three factor model are discussed.

Three Factor Models

Given the annual forecasting regressions, probably the most surprising result of the Kalman filter estimation is that the $AP(3,1)$ model already contains all of the predictability of long-only bond returns (see the third line of panel B in table 4.2). This is actually the three factor version of the Cochrane-Piazzesi model. R^2 for the level factor is 0.25 and thus slightly higher than for the $AP(4,3)$ and also the highest value among all models estimated here.

This result is in strong opposition to the annual regression. It reveals that, once properly smoothed, the conventional level, slope and curvature factors capture all of the predictability of excess returns on long bonds.

Figure 4.4: Forecasting Regression Parameters Implied by the Kalman Filter



Note: Coefficients of a regression of annual excess returns on $y^{(1)}$, $f^{(3)}$, $f^{(5)}$, $f^{(5)}$ implied by the parameters of the $AP(4,3)$ model estimated by the Kalman filter. Quarterly extended DL data set (1970:2006). The constant term is omitted.

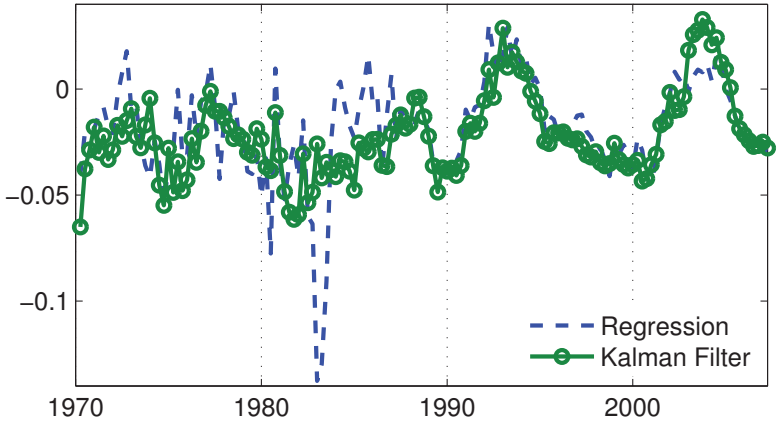
Finally, $AP(3,2)$ is not able to improve the forecasting power for single bonds in the longer data set.

4.3.4 Predicting Slope Portfolios

Four Factor Models

As in the previous chapter, I predict $\widetilde{sr} = rx^{(10)} - 6.6rx^{(2)}$, which is a difference of log returns and thus not strictly speaking a portfolio. As with level returns, I estimate R^2 using (4.24). The $AP(4,3)$ model estimated by the Kalman filter is able to improve R^2 which is now 0.16 (see panel A of table 4.1). Both the Kalman filter forecast and the regression forecast are plotted in figure 4.5. They are relatively similar in the second half of the sample. In the first half, however, the regression forecast is much more volatile. The largest deviation between the two forecasts is again in March and June 1983.

Figure 4.5: Expected 1-year Excess Return on a Slope Portfolio: 1970-2006



Note: Quarterly data based on the extended DL data set. Parameters of the regression approach are estimated from monthly overlapping data.

Three Factor Models

In general, the forecasting power for the slope portfolio is reduced in the three factor specification when compared with the four factor models. The $AP(3,1)$ model has no power at all, as we can see from the fourth line of each panel in table 4.2. $AP(3,2)$ has an R^2 of 0.09 at least which, however, is considerably below the 0.16 of $AP(4,3)$.

We see that the four factor model produces better forecasts than either the annual regression or as the three factor model. Apparently, the fourth factor considerably improves R^2 for the slope return regression (the PCA shows that level, slope and curvature explain 73 percent). This is because on the one hand, the three factor model cannot use the fourth factor and thus its fit is lower; on the other hand, the regression analysis is not able to disentangle the signal (the fourth factor) from the measurement noise and thus is also unable to use its information. Only Kalman filtering with four factors reveals the proper signal and uses it for a better forecast of slope portfolios.

4.3.5 Analysis of the Measurement Error

As above, I discuss the properties of the four factor model with three predicability factors. If the results are qualitatively different, I also discuss the three factor model with rank λ_1 equal to 2. As it is common practice, I ignore the fact that the measurement errors are based on estimated parameters and thus the usual test statistics are slightly biased. All results are for the extended DL data set.

Bias at Different Maturities

Although the mean of the measurement error for the 6-year bond is positive and highly significant (t-value above 4), there is no pattern visible for the mean. For the three factor model other maturities are significant, but there is no pattern either.

Variance at Different Maturities

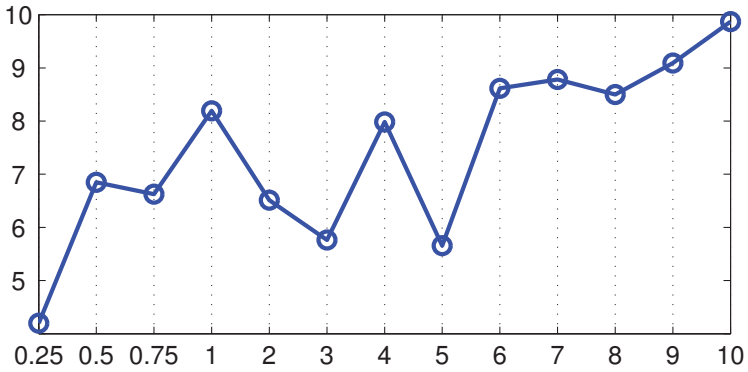
Figure 4.6 shows that the short rate has clearly the lowest variance and that the standard deviation is highest for the 10-year yield. Regressing standard deviations against maturity (and an intercept) provides strong, highly significant results with a t-value of 3.6. In contrast to this, using smoothed yields, Brennan and Xia (2002) find that standard deviations of the measurement error declines for maturities over 5 years. Finally, for the three factor model, there is no significant trend in the variance.

The principal component analysis clearly explains the difference between the three and the four factor models: the fourth factor captures important information at the short end of the yield curve. Thus, the variance of the measurement error at the short end is substantially higher in the three factor model compared to $AP(4,3)$ (see also the discussion of cross-sectional correlation below).

Heteroscedasticity in the Time-Series

Figure 4.7 plots the cross-sectional variance for each date. We can clearly see that during, and for some quarters after the disinflationary period of the early eighties, the measurement error is significantly higher. Before and after this period, the variance is relatively stable, although it seems that it is slightly

Figure 4.6: Standard Deviation of Measurement Error in the Cross-Section



Note: Standard deviation of the estimated measurement errors form the $AP(4,3)$ model. Quarterly extended DL data, 1970:2006.

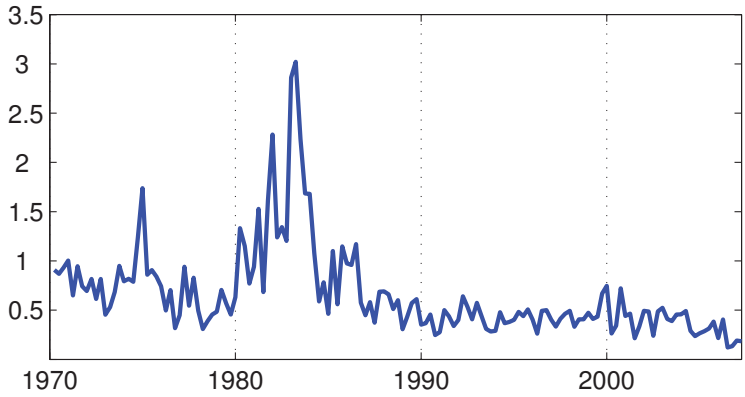
higher in the seventies than at the end of the sample. There is also a very high value in 1975.

Correlation in the Cross-Section

The measurement error of the three-month rate is significantly negatively correlated with the next two bonds. This is much more pronounced with the three factor model. Here, the correlation between the 3-month bond and the 3-quarter bond is as high as 0.86. The principal component analysis in figure 4.1 shows that the fourth factor captures some extra dynamics at the short end of the yield curve, so this does not come as a surprise. The three factor model is not able to capture this factor and thus generates correlated errors.

A similar pattern is also visible at the long end of the yield curve. However, this only holds for the four factor model. Besides this, there are also quite a few other correlations that are significant. Given the level of heteroscedasticity we observe, that is really what we expect. These values, however, occur randomly and cannot be attributed to a common factor.

Figure 4.7: Standard Deviation of Measurement Error in the Time-Series



Note: Standard deviation for measurement errors of all 13 bond yields at each data point based on the $AP(4,3)$ model. Quarterly extended DL data, 1970:2006.

Autocorrelation

Probably the highest concerns about the specification of the model is the autocorrelation of the measurement error for longer maturities. It peaks for the 6-year bond where its value is 0.70. Two lag autocorrelation is also significant for longer maturities.

A possible economic interpretation of the autocorrelation for six-year bonds is that these bonds are exceptionally illiquid (who is trading six-year bonds?). Due to their illiquidity, their measurement error is autocorrelated on the one hand, on the other hand, if they are traded, they are traded at a discount because investors prefer liquid assets. This also generates the higher yield we have discussed just above in the bias paragraph.

The cross-autocorrelation, however, is only significant for a few values that occur randomly.

4.4 Infinite Maturity

The behavior of the yield curve when maturity approaches infinity is not determined by the affine set-up. In principal, it allows the long end to vary with the state variables (when the maximal eigenvalue of Φ^* is equal to 1) or remain fixed (all eigenvalues of Φ^* have absolute values below one).

We can, however, exclude an eigenvalue equal or greater than 1 by inspection of equation (4.15)

$$A_n^f = -B_{n-1}^T \alpha^* - \frac{1}{2} B_{n-1}^T \Sigma B_{n-1}. \tag{4.25}$$

Unlike in the case of eigenvalues below one, B_{n-1}^T is not bounded if Φ^* has eigenvalues greater than one. In this case, B_n grows linearly with n (see equation 4.8). The second term on the right hand side thus grows with n^2 and cannot be compensated by the first term as it grows only with n. Finally, the second term is not zero since Σ is positive-definite. The forward rate therefore approaches minus infinity when n approaches infinity. Consequently, eigenvalues of Φ^* should be bounded below 1.

Empirically, the largest eigenvalue can be as high as 0.997 or 0.998 and these values are not statistically different from one. Furthermore, there are no signs of convergence as long as the yield curve is observable. In fact, we are very close to a model with a unit root empirically, but we can exclude it for theoretical reasons.

So far we know, the eigenvalues of Φ^* are bounded, implying that B_{n-1} is also bounded. Still, this expression can become very large (in absolute terms) as we are very close to a unit root model. A very large B_{n-1} allows the infinite maturity forward rate in (4.25) to take almost any value. Empirically, I find that the model implies implausible and very negative forward rates for very long maturities. The yields remain in a plausible range only for about the first few hundred quarters. After that, they decline strongly and finally converge to a value below -1. It seems that the model does not get the very long end of the term structure right when it is estimated from bonds with only limited maturities. Economically, the empirically estimated parameters of the model imply some sort of bubble[4].

[4]Dybvig, Ingersoll, and Ross (1996) claim that a falling long forward or zero-coupon rate, as we observe it here, implies an arbitrage opportunity. The present approach, however, is free of arbitrage as it is based on a pricing kernel, but still the long forward rate falls. The

Table 4.3: Kalman Filter Estimates with Restrictions on A_∞^f

A_∞^f	free	0.00	0.05	0.10	0.15	0.20
	Panel A: Max. Value of the Likelihood Function					
$AP(4,0)$	9659.0	9652.4	9649.9	9642.3	9622.7	9591.7
$AP(4,1)$	9673.2	9667.0	9664.1	9657.5	9638.7	9607.4
$AP(4,2)$	9684.3	9679.0	9675.8	9668.7	9650.3	9616.5
$AP(4,3)$	9692.3	9686.7	9683.8	9676.4	9658.4	9624.1
$AP(4,4)$	9692.4	9687.0	9684.0	9677.0	9658.4	9624.5
	Panel B: R^2 of Level Returns					
$AP(4,0)$	-	-	-	-	-	-
$AP(4,1)$	-0.026	-0.002	0.000	-0.017	-0.025	-0.021
$AP(4,2)$	0.234	0.240	0.242	0.263	0.251	0.257
$AP(4,3)$	0.234	0.240	0.239	0.255	0.250	0.238
$AP(4,4)$	0.237	0.238	0.240	0.245	0.250	0.239
	Panel C: R^2 of Slope Returns					
$AP(4,0)$	-	-	-	-	-	-
$AP(4,1)$	0.036	0.039	0.039	0.044	0.044	0.041
$AP(4,2)$	0.056	0.057	0.060	0.062	0.065	0.055
$AP(4,3)$	0.161	0.169	0.171	0.167	0.169	0.169
$AP(4,4)$	0.183	0.180	0.178	0.175	0.169	0.178
	Panel D: Correlation between Level and Slope					
$AP(4,0)$	-	-	-	-	-	-
$AP(4,1)$	-	-	-	-	-	-
$AP(4,2)$	0.390	0.300	0.277	0.256	0.192	0.184
$AP(4,3)$	0.007	-0.049	-0.082	-0.154	-0.103	0.001
$AP(4,4)$	0.070	-0.140	-0.128	-0.123	-0.103	0.104

Note: Quarterly extended DL data (1970-2006). The first line indicates the (restricted) value of A_∞^f. The first column shows the model. R^2 based on smoothed return forecasts. For more details see also table 4.1.

In view of this result, we should try to implement a no bubble condition. Unfortunately, there is no theoretical guideline as to how this should be done.

apparent contradiction stems from an overly general definition of arbitrage in this paper (see McCulloch (2000)).

The only thing we know, is that A_∞^f should take some reasonable value. Therefore, I re-estimate the model by restricting A_∞^f to annual values of 0 to 8 percent and compare the results with the original estimate. Doing this, I find that the restriction does not influence the predictability pattern. The model preserves its forecasting ability for level and slope portfolios and the correlation between the two forecasting factors also remains similar (see panels B, C, and D of table 4.3).

The maximum value of the likelihood function, however, falls dramatically within the plausible range of A_∞^f (panel A of table 4.3). Whereas restricting A_∞^f to 0 lowers the likelihood by about 5, the decrease is nearly 70 for $A_\infty^f = 0.02$ (an annual rate of 8 percent). A likelihood ratio test would thus reject the restriction. The decrease in the likelihood is also reflected in an increase in the measurement error (not reported).

4.5 Summary of the Empirical Findings

In this chapter I, have reviewed the predictability pattern detected in the annual regression analysis of the previous chapter by means of maximum likelihood estimation using the Kalman filter. The key findings are:

- $AP(4,3)$ is the dominant model.
- Both three and four factor models predict annual returns on level portfolios with an R^2 of up to 0.25.
- The reduction of R^2 compared with the annual regression can be attributed to a selection bias implied by outliers in an environment with high measurement error.
- The conventional level, slope and curvature capture all of the remaining predictability of level portfolios.
- Slope portfolios are also predictable with an R^2 of up to 0.18.
- The fourth factor revealed in the four factor model is important for predicting slope returns. Regression analysis cannot use it because it is hard to disentangle from measurement noise.

4.A Appendix: Derivatives

This appendix summarizes the derivatives used to calculate the gradient of the likelihood function for the Kalman filter estimation method. Once one is familiar with matrix calculation, most derivatives are fairly easy. A comprehensive online formulary of matrix derivatives is the matrix cookbook of Kaare Brandt Peterson and Michael Syskind Peterson[5].

I deduce formulas for prices; results for yields follow from (4.10). If the rank of λ_1 is smaller than K, we can apply the chain rule to find derivatives for the restricted parameter vector. This is easiest by parameterizing $\Sigma\lambda_1$ and then putting $\Phi = \Phi^* + \Sigma\lambda_1$ implied by (4.6a).

We only need derivatives with respect to A_n and B_n. Then we can construct the derivative by the chain rule using the derivatives of the Kalman filter in section 2.3.

4.A.1 Derivatives of B

Write B in the form given in (4.8):

$$B_n^T = -\delta_1^T \sum_{i=0}^{n-1} \Phi^{*i} \tag{4.26}$$

and find the derivative with respect to c:

$$\frac{\partial B_n^T}{\partial c_k} = -\sum_{j=0}^{n-2} \Phi_{1,K}^{*j} \cdot \left(B_{k+j}^T - B_{k-1}^T\right). \tag{4.27}$$

The proof is as follows:

$$\begin{aligned}
\frac{\partial B_n^T}{\partial c_k} &= -\delta_1^T \sum_{j=0}^{n-2} \sum_{i=j+1}^{n-1} \Phi^{*j} \frac{\partial \Phi^*}{\partial c_k} \Phi^{*(i-1-j)} \\
&= -\sum_{j=0}^{n-2} \delta_1^T \Phi^{*j} \frac{\partial \Phi^*}{\partial c_k} \sum_{i=j+1}^{n-1} \Phi^{*(i-1-j)} \\
&= -\sum_{j=0}^{n-2} \delta_1^T \Phi^{*j} \frac{\partial \Phi^*}{\partial c_k} \sum_{i=0}^{n-j-2} \Phi^{*i}
\end{aligned} \tag{4.28}$$

[5] The matrix cookbook is available at http://matrixcookbook.com/.

The only element of the vector $\delta_1^T \Phi^{*j} \frac{\partial \Phi^*}{\partial c_k}$ which is different from zero is the kth element which is equal to $\Phi_{1,K}^{*j}$. For $k = 1$, we find $\delta_1^T \Phi^{*j} \frac{\partial \Phi^*}{\partial c_k} = \Phi_{1,K}^{*j} \delta_1^T$ and the result follows since $B_0 = 0$. For $k > 1$, we can use $\Phi_{k,[:]}^{*i} = \Phi_{k-l,[:]}^{*i+l}$ $(l < k)$ which holds since Φ^* is a companion matrix. Thus,

$$\begin{aligned}
\frac{\partial B_n^T}{\partial c_k} &= -\sum_{j=0}^{n-2} \Phi_{1,K}^{*j} e_k^T \sum_{i=0}^{n-j-2} \Phi^{*i} \\
&= -\sum_{j=0}^{n-2} \Phi_{1,K}^{*j} e_1^T \sum_{i=0}^{n-j-2} \Phi^{*i+k-1} \\
&= -\sum_{j=0}^{n-2} \Phi_{1,K}^{*j} \delta_1^T \left(\sum_{i=0}^{n-j-2} \Phi^{*i} - \sum_{i=0}^{k-1} \Phi^{*i} \right)
\end{aligned}$$

where e_k is the k^{th} column of the identity matrix. This completes the proof since the derivative (4.27) now follows from (4.26).

An alternative way to estimate the derivative is to take one of the equations in (4.28). This is possible for any Φ^*. For a companion matrix, however, using equation (4.27) is computationally much more efficient.

4.A.2 Derivatives of A

In order to calculate the derivatives of A, we rewrite (4.5d) as

$$A_n = \left(\sum_{i=0}^{n-1} B_i^T \right) \alpha^* + \sum_{i=0}^{n-1} B_i^T \Sigma B_i.$$

Immediately, we see that the derivative with respect to α^* is the sum in the brackets:

$$\frac{\partial A_n}{\partial \alpha^*} = \sum_{i=0}^{n-1} B_i^T.$$

The derivatives with respect to c can be calculated through the derivative with respect to B and then applying the chain rule:

$$\frac{\partial A_n}{\partial B_k} = \alpha^* + 2\Sigma B_k.$$

Finally, the derivative with respect to Σ is:

$$\frac{\partial A_n}{\partial \Sigma} = \frac{1}{2} \sum_{i=0}^{n-1} B_i B_i^T.$$

To keep the covariance matrix positive-definite, it is often parameterized by the Cholesky decomposition of $\Sigma = CC^T$:

$$\frac{\partial A_n}{\partial C} = \sum_{i=0}^{n-1} B_i B_i^T C$$

These two formulas are given in the matrix cookbook.

We are done. We can now use the standard results for the gradient of the Kalman filter together with the chain rule to estimate derivatives of the likelihood function.

Write the log-likelihood function in terms of the new variables:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}\left(\Phi, \alpha, \Sigma, B(c), A(B(c), \alpha_K^*, \Sigma), H\right) \\ &= \mathcal{L}'\left(\Phi, \alpha, \Sigma, c, \alpha_K^*, H\right).\end{aligned}$$

Putting things together yields the gradient as a function of the new parameters:

$$\begin{aligned}\frac{\partial \mathcal{L}'}{\partial \Phi} &= \frac{\partial \mathcal{L}}{\partial \Phi} \\ \frac{\partial \mathcal{L}'}{\partial \alpha} &= \frac{\partial \mathcal{L}}{\partial \alpha} \\ \frac{\partial \mathcal{L}'}{\partial \Sigma} &= \frac{\partial \mathcal{L}}{\partial \Sigma} + \sum \frac{\partial A_n}{\partial \Sigma} \frac{\partial \mathcal{L}}{\partial A_n} \\ \frac{\partial \mathcal{L}'}{\partial c} &= \frac{\partial \mathrm{vec}(B)^T}{\partial c} \left(\frac{\partial \mathcal{L}}{\partial \mathrm{vec}(B)} + \frac{\partial A^T}{\partial \mathrm{vec}(B)} \frac{\partial \mathcal{L}}{\partial A}\right) \\ \frac{\partial \mathcal{L}'}{\partial \alpha_K^*} &= \frac{\partial A^T}{\partial \alpha_K^*} \frac{\partial \mathcal{L}}{\partial A} \\ \frac{\partial \mathcal{L}'}{\partial H} &= \frac{\partial \mathcal{L}}{\partial H}.\end{aligned}$$

Chapter 5

More about Affine Models

Self consistent yield curve models imply certain features that are in obvious conflict with the data or economic theory.

First, do we believe that the entire state of the economy is described by the three or four yield curve factors? We do not, because this would imply that all aggregate or macroeconomic quantities are, up to some measurement error, described by these factors. This, of course, is strongly rejected by the data. Thus, if we intend to describe an economy that does not consist exclusively of the bond market, we must introduce additional state variables that are not reflected by yields. Such an enlarged model also entails an adaption of the yield curve model since, by definition, the self-consistent set-up of the previous section implies that all state variables can be represented as affine functions of yields. This adaption is the subject of the second section of this chapter, where it is shown how variables that do not affect the current yield curve can be integrated into an affine framework. The section also describes its consequences for predictability and contains an example that highlights the ability of these unspanned macro factors to improve bond return forecasts based solely on yields.

Secondly, economic theory suggests that persistent shocks are priced. For instance, the CAPM implies that wealth shocks are priced. Similarly, utility theory suggests that aggregate consumption shocks are priced. In contrast to these models, affine yield curve models associate all risk compensation with shocks to mean-reverting state variables. In order to circumvent this

discrepancy, the third section shows how to integrate persistent shocks into the affine framework.

Before these two points are discussed, I derive a linear relationship between risk and return that is the log version of Ross's (1976) arbitrage pricing theory (APT). It can be used to enlarge the cross-section, and thus increase the precision of parameter estimates, particularly those related to returns forecasting. The section also specifies a test as to whether the model correctly explains cross-sectional variation in unconditional returns.

5.1 Expected Returns

Let rx_t be a log excess return joint normally distributed with the pricing kernel (4.2). Then the expected excess return $\mathrm{E}_{t-1}[rx_t]$ satisfies:

$$\mathrm{E}_{t-1}[rx_t] + \frac{1}{2}\mathrm{var}[rx_t] = \mathrm{cov}[rx_t, \varepsilon_t^{x,T}]\lambda_{t-1} \tag{5.1}$$

which says that the loadings λ_t are also the market prices of risk (see Cochrane and Piazzesi (2005b)).

The proof is not straightforward so it is outlined here: it starts with the general pricing equation $\mathrm{E}_{t-1}[M_t R_t] = 1$, which holds for gross returns $R_t = \exp(r_t)$ on any asset. Using $m_t = \log(M_t)$, we can write

$$\begin{aligned} 1 &= \mathrm{E}_{t-1}[M_t R_t] = \mathrm{E}_{t-1}[\exp(m_t + r_t)] \\ &= \mathrm{E}_{t-1}[m_t] + \mathrm{E}_{t-1}[r_t] + \frac{1}{2}\mathrm{var}[m_t] + \frac{1}{2}\mathrm{var}[r_t] + \mathrm{cov}[r_t, m_t] \end{aligned}$$

since m_t and r_t follow a joint log-normal distribution. The same expression for the one-period bond is

$$1 = \mathrm{E}_{t-1}[m_t] + y_{t-1}^{(1)} + \frac{1}{2}\mathrm{var}[m_t].$$

Subtracting the two equations we find:

$$\mathrm{E}_{t-1}[rx_t] + \frac{1}{2}\mathrm{var}[r_t] = -\mathrm{cov}[r_t, m_t]$$

which implies (5.1).

The connection with the APT can be understood by writing unexpected returns of an asset as a linear combination of the innovations to the state

variables (or the APT factors) plus an idiosyncratic part u_t

$$rx_t - \mathrm{E}_{t-1}[rx_t] = C_{rx}^T \varepsilon_t^x + u_t. \tag{5.2}$$

ε_t^x and u_t follow a joint normal distribution with $\mathrm{cov}[u_t, \varepsilon_t^x] = 0$. Equation (5.1) renders the following restrictions:

$$\begin{aligned}\mathrm{E}_{t-1}[rx_t] &= C_{rx}^T \Sigma^x \lambda_{t-1} - \frac{1}{2}\mathrm{var}[rx_t] \\ &= A_{rx} + B_{rx}^T x_{t-1}.\end{aligned}$$

Bond returns in paragraph 4.1.3 are, of course, a special case of this formula. Using the method of undetermined coefficients and expanding the variance of rx_t yields:

$$A_{rx} = C_{rx}^T \Sigma^x \lambda_0 - \frac{1}{2} C_{rx}^T \Sigma^x C_{rx} - \frac{1}{2}\sigma_u^2 \tag{5.3a}$$

$$B_{rx}^T = C_{rx}^T \Sigma^x \lambda_1. \tag{5.3b}$$

These two equation express that given the factor representation (5.2), returns are affine in the state variables and the shocks. Thus, they can be straightforwardly integrated into the affine set-up. Estimation using the Kalman filter remains possible, although we need to augment the state vector since $\mathrm{E}_{t-1}[rx_t]$ depends on x_{t-1} (5.3.2 below for the details see paragraph).

We can also define the risk premium ζ^i associated with an ε_i-shock as:

$$\zeta_t^i = e_i^T \Sigma^x \lambda_0 + e_i^T \Sigma^x \lambda_1 x_t$$

where e_i is a vector of zeros with a 1 at the i^{th} entry. This follows by setting $C_{rx} = e_i$ and omitting Jensen's inequality terms in (5.3a). The APT representation is completed by writing excess returns in terms of the factor loadings and the risk premia:

$$\mathrm{E}_{t-1}[rx_t] + \frac{1}{2}\mathrm{var}[rx_t] = C_{rx,1}\zeta_1 + \cdots + C_{rx,K}\zeta_K. \tag{5.4}$$

Finally, we can test the parameter restriction (5.3a) by comparing the likelihood of the constraint model with one leaving some A_{rx} as free parameters. If the restriction is rejected, one (or a subgroup) of assets has an unconditional return that is not explained by the model. This is the likelihood ratio version of the Gibbons, Ross, and Shanken (1989) test on the cross-sectional pricing ability of a model.

5.2 Unspanned Macro Factors

In this section, I show how an exponential affine pricing kernel together with a K-dimensional VAR, that describes the evolution of the state variables must be restricted such that the resulting yield curve model has only $K - R$ factors. The remaining R state variables, referred to as unspanned macro factors, affect the physical probability of the yield curve, but not the risk-neutral probability. This allows the model to have state variables that are relevant to explain the bond risk premium, yet are irrelevant to explain the cross-section of bond prices. Note that they can still influence prices and returns of other assets or actual values of macroeconomic quantities.

In general, the latent yield curve factors have no straightforward economic interpretation. They are rather the result of some particular rotation which forces the latent factors to have opposite effects on expected future yields and the bond risk premium. More precisely, the two effects offset each other such that they leave the current term structure - but not the expected future term structure - unaffected. Duffee (2008b) gives an example: consider economic news that increases risk premia and simultaneously leads investors to believe the Federal Reserve Bank will soon cut short-term interest rates. The increase in risk premia induces an immediate increase in long-term bond yields, while the expected drop in short rates induces an immediate decrease in these yields.

Although these non-self-consistent models are mathematically a restriction of the original model, economically, they are a generalization, since factors that are not related to the yield curve can become state variables affecting future yields.

This section contains two paragraphs. The first describes the inclusion of unspanned factors. The second contains an example that provides strong evidence for the existence of unspanned factors.

5.2.1 The Model

Duffee (2008b) and Joslin, Priebsch, and Singleton (2009) are the only two sources I am aware of that systematically use unspanned factors. Duffee assumes that the risk-neutral VAR (4.7) of the Ang-Piazzesi model can be rotated such that Φ^* is diagonal. With a diagonal Φ^*, imposing a restriction such that the k^{th} variable does not affect the current yield curve. Simply set the k^{th} value of δ_1 equal to zero. In this case, forward rates $f_t^{(n)} =$

$A_n^f + \delta_1^T \Phi^{*n} x_t$, and hence yields and prices, do not depend on the k^{th} state variable.

My approach is slightly more general as it is not based on the restrictive assumption that Φ^* is diagonal, which excludes complex eigenvalues. I only assume that Φ^* has a staircase form and $\delta_1 = e_1$. This parametrization simply imposes the intended restriction that the actual values of the variables in the second block do not affect the current yield curve but leaves all other features of the model unaffected. Because Φ, and thus the physical VAR (4.1), does not have a restricted structure, expected future yields depend in general on all state variables.

More formally, we can always rotate the factors so that the risk-free rate is the first factor and that only the first $K - R$ factors of the first block affect the yield curve. In order to see this, write the $K \times K$ matrix Φ^* as follows:

$$\Phi^* = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{5.5}$$

where A_{22} is a $R \times R$ matrix (which also determines the dimension of the other variables).

We want to construct a model with $\delta_1 = e_1$ and the property that the forward rates $f_t^{(n)} = A_n^f + \delta_1^T \Phi^{*n} x_t$ only depend on the first $K - R$ variables. This is equivalent to requiring that for all n, the first row of Φ^{*n} is of the following form:

$$\Phi^{*n}_{1,[:]} = \begin{bmatrix} \underbrace{* \cdots *}_{K-R \text{ times}} & 0_{1\times R} \end{bmatrix}. \tag{5.6}$$

$*$ denotes that the corresponding values or matrices can take non-zero values. Straightforward matrix calculation shows that this condition holds if $A_{12} = 0$, since in this case

$$\Phi^{*n} = \begin{bmatrix} A_{11}^n & 0_{K-R\times R} \\ * & A_{22}^n \end{bmatrix} \tag{5.7}$$

for all powers ($n = 1, 2 \ldots$). This implies that the risk-neutral process of the first $K - R$ factors is Markovian:

$$x_{t+1,[1\ldots K-R]} = \alpha^* + A_{11} x_{t,[1\ldots K-R]} + \varepsilon_{t+1,[1\ldots K-R]}.$$

In fact, it is equivalent to the case of self-consistent models. Consequently, we can rotate the yield curve factors in the same way as in the previous chapter.

That is to say that the first $K - R$ factors are the risk-free rate and its $1 \ldots (K - R - 1)$-period-ahead expectations under the risk-neutral measure.

For the remaining factors, none are obviously analogous to the short rate.[1] Therefore, in order to complete the VAR, I usually select $K - R$ meaningful variables besides the yield curve factors. Consequently, equation (5.5) takes the following staircase form:

$$\Phi^* = \begin{bmatrix} C_1 & 0 \\ A_{21} & A_{22} \end{bmatrix} \tag{5.8}$$

where C_1 is a companion matrix. Models constructed this way are generally identified.

For $R = 1$, however, there is a general parametrization. In this case, block-diagonalizing Φ^* identifies the last state variable up to a constant and to a volatility level. If C_1 is not degenerate, we can rotate the model such that

$$\Phi^* = \begin{bmatrix} C_1 & 0 \\ 0 & c \end{bmatrix} \tag{5.9}$$

where c is a constant. Identification is achieved up to a constant (e.g. α_K) and to a volatility level (Σ_{KK}). In a specific model, these two parameters are identified through restrictions related to the measurement equation. I give an example in the next paragraph.

Finally, we can generalize the classification of essential affine models. A model with K state variables, $K - R$ yield curve factors and $\text{rank}(\lambda_1) = L$ is denoted as an $AP(K, K - R, L)$ model.

5.2.2 Integrating Inflation Forecasts into an Affine Model

Expected inflation is not spanned by the yield curve factors. In particular, past inflation and expected inflation from the Livingston Survey improves inflation forecasts solely based on yields (see table 5.1). This result is robust against changes to the time period and it is valid for different consumer price indices: specifically for core inflation, a less volatile measure of inflation that

[1] A parametrization that in theory always works is to use the expected return of a 2-period bond as the $k + 1^{\text{th}}$ state variable. Then the model is completed by taking its $1 \ldots R-1$ period ahead expectations under the risk neutral measure as the remaining state variables. While this parametrization has the beauty that it is unequivocally identifies the model, it has the short-coming that it is algebraically difficult.

excludes food and energy (not reported)[2]. If expected inflation is a state variable, this simple observation is evidence against the self-consistent set-up because self-consistency implies that all state variables are spanned by the yield curve.

Clearly, expected inflation is a plausible state variable. On inspection of the data, it is obvious that yields and expected inflation share a lot of common movement. Therefore, at least in the medium-term, either expected inflation must adjust to yields or vice versa. Clearly, a joint VAR can capture this common movement.

We can also look at the real rate, which is essentially the difference between the nominal yield and expected inflation. According to utility theory, the real rate is an important state variable as it measures the actual price of intertemporal substitution. If expected inflation is not spanned by yields neither is the real rate.[3] In this case, self-consistent yield curve models do not fully incorporate the price of intertemporal substitution. This is definitely not a desired property of a yield curve model and thus economically motivates us to abandon the self-consistent set-up.

This paragraph can be summarized as follows: firstly, I run the inflation forecasting regressions; later, non-self-consistent yield curve models that include expected inflation data from the Livingston survey are estimated. A likelihood ratio test provides strong evidence against the self-consistent set-up.

Regression Analysis

In this section, I discuss regression based inflation forecasts. Specifically, the four factors from the $AP(4,3)$ model together with realized inflation and the inflation forecast from the Livingston Survey are used to forecast six-month-ahead inflation. I choose a six-month forecast horizon because the survey records the corresponding data. Results are reported for the more robust

[2]A similar observation is that yields of inflation protected securities (tips) are not spanned by nominal bonds yields. As tips are far less liquid than nominal bonds this difference, however, could also stem from a time-varying liquidity premium.

[3]Of course non yield curve based variation of expected inflation could exactly offset non yield curve based movements of the expected real rate. Mathematically, this means that the partial correlation, given yields, of expected inflation and the expected real rate is minus one. There is barely any economic reasoning supporting such an entirely antagonistic movement of the two quantities.

time period 1952-2006. When the first five years are included, R^2 drops considerably in all regressions, since inflation is extremely volatile during these five years. The main conclusions, however, remain unchanged. They are very similar for different time periods, forecasting horizons, and various definitions of inflation.

Table 5.1: Inflation Forecasting Regression

	cte	f_1	f_2	f_3	f_4	past	π^{liv}	R^2
1	0.00	-44.80	230.35	-334.77	150.31			
	0.56	-3.97	4.08	-4.02	3.94			0.51
2	0.00	-25.64	133.85	-194.81	87.14	0.44		
	1.07	-2.78	2.93	-2.90	2.84	5.42		0.59
3	0.01	3.61	-22.46	45.23	-27.05		1.26	
	4.76	0.38	-0.47	0.64	-0.84		9.66	0.70
4	0.01	3.68	-22.87	45.97	-27.45	-0.01	1.28	
	4.41	0.38	-0.46	0.63	-0.82	-0.09	5.87	0.70
5	0.03						1.03	
	2.25						9.16	0.58

Note: Forecast of 6-month inflation ($cpi_{t+2} - cpi_t$) using a constant (cte), the four yield curve factors at time t ($f_1 \ldots f_4$), past 6-month inflation (past), and the inflation forecast from the Livingston Survey (π^{liv}) for the corresponding period known at time t. The first line of each regression contains the loadings and the second line the, corresponding t-values using 3 Newey-West lags. The last column shows the unadjusted R^2. Quarterly data (1952-2006).

First, I run a regression of the six-month realized inflation on the four yield factors (see Table 5.1, regression 1). The corresponding R^2 is 0.51. Including past inflation (regression 2) is clearly significant and raises R^2 to 0.59[4]. The forecast from the survey data (π^{liv}) in the third regression is even more significant (t-stat of 9.66) and boosts R^2 to 0.70. It is regression 3,

[4]I run these regressions for a whole bunch of other measures of inflation (not reported). They all have the same property: their past values contain significant information about future inflation that is not already contained in the yield curve. I also tried the GDP-deflator and the producer prices index, the general result does not alter, though it is less severe for producer prices.

and to a lesser extend regression 2, that provides the strong evidence that expected inflation is not spanned by the yield curve factors and we thus need at least one additional state variable. Continuing with the description of the results, we see that including past inflation in addition to the survey data is no longer significant (regression 4). Finally, the last regression shows that the R^2 drops to 0.58 when we exclude the yield curve factors. An F-test[5] shows that the yield factors are jointly significant (F-value is 10.8 with a 99 percent critical value of 3.4), although the high correlation between the factors entails insignificant individual t-statistics. The yield curve thus contains additional information about expected inflation not contained in the survey data.

Expected Inflation as an Unspanned Macro Factor

I work out an $AP(4,3,4)$ as well as an $AP(5,4,5)$ model, both including expected inflation as an extra variable besides the original yield curve factors. These models are estimated by maximum likelihood using the Kalman filter. Empirically, the additional variable significantly improves the estimation of the bond risk premia.

The description begins with the definition of the state equation. In order to do this, we need only normalize the last state variable, since we can adopt the specification given in (5.9). The variable is defined such that the expected inflation π_t is linear in the state variables

$$\pi_t = B_\pi^T x_t \tag{5.10}$$

with $B_{\pi,K} = 1$, while also satisfying (5.9). x_K is therefore the sum of the residual and the intercept of a regression of π onto the yield curve factors. It is unequivocally defined and has a clear economic interpretation; therefore it globally identifies the model. This choice is algebraically easier than taking π as the last factor, because imposing the identifying restrictions below is more cumbersome in the latter case. For the four factor model, the implied parameters of the risk-neutral dynamics are:

$$\alpha^* = \begin{pmatrix} 0 \\ 0 \\ \alpha_3^* \\ \alpha_4^* \end{pmatrix} \qquad \Phi^* = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ c_1 & c_2 & c_3 & 0 \\ 0 & 0 & 0 & c_4 \end{pmatrix}. \tag{5.11}$$

[5]In order to circumvent the overlapping data problem, this test is based on non-overlapping semi-annual data.

The five factor case is completely analogous using a four dimensional companion matrix in the upper left block of Φ^*. The rest of the parametrization follows the general specification of non-self-consistent models in paragraph 5.2.1.

Estimation of the Models with an Unspanned Factor

I estimate the model with bonds and the inflation forecast from the survey. By construction, bond returns are fully explained by shocks to the first $K-1$ state variables. Because no other asset return is used for calibrating, the price of risk associated with the last state variable remains undetermined. Its value is thus restricted to zero using equation (5.3). Specifically, $\lambda_{0,K} = 0$ and the last row of $\Sigma^X \lambda_1$ is set to zero. For purposes of estimation, the first restriction can also be substituted by setting $\alpha_K^* = 0$, since this quantity does not enter the likelihood function.

As in Brennan, Wang, and Xia (2004), expected inflation π is identified by a measurement equation based on the Livingston Survey data on the expected rate of inflation over the next six months (π^{liv}):

$$\begin{aligned} \pi_t^{liv} &= A_{liv} + B_{liv}^T x_t + \eta_t^{liv} \\ &= B_\pi^T \alpha + B_\pi^T (1+\Phi) x_t + \eta_t^{liv}. \end{aligned}$$

The second line is derived in the appendix to this paragraph. η^{liv} is a pure measurement error that is not correlated with any other shock. Its standard deviation is h^{liv}. The measurement equation is completed by presuming that the measurement of bonds remains in its original form.

Note that this specification does not account for the fact that yields add information to the inflation forecast. Since the model includes all state variables and thus all the interesting specification for return prediction, this, however, is of minor interest here. If we change the focus towards inflation forecasting, we should enlarge the model and include realized inflation in the measurement equation.

We can test whether the new variable helps to predict bond returns by restricting the last column of $\Sigma^X \lambda_1$ to zero and performing a likelihood ratio test. This is most obvious from equation (5.3b), which shows that in the constraint model $x_K = \pi^{liv}$ is without influence on the bond risk premia. Table 5.2 shows that restricting $\Sigma^X \lambda_1$ is indeed significant. For the four factor model, the decrease in the maximum of the log-likelihood function between

Table 5.2: Affine Models Including Expected Survey Inflation

Model	Likelihood	R^2 Level	R^2 Slope	corr	para
$AP(4,3,3)$, ur.	16557.6	0.315	0.097	-0.12	36
$AP(4,3,3)$, rest.	16548.9	0.243	0.079	0.15	33
$AP(3,3)$		0.228	0.090	0.11	
$AP(5,4,4)$, ur.	16980.2	0.316	0.190	-0.33	52
$AP(5,4,4)$, rest.	16973.2	0.245	0.175	-0.39	48
$AP(4,4)$		0.219	0.181	0.05	

Note: Model estimated based on quarterly data (1947-2006) of the following yields: $y^{(1/4)}$, $y^{(1/2)}$, $y^{(3/4)}$, $y^{(1)}$, $\dots y^{(10)}$ and the expected inflation from the Livingston Survey. The second column shows the maximum value of the log-likelihood function for the model indicated on the left hand side. The next two columns give the R^2, as defined by equation (4.24), of the implied return forecast on $rx^{(10)}$ and a $\widetilde{sr}$ (using the filtered series). The fifth column shows the correlation between the two expected returns. The last column indicates the number of free parameters.

the unrestricted model in the first line and the restricted model in the second line function is $8.7 = 16557.6 - 16548.9$ compared with a 1 percent critical value of 5.7. For the five factor models, the decrease is 7.0 and thus also above the 1 percent critical value of 6.6.

There is also a dramatic increase in R^2 for the return forecast for 10-year bonds. It is now 0.32 for both models that allow expected inflation to predict bond returns, compared with values below 0.25 for the restricted models. R^2 for the slope portfolio are also higher, though the effect is much weaker. Finally, the table also repeats the estimates of the self-consistent models. As we expect, they are very similar to the restricted model.

Efficiency

As in the case of self-consistent models, the Kalman filter approach does not improve predictability of yield curve models with unspanned factors if there is no measurement error. The reasoning is still the same as in paragraph 4.3.1 and we can also apply the same regression-based parameter estimation procedure. Still analogously to the self-consistent case, estimation is improved if measurement error is present. Consequently, the Kalman filter approach improves parameter estimates in an environment with noisy data as it properly

disentangles economically meaningful variables from measurement noise.

An undesired feature of models with unspanned macro factors is that selection bias returns. Whereas, in the self-consistent case, the researcher has no subjective judgement, there is no guideline as to how macro variables should be selected in the non-self-consistent case. Virtually any macro variable can serve as a plausible state variable. Consequently, there should be a good motivation for incorporating further variables into the state space system in order to reduce the danger of overfitting the predictability pattern.

Appendix: Multi-Period Expectations in the Measurement Equation

Inflation forecasts from the Livingston Survey predict two-quarters-ahead inflation. If we want to use them in the measurement equation, we must extract expected two-period inflation from the one-period model. Let $\pi_t = A_\pi + B_\pi^T x_t$, then expected two-period inflation at time t is

$$\begin{aligned} \pi_t + \mathrm{E}_t[\pi_{t+1}] &= A_\pi + B_\pi^T x_t + A_\pi + B_\pi^T(\alpha + \Phi x_t) \\ &= 2A_\pi + B_\pi^T \alpha + B_\pi^T(1+\Phi)x_t = A_{liv} + B_{liv} x_t \end{aligned}$$

with the obvious definition of A_{liv} and B_{liv}. We can often parameterize the model directly with A_{liv} and B_{liv} and then solve for A_π and B_π.

5.3 Shocks to Persistent Variables

As already mentioned in the introduction to this chapter, utility theory, in particular Breeden's (1979) consumption-based asset pricing model suggests that aggregate consumption shocks are priced. For example, with constant relative risk aversion the pricing kernel reads

$$M_{t+1} = \gamma \left(\frac{C_{t+1}}{C_t} \right) = \exp(\gamma \ln \left(\frac{C_{t+1}}{C_t} \right))$$

which is a very simple version of the exponential affine pricing kernel. In many refinements of this basic model consumption growth plays a similar role (e.g. Söderlind (2006)).

Similarly, the CAPM suggests that (only) wealth shocks are priced. Both consumption and wealth are usually modeled as non-stationary variables affected by unexpected and persistent shocks. In this section, I therefore outline

how these persistent shocks, or simply any shock, can be incorporated into an exponential affine pricing kernel.

5.3.1 Incorporating Persistent Shocks into an Exponential Affine Pricing Kernel

As will be shown below, it is possible to write the new pricing kernel as a special case of the original one in a straightforward manner. Consequently, the main purpose of this section is to clarify the notation and to integrate the model into the Kalman filter framework.

Let v_t be an L-dimensional random vector that is not spanned by innovations to the state variables. Furthermore, recall the K-dimensional VAR (4.1) that defines the state variables x_t

$$x_{t+1} = \alpha + \Phi x_t + \varepsilon^x_{t+1}. \tag{5.12}$$

where the possibly correlated vectors v_{t+1} and ε^x_{t+1} follow an iid joint normal distribution. We can write the two processes in a single, degenerated VAR. Define

$$X_t = \begin{pmatrix} x_t \\ v_t \end{pmatrix} \qquad \varepsilon^X_t = \begin{pmatrix} \varepsilon^x_t \\ v_t \end{pmatrix} \tag{5.13}$$

as well as

$$\alpha^X = \begin{pmatrix} \alpha \\ 0_{R\times 1} \end{pmatrix} \qquad \Phi^X = \begin{pmatrix} \Phi & 0_{K\times R} \\ 0_{R\times K} & 0_{R\times R} \end{pmatrix}. \tag{5.14}$$

which have the dimension $(K+L)\times 1$ and $(K+L)\times(K+L)$ respectively. Moreover, the covariance matrix of ε^X is written as Σ^X. The completed VAR including all priced shocks reads

$$X_{t+1} = \alpha^X + \Phi^X X_t + \varepsilon^X_{t+1} \tag{5.15}$$

and the pricing kernel remains in its original form:

$$\begin{aligned} M_{t+1} &= \exp(-\delta_0 - \delta_1^T x_t - \frac{1}{2}\lambda_t^T \Sigma^X \lambda_t - \lambda_t^T \varepsilon^X_{t+1}) \\ \lambda_t &= \lambda_0 + \lambda_1^X X_t = \lambda_0 + \lambda_1 x_t. \end{aligned} \tag{5.16}$$

λ_t and λ_0 are $(K+L)$-vectors and λ_1 is a $(K+L)\times K$-matrix. The $(K+L)\times(K+L)$ matrix $\lambda_1^X = \begin{bmatrix} \lambda_1 & 0 \end{bmatrix}$ is only introduced to highlight that (5.16)

is, in fact, a special case of the original exponentially affine pricing kernel (4.2). The equivalence between the two notations in the second line follows from a trivial block matrix multiplication. I will use the simpler expression $\lambda_t = \lambda_0 + \lambda_1 x_t$ throughout the rest of this thesis.

In order to get a vector of the right dimension, it is sometimes easier to write prices, or other affine functions of the state variables, as a function of X_t. In this case, I write $A_{(\cdot)} + B_{(\cdot)}^T x_t = A_{(\cdot)} + B_{(\cdot)}^{X,T} X_t$, where the first K entries of $B_{(\cdot)}^X$ are as in $B_{(\cdot)}$ and the last L entries are zero.

Similar notation is used for $\Phi^{*,X}$ and $\alpha^{*,X}$ which can be defined as in (4.6). If we only need the state variables we can write

$$\Phi^* = \Phi - \Sigma^{xX}\lambda_1 \tag{5.17a}$$

$$\alpha^* = \alpha - \Sigma^{xX}\lambda_0 \tag{5.17b}$$

where $\Sigma_{ij}^{xX} = \text{cov}[x_i, X_j]$.

Finally we can denote a model with $K+L$ prices shocks, K state variables, $K-R$ yield curve factors and M independent predictability factors as an $AP(K+L, K, K-R, M)$ model.

5.3.2 Example: A Simple Stock Bond Model

I construct a model that combines stock returns (r_t) and bonds in a very simple way. The bond model remains in its original self-consistent form. In order to price stocks, we add homoscedastic unexpected stock returns (ε_t^r) to the pricing kernel as described in the previous paragraph. The yield curve factors follow (5.12). Defining $v_t = \varepsilon_t^r$ defines the degenerated VAR (5.15) and the pricing kernel (5.16) using definitions in (5.13) and (5.14).

The equation for the factor risk premia (5.3) implies that expected stock returns (μ_t) are affine in the state variables:

$$\text{E}_{t-1}[r_t] = \mu_{t-1} = A_\mu + B_\mu^T x_{t-1} \tag{5.18}$$

with

$$A_\mu = e_{K+1}^T \Sigma^X \lambda_0 - \frac{1}{2}\Sigma_{K+1,K+1}^X \tag{5.19a}$$

$$B_\mu^T = e_{K+1}^T \Sigma^X \lambda_1. \tag{5.19b}$$

Taking the expected and the unexpected part together, the return process can be written as

$$r_t = A_\mu + B_\mu^T x_{t-1} + \varepsilon_t^r. \tag{5.20}$$

Inconveniently, returns at time t are a function of x_{t-1}, whereas contemporaneously observed yields are a function of x_t. Using the two types of equation in a joint system of measurement equation therefore needs an adaption of the original state equation (5.15). One possible solution is to use the following augmented state vector:

$$W_t = \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}$$

With this definition, the augmented state equation becomes

$$W_{t+1} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} \Phi & 0 \\ 1 & 0 \end{pmatrix} W_t + \begin{pmatrix} \varepsilon_{t+1}^x \\ 0 \end{pmatrix}.$$

Stock return innovations can be correlated with the innovations in the state variables. We write

$$\varepsilon_t^r = C_r^T \varepsilon_t^x + \eta_t^r$$

where the residuum η_t^r is an iid normally distributed error uncorrelated with the state variables. Because η_t^r is independent of ε_t^x, it can be used as an error term in the measurement equation. More precisely, using

$$\varepsilon_t^x = x_t - \Phi x_{t-1} - \alpha$$

we can write

$$\varepsilon_t^r = C_r^T x_t - C_r^T \Phi x_{t-1} - C_r^T \alpha + \eta_t^r.$$

This implies the following measurement equation:

$$\begin{bmatrix} y_t^{obs} \\ r_t \end{bmatrix} = \begin{bmatrix} A_y \\ A_\mu - C_r^T \alpha \end{bmatrix} + \begin{bmatrix} B_y & 0 \\ C_r^T & B_\mu^T - C_r^T \Phi \end{bmatrix} W_t + \begin{bmatrix} \eta_t^y \\ \eta_t^r \end{bmatrix}.$$

The error term in the second line (η_t^r) of the measurement equation is now an economically meaningful shock. So the interpretation of the etas as pure measurement error is lost. Mathematically, this is, of course, not a problem as long as they remain independent of the shocks to the state variables, as is the case here.

The model is considerably simplified by parameterizing $\widetilde{A}_\mu = A_\mu - C_r^T \alpha$ and $\widetilde{B}_\mu^T = B_\mu^T - C_r^T \Phi$ instead of the last row of λ_0 and λ_1. It is completed by assuming η_t^y and η_t^r to be uncorrelated and by forcing the covariance matrix of η_t^y to be a scaled identity as in the case of pure yield curve models.

Efficiency

Due to the simplicity of the model, its empirical content is necessarily limited. Disregarding measurement error, equation (5.18) says that we are running a stock return forecasting regression with yields as explanatory variables, besides the estimation of a yield curve model which remains in its original form. The two parts only interact in the same way as in a system of regression equations. In particular, (5.19) does not imply a parameter restriction that could improve parameter estimates. Thus, as with previous models, the advantage of the Kalman filter estimates is that it cleans the noisy bond data. If we can measure yields without error, the OLS forecast and the Kalman filter estimates are both maximum likelihood estimators and thus equivalent (see paragraph 4.3.1 and the references there).

Empirical Results

Table 5.3 shows the results of the implied forecast of the simple stock-bond model. Using an unrestricted, self-consistent four factor model for bonds, the implied R^2 for the quarterly stock return (r) forecast is 0.013 and excess returns (rx) are predicted with an R^2 of 0.037. Annual excess returns for ten-year bonds ($rx^{(10)}$) and the slope portfolio ($\widetilde{sr}$) are predicted with an R^2 of 0.22 and 0.18 respectively and thus nearly identical to the results of the $AP(4,4)$ model in table 4.1. Table 5.3 also contains the results for the models based on a three factor yield curve model.

Table 5.3: Forecasts in the Simple Stock Bond Model

Model	$rx^{(10)}$	$\widetilde{sr}$	r	rx
AP(4,3,3,3)	0.228	0.090	0.010	0.034
AP(5,4,4,4)	0.218	0.180	0.013	0.037

Note: R^2 for the (excess) return forecasts based on equation (4.24). The bond returns $rx^{(10)}$ and $\widetilde{sr}$ are overlapping annual returns, the stock returns r and rx are quarterly returns. Model estimated with quarterly data (1947-2006) of the following yields: $y^{(1/4)}$, $y^{(1/2)}$, $y^{(3/4)}$, $y^{(1)}$, $\ldots y^{(10)}$ and aggregate stock returns.

5.4 Summary of the Empirical Findings

In this chapter, I have extended the affine set-up to unspanned macro factors and persistent shocks. The empirical results are:

- Expected inflation is not spanned by yields.
- Expected inflation significantly improves bond return forecasts. Especially, excess returns of 10-year bonds are much better predicted when expected inflation is added as an unspanned macro factor.
- The four yield curve factors predict aggregate stock excess returns with an R^2 of 3.7 percent.

Chapter 6

Dividends and Returns in a State Space Model

In this chapter, I show how a state space model with latent variables can be combined with the approximate present value relation of Campbell and Shiller (1988). In this framework, expected dividends, expected returns, and the dividend yield are affine functions of the state variables. Moreover, unexpected returns are an affine function of innovations in the state variables plus unexpected dividend growth. Due to its linearity, the entire model can be written in the state-space form and thus be estimated by maximum likelihood using the Kalman filter.

This is important because there is a material difference between the approach presented here and the VAR model in Campbell (1991, 1993) and Campbell and Vuolteenaho (2004). Here, latent rather than observable variables are used; consequently, the model cannot be estimated by OLS or another regression technique.

Similar state space models have recently been analyzed by van Binsenbergen and Koijen (2008), Rytchkov (2007), and Cochrane (2008b). This chapter contributes to this literature by analyzing a quarterly, instead of an annual framework. Moreover, following the advice of Cochrane (2008b), I use returns instead of the observed dividend growth rate in the measurement equation.

6.1 The Campbell-Shiller Approximation with Latent Variables

I derive the Campbell and Shiller (1988) (hereafter CS) approximation in a latent state space framework where expected returns and the expected dividend growth rates are assumed to be affine in the state variables. As will be shown, the resulting price-dividend ratio is also affine in the state variables.

Suppose that the expected return on a equity portfolio μ_t, and its expected dividend growth rate g_t can be written as

$$\mu_t = A_r + B_r^T x_t \tag{6.1}$$
$$g_t = A_d + B_d^T x_t \tag{6.2}$$

where x_t follows a VAR with normally distributed error $\varepsilon_{t+1} \sim N(0, \Sigma)$:

$$x_{t+1} = \alpha + \Phi x_t + \varepsilon_{t+1} \tag{6.3a}$$
$$= \overline{x} + \Phi(x_t - \overline{x}) + \varepsilon_{t+1}. \tag{6.3b}$$

In general, the vector x_t summarizes the state of the economy as implied by a particular model and $\overline{x}$ is its unconditional mean. Write returns as

$$r_{t+1} = \log\left(\frac{P_{t+1} + D_{t+1}}{P_t}\right) \tag{6.4}$$

where P_t denotes the price of the portfolio and D_t its period t dividend. The dividend process

$$\Delta d_{t+1} = \log\left(\frac{D_{t+1}}{D_t}\right) = g_t + \varepsilon_{t+1}^d \tag{6.5}$$

completes the definition of the model. ε_{t+1}^d is a normally distributed shock that can be correlated with the VAR shocks ε_{t+1}.

Let $pd_t = \log(P_t/D_t)$ be the log price-dividend ratio. Following CS, we can write log linearized returns as:

$$\begin{aligned} r_{t+1} &= \log(1 + \exp(pd_{t+1})) + \Delta d_{t+1} - pd_t \\ &\approx \log(1 + \exp(\overline{pd})) + \frac{\exp(\overline{pd})}{1 + \exp(\overline{pd})}(pd_{t+1} - \overline{pd}) + \Delta d_{t+1} - pd_t \\ &= \kappa + \rho\, pd_{t+1} + \Delta d_{t+1} - pd_t \end{aligned} \tag{6.6}$$

by taking a first-order Taylor expansion around the unconditional mean of the log price-dividend ratio $\overline{pd}$. Solving for the log price-dividend ratio we have

$$pd_t = \kappa + \rho\, pd_{t+1} + \Delta d_{t+1} - r_{t+1}. \tag{6.7}$$

The linearization parameters are defined as

$$\kappa = \log(1 + \exp(\overline{pd})) - \rho\, \overline{pd} \tag{6.8}$$

$$\rho = \frac{\exp(\overline{pd})}{1 + \exp(\overline{pd})}. \tag{6.9}$$

Iterating forward, renders the approximate price-dividend ratio:

$$\begin{aligned} pd_t &= \kappa + \rho\, pd_{t+1} + \Delta d_{t+1} - r_{t+1} \\ &= \frac{\kappa}{1-\rho} + \sum_{j=1}^{\infty} \rho^{j-1}(\Delta d_{t+j} - r_{t+j}), \end{aligned} \tag{6.10}$$

assuming that the usual no-bubbles condition $\lim_{j\to\infty} \rho^j pd_{t+j} = 0$ holds. Plugging in the VAR (6.3) and taking expectation yields the price-dividend ratio as a function of the state variables:

$$\begin{aligned} pd_t &= \frac{\kappa}{1-\rho} + \sum_{j=1}^{\infty} \rho^{j-1} \mathrm{E}_t[\Delta d_{t+j} - r_{t+j}] \\ &= \frac{\kappa}{1-\rho} + \sum_{j=1}^{\infty} \rho^{j-1} \mathrm{E}_t[g_{t+j-1} - \mu_{t+j-1}] \\ &= \frac{\kappa}{1-\rho} + \sum_{j=0}^{\infty} \rho^{j} \mathrm{E}_t[g_{t+j} - \mu_{t+j}] \\ &= \frac{A_d - A_r + \kappa}{1-\rho} + (B_d - B_r)^T \sum_{j=0}^{\infty} \rho^j \overline{x} + \rho^j \Phi^j (x_t - \overline{x}) \\ &= \frac{\kappa + \overline{g} - \overline{\mu}}{1-\rho} + (B_d - B_r)^T (1 - \rho\Phi)^{-1} (x_t - \overline{x}) \\ &= A_{pd} + B_{pd}^T x_t \end{aligned} \tag{6.11}$$

where

$$A_{pd} = \overline{pd} - B_{pd}^T \overline{x} \tag{6.12a}$$

$$B_{pd}^T = (B_d - B_r)^T (1 - \rho\Phi)^{-1}. \tag{6.12b}$$

The first three lines follow from the definition of the state variables and rewriting indices, the fourth line follows by plugging in the multiperiod forecast of the state variables generated by the VAR:

$$\mathrm{E}_t[x_{t+j}] = \overline{x} + \Phi^j(x_t - \overline{x}).$$

The fifth line uses well-known properties of geometric series and plugs in unconditional expectation on g and μ. Finally, the last lines separates terms multiplying x_t. It says that the price-dividend ratio is affine in the state variables and implicitly defines the corresponding coefficients.

For the derivations below, the following variants of the linearization parameters are useful:

$$\rho = \exp(\overline{g} - \overline{\mu}) \tag{6.13}$$

$$\overline{pd} = \log \frac{\rho}{1-\rho} \tag{6.14}$$

$$\kappa = (\rho - 1)\log(1-\rho) - \rho \log \rho \tag{6.15}$$

Finally, we can rewrite returns. Plugging in the price-dividend ration (6.11) into the approximate return equation (6.6) gives

$$\begin{aligned} r_{t+1} &= \kappa + \rho A_{pd} + \rho B_{pd} x_{t+1} + \Delta d_{t+1} - pd_t \\ &= \mu_t + \rho B_{pd}^T \varepsilon_{t+1} + \varepsilon_{t+1}^d \\ &= \mu_t + \rho B_{pd}^T \varepsilon_{t+1} + C_d^T \varepsilon_{t+1} + \eta_{t+1}^d \\ &= \mu_t + C_r^T \varepsilon_{t+1} + \eta_{t+1}^d \end{aligned} \tag{6.16}$$

where unexpected dividend growth is written as $\varepsilon_t^d = C_d^T \varepsilon_t + \eta_t^d$ with η_t^d iid normal and independent of ε_t and $C_r = \rho B_{pd} + C_d$.

6.2 State Space Framework with Noisy Dividends

The aggregate dividend stream is a very noisy measure of aggregate cash-flow payments, especially when we analyze monthly or quarterly data. There are at least three fundamental sources of this measurement noise: firstly, and most importantly, dividend payout policy is strongly affected by tax issues and trends. In particular, since the 1980's, cash-flow distribution in the form

of share repurchases has become more and more important. While share repurchases were virtually non-existent in the seventies, they account now for more than 50 percent of the total cash-flow firms paid to equity holders. This disturbs the dividend stream because traditional dividends are no longer the right cash-flow measure. In order to account for this trend in the pay-out policy, I include a measure of equity repurchases in the dividend stream (see the data appendix for details). Unfortunately, this measure lacks precision, since the price and the date of the repurchase are not easily available. This imprecision is captured by introducing an explicit measurement error. This is relatively simple in the state space framework (details are outlined in the reminder of this section). The measurement error is also assumed to account for changes in the tax regime, which can influence the dividend policy because taxes are lower in either the new or the old regime. In the former case, dividends are postponed, in the latter case, they are paid earlier.

Secondly, dividend payments display a seasonal pattern. Even worse, the pattern varies greatly over time (see also the data appendix). The standard solution to account for the seasonality is to follow Hodrick (1992) and use one-year trailing dividends. This implies a strong structural autocorrelation in the dividend series when we use quarterly or monthly data. Divided growth is thus not directly observable at higher than annual frequency. In a state space framework, however, it is possible to jointly model the dynamic of the state variables and the seasonality pattern of dividends in a very general manner (see section 6.2.2 below).

Thirdly, I estimate dividends from return differences in the total return index and the price index. This method of calculating the dividend stream implicitly assumes that dividends paid early in the period are reinvested at the market return. As returns are more volatile than dividend growth rates, this substantially increases the volatility of the dividend stream (Cochrane (2008a)). Furthermore, the resulting measurement error is correlated with returns (see van Binsenbergen and Koijen (2008)). Measurement error partly capture this problem. In particular, it induces the volatility increase. The form of measurement I use, however, is not able to model the implied correlation with the return and is therefore ignored throughout this thesis.

6.2.1 Observable Variables

As outlined above, it is assumed that observable log dividends are affected by measurement error:

$$d_t^{obs} = d_t + \varepsilon_t^M.$$

where ε_t^M is an error term more precisely specified in (6.18) below. On the other hand, returns are assumed to be observed without measurement error. In this case, we have three observable quantities:

$$r_{t+1}^{obs} = r_{t+1} = \mu_t + \rho B_{pd}^T \varepsilon_{t+1} + \varepsilon_{t+1}^d \tag{6.17a}$$

$$pd_t^{obs} = A_{pd} + B_{pd} x_t - \varepsilon_t^M \tag{6.17b}$$

$$\Delta d_t^{obs} = g_{t-1} + \varepsilon_{t+1}^D + \varepsilon_t^M - \varepsilon_{t-1}^M \tag{6.17c}$$

$(t = 1 \dots \tau)$. The exact versions of these three equations are not independent, rather they are mechanically linked by the return approximation (6.7). Consequently, either the return equation or the dividend equation is redundant and can be omitted.

Note that the price-dividend ratio cannot be omitted in a limited sample. Due to the recursive structure of the link between the three variables, we need at least one observation of the price-dividend ratio to initialize the recursion. Once this observation is available, we could even estimate the model using realized dividend growth and returns only.

From the remaining two equations, I decided to use the return equation. There are mainly three reasons for this. Firstly, the return equation is not seasonal and thus we do not need to apply seasonal adjustments. Secondly, we are fundamentally interested in returns and making inference about returns that are not exactly returns is dangerous (Cochrane (2008b)). If we use dividend growth instead, the return series implied by the framework must be reconstructed using the CS approximation and is therefore not exactly a return. In this case, we should work out the consequences of the approximation for returns more precisely. Conversely, if we assume that returns are correctly mapped by the model, only dividends which are less important for asset pricing must be reconstructed. The last advantage of utilizing returns is that they imply a smaller measurement equation as we do not need to include ε_{t-1}^M. This will be outlined in the next paragraph.

6.2.2 Specification of the Measurement Error

Let dividends that are postponed or paid earlier in period t be denoted as $\xi_t \sim N(0, \sigma_\xi)$. Given this process, the period t measurement error for dividends is defined as

$$\varepsilon_t^M = \xi_t - \frac{1}{3}(\xi_{t-1} + \xi_{t-2} + \xi_{t-3}) \tag{6.18}$$

which says that the retained dividends are paid out in equal parts over the next three quarters. This specification of measurement error follows most closely the advice of Hodrick (1992) because it imitates one-year trailing dividends and thus removes seasonality in a very robust manner. It is by far the easiest and most parsimonious filter that captures the irregular and time-varying seasonality pattern we can see in the data. Alternatively, we must specify how the pattern changes. This would involve more parameters, make the model even more complicated and increase the danger of mis-specification.

Another important property of (6.18) is that it does not cause persistent shocks to the log level of dividends, since each shock ξ_t is fully compensated over the next three quarters. This is of special importance when dealing with variables that are cointegrated with dividends. If it is not satisfied, which simply means that measurement error aggregates over time, the growing error drives a wedge between the observable variables and the cointegrating relationship between the true variables remains obscured.

An unexpected feature of the model is that retained dividends are not free of risk. Rather, they move one to one with regular dividends. This feature is needed to circumvent a retained dividend (as it theoretically does) lowering the leverage ratio of an asset and thus altering its risk characteristics. In the context of our models, this would imply that the models (affine) parameters would become time-varying. As the amount of dividends is small, the effect is negligible and we can simply ignore it. Under the assumption of invariable parameters, retained dividends must have the same risk characteristics as regular dividends because otherwise the no-arbitrage condition would be violated.

6.2.3 Augmented State Equation

We can put the measurement equation (6.18) and the VAR into a state space framework as follows. Firstly, define an augmented state vector that includes the state variables at time t and $t-1$ and the postponed dividends ξ_t plus

three of its lags:

$$W_t = \begin{bmatrix} x_t \\ \xi_t \\ \xi_{t-1} \\ \xi_{t-2} \\ \xi_{t-3} \\ x_{t-1} \end{bmatrix}$$

which has $2 \times K + 4$ elements. Define the transition equation

$$W_t = c + TW_{t-1} + R\varepsilon_t^W \tag{6.19}$$

where the vector c and the matrix T are defined as

$$c = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \qquad T = \begin{bmatrix} \Phi & 0 & 0 \\ 0 & \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

as well as

$$\varepsilon_t^W = \begin{bmatrix} \varepsilon_t^x \\ \xi_t \end{bmatrix}, \qquad R = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The identity matrix in the center of T has dimension 3. Finally, the covariance matrix of ε_t^W is

$$\text{cov}[\varepsilon_t^W] = \begin{bmatrix} \Sigma^x & 0 \\ 0 & \sigma_\xi^2 \end{bmatrix}.$$

6.2.4 Measurement Equation

The measurement equation for the observable price-dividend ratio at time t is

$$pd_t^{obs} = A_{pd} + B_{pd}^T x_t - \varepsilon_t^M. \tag{6.20}$$

Returns are exact and given by equation (6.16):

$$\begin{aligned} r_t &= \mu_{t-1} + C_r^T \varepsilon_t + \eta_t^d && (6.21) \\ &= A_r + B_r^T x_{t-1} + C_r^T x_t - C_r^T \Phi x_{t-1} - C_r^T \alpha + \eta_t^d && (6.22) \end{aligned}$$

Using $y_t = \begin{pmatrix} r_t \\ pd_t^{obs} \end{pmatrix}$ as the observable vector, the matrices of the state equation

$$y_t = d + ZW_t + \eta_t^W \tag{6.23}$$

have the following forms:

$$d = \begin{bmatrix} A_r - C_r^T \alpha \\ A_{pd} \end{bmatrix} \qquad \eta_t^W = \begin{bmatrix} \eta_t^d \\ 0 \end{bmatrix}$$

and

$$Z = \begin{bmatrix} C_r & 0 & 0 & 0 & 0 & B_r - C_r \Phi \\ B_{pd}^T & -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}.$$

Note that the second measurement equation does not contain an error, since all measurement error is in the state equation. This is unavoidable given the measurement error pattern defined in (6.18). A singular covariance matrix of the measurement noise does not present any theoretical problems. In fact, Kalman's (1960) original paper contains an example that uses perfect measurement.

6.3 The Basic Model

The basic model explored in this section defines a VAR:

$$\begin{pmatrix} \mu_{t+1} \\ g_{t+1} \end{pmatrix} = \begin{pmatrix} \overline{\mu} \\ \overline{g} \end{pmatrix} + \Phi \begin{pmatrix} \mu_t - \overline{\mu} \\ g_t - \overline{g} \end{pmatrix} + \varepsilon_{t+1},$$

together with a process for log dividends as in (6.5):

$$d_{t+1} = d_t + g_t + \varepsilon_{t+1}^d.$$

It is a property of the unrestricted version of this framework that the price-dividend ratio summarizes all the information available to the system that is useful to forecast returns or dividends. This means that estimates for μ_t and g_t are necessarily perfectly correlated. In fact, in the absence of measurement error regressing realized returns onto the price-dividend ratio produces the same forecasts as the Kalman filter approach. Cochrane (2008b) gives an analytical proof of this claim.[1]

We need some extra information to identify the independent dynamics of μ and g. The easiest way to separate the two variables is to impose a diagonal

[1] Another way to understand this is to use the property that d_t and p_t are a cointegrated system. Without further structure $p_t - d_t$ is an unbiased predictor of Δp_{t+1} and thus of r_{t+1} as well as of Δd_{t+1} by the Granger representation theorem.

VAR:

$$\begin{aligned} \mu_{t+1} &= \overline{\mu} + \Phi_\mu(\mu_t - \overline{\mu}) + \varepsilon^\mu_{t+1} \\ g_{t+1} &= \overline{g} + \Phi_g(g_t - \overline{g}) + \varepsilon^g_{t+1}. \end{aligned}$$

In this case the whole history of returns and the history of dividend growth rates contain information about the expected return and the expected dividend growth. For example when the two shocks are independent, we can find a constant a such that

$$r_{t+1} = \overline{r} + a\sum_{i=0}^{\infty}(\Phi - a)^i(r_t - \overline{r}) + \nu_{t+1}$$

is a forecasting regression that does not even use the price-dividend ratio. It says that the slow moving conditional mean of the expected return can be forecasted by an exponentially weighted mean of the entire history of past returns. The model in this section implicitly combines an analogous forecasting instrument with the information contained in the dividend yield. It is thus potentially able to improve the forecast of the simple regression. Moreover, it can separate expected dividend growth and expected returns.

Independent of the specification of Φ, the covariance matrix of the observable state space model is degenerate. This is because the three observable variables (realized returns, realized dividend growth, and the price-dividend ratio) are mechanically linked by the present value relation (see 6.17 and the discussion following it). Thus, two variables already contain all the information. We therefore only actually observe two processes in order to describe three shocks. Necessarily, the estimated covariance matrix of the three error terms ($\varepsilon^\mu_{t+1}, \varepsilon^g_{t+1}$, and ε^d_{t+1}) only has rank 2. This also means that it is not identified by the data and many covariance matrices produce the same maximum value for the likelihood function (see Rytchkov (2007)). This paper also shows that we can achieve identification by imposing a single restriction on the covariance matrix. I impose the restriction that ε^g_{t+1}, and ε^μ_{t+1} are uncorrelated.

6.3.1 Judging the Growth Rate Forecast

Whereas R^2 for the return forecast can be estimated as in (4.24), judging the ability of the model to forecast dividends is severely disturbed by the

measurement error and the seasonal pattern. I therefore compare annual changes in quarterly dividends:

$$\Delta_4 d_{t+4}^{obs} = d_{t+4}^{obs} - d_t^{obs}$$

which is reasonable for the slow-moving seasonality pattern we see in the data. In order to further reduce the influence of seasonality and measurement error, I also consider changes in four quarter averages:

$$\Delta_4 \bar{d}_{t+4}^{obs} = \sum_{i=1}^{4} d_{t+i}^{obs} - \sum_{i=0}^{3} d_{t-i}^{obs}.$$

The models forecast for this quantity is

$$\mathrm{E}_t\left[\Delta_4 \bar{d}_{t+4}\right] = 4g_{t|t} + 3(g_{t+1|t} + g_{t-1|t}) + 2(g_{t+2|t} + g_{t-2|t}) + g_{t+3|t} + g_{t-3|t}$$

where $\Delta_4 \bar{d}_{t+4}$ is defined in the same way as $\Delta_4 \bar{d}_{t+4}^{obs}$ (without the *obs* superscript) and $g_{t|s}$ is the expectation of g_t using information up to time s. The influence of measurement error can be even further reduced by considering two-year changes:

$$\Delta_8 \bar{d}_{t+8}^{obs} = \sum_{i=5}^{8} d_{t+i}^{obs} - \sum_{i=0}^{3} d_{t-i}^{obs}.$$

and the corresponding model implied quantity:

$$\begin{aligned}\mathrm{E}_t\left[\Delta_8 \bar{d}_{t+8}\right] = {} & 4\left(g_{t|t} + g_{t+1|t} + g_{t+2|t} + g_{t+3|t} + g_{t+4|t}\right) \\ & + 3(g_{t+5|t} + g_{t-1|t}) + 2(g_{t+6|t} + g_{t-2|t}) + g_{t+7|t} + g_{t-3|t}.\end{aligned}$$

In all these cases, R^2 is estimated as in (4.24).

6.3.2 Empirical Results

I use post-war data starting in June 1947 and ending in June 2006 (details are given in the data appendix). I find that the nominal version of this framework predicts quarterly aggregate stock returns with an R^2 of 2.7 percent. The implied excess returns can be predicted with an R^2 of 4.6 percent. The coefficients of the VAR matrix show that both expected returns and expected dividend growth are highly persistent. The corresponding coefficients are $\Phi_\mu = \Phi_{11} = 0.971$ and $\Phi_g = \Phi_{22} = 0.986$. There is, however, an unpleasant

result: the variance of ε_g converges to zero; thus dividend growth remains unpredictable and the corresponding R^2 are all below zero for all measures defined in the previous paragraph.

I also estimate a real version of the model. Real equity returns are predicted with an R^2 of 1.98 percent. This model implies $\Phi_g = 0.975$ and $\Phi_g = 0.743$. It is also rather unrealistic because the estimate of σ_{η^d} converges to zero. This means that all short term variability of the unexpected dividend growth is attributed to measurement error and innovations to the state variables. As in the nominal case, dividend growth remains unpredictable.

Even though we have already imposed a restriction through the diagonal VAR, the two measurement equations we have available cannot properly disentangle the three shocks (ε_t^g, η_t^d and ξ_t) that influence the dividend stream. Moreover, the likelihood function has also several other (local) maxima. All of them are border solutions or imply an unrealistic sawtooth-like impulse response function with $\Phi_g < 0$ (see van Binsenbergen and Koijen (2008) for similar problems). In order to estimate the model, we need more structure. This is the subject of the next chapter.

Chapter 7

Joint Stock-Bond Market Models

So far we have understood how an affine yield curve model can be integrated into a state space model (chapter 4) and how macro variables and returns can be integrated into this framework (chapter 5). Then, in the last chapter, we have seen how the Campbell-Shiller linearization can be used to put the price-dividend ratio and equity returns into the state space form. In this chapter, insights of the previous three chapters are combined in a joint stock-bond market model. I explore predictability and the ability of these models to explain the cross-section of returns.

7.1 Unspanned Factors Again

Annual dividend growth has a volatility of about 8 percent, implying an annualized quarterly dividend volatility of below 12 percent[1]. Stock return volatility is 16 percent. The gap between return volatility and dividend volatility must be filled by movements in the state variables. In a self-consistent set-up this means that yield changes are able to fill the gap.

[1]The difference between the volatility of quarterly and annual dividends stems from the fact that annual dividends are an average over four quarters. See Breeden, Gibbons, and Litzenberger (1989) for similar aggregation problems.

A closer look at the data, however, shows that self-consistent models produce a link between stocks and bonds that is too strong because they require that bond returns and changes in the price-dividend ratio are necessarily explained by the same shocks. Equivalently, unexpected dividend growth is the only source of equity returns that is not mimicked by bond returns. As the model is linear, a rough estimate shows that this means the three or four bond factors should explain at least 50 percent of aggregate equity returns. Empirically, however, they only explain about 8 percent (in terms of R^2) of equity returns.[2] Consequently, we need at least one state variable that does not influence bond prices but whose innovations explain a potentially large fraction of equity returns.

Of course, the theory of unspanned factors developed in section 5.2 is the key to relaxing the relationship between bond and equity returns. In the remainder of this section, I therefore develop and estimate models where the dividend yield is an unspanned factor.

7.2 The Campbell-Shiller Approximation and the Exponential Affine Pricing Kernel

General asset pricing says that given a pricing kernel, the dividend stream defines the price-dividend ratio as well as expected and unexpected returns. The exact solution of this problem is always non-linear to some extent (see appendix C, for the continuous time case). Under the exponential affine pricing kernel (5.16), however, a linear solution can be found using the CS approximation. In this case, a homoscedastic dividend stream with affine expected growth rates implies (approximate) expected returns and price-dividend ratios that are also affine in state variables. In order to see this, write the dividend stream as

$$\Delta d_t = A_d + B_d^T x_{t-1} + C_d^T \varepsilon_t^X + \eta_t^d \tag{7.1}$$

where the process of all priced variables X_t follows the degenerated VAR defined in (5.14). Recall that X_t can include arbitrary shocks as well as the state variables x_t. The residuals $\eta_t^d \sim N(0, \sigma_{\eta^d})$ are uncorrelated with ε_t^X.

[2]Note that the correlation between daily stock and bond returns varies considerably over time. The importance of this finding for a quarterly model is difficult to determine as time varying correlation is hard to measure at quarterly frequency.

Given these definitions, assume for the moment that the CS approximation holds and we can write returns as

$$r_{t+1} = \mu_t + (\rho B_{pd}^X + C_d)^T \varepsilon_{t+1}^X + \eta_t^d. \tag{7.2}$$

As in (6.16), let $C_r = \rho B_{pd}^X + C_d$. Due to the affine structure of this equation and the assumption of an exponential affine pricing kernel, we can use the risk return relationship (5.3) to derive expected returns which are indeed affine in the state variables

$$\mu_t = A_r + B_r^T x_t. \tag{7.3}$$

where

$$B_r^T = \rho B_{pd} \Sigma^{xX} \lambda_1 + C_d^T \Sigma^X \lambda_1 + \delta_1^T \tag{7.4}$$

and A_μ is given in (5.3a). We can now go through (6.11) to see that equation (7.2) was correctly assumed. This demonstrates that returns, dividends and the price-dividend ratio are all affine in the state variables and the shocks.

We can solve for the parameters of the price-dividend ratio:

$$\begin{aligned} B_{pd}^T &= (B_d^T - C_d^T \Sigma^X \lambda_1 - \delta_1^T)(1 - \rho\Phi + \rho\Sigma^{xX}\lambda_1)^{-1} \\ &= (B_d^T - C_d^T \Sigma^X \lambda_1 - \delta_1^T)(1 - \rho\Phi^*)^{-1}. \end{aligned} \tag{7.5}$$

To see this, plug B_μ in (7.4) into (6.12b) and solve for B_{pd}. Then use the definition of Φ^* (5.17a) for the second line. Finally, A_{pd} can be taken from (6.12a).

Note that this is not an explicit solution, since ρ is only implicitly defined. In general, solving for ρ turns out to be difficult and I am not aware of a general explicit solution.

7.3 The Stock-Bond Framework

Now we have all the ingredients for a joint stock-bond model. Simply assume that the yield curve model and the stock market model are based on the same VAR and priced with the same pricing kernel. In a self-consistent model we could use the yield curve factors to define the dividend stream and thus the equity model. In a model with unspanned factors, the state variables of the second block of (5.8) must also be defined. I use this indeterminacy to find an explicit solution for ρ in the next paragraph.

7.3.1 Parametrization

Many rotations of the model allow to solve for ρ. One of the most obvious - taking the price-dividend ratio as an additional state variable - turns out to be especially simple and is therefore used for all the models that follow. I provide details for $R = 1$, yet the procedure is very similar for more than one latent yield curve factor.

Taking the yield curve factors used above, plus the price-dividend ratio as an additional state variables, makes the affine parameters of the price-dividend ratio trivial ($A_{pd} = 0$ and $B_{pd} = e_K$) and the parameters of the risk-neutral VAR (Φ^* and α^*) remain in the staircase form of equation (5.8). Furthermore, $\delta_0 = 0$, $\delta_1 = e_1$, and Σ^X, $\Sigma^X\lambda_1$ are unrestricted as with previous models. The only minor adaptation of the yield curve part is that $\overline{x}$ is parameterized instead of α, since this is algebraically easier. It allows for a straightforward solution of ρ using (6.9). By definition, $\overline{x}_K = \overline{pd}$ and thus

$$\rho = \frac{\exp(\overline{x}_K)}{1 + \exp(\overline{x}_K)}$$

which is then used to calculate A_r and B_r using (7.2) and (5.3).

While in the previous section, A_{pd} and B_{pd} are functions of the dividend stream and the pricing kernel, they are now fixed and thus entail $K + 1$ parameter restrictions elsewhere. They are revealed by solving (7.5) for

$$B_d^T = e_K^T - e_K^T\rho\Phi^* + C_d^T\Sigma^X\lambda_1 + e_1^T \tag{7.6}$$

which is plugged into (6.13) to determine

$$A_d = \log(\rho) + A_r + (B_r - B_d)\overline{x}. \tag{7.7}$$

The specification of the economic part of the model is completed by allowing C_d to be unrestricted.

Note that this parametrization involves the minimum number of parameters necessary to describe the model without imposing (implicit) restrictions. Furthermore, all state variables have an unequivocal economic interpretation, thus the model is globally identified.

7.3.2 Definition of the State Space Model

With these definitions, the augmented state equation developed in section 6.3 remains unchanged. The measurement equation for bonds will also be used in

its original form defined in chapter 4. Adopting the seasonality filter given in (6.18), the same holds for the measurement of the price-dividend ratio and the returns defined in (6.23). With these measurement equations, we only have K independent returns ($K-1$ independent bond returns and the equity return) to price $K+1$ shocks (the K state variables plus η^d). Therefore, it is assumed that the idiosyncratic dividend shocks η^d are not priced, i.e. $X_t = x_t$, in the definition of the pricing kernel (5.16). This identifies the model and completes the parametrization.

Another advantage of choosing the price-dividend ratio as a state variable becomes visible by inspection of the corresponding measurement under the chosen parametrization:

$$pd_t^{\text{obs}} = x_{t,K} - \xi_t + \frac{1}{3}\xi_{t-1} + \frac{1}{3}\xi_{t-2} + \frac{1}{3}\xi_{t-3}.$$

This is the equation with the parameters whose derivatives relative to the likelihood function cannot be estimated by the method developed in chapter 2 because there is no measurement error term involved. This is no longer a problem since all parameters are now fixed. Consequently, the entire derivative can be estimated analytically, delivering fast and reliable numerical optimization.

7.3.3 Preliminary Results

A joint stock-bond model based on an unrestricted three factor yield curve model with the price-dividend ratio as the only unspanned macro factor is analyzed. As already mentioned, in order to identify the model, only shocks to the state variables are priced.

The first line of table 7.1 shows that the model predicts quarterly returns for the aggregate stock market with an R^2 of 0.043. Excess returns are forecasted with an R^2 of 0.062 and annual dividend growth is fitted with slightly more than two percent. The two-year forecast explains a larger fraction and reaches an R^2 of 0.101. Finally, the bond risk premium is essentially the same as in the $AP(3,3)$ model.

7.3.4 Enlarging the Cross-Section of Returns

In the previous paragraph, the pricing kernel is not identified because there are less returns than priced shocks. Clearly, this problem can be overcome

Table 7.1: Estimates for Four-Factor Stock-Bond Models

	rx^{10}	$\widetilde{sr}$	r	rx	$\Delta_4 d_{t+4}$	$\Delta_4 \bar{d}_{t+4}$	$\Delta_8 \bar{d}_{t+8}$
1	0.215	0.096	0.043	0.062	0.023	0.025	0.101
2	0.216	0.094	0.049	0.068	0.018	0.032	0.061
3	0.217	0.093	0.052	0.071	0.017	neg.	0.083
4	0.218	0.093	0.053	0.071	0.022	0.014	0.087

Note: in model 1, dividend shocks η_t^d are unpriced. For models 2-4, these shocks are priced. The second model includes the sector portfolio returns, the third model, the size and book-to-market sorted returns. Finally, model 4 adds the price-earnings ratio as an additional measurement equation to the third model. All figures are R^2 of the implied forecast based on the filtered series and estimated using (4.24). Bond returns are overlapping annual returns, stock returns are quarterly. Dividend growth rates are defined in paragraph 6.3.1. Yields are as in table 4.1. Quarterly data, June 1947 - June 2006.

by simply adding further asset returns to the system. The corresponding technique has already been derived in section 5.1.

Although one additional asset is already sufficient for identification, more than one series is added. This generates (testable) overidentifying restrictions. More precisely, two different set of returns are used. Firstly, the six size and book-to-market sorted Fama-French portfolios and, secondly, 9 of the 10 industry portfolios ('others' is excluded) from the same source as described in the data appendix. It is assumed that these returns obey (5.2)

$$rx_t^i - \mathrm{E}_{t-1}[rx_t^i] = C_{rx}^{i,T} \varepsilon_t^x + u_t^i$$

where each u_t^i is an idiosyncratic shock. In this case, expected returns follow from (5.3) and returns are therefore homoscedastic and affine. Consequently, they can be included in the measurement equation following the method described in paragraph 5.3.2. Finally, let the state equation that describes the economy be defined as in the previous paragraph.

Table 7.1 shows that these extensions have similar prediction power as the models without the extra returns. There is a small shift towards a better return prediction, while the ability to forecast dividends is slightly reduced.

7.3.5 Overidentifying Macro Variables

We can also add macroeconomic quantities (referred to as y_t^m) to the measurement equation as long as these variables can reasonably be modeled as affine functions of the state variables. This implies extra measurement equations of the following form:

$$y_t^m = A_m + B_m x_t + \eta_t^m. \tag{7.8}$$

In general, A_m and B_m are unrestricted parameters. η_t^m is usually assumed to be an idiosyncratic measurement error. I estimate an example of this model using the price-earnings ratio. In this case, the affine form can be justified by assuming that prices, dividends and earnings are all cointegrated. This holds under very mild conditions. Such a system is explored by Lamont (1998).

The inclusion of this variable slightly improves predictability. Of course, the price-earnings ratio could also be included as an additional unspanned factor. Utilizing another set of instruments, this extension is the subject of the next section.

7.4 Increasing the Cross-Section of Prices

7.4.1 Specification

Under very general conditions, the price-dividend ratio of a dividend paying asset is approximately affine in the state variables. For instance, if the dividend stream is of the general affine form defined in (7.1), this follows from the CS approximation derived in second section of this chapter. There are even more involved processes that lead to an affine structure. For example, we can generalize (7.1) as follows:

$$\mathrm{E}_t[g_{t+s+1}] = A_d^{\infty} + \rho^s \widehat{A}_d + B_d^T \mathrm{E}_t[x_{t+s}].$$

This specification captures the feature that dividends, e.g. of growth stocks, temporarily grow faster. It is also possible to define B in a similar way. Even in this case, the resulting price-dividend ratio is affine in the state variables.[3] Note that there is, in general, no restriction between the affine parameters of returns (C_r) and those of the price-dividend ratio (A_{pd} and B_{pd}), since the

[3]Details are available from the author upon request.

above dividend processes are parameterized with at least $2 \times K + 1$ parameters. Unless there is more structure, we therefore need not specify the dividend stream exactly, rather we can simply assume the price-dividend ratios to be an unrestricted affine functions of the state variables. Consequently, they are used as measurement equations in exactly the same ways as the macro factors in equation (7.8). Seasonality of the payments is removed by using trailing dividends as specified in (A.3).[4] Error terms are interpreted as measurement errors and thus assumed to be uncorrelated with one another and with economic shocks. They are assumed to have the same standard deviation; their covariance matrix is thus a scaled identity as in the case of bonds.

Finally, Φ^* and α^* are parameterized as

$$\Phi^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ c_1 & c_2 & c_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \phi^*_{51} & \phi^*_{52} & \phi^*_{53} & \phi^*_{54} & \phi^*_{55} \end{bmatrix} \qquad \alpha^* = \begin{bmatrix} 0 \\ 0 \\ \alpha^*_5 \\ 0 \\ \alpha^*_5 \end{bmatrix},$$

which means that the first three factors are the standard yield curve factors. The fourth factor is the price-dividend ratio and the fifth, its one-step-ahead expectation under the risk-neutral measure.

This model is estimated using price-dividend ratios and returns of the aggregate market and the six Fama-French portfolio as well as the 13 yields already used before.

7.4.2 The Predictability Pattern

I perform likelihood ratio tests for the rank of λ_1 in the same way as in chapter 4. These results are displayed in panel B of table 7.2.[5] The test

[4] It is also possible to assume that the payout policy for all equity portfolios follows the same seasonality pattern as that of the aggregate market defined in (6.18). Although the ability of this model to predict dividends, and to a lesser extend returns, is reduced, the qualitative properties of this model are similar.

[5] Due to the large size of these models, estimation is no longer fully stable. The gradient-based optimization algorithm was, in certain cases, unable to improve the estimate of a constraint model because the likelihood function was so flat that even the analytical gradient was unstable. Another starting point was then found by restricting the search space to the previously restricted part or using the Nelder-Mead algorithm. This was possible in every case. Once the gradient method began to improve from a reasonable starting point, it

strongly rejects the null of rank zero against rank one and the null of rank 1 against rank 2 at all conventional significance levels. The null of rank 2 is rejected against the alternative of rank 3 at 5 percent but not at 1 percent. The increase in the likelihood function is clearly insignificant for higher ranks. In the remainder of this chapter, the rank-two model is therefore discussed.

The distinguishing property of the five factor models is that they imply much stronger cash-flow predictability. The R^2 for annual changes of quarterly dividends is now almost 10 percent. This value increases to 0.143 for yearly changes in annual dividends and reaches a striking value of 0.219 for two-year changes. These values are more than four times higher than those of the previous model. Bi-annual figures have also more than doubled.

The slope return forecast has also improved and is now 0.126. The R^2 for the stock market excess return, however, has diminished to 0.048. Moreover, the forecast for the term premia is also lower, at least for the rank two model. All these results are summarized in panel B of table 7.2, which also contains results for the six size an book-to-market sorted portfolios.

7.5 Cross-Section of Returns

7.5.1 Hypothesis Tests

At the end of section 5.1, a test for the cross-sectional pricing ability of an affine model is described: we can test the parameter restriction on A_{rx} implied by equation (5.3a) by comparing the likelihood of the constraint model with one leaving some A_{rx} as free parameters. For instance, the model implied null can be tested against the alternative that one single equity portfolio return is not correctly specified. The one and five percent critical values are 3.3 and 1.9 respectively. These test statistics are displayed in the first row of panel B of table 7.3. The null is rejected for the small growth portfolio and also for the large value portfolios: in both cases, even for the one percent critical value. For the other four portfolios, the null cannot be rejected at the five percent level.

The joint significance of all restriction can also be tested. In this case, there are six degrees of freedom and the null is rejected for an increase in

reached the same optimal value with reliable consistency. Still, all the results provided here are reliable because there is no evidence for other (local) maxima.

the likelihood function of 6.3 and 8.4 (five and one percent critical values respectively). The test statistic is 8.1 and the null is thus rejected at the five percent value only.

In absolute values, the pricing error of the model is -0.47 percent per quarter for the small growth portfolio and -0.37 for the large value portfolio. All other portfolios have a much smaller pricing error as can be seen in the

Table 7.2: Estimates for Five Factor Stock-Bond Models

	Panel A: Prediction					
Rank of λ_1	0	1	2	3	4	unrestr.
$rx^{(10)}$	-	0.008	0.192	0.218	0.220	0.220
$\widetilde{sr}$	-	0.125	0.126	0.147	0.169	0.178
r	-	0.008	0.030	0.028	0.042	0.043
rx	-	0.027	0.048	0.047	0.060	0.062
$\Delta_4 d_{t+4}$	-	0.098	0.093	0.093	0.102	0.101
$\Delta_4 \bar{d}_{t+4}$	-	0.170	0.143	0.147	0.158	0.159
$\Delta_8 \bar{d}_{t+8}$	-	0.229	0.219	0.226	0.252	0.251
rx, s/g	-	0.021	0.038	0.036	0.055	0.055
rx, s/m	-	0.010	0.036	0.040	0.051	0.051
rx, s/v	-	0.008	0.039	0.047	0.057	0.057
rx, l/g	-	0.027	0.040	0.038	0.047	0.048
rx, l/m	-	0.027	0.047	0.045	0.058	0.060
rx, l/v	-	0.013	0.034	0.037	0.054	0.055
	Panel B: Likelihood Ratio Test					
Log Likelihood	18069.1	18093.3	18113.2	18119.3	18122.2	18122.5
Number of Para.	116	126	134	140	144	146
χ^2_{95}	18.3	15.5	12.6	9.5	6.0	
χ^2_{99}	23.2	20.1	16.8	13.3	9.2	

Note: the prediction section contains R^2 (based on equation (4.24)) for the filtered forecasts from the quantities indicated in the first column. The specification of the model is indicated in the first row. Bond returns are overlapping annual returns, while stock returns are quarterly. Dividend growth rates are defined in paragraph 6.3.1. The same yields as in table 4.1 are used. The last two lines of the second section contain critical values for the likelihood ratio test between the model in the corresponding column and that of the next column. Quarterly data (June 1947 - June 2006).

third row of panel B.

7.5.2 Loadings to the Priced Shocks

The first section of table 7.3 displays the loadings of the style portfolios relative to the priced shocks (C_{rx}) and the corresponding prices of risk. The most interesting outcome is that value stocks have much higher loadings to η_d as can be seen in the last row of the panel. For small and large caps growth portfolios, the loading is 0.75 and 0.78 respectively. For value portfolios, the corresponding values are equal to 1.27 and 1.34 respectively. The Sharpe ratio of this shock is by far the highest and thus the different loadings to η_d explain the value premium to a great extent.

Economically, this is a sensible interpretation of the value premium. η_d affects prices and dividends in much the same way. More precisely, a one percent η_d shock unexpectedly increases both prices and dividends by one percent, but leaves the entire state of the economy unchanged. It is thus a pure and persistent growth shock that increases the level of the entire economy, or the cointegrated system consisting of dividends and stock prices, permanently by one percent.

A concern about this interpretation is that η_d is a nominal shock rather than a real shock. This should not be a large problem for two reasons. Firstly, it is an unexpected growth shock and unexpected inflation is not very important at quarterly frequency. The implicit assumption that unexpected inflation is not priced is thus not crucial (see also Brennan, Wang, and Xia (2004) for another model with this feature). Secondly, we can reparameterize the model as in panel A2 of table 7.3. In this case, the priced shocks are the excess return of the stock market and innovations to the state variables. This is the usual way to write Merton's ICAPM. Still, including expected and unexpected inflation into the model would be an interesting and potentially fruitful extension.

The interpretation of the state variables, however, depends strongly on the rotation of the model. If we rotate the model such that the market excess return, the equity premium and the expected real growth rate (with the short rate as a crude proxy for expected inflation) as the first three factors and othogonalized yield curve factors as the remaining factors, we see the usual market beta pattern with small and growth stocks having higher betas (table 7.3, panel A2). There is not much variation in the loadings to the equity

premium. Moreover, the exposure to the shocks to dividend growth is higher for small and value stocks. This difference and the influence of the yield curve explain the value premium under this parametrization.

Panel A3 contains a third rotation. It uses the η_d shock instead of the market excess return but is otherwise constructed in the same way as the previous rotation. The η_d loadings are exactly the same as in the unrotated model (because this is an orthogonal shock). The exposure to the equity premium is slightly more negative for growth and small stocks whereas the coefficients for the expected growth rate are clearly higher for small caps.

The last four rows of panel B also provide a model-implied attribution of filtered unconditional returns using the APT-like relationship (5.4). The variation in the loadings to η_d alone implies a value premium of more than one percent per quarter. Moreover, this attribution also shows that small caps have a higher premium due to an increased exposure to yield-related risk and to a lesser extend, to the dividend growth rate. Neither the filtered expected return nor the sample mean, however, provide evidence for a small cap premium.

As already mentioned, a shortcoming of the model is that it is not able to explain the poor performance of the small growth portfolio and it overestimates the large value premium. This is due to the fact that the value premium is much more pronounced for small caps, but the variation in the exposure to the factor that explains it (η_d) is nearly identical for small and large caps. The model is thus not able to explain the large difference between the value premium of small caps relative to the smaller value premium of large caps.

7.6 Statistical and other Problems

7.6.1 Efficiency

If the state variables are observable, even this model provides no efficiency gain for predictability relative to the regression approach. The reasoning is the same as for the yield curve models. We can start the estimation with the OLS estimates of the VAR of the state variables and the return equation (7.2). Given these regression coefficients, B_d can be estimated from the approximate return equation (6.6), which determines Φ^* through (7.6) and the yield curve model. The mechanism is quite obvious: there is no efficiency gain from

the state space approach, as long as we add K parameters (by enlarging λ_1) for each new shock as we do in the general case. Only restrictions implied by economic theory or rank-reduction techniques will improve efficiency and thus lower the standard deviation of the estimated forecasting parameters.

This raises another question: if the Kalman filter forecast is the same as the OLS forecast, how can it be that the R^2 in the five factor model is lower than in the smaller model with fewer variables? This feature stems, of course, from the over-identifying restrictions implied by the larger cross-section of returns. If we have as many returns as priced shock, a larger unrestricted model would always imply a higher in-sample R^2 as with OLS.

7.6.2 Stability

Stability can be a problem for large models because small mis-specifications of one part of the model can be amplified to completely implausible estimates of other aspects of the model due to implied restrictions or other connections. This is partially true here. As the model explains 98.6 percent of some portfolios (see panel C of table 7.3), an economically small pricing error of such an asset can become statistically very important because the idiosyncratic risk is very small. This makes the cross-sectional pricing sensitive to outliers.

Stability, however, does not depend on the parametrization. Although the way the state variables are defined clearly causes multi-collinearity problems, the economically meaningful variables, such as yields and filtered (expected) returns are numerically stable and thus independent of the parametrization. At the maximum value of the likelihood function, all covariance matrices have an inverse of the conditional number that is at least 10^6 times higher than Matlab's numerical precision. Note that the conditional number does not depend on the parametrization.

Sometimes, the multicollinearity problem harms optimization because we are close to an instable model. Small deviations caused by the numerical search algorithm can lead to a point were the analytical gradient is no longer appropriate and the algorithm thus stops. This is not a fundamental problem and the way it is solved is already described in footnote 5 on page 107 and in appendix B.

A few parameters are also very unstable. Most importantly, these are the means of the short rate and the dividend yield. As these two parameters considerably influence unconditional risk prices, the model's estimate of the

unconditional equity premium is not stable either. This instability, however, barely affects the filtered return expectations. For instance, imposing the historical average for the short rate and the dividend yield lowers the unconditional equity premium by almost one percent, however, does not reduce the average of the filtered return expectations. Because of their higher stability, averages of filtered state variables are used in panel B of table 7.3. Finally, note that the two parameter vectors are not statistically different from each other. The reduction of the maximum value of the likelihood function implied by the constraint is only 0.6.

I also explored different specifications of the model (not reported); specifically, the version of the model with seasonally adjusted dividend-price ratios (see footnote 4 on page 106), and different versions with more assets (sectors) or where some assets are omitted. For all these specifications, there is a substantial predictability of dividends and the value effect is greatly explained by different exposures to the persistent shock. This is not the case if the 25 size and book-to-market sorted Fama-French returns are added to the measurement equation. This model is numerically unstable and there seem to be different local maxima of the likelihood function. For the model with the largest likelihood (that I found), η_d was still the shock with the highest risk-return ratio. This estimate also produces increasing loadings when we move from growth to value stocks, at least for small and mid caps. This effect vanishes for the largest size quintile.

7.6.3 Analysis of the Measurement Error

An analysis of the measurement error delivers similar results as those of paragraph 4.3.5. Autocorrelation of the noise term in the measurement error has, however, become even more dramatic. The autocorrelation of the price-dividend ratios of the size and book-to-market sorted portfolios is as high as 0.85 on average. This does not come as a surprise because seasonality is removed by using four-quarter averages and not using a filter as for the aggregate market. There is also substantial multi-lag autocorrelation of these quantities as well. For instance, for the fourth lag, the autocorrelation is still 0.39. This value can no longer be explained by the construction of the dividend yield using trailing dividends.

7.7 Summary of the Empirical Findings

In this chapter, joint stock-bond models have been explored. The main empirical results are:

- Quarterly equity excess returns are predictable with an R^2 of approximately five percent.
- There are at least two independent predictability factors: the term premium and the equity premium.
- Annual and bi-annual dividend growth is predictable with an R^2 of 0.14 and 0.22 respectively.
- A large fraction of the value premium can be attributed to a higher exposure of value stocks to the single persistent shock of the system.
- Small stocks have a higher exposure to interest rate risk and to shocks to the dividend growth rate.

Table 7.3: Five Factor Stock-Bond Model with Rank $\lambda_1 = 2$

	Sharpe	s/g	s/m	s/v	l/g	l/m	l/v
				Panel A1: Not Rotated			
$y^{(1)}$	-0.16	283.69	337.68	409.41	-130.91	77.92	146.52
$f^{(2)}$	-0.12	-686.80	-796.95	-959.89	291.68	-183.49	-339.90
$f^{(3)}$	-0.09	441.71	493.99	590.23	-162.51	114.61	209.88
dp	0.09	15.04	19.19	23.56	-6.78	3.19	8.84
$\mathrm{E}^*[dp]$	0.08	-13.35	-17.96	-22.41	7.82	-2.42	-8.02
η_d	0.68	0.75	1.03	1.27	0.78	1.07	1.34
				Panel A2: Rotation 1			
$r - y^1$	0.34	1.46	1.09	1.03	1.09	0.81	0.87
$\mu - y^1$	0.12	-1.29	-0.45	-0.35	-0.18	-0.38	-0.18
$g - y^1$	0.05	2.59	4.09	4.65	-1.46	2.08	1.95
				Panel A3: Rotation 2			
η_d	0.68	0.75	1.03	1.27	0.78	1.07	1.34
$\mu - y^1$	0.12	-5.62	-3.68	-3.41	-3.39	-2.78	-2.76
$g - y^1$	0.05	10.95	10.35	10.56	4.75	6.72	6.93
				Panel B: Cross-Sectional Pricing			
Δ Log Likelihood		4.8	0.6	0.6	0.1	1.5	4.1
Mean Sample Return		1.34	2.97	3.47	2.16	2.66	3.09
Pricing Error		-0.47	0.07	0.06	-0.03	-0.18	-0.37
$\mathrm{E}[rx]$		1.29	2.48	3.00	1.79	2.48	3.09
η_d Contr.		1.38	1.88	2.33	1.44	1.96	2.47
$\mu - y^1$ Contr.		-1.40	-0.92	-0.85	-0.84	-0.69	-0.69
$g - y^1$ Contr.		0.32	0.31	0.31	0.14	0.20	0.21
Yld. Contr.		0.89	0.74	0.75	0.46	0.28	0.45
				Panel C: Explanation of Returns			
h		7.591	3.376	3.398	2.934	4.281	5.674
Std. Dev.		26.407	20.101	20.175	17.386	14.099	15.936
R^2		0.958	0.986	0.986	0.986	0.953	0.934
R^2 (CAPM)		0.813	0.799	0.719	0.934	0.868	0.775

Note: panel A1 contains C_{rx} as defined in the text and prices of risk (expected log returns divided by their standard deviation). Panels A2 and A3 do the same for rotated parametrizations: yield curve shocks are orthogonalized and not displayed, other variables are indicated on the left. The first row of panel B contains the increase of the log likelihood leaving the intercept of the corresponding return unrestricted; then the sample mean and the pricing error of the model, followed by the average of filtered expectations and an attribution using the parametrization of panel A3 are displayed. Panel C contains the model-implied idiosyncratic risk and the return volatility followed by the fit of the model and that of the CAPM. The six size and book-to-market sorted Fama-French returns are indicated in the first row. Quarterly data June 1947 - June 2006 using the same yields as in table 4.1.

Chapter 8

Conclusion

Solely based on prices of some of the most important asset classes, this thesis has provided strong evidence in favor of return predictability that is essentially free of any selection bias. The detected pattern is very rich and even implies independent movement in expected returns across different risk factors and asset classes: there are time periods where some risk premium - e.g. the equity premium - is high while another - e.g. the term premium - is low or even negative. This feature of the data is extremely hard to understand in a utility-based framework and cannot even be explained by simple time-varying risk aversion. Why should an investor fear equity risk when he has no aversion to interest rate risk?

There are essentially two explanations of this puzzle. The first is strong time-varying covariance between macroeconomic variables and asset returns (see Campbell, Sunderam, and Viceira (2008) for a model with that feature). The second is that risk-sharing is not fully efficient across asset classes.

In both cases, passive investment strategies, such as a simple buy and hold strategy, are clearly not efficient. This is evident in the case of inefficient risk-sharing. In the case of time-varying covariances, constant portfolio holdings are inefficient, since individual risk aversion does not necessarily move one-for-one with aggregate risk aversion. For example, a long-term investor who is not exposed to macroeconomic risk should increase his holdings in risky assets when the premium is high. Conversely, the exposure should be reduced when the risk premium is low. Since there is independent movement in the

prices of risk across different assets, this investor should even actively select where to allocate the risk budget.

An essential feature of this thesis is that the economy is modeled as a dynamic state space system. This is the standard approach for modeling yield curves, but only in recent years has it become popular to describe equity markets. I expect this trend to continue because this thesis has highlighted many advantages of the state space analysis relative to the standard factor approach. Probably the most distinguishing feature is that returns are linked to movements of meaningful state variables, particularly expectations.

In practice, this connection helps to estimate actual expected returns and thus provides the basis for an active investment strategy that accounts for the rich time-varying predictability pattern. More precisely, once the relationship between movements in state variables and returns is understood, this connection can also be reversed. Given the sensitivity of an asset to changes of some state variables, a cross-sectional regression of returns already provides an estimate of changes in the state variables. If the cross-section is sufficiently large, these estimates will be accurate. We therefore do not need to observe all state variables immediately; returns are sufficient to provide an appropriate update of expectations and to adjust the portfolio to unexpected changes of the state of the economy. Once macroeconomic quantities are observed, we can update the expectation and possibly reallocate the portfolio again. Using Kalman filter terminology, these steps are, of course, (partial) measurement updates with potentially lagged measurement.

The state space approach is also appropriate for portfolio simulation used for asset and liability management (ALM) or to define the (strategic) asset allocation (see e.g. Keel and Müller (1995) for an introduction to the topic). This has some obvious advantages relative to the standard random walk hypothesis or even to a model with predictable factors. Most importantly, it links current and future returns. For instance, an increase in the equity premium produces an immediate decline in prices and thus connects the current price shock to the expectation of future returns. An investor who erroneously assumes that prices essentially follow a random walk and whose ability to take risks depends on current wealth will reduce his risk exposure. If the same investor understands the relationship between future returns and current prices, there are two offsetting effects. On the one hand, his willingness to take equity exposure has increased because of the higher premium; on the

other hand, his ability to take risks is lower because of the lower reserves. This thesis has provided evidence that shocks to the state variables explain more than 50 percent of aggregate equity returns. Consequently, the difference of the two approaches may be substantial and possibly even more pronounced in the long run.

Another advantage is that it is possible to include important macroeconomic quantities such as inflation and GDP growth and to properly link them to asset returns. The long-term relationship between dividends (or equity prices) and macroeconomic quantities such as aggregate labor income can be particularly important in this context. For instance, they can be modeled as a cointegrated system sharing a common macroeconomic trend as in Ribeiro (2004). This will further reduce the random walk component of equity returns.

If the portfolio management decides to follow an optimal, and thus active, asset management approach, this decision should also be fully reflected in the ALM framework. More precisely, we should not only simulate constant portfolio holdings, rather the simulation should be based on the intended active portfolio strategy that is to be implemented. This implies that active tactical portfolio allocation and long-term strategic asset and liability matching should be integrated in a single optimization framework that delivers decision rules based on the actual state of the economy and the investor-specific liabilities. The conventional separation between a beta-producing strategic asset allocation and alpha-producing active management is no longer advisable because it is clearly a suboptimal investment policy. It is founded on the presumably wrong assumption that beta returns follow a random walk and alpha returns are produced by management skills that are independent of the state and the evolution of the economy. The recent performance of active managers, particularly hedge funds, has produced a lot of evidence for this claim.

As already indicated, ALM studies should potentially include more assets and macroeconomic quantities than the joint stock-bond models of the previous section. The Kalman filter technique used in this thesis has, however, already reached its limits in terms of numerical stability and computation time. In order to estimate these larger models, we therefore need a more robust and faster estimation technique. A natural way to simplify estimation is to separate the two steps of the Kalman filter and to apply a two-stage procedure. In the first stage, the latent state variables are estimated based on

some rank-reduction technique such as the principal component analysis. The predictability pattern is then fitted as a (restricted) VAR. This is not a big loss because the Kalman filter approach does not provide better forecasts per se. The main advantage of this technique is that it produces cleaner data. If, however, the cross-section is sufficiently large, a cross-sectional regression already provides accurately fitted state variables. A two-stage approach might therefore be almost as efficient as the fully efficient maximum likelihood estimation based on the Kalman filter.

Clearly, this approach can also be used to incorporate those findings that I was not able to include in the joint stock-bond market models of the previous chapter. Specifically, these are the ability of expected inflation to improve the level return forecast and the better predictability of slope returns in four factor models.

Still, the Kalman filter is a natural way to estimate state space systems that has some important advantages for the purposes of this thesis. Specifically, one goal of this thesis was to explore the predictability pattern by means of rank-reduction techniques, using only the most important assets in order to reduce the selection bias. This, of course, requires testable hypotheses. Statistical tests, however, are easier and more accurate for one-step estimators.

To conclude, one final advantage of the state space framework should be highlighted again: state space models define the dynamics of the entire system, which can therefore be simulated as a whole. Consequently, the small sample properties of the test statistics can be simulated. This and out-of-sample tests of return prediction are left for future research.

Appendix A

Description of the Data Sets

Dividend yields are estimated as in Ang and Liu (2004). They compute monthly dividend yields as the difference between the return with dividends and capital gains, and the return excluding dividends. I use the same technique for quarterly data:

$$\frac{\overline{D}'_{t+1/4}}{P_t} = \frac{D'_{t+1/4} + P_{t+1/4}}{P_t} - \frac{P_{t+1/4}}{P_t} \tag{A.1}$$

$$\overline{D}'_{t+1/4} = \left(\frac{D'_{t+1/4} + P_{t+1/4}}{P_t} - \frac{P_{t+1/4}}{P_t} \right) P_t \tag{A.2}$$

where the frequency 1/4 refers to quarterly data. The bar superscript indicates a quarterly, as opposed to annual, dividends. Cochrane (1991) shows that this procedure implies that dividends paid early in the month are reinvested at the portfolio return to the end of the month. This leads to an overestimation of dividends in high-return months, and to an underestimation in months with low returns.

The standard practice to remove seasonality is to sum up the dividends over the past four quarters (see Ang and Liu (2004) and Hodrick (1992)):

$$D_t = \sum_{i=0}^{3} \overline{D}_{t-i/4}. \tag{A.3}$$

Other seasonally adjusted dividends cannot be constructed because the seasonal pattern of dividends changes dramatically in the 1960's.

In recent years, a growing share of cash-flow distribution to investors was in form of equity repurchase instead of dividend payments. The implied cash-flow induced by these repurchases can be written as

$$\overline{C}_{t+1/4} = \frac{\sum \max(\Delta N, 0) \cdot \frac{P_t}{P_{t+1/4}} \mathrm{MV}_{t+1/4}}{\sum \mathrm{MV}_{t+1/4}}$$

where

$$\Delta N = \frac{\mathrm{NOSH}_t \cdot \mathrm{AF}_t - \mathrm{NOSH}_{t+1/4} \cdot \mathrm{AF}_{t+1/4}}{\mathrm{NOSH}_t \cdot \mathrm{AF}_t}.$$

NOSH is the number of shares. AF is a factor that adjusts for share splits and other transactions such as face value reduction. The dividend yield at time t is then defined as

$$\frac{\overline{D}_t + \overline{C}_t}{P_t}. \tag{A.4}$$

The annual dividend yield is estimated using (A.3) and the analog definition for C_t.

Returns with dividends and capital gains as well as break points for the size- and value-sorted portfolios are from Kenneth French's web site.[1] Equity repurchases are estimated from Datastream stock data (all US stocks) and therefore do not cover exactly the same universe as the Fama-French data. They only start in 1980 for the size and value sorted portfolios and in 1973 for the aggregate market. The later start of these data is a minor problem because equity repurchases were virtually non-existent before 1980.

The earnings-price ratio is defined as the corresponding value of the Standard & Poors (S&P) Composite Index. S&P earnings are also used by Lamont (1998) and Shiller (2000). This data is made available by both authors on their web pages[2] and by Datastream. Robert Shiller's data set is regularly updated. Following Lamont (1998), I use single quarter earnings without seasonal adjustment. Clearly, earnings for the current quarter are not known at the end of the quarter, since they are published only during the next quarter. To avoid any forward-looking bias implied by this fact, I use earnings for

[1] Available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french

[2] See http://www.som.yale.edu/faculty/oal4/research/earnpublic.prn for Owen Lamont's data set and http://www.econ.yale.edu/~shiller/data/ie_data.xls for that of Robert Shiller.

the last quarter divided by the current price as the definition of quarter-end price-earnings ratio.

CPI data is from the Federal Reserve Economic Data (FRED) and is made available by the Federal Reserve Bank of St. Louis[3]. The CPI is the seasonally-adjusted CPI for all urban consumers (CPIAUCSC).

[3]http://research.stlouisfed.org/fred2/

Appendix B

Optimization

2.1 An Alternative Optimization Procedure

I also tested the following optimization routine:

1. For each element of the parameter vector estimate, the forward and the backward derivative by the finite differencing method using an initial step length h_0.

2. If the ratio of the two derivatives is between ζ and $\frac{1}{\zeta}$, the average of the two is taken as the (partial) derivative. ζ is a number close to 1. Otherwise, the step length is divided by ψ and the forward and backward difference is reestimated with the new step length.

3. Step 2 is repeated until the step length of the finite difference method is $h_{\min}$. If the ratio of the two derivatives is still outside the interval, it cannot be estimated accurately and is thus set to 0.

Typically, I use $h_0 = 10^{-6}$, $\zeta = 0.9$, $\psi = 10$, and $h_{\min} = 10^{-12}$ but sometimes h_0 must be greater. This is because for large parameters a step length that is too small can also be problematic, due to rounding errors.

The algorithm works well for small models, but it is extremely slow for the bigger ones, since each derivative needs at least three runs of the Kalman filter. Consequently, all the numerical optimizations in this thesis are, whenever possible, based on analytical derivatives. Unfortunately, as described

in section 2.3, not all derivatives can be calculated analytically when one of covariance matrices (Q or H) is degenerate. In this case, the very high precision of this numerical algorithm is very useful. In particular, it can be used to estimate those partial derivatives of the likelihood function relative to the Kalman filter parameters that cannot be calculated analytically. The resulting numerical derivatives are then used to calculate the derivatives of the log likelihood function relative to the model's parameter vectors by utilizing the chain rule in exactly the same way as in the case of analytical derivatives. Note that the application of the chain rule requires very high precision of the derivatives and that this is generally not provided by standard finite differencing. This is because small errors can become very big through the matrix multiplications involved in the chain rule. Consequently, I use a value of the precision parameter ζ as high as 0.98 or higher, in this case.

Another useful application of the algorithm is based on the fact that it also provides an upper bound for the error of the numerical derivative. This makes it perfectly suited to check the implemented likelihood function and its derivatives.

2.2 Numerical Stability and Convergence

I also utilize the derivative-free Nelder-Mead minimization algorithm (Matlab routine fminsearch) for two purposes. Firstly, it has proven to be a robust method to generate a crude parameter vector relatively close to the optimum when I check convergence using different starting points. This is necessary because the gradient is not accurate if the covariance matrices H and Q are numerically unstable. The analytical gradient method thus sometimes stops far from the optimal parameter vector. Secondly, I use the Nelder-Mead algorithm as well as the numerical method described in the previous section to check convergence.

Unless discussed in the text, all results presented in this thesis are robust in the sense that they do not depend on the starting point. Computation time, however, can be high when the starting point is far from the optimum. This is far better if the optimization is started with a parameter vector based on a two-stage regression or a nested model. In these cases, the analytical gradient method always leads to quick and reliable convergence.

Appendix C

State Variables and Dividends in Continuous Time

Originally, this thesis was planned in a continuous-time framework. This makes modeling easier and also more flexible. However, empirical work turns out to be much more demanding in this case. The price-dividend ratios implied by these models are non-linear functions of the state variables and thus the state space system cannot be estimated using the Kalman filter. We need to apply simulation methods or the extended Kalman filter. Both techniques are much more involved than the present approach based on the Campbell-Shiller approximation. As the limits of estimation in terms of numerical stability and reliability have already been reached with the simpler approach, there is not much hope that a large continuous-time model can be estimated reliably using filtering techniques.

From this work, there is some remaining material that is worth including in this thesis. Most importantly, this is a tractable generalization of a theorem proofed in Ang and Liu (2007). It relates dividends and state variables in a very general manner. In fact, it is the exact generalization of the Campbell-Shiller approximation in continuous time.

3.1 Definition of the Model

Suppose the state of the economy is described by a K-dimensional state vector $\mathbf{x}$, which follows the diffusion process:

$$dx_k = \mu_k(\mathbf{x})dt + \sigma_k(\mathbf{x})dz_k. \tag{C.1}$$

μ_k and σ_k are scalar functions of the state variables. This is a very general definition of an economy. As prices should only depend on the state of the economy, it also means that the value of any cash-flow stream is fully determined by x. Mathematically, it means that bond yields, dividend yields and all other prices and valuation ratios can be written as functions of x.

A stock or a portfolio of stocks is an asset that pays the dividend stream D_t which follows the process:

$$d\ln D = \alpha_D(\mathbf{x})dt + \sigma_D(\mathbf{x})dz_D \tag{C.2}$$

or

$$\frac{D_t}{D_0} = \exp(\int_0^t \mu_D(\mathbf{x}_s)ds + \sigma_d(\mathbf{x}_s)dz_D). \tag{C.3}$$

Shocks to the state variables may be correlated. The correlation between shocks to the k'th and the l'th state variable is denoted by ρ_{kl}. Accordingly, ρ_{kD} is the correlation between dz_k and dz_D.

By definition, the instantaneous total return is

$$dR_t = \frac{dP_t + D_t dt}{P_t}. \tag{C.4}$$

Taking expectation one finds the following expression for the expected return μ_R:

$$\mu_R dt = \mathrm{E}_t\left[\frac{dP_t}{P_t}\right] + \frac{D_t}{P_t}dt. \tag{C.5}$$

By iterating (C.5), the price becomes

$$P_t = \mathrm{E}_t\left[\int_t^T \exp(-\int_t^s \mu_R du)D_s ds + \exp(-\int_t^T \mu_R du)P_T\right]. \tag{C.6}$$

Imposing the transversality condition

$$\lim_{T\to\infty} \mathrm{E}_t\left[\exp(-\int_t^T \mu_R du)P_T\right] = 0 \tag{C.7}$$

rules out bubbles in which stock prices grow so rapidly that people will buy just to resell at higher prices later, even if there are no dividends paid (see Cochrane (2001), chapter 1.4). Imposing this condition is not restrictive in an economic sense, since all feasible descriptions of the economy satisfy this condition.

The transversality condition simplifies the price formula (C.6), which becomes:

$$P_t = \mathrm{E}_t \left[\int_t^\infty \exp(-\int_t^s \mu_R du) D_s ds \right]. \tag{C.8}$$

With these definitions the following theorem can be stated:

Theorem C.1 *Suppose the state of the economy is described by* $\mathbf{x}$*, which follows (C.1), and the dividend process is given by (C.2). Assume that the price-dividend ratio* $\frac{P}{D}$ *is a function* $f = f(\mathbf{x})$*, and that the no bubble condition (C.7) holds. Then the instantaneous total return satisfies the following equation:*

$$\begin{aligned} dR_t &= \frac{1 + \sum_k f_k \mu_k + \frac{1}{2}\sum_{kl} f_{kl}\rho_{kl}\sigma_k\sigma_l + \sum_k f_k \rho_{kd}\sigma_k\sigma_D}{f} dt \\ &\quad + \mu_D\, dt + \frac{1}{2}\sigma_D^2\, dt + \frac{\sum_k f_k \sigma_k dz_k}{f} + \sigma_D dz_D. \end{aligned} \tag{C.9}$$

Conversely, if the total return follows

$$dR_t = \mu_R(\mathbf{x}_t)dt + \sum_k \sigma_{Rk}(\mathbf{x}_t)dz_k + \sigma_{RD}(\mathbf{x}_t)dz_D \tag{C.10}$$

then f *satisfies:*

$$\sum_k f_k \mu_k + \frac{1}{2}\sum_{k,l} f_{kl}\rho_{kl}\sigma_k\sigma_l + \sum_k f_k \rho_{kD}\sigma_k\sigma_D + f(\mu_D + \frac{1}{2}\sigma_D^2 - \mu_R) = -1 \tag{C.11}$$

together with

$$\sigma_{Rk} = \frac{f_k}{f}\sigma_k \tag{C.12}$$

$$\sigma_{RD} = \sigma_D. \tag{C.13}$$

Equation (C.12) holds for $k = 1, ..., K$*.*

Theorem C.1 says that if the dynamics of the dividend process and the price-dividend ratio are given as functions of the state variables, then the return process is unequivocally determined by equation (C.9). Conversely, if the return process is given together with the price-dividend ratio, then the dividend process results as a function of these two from (C.11, C.12, and C.13).

More generally, theorem C.1 says that there are two restrictions among the following five variables: the expected dividend growth rate, the dividend volatility, the expected stock return, the stock return volatility, and the price-dividend ratio. The first is induced by the definition of the return in (C.6) together with the transversality condition (C.7) through equation (C.10). The second stems from the definition of f, the price-dividend ratio.

Proof. Under the transversality condition (C.7), the conditions to apply Ito's lemma are satisfied. Applying the product rule to $P = f \cdot D$ gives $dP = df \cdot D + dD \cdot f + df \cdot dD$. A second application of Ito's lemma, this time to the definition of $f = f(x)$ yields:

$$\begin{aligned} df &= \sum_k f_k \, dx_k + \frac{1}{2} \sum_{k,l} f_{kl} dx_k \, dx_l \\ &= \sum_k f_k \mu_k \, dt + \frac{1}{2} \sum_{kl} f_{kl} \rho_{kl} \sigma_k \sigma_l \, dt + \sum_k f_k \sigma_k \, dz_k. \end{aligned}$$

To deduce (C.11), first plug in dP into the definition of the return

$$dR = \frac{dP + Ddt}{P} = \frac{df}{f} + \frac{dD}{D} + \frac{1}{f} dt + \frac{df \cdot dD}{f \cdot D}$$

and then expand this expression using df:

$$\begin{aligned} dR &= \frac{1 + \sum_k f_k \mu_k + \frac{1}{2} \sum_{kl} f_{kl} \rho_{kl} \sigma_k \sigma_l + \sum_k f_k \rho_{kD} \sigma_k \sigma_D}{f} \, dt \\ &\quad + \mu_D \, dt + \frac{1}{2} \sigma_D^2 \, dt + \frac{\sum_k f_k \sigma_k dB^k}{f} + \sigma_D \, dz_D. \end{aligned}$$

Equating coefficients completes the proof. ■

Bibliography

Ang, A., and J. Liu (2004): "How to Discount Cashflows with Time-Varying Expected Returns," *Journal of Finance*, 59(6), 2745–2784.

——— (2007): "Risk, Return and Dividends," *Journal of Financial Economics*, 85(1), 1–38.

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.

Bansal, R., and A. Yaron (2004): "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance*, 59, 1481–1509.

Beckaert, G., and S. Grenadier (2001): "Stock Bond Pricing in an Affine Economy," Mimeo.

Bernhard, J. (2008): "Long-Run Risks in International Stock Markets," Ph.D. thesis, University of St. Gallen.

Breeden, D. T. (1979): "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, 7, 265–296.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger (1989): "Empirical Tests of the Consumption-Oriented CAPM," *Journal of Finance*, 44, 231–262.

Brennan, M. J., A. W. Wang, and Y. Xia (2004): "Estimation and Test of a Simple Model of Intertemporal Capital Asset Pricing," *Journal of Finance*, 59, 1743–1776.

Brennan, M. J., and Y. Xia (2002): "Dynamic Asset Allocation under Inflation," *Journal of Finance*, 57(3), 1201–1236.

——— (2006): "Risk and Valuation Under an Intertemporal Capital Asset Pricing Model," *Journal of Business*, 79, 1–36.

Campbell, J., A. Sunderam, and L. Viceira (2008): "Inflation Bets or Deflation Hedges? The Changing Risks of Nominal Bonds," working paper, Harvard University.

Campbell, J. Y. (1991): "A Variance Decomposition for Stockreturns," *The Economic Journal*, 101, 157–179.

——— (1993): "Intertemporal Asset Pricing Without Consumption Data," *The American Economic Review*, 83, 487–510.

Campbell, J. Y., and R. R. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1, 195–227.

Campbell, J. Y., and T. Vuolteenaho (2004): "Bad Beta, Good Beta," *American Economic Review*, 94, 1249–1275.

Cochrane, J. H. (1991): "Explaining the Variance of Price-Dividend Ratios," *Review of Financial Studies*, 5, 243–280.

——— (2001): *Asset Pricing*. Princeton University Press.

——— (2008a): "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21, 1533–1575.

——— (2008b): "State-Space vs. VAR Models for Stock Returns," Graduate School of Business, University of Chicago.

Cochrane, J. H., and M. Piazzesi (2005a): "Bond Risk Premia," *The American Economic Review*, 95, 138–160.

——— (2005b): "Web Appendix to Bond Risk Premia," faculty.chicagogsb.edu/monika.piazzesi/research/cpapp.pdf.

——— (2008): "Decomposing the Yield Curve," Working Paper, Graduate School of Business, University of Chicago.

Collin-Dufresne, P., R. S. Goldstein, and C. S. Jones (2008): "Identification of Maximal Affine Term Structure Models," *Journal of Finance*, 63(2), 743–795.

D'Addona, S., and A. H. Kind (2006): "International Stock-Bond Correlations in a Simple Affine Asset Pricing Model," Journal of Banking and Finance, forthcoming.

Dai, Q., and K. J. Singleton (2000): "Specification Analysis of Affine Term Structure Models," *The Journal of Finance*, 50(5), 1943–1978.

Diebold, F. X., and C. Li (2006): "Forecasting the term structure of government bond yields," *Journal of Econometrics*, 130, 337–364.

Duffee, G. R. (2002): "Term Premia and Interest Rate Forecasts in Affine Models," *The Journal of Finance*, 57(1), 405–443.

——— (2008a): "Forecasting with the Term Structure: The Role of No-Arbitrage," working paper, Haas School of Business, University of California at Berkeley.

——— (2008b): "Information in (and not in) the Term Structure," working paper, Haas School of Business, University of California at Berkeley.

Duffee, G. R., and R. H. Stanton (2004): "Estimation of Dynamic Term Structure Models," working paper, Haas School of Business, University of California at Berkeley.

Duffie, D. J., and R. Kan (1996): "A Yield Factor Model of the Term Structure of Interest Rates," *Mathematical Finance*, 6, 379–406.

Durbin, J., and S. J. Koopman (2001): *Time Series Analysis by State Space Methods*. Oxford Statistical Science Series.

Dybvig, P. H., J. E. Ingersoll, and S. A. Ross (1996): "Long Forward and Zero-Coupon Rates Can Never Fall," *Journal of Business*, 69, 1–25.

Fama, E., and R. R. Bliss (1987): "The Information in Long-Maturity Forward Rates," *American Economic Review*, 77(4), 680–692.

Fama, E. F., and K. French (1992): "The Cross-Section of Expected Returns," *The Journal of Finance*, 47(2), 427–465.

Fama, E. F., and K. R. French (1988): "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 25, 3–24.

Gibbons, M., S. A. Ross, and J. Shanken (1989): "A test on the Efficiency of a Given Portfolio," *Econometrica*, 57, 1121–1152.

Greene, W. H. (1997): *Econometric Analysis*. Prentice-Hall, 3. edn.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2006): "The U.S. Treasury Yield Curve: 1961 to Present," Discussion paper, Federal Reserve System.

Hahn, J., and H. Lee (2006): "Yield Spreads as Alternative Risk Factors for Size and Book-to-Market," *Journal of financial and quantitative Analysis*, 41(2).

Harvey, A. C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hodrick, R. J. (1992): "Dividend Yields and Expected Stock Returns, Alternative Procedures for Inference and Measurement," *Review of Financial Studies*, 5, 357–386.

Joslin, S., M. Priebsch, and K. J. Singleton (2009): "Risk Premium Accounting in Macro-Dynamic Term Structure Models," Working Paper, Stanford University.

Joslin, S., K. J. Singleton, and H. Zhu (2010): "A New Perspective on Gaussian DTSMs," Working Paper, Stanford University.

Kalman, R. E. (1960): "A New Approach to Linear Filtering and Prediction Problems," *AMSE Journal of Basic Engineering*, 82, 35–45.

Keel, A., and H. Müller (1995): "Efficient Portfolios in the Asset Liability Context," *Astin Bulletin*, 25, 33–48.

Koopman, S. J., and N. Shepard (1992): "Exact Score for Time Series Models in State Space Form," *Biometika*, 79, 823–826.

Lamont, O. (1998): "Earnings and Expected Returns," *The Journal of Finance*, 53(5), 1563–1587.

Lemke, W., and T. Werner (2009): "The Term Structure of Equity Premia in an Affine Arbitrage-Free Model of Bond and Stock Market Dynamics," European Central Bank, Working Paper Series.

Mamaysky, H. (2002): "On the Joint Pricing of Stocks and Bonds: Theory and Evidence," Yale ICF, Working Paper.

McCulloch, J. H. (2000): "Long Forward and Zero-Coupon Rates Indeed Can Never Fall, but Are Indeterminate: A Comment on Dybvig, Ingersoll and Ross," Discussion paper, Ohio State University.

McCulloch, J. H., and H.-C. Kwon (1993): "U.S. Term Structure Data, 1947-1991," Discussion Paper 93-6, Ohio State University.

Menzly, L., T. Santos, and P. Veronesi (2004): "Understanding Predictability," *Journal of Political Economy*, 112, 1–47.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867–887.

——— (1995): *Continuous-Time Finance*. Blackwell.

Petkova, R. (2006): "Do The Fama-French Factors Proxy for Innovations in Predictive Variables?," *The Journal of Finance*, 61(2), 581–612.

Ribeiro, R. M. (2004): "Predictable Dividends and Returns: Identifying the Effect of Future Dividends on Stock Prices," Discussion paper, University of Pennsylvania.

Rosenbaum, T., and A. Zetlin-Jones (2006): "The Kalman Filter and the EM Algorithm," Working Paper: University of Minnesota, available at: www.econ.umn.edu/ zetli001/files/kalman.pdf.

Ross, S. A. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

Rytchkov, O. (2007): "Filtering out Expected Dividends and Expected Returns," Working paper: Sloan School of Management.

Shiller, R. J. (2000): *Irrational Exuberance*. Princeton University Press.

Söderlind, P. (2006): "C-CAPM Refinements and the Cross-Section of Returns," *Financial Markets and Portfolio Management*, 20, 49–73.

Svensson, L. E. O. (1994): "Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994," International Money Fund, IMF Working Paper, 1994/114.

van Binsenbergen, J. H., and R. S. Koijen (2008): "Predictive Regression: A Present-Value Approach," Working paper: Stanford GSB and Chicago GSB.

Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, 5, 177–188.

Wäger, L. (2009): "Analysis of Bond Risk Premia: Extensions to Macro-Finance and Multi-Currency," Vorstudie zum Doktorat, University of St. Gallen.

Curriculum Vitae

Education

2004 - 2011	Doctoral Student, University of St. Gallen
1998 - 2001	Study in Economics and Finance, University of St. Gallen
1995 - 1998	Study in Mathematics, ETH Zürich
1993 - 1995	Study in Mathematics, EPF Lausanne

Experience

Since 2011	Senior Investment Strategist, Swiss National Bank, Zürich
2010	Head Quant Research, Zurich Cantonal Bank, Zürich
2004 - 2009	Managing Partner and Head Investment Management, Kraus Partner Investment Solutions AG, Zürich
2002 - 2004	Research Assistant, Institute for Operation Research, University of St. Gallen
2000 - 2001	Project Assistant, Zurich Financial Services, Zürich