Model uncertainty and expected return proxies

Christoph Jäckel

Technical University Munich

5 December 2013

Model uncertainty and expected return proxies

Christoph Jäckel*

First draft: November 18, 2013
This draft: December 5, 2013

Abstract

Over the last two decades, alternative expected return proxies have been proposed with substantially lower variation than realized returns. This helped to reduce parameter uncertainty and to identify many seemingly robust relations between expected returns and variables of interest, which would have gone unnoticed with the use of realized returns. In this study, I argue that these findings could be spurious due to the ignorance of model uncertainty: because a researcher does not know which of the many proposed proxies is measured with the least error, any inference conditional on only one proxy can lead to overconfident decisions. As a solution, I introduce a Bayesian model averaging (BMA) framework to directly incorporate model uncertainty into the statistical analysis. I employ this approach to three examples from the implied cost of capital (ICC) literature and show that the incorporation of model uncertainty can severely widen the coverage regions, thereby leveling the playing field between realized returns and alternative expected return proxies.

Keywords: Time-varying expected returns, implied cost of capital, asset pricing, model averaging, model selection
JEL Classifications: G12, C11

1 Introduction

Realized returns are an unbiased, but noisy, estimator of expected returns. The noise, driven by changes in expectations about future cash flows and discount rates, is typically assumed to be an order of magnitude larger than the variation in expected returns. These information surprises make it notoriously hard to

*Department of Financial Management and Capital Markets, Technische Universität München, Arcisstr. 21, 80333 Munich, Germany, Phone: +49 89 289-25487, christoph.jaeckel@tum.de.

uncover a relation between a variable of interest and latent expected returns, if realized returns are used as proxy.

In his presidential address, Elton (1999) highlighted these points and challenged the profession to propose alternative ways to estimate expected returns. This request has been followed by a large number of studies that propose alternative proxies. They are forward looking and therefore, at least in theory, unaffected by any news. They rely on earnings forecasts (see, e.g., Claus and Thomas 2001), CDS spreads (see Friewald et al. 2013), and corporate bond yields (see Campello et al. 2008). Due to their substantially lower standard deviation in comparison to realized returns, they allow researchers much sharper statistical inferences and provide them with very robust results. In other words, parameter uncertainty is greatly reduced, which has helped to identify relations between expected returns and factors that are overshadowed by noise when realized returns are used.[1]

In this paper, I argue that the seemingly robust results of these alternative expected return estimates could be driven, at least partly, by the ignorance of model uncertainty. As a solution, I introduce a Bayesian model averaging (BMA) approach that directly incorporates model uncertainty into the statistical inference. This levels the playing field between realized returns, with large parameter uncertainty and no model uncertainty, and alternative proxies, with typically modest parameter uncertainty, but potentially large model uncertainty. I apply BMA to three research questions from the implied cost of capital (ICC) literature and show that the incorporation of model uncertainty can lead to sampling uncertainties in estimated parameters that are no better than those based on realized returns.

There are two channels through which model uncertainty of expected return proxies is introduced. First, every proxy is based on an underlying theoretical framework such as the Merton-model that links debt and equity returns or the dividend discount model that links the current stock prices with future expected dividends. Proxies based on any such model inherit the model's shortcomings. Second, and more importantly, to get empirically traceable versions of the theoretical models additional simplifying assumptions have to be made. These problems are best seen in the case of the most successful proxy, the implied cost of capital, which has seen widespread use in both asset pricing and corporate finance. Theoretically, the ICC is just the IRR of the classical dividend

1. For example, in the time-series Pástor, Sinha, et al. (2008) are able to find a positive risk-return tradeoff with the help of the ICC. In simulations, Lundblad (2007) shows that it takes a very long sample to identify this relation if realized returns are used. In the cross-section, Hail and Leuz (2009) are able to quantify the reduction in the cost of capital triggered by a U.S. cross-listing to lie between 70 and 120 basis points, while studies using realized returns report unreasonable estimates of up to 1,000 basis points.

discount model. However, expected dividends are as unobservable as expected returns. As such, proxies, such as analyst earnings forecasts, have to be used and terminal value assumptions have to be made about future periods for which those proxies are not available anymore. This has led to the availability of a few dozens alternative ICC implementations, with minor and major modifications. All alternatives want to measure expected returns, yet all alternatives differ, which directly implies that, at the most, one measures expected returns correctly. Therefore, model uncertainty is caused by measurement error. However, there is ample evidence – both theoretical and empirical – that this measurement error is persistent and systematic.[2] Consequently, this measurement error can systematically bias the results in those cases where it is correlated with the variable of interest. And even if one is willing to assume that measurement error is white noise, there is an increasing chance of finding significant results due to an increasing number of alternative proxies, all with a different measurement error process. This increases the danger of data snooping.

The current approach to deal with this problem is to run robustness tests in an ad-hoc manner. To the best of my knowledge, there is no study that takes the evidence from different proxy classes, i.e. proxies derived from different underlying theoretical models, into account. Additionally, most studies that only focus on one proxy class, in almost all cases the ICC, only vary minor specifications of a specific ICC method. Of course, this approach can be justified by the large number of possible specifications that make it hard for the researcher to check for and for the reader to comprehend. Yet, it is well known from the model selection literature that results conditional on only a few models lead to the understatement of predictive and inferential uncertainty about a parameter of interest, "leading to inaccurate scientific summaries and overconfident decisions that do not incorporate sufficient hedging against uncertainty" (Draper 1995, p. 45).

However, a solution to this problem is available: Bayesian model averaging, in which an econometrician averages across the evidence conditional on each model. Hereby, the weight of each model is determined by the prior beliefs of the econometrician about the model probability and the parameters, on the one hand, and the evidence in the data, on the other hand. Herein lies the key difference of my implementation of BMA to other studies. In those studies, a model

2. In the case of the ICC, Pástor, Sinha, et al. (2008) are able to establish a perfect correlation between the ICC and expected returns only under certain simplifying assumptions in which both dividend growth and expected returns follow AR(1) processes. In particular, it is assumed that expected cash flows are measured without error, although there is ample empirical evidence to the contrary. In particular, the most prominent cash flow proxy, analyst forecasts, are systemically biased (for an overview of the literature on analyst forecasts, see Ramnath et al. 2008). Furthermore, Hughes et al. (2009) show that the difference between the ICC and true expected returns is driven by volatilities in, as well as correlations between, expected returns and cash flows, growth in cash flows, and leverage.

gets the higher posterior model weight, the better it is able to explain the data at hand, which is a faulty approach in the case of measurement error. Suppose we are interested in the relation between expected returns and any variable of interest x, but only have a set of different proxies for expected returns available. Suppose further that there is no relation between expected returns and x, but between the measurement error of one of the proxies and x. Obviously, this proxy will have the strongest relation with x asymptotically and hence, given the data, it would get the highest weight.

We, therefore, need an *external* validation of a proxy's quality. Previous research, such as Easton and Monahan (2005) and Lee, So, et al. (2011), argue that, since realized returns over period $t+1$ are an unbiased estimator of returns for that period, expected at time t, any alternative proxy eventually has to explain subsequent realized returns. However, they only use such predictive regressions to identify the best proxy. In contrast, I apply a BMA setup to obtain posterior model weights from predictive regressions. I thereby closely follow Avramov (2002), Cremers (2002), and Binsbergen et al. (2013), who also combine BMA and predictive regressions, but with different goals. Avramov (2002) and Cremers (2002) want to improve the predictive accuracy by combining signals from many predictors. Binsbergen et al. (2013) evaluate the quality of their equity yields in comparison to other predictors. So their approach is similar to mine in that they differentiate between different predictors (in their case) or proxies (in my case). However, they do not use these model weights in a follow-up step. In contrast, the computation of model weights is only an intermediate step in my approach. In a final step, I use the model weights to average across the parameters of interest over all proxies under consideration. If these parameters differ between proxies, this results in larger coverage regions that take model uncertainty into consideration.

Of course, predictive regressions rely on realized returns, the proxy that alternatives are meant to replace. The model averaging approach illustrates the circularity here: if we want to evaluate the quality of any proxy over another, we have to rely on tests based on realized returns and those tests are subject to the same points of criticism brought forward against realized returns as a proxy of expected returns. In particular, I show for an U.S. sample ranging from 1985 to 2011 that in the case of the ICC we cannot reliably identify a superior specification. Furthermore, the model weights may be biased because the information surprises driving the realized returns could be correlated in sample and bias the results. However, if we decide to dismiss the evidence in the predictive regressions, this directly implies that there is only one way we can differentiate between the multitude of available proxies: we have to argue on prior grounds why one proxy is superior to the others. I am not aware of

any study within the ICC literature that has provided conclusive arguments to favor one specification over another. The same is true for comparisons across proxy classes.

The approach also highlights another problem of alternative expected return proxies. If the true underlying expected return process is not within the set of proxies under examination, results will be biased, even asymptotically. Hence, a model averaging approach can only reduce model uncertainty issues, not solve them.

In summary, the contribution of this paper is twofold: First, it provides a general framework to incorporate model uncertainty into the inference based on expected return proxies. This framework is also very instructive in pointing out the underlying weaknesses of alternative expected return measures other than realized returns: in particular, the inability for a researcher to establish their unbiasedness and the dependency on realized returns to evaluate them. Second, I apply the model averaging framework to three examples based on the time-series of the ICC and show that ignorance of model uncertainty can lead to overconfident decisions. In particular, I find that the recent results of Chen et al. (2013), who use the ICC to answer the question of whether stock price movements are driven by cash flow or discount rate news, are not robust when evidence is based on alternative, reasonable specifications of the ICC.

The rest of the paper is organized as follows. Section 2 shows the problems that are caused by the uncertainty inherent in the proxy selection process and introduces BMA as a possible solution to this problem. Section 3 describes the data set that is used in Section 4 to apply BMA to three examples from the ICC literature. Section 5 concludes.

2 Theoretical considerations

In the following, I show in a simple setup how the ad-hoc selection of one proxy from a set of proxies that are potentially measured with error can systematically bias and overstate the significance of the results. I show how the bias problem can be improved on by applying an external validation to the proxies under consideration and how the overstatement problem can be improved on by incorporating the uncertainty in choosing the right proxy with a model averaging approach.

2.1 Theoretical framework

Suppose we are interested if the time-series of variable x_t is related to expected returns $E_t[r_{t+1}] \equiv \mu_t$. Concretely, suppose the following linear relation:[3]

$$\mu_t = \gamma x_t + \epsilon_t, \qquad t = 1, \ldots, T. \tag{1}$$

Using a univariate, classical regression setting with normally distributed mean zero errors ϵ_t, we can regress μ_t on x_t and test if the estimator $\widehat{\gamma}$ is significantly different from zero. However, since μ_t is unobservable, the researcher has to identify observable proxies, one of which is realized returns r_{t+1}. Realized returns over a period $t+1$ are the sum of expected returns at time t and unexpected returns u_{t+1} that have a mean of zero conditional on all of the information available at time t:[4]

$$r_{t+1} = \mu_t + u_{t+1}, \qquad t = 1, \ldots, T. \tag{2}$$

Given equation (2), it is easily shown that the unconditional mean of r_{t+1} is equal to the unconditional mean of μ_t. In other words, realized returns are an unbiased estimator of expected returns. Despite this advantageous characteristic, realized returns have come under criticism due to the fact that the variation in u_{t+1} is an order of magnitude larger than the variation in μ_t. This makes statistical inference notoriously hard, as highlighted by Elton (1999). Plugging equation (2) into (1), the sampling variance of the estimator based on realized returns, $\widehat{\gamma}_{rr}$, is given by:

$$Var(\widehat{\gamma}_{rr}) = \frac{Var(\epsilon_t) + Var(u_{t+1})}{\sum_{t=1}^{T}(x_t - E[x])^2}. \tag{3}$$

In small samples, the large variance of u_{t+1} results in a large sampling variance for $\widehat{\gamma}_{rr}$ that can hinder the detection of an existing relation between μ_t and x_t. This insight has led to a "proxy variable search" with the hope of identifying alternative proxies for expected returns that are not plagued with the large noise inherent in realized returns.

However, these proxies are subject to measurement error. In line with previous research (see, for example, Easton and Monahan (2005) and Lee, So, et al. (2011)), I assume that a proxy $\widehat{\mu}_{t,k}$, measured at time t, is tracking μ_t with an additional, additive proxy-specific error term $w_{t,k}$:

$$\widehat{\mu}_{t,k} = \mu_t + w_{t,k}, \qquad t = 1, \ldots, T. \tag{4}$$

3. For simplicity, I assume the intercept to be zero.

4. For studies that explore this relation between expected and realized returns, see for instance Pástor and Stambaugh (2009) or Sadka and Sadka (2009).

If we run regression (1) with a proxy as defined in equation (4), it can be shown that the resulting regression coefficient $\widehat{\gamma}_k$ is equal to the sample estimate that we would obtain if μ_t were observable, $\widehat{\gamma}$, and an additional bias term:

$$\widehat{\gamma}_k = \widehat{\gamma} + \frac{Cov(w_{t,k}, x_t)}{Var(x_t)} = \widehat{\gamma} + Bias_k. \tag{5}$$

Obviously, the hope of a researcher who applies proxy k is that the mean and variance of $w_{t,t}$ are close to zero. In this case, the researcher is able to detect a relation between x_t and μ_t much more precisely, compared to the analysis that employs realized returns. The danger, however, is quite obvious as well: if $w_{t,k}$ is systematically correlated with x_t, or if by chance the two comove in sample, then one might incorrectly deduce a relation between μ_t and x_t from the data, although this relation is solely due to the specific measurement error of the proxy under consideration.

Studies such as Easton and Monahan (2005) and Lee, So, et al. (2011) offer a first step to address this problem. They realize that we need an *external* validation of the quality of alternative proxies. If we can establish that some of the proxies are unsuitable to track expected returns, we can dismiss them from the beginning. Fortunately, such an external validation is available. Since realized returns over a period are an unbiased estimator of the expected return at the beginning of this period, every reasonable alternative proxy has to be able to predict them eventually. Consequently, predictive regressions such as in Li et al. (2013) provide evidence of the quality of expected return proxies. Therefore, these studies recommend selecting proxies that perform best in such predictive regressions.

However, it is well known from the return predictability literature that predicting subsequent realized returns is notoriously hard. Until now, there is still a heated debate if *any* predictor, not just expected return proxies, can actually predict realized returns.[5] In summary, we have a large number of proxies to choose from and we are unable to precisely determine which of these proxies is best. Therefore, we face large model uncertainty and the main contribution of this paper is to introduce model averaging techniques, that have been used with great success in different settings, to the expected return proxy literature. The next section shortly introduces model averaging and applies it to the problem at hand.

5. For recent surveys of the literature, see Koijen and Van Nieuwerburgh (2011) and Cochrane (2011).

2.2 Model averaging approach

Within a Bayesian framework, Leamer (1978) shows that the posterior distribution of a quantity of interest Δ can be computed, given data D, as the average of the posterior distributions under each model, weighted by their posterior model probability.[6] If we interpret each proxy as a separate model, we therefore get:

$$p(\Delta|D) = \sum_{i=1}^{k} p(M_i|D)p(\Delta|D, M_i), \tag{6}$$

where $M_1, \ldots, M_k$ are the models considered. Equation (6) implies that the marginal distribution of the parameter of interest is a mixture distribution. The mixture probabilities are the posterior model weights, $p(M_i|D)$, and the individual distributions are the distributions of the parameter of interest, conditional on a specific model.

In this simple framework, the only difference between two models M_i and M_j is that the proxy $\widehat{\mu}_i$ is replaced with $\widehat{\mu}_j$.[7] Additionally, D is split into two parts here: the part D_{RQ} is the part of the data needed to answer the research question at hand. In the simple setup introduced in the previous section, D_{RQ} consists of the matrix of expected return proxies and x_t. D_P is the part of the data needed to compute the posterior probability of each proxy measuring true expected returns, given that one of the proxies is indeed correct. As argued before, this data consists of the set of proxies under consideration and subsequent realized returns. The separation of the data set is a direct consequence of the previous discussion: if we evaluate the quality of a proxy in terms of how well it explains the research question at hand, we run the risk of finding spurious relations driven by measurement error, and not by true expected returns. This is the main differentiation between my approach and other studies that apply BMA: Since measurement error is not such an obvious problem in other studies, a model is considered to be superior if it is able to explain the research question at hand better.[8] In contrast, I use subsequent realized returns to infer the posterior

6. The most popular derivative of model averaging is the one based on Bayes' theorem, which is also the foundation in Leamer's work. This approach subsumes under the name of Bayesian model averaging. However, there are also frequentist alternatives based on information-theoretical criteria such as AIC or BIC (see Claeskens and Hjort 2008). As I show later on, due to the simplicity of my setup both approaches yield identical results. Therefore, the debate about differences between the two approaches is not an issue for the purpose of this paper. Still, in line with most studies that focus on model averaging techniques, I use a Bayesian setting to motivate my approach.

7. This is important for the reader to keep in mind because I use the words proxies and models somewhat interchangeably. The latter is often used to conform with the language of the model selection and averaging literature.

8. For example, Fernandez et al. (2001) and Sala-I-Martin et al. (2004) employ BMA to test the robustness of explanatory variables in cross-country economic growth regressions. Since the literature has come up with a multitude of possible regressors, the question arises which combinations of those regressors help in explaining cross-country economic growth and

model weights $p(M_i|D_P)$, which is a measure of the quality of the proxy.[9] This computation is independent of the specific research question. Results for each proxy are then obtained for the research question at hand and averaged across the proxies based on the model weights.

To emphasize the separation between model computation and subsequent statistical inference, equation (6) can be rewritten as:

$$p(\Delta|D_{RQ}, D_P) = \sum_{i=1}^{k} p(M_i|D_P)p(\Delta|D_{RQ}, M_i). \tag{7}$$

In principle, the posterior distribution of the parameter of interest, $p(\Delta|D_{RQ}, M_i)$, can be derived from a Bayesian perspective. That is, one would specify the prior for this parameter, $p(\Delta|M_i)$, and the likelihood conditional on the parameter, $p(D_{RQ}|\Delta, M_i)$. The posterior is proportional to the product of likelihood and prior. While this seems to be a daunting task at first glance, it is greatly simplified by the fact that each model has the same interpretation in the case of expected return proxies. Since each proxy wants to measure the same thing without error, all parameters should have the same interpretation. In this paper, I choose a simpler approach that is more in line with current practice that mostly applies frequentist approaches: For the posterior distribution of the parameter of interest I use the sampling distribution from the frequentist approach. For example, if we run a time-series regression as specified in (1) and are interested in the slope coefficient, $p(\Delta|D_{RQ}, M_i)$ is just the sampling distribution of the slope coefficient from a regression of a specific proxy on x_t. These distributions are easily adjusted to incorporate heteroskedastic or autocorrelated error structures, as will be shown in the empirical examples later on.

Coming back to the research question from the previous section, the first two moments of $\widehat{\gamma}$ can then be calculated as:[10]

$$E[\widehat{\gamma}_{BMA}|D_{RQ}, D_P] = \sum_{i=1}^{k} p(M_i|D_P)\widehat{\gamma}_i, \tag{8}$$

how to take issues of model uncertainty into consideration. In this application problems of measurement error are ignored. As a consequence, the better a model explains the dependent variable, the higher its posterior model probability will be.

9. The computation of the weights is discussed further below.

10. Leamer (1978) provides a derivation of these results.

and

$$Var(\widehat{\gamma}_{BMA}|D_{RQ}, D_P) = \sum_{i=1}^{k} p(M_i|D_P) Var(\widehat{\gamma}_i|D_{RQ}, M_i) + \\ + \sum_{i=1}^{k} p(M_i|D_P)(\widehat{\gamma}_i - E[\widehat{\gamma}_{BMA}|D_{RQ}, D_P])^2. \quad (9)$$

While the mean estimate across all models is simply a weighted average across the estimate of each model, the variance of the combined estimate $\widehat{\gamma}_{BMA}$ exceeds a weighted average of the variances of the estimates *within each model* by an amount that depends on the variability of the estimates *across models*. Consider a case in which $Var(\widehat{\gamma}_i|D_{RQ}, M_i)$ is quite small for all models, i.e. conditional on a certain model the regression coefficient is measured accurately, but across models, the coefficients vary widely. In such a case, one would severely underestimate the variability of the parameter of interest if one only focuses on one model.

In the case of the proxy literature, it is an apparent advantage (see, for instance, Lee, Ng, et al. (2009)) that the statistical inference is much more robust due to much lower standard errors. This statement, however, is often based on evidence for one proxy. Equation (9) shows that the variance could indeed be much larger if the coefficients between different proxies differ substantially.

By plugging equation (5) into equation (8) and (9) and rearranging, we can express $E[\widehat{\gamma}_{BMA}|D_{RQ}, D_P]$ and $Var(\widehat{\gamma}_{BMA}|D_{RQ}, D_P)$, respectively, as follows:

$$E[\widehat{\gamma}_{BMA}|D_{RQ}, D_P] = \widehat{\gamma} + Bias_{BMA}, \quad (10)$$

and

$$Var(\widehat{\gamma}_{BMA}|D_{RQ}, D_P) = \sum_{i=1}^{k} p(M_i|D_P) Var(\widehat{\gamma} + Bias_i) \\ + \sum_{i=1}^{k} p(M_i|D_P)(Bias_i - Bias_{BMA})^2, \quad (11)$$

where $Bias_{BMA} = \sum_{i=1}^{k} p(M_i|D_P) Bias_i$. Equation (10) and (11) are instructive representations of the discussion above. First, if one of the proxies is measured without error, we want the posterior model probability of this proxy to approach unity. In this case, $E[\widehat{\gamma}_{BMA}|D_{RQ}, D_P]$ and $Var(\widehat{\gamma}_{BMA}|D_{RQ}, D_P)$ will converge to $\widehat{\gamma}$ and $Var(\widehat{\gamma})$, respectively. Second, if the bias over all models varies randomly around zero and all proxies get equal support in the data, the average estimate across the models, $\widehat{\gamma}_{BMA}$, will be unbiased, but there is considerable

model uncertainty that is automatically incorporated into the BMA analysis. In contrast, if an econometrician only examines a subset of the proxies, one might end up with biased estimates. Third, if all proxies under consideration are systematically biased, BMA will fail.[11] Finally, all approaches that base their results on only one proxy, whether this proxy is chosen ad-hoc or by its ability to predict subsequent realized returns, ignore the variability of the parameters from different models. This leads to overoptimistic decisions and can result in the false identification of seemingly robust relations.[12]

2.3 Computation of posterior model weights

I follow Avramov (2002), Cremers (2002), and Binsbergen et al. (2013), which are all studies that run predictive regressions in a BMA framework to compute posterior model weights. Consider a set of k linear univariate models $M_1, \ldots, M_k$. Let the ith model be given by:

$$r_{t+1} = \beta_0 + \beta_1 \widehat{\mu}_{t,i} + \varepsilon_{t+1}, \qquad t = 1, \ldots, T, \tag{12}$$

where ε_{t+1} is assumed to be identically, independently, and normally distributed with mean zero and unknown variance σ^2. In general, the posterior model probability for model i is computed, given data D_P, via Bayes' theorem as:

$$p(M_i|D_P) = \frac{p(D_P|M_i)p(M_i)}{\sum_k p(D_P|M_k)p(M_k)}, \tag{13}$$

where

$$p(D_P|M_i) = \int\int p(D_P|\beta_i, \sigma^2, M_i)p(\beta_i|\sigma^2, M_i)p(\sigma^2)d\beta_i d\sigma^2. \tag{14}$$

Therefore, we have to specify two priors. First, a prior about the probability of each model, $p(M_i)$. Second, priors about the parameters $\beta = (\beta_0, \beta_1)$ and σ^2. Both cases can be tricky if the number of explanatory variables differs

11. Therefore, it is a commonly made assumption in the BMA literature that the true model is part of the set of models considered.

12. In the internet appendix of this paper, I simulate the relations between a variable of interest, latent expected returns, and expected return proxies for many different specifications in a simple setup. The simulation results confirm the statements here. In short samples, BMA can severely decrease the bias in estimates and increase the coverage, i.e. result in confidence regions that include the true underlying parameter. Alternative approaches, such as using the proxy that shows the strongest relation with the variable of interest or averaging across the proxies before the analysis perform mostly worse, sometimes equally well, and almost never better. The cases in which they perform better are cases of random measurement error added to the proxies and in which there is an actual relation between the variable of interest and true expected returns. In these cases, the attenuation bias is compensated by the overestimation due to selecting the best proxy. Hence, in these cases the BMA approach is more conservative. The simulation results can be found at this link.

between models and if the parameters' interpretation changes from model to model.[13] In my case, however, this is not an issue because each model has only one explanatory variable and the interpretation in each model is the same. The default assumption about $p(M_i)$ is to give each model the same weight a priori, i.e.:

$$p(M_i) = \frac{1}{k}. \tag{15}$$

I use the same priors as Wright (2008) and Binsbergen et al. (2013). Concretely, I make the assumption that β takes the natural conjugate g-prior specification proposed by Zellner (1986). The prior on β conditional on the variance of the error term σ^2 is therefore given as $N(0, \phi\sigma^2(X_i'X_i)^{-1})$, where ϕ is a shrinkage parameter that controls the informativeness of the prior and X_i is the $T \times 2$ matrix of a T vector of ones and the T vector $\widehat{\mu}_i$. Since σ^2 is identical across models, we can use an improper prior of an inverse gamma (0,0) that is proportional to $1/\sigma$. Then, the posterior model weights can be computed from:[14]

$$p(M_i|D_P) \propto \left(r'r - \left(\frac{\phi}{1+\phi}\right) r'X_i(X_i'X_i)^{-1}X_i'r\right)^{-T/2}, \tag{16}$$

where $r \equiv (r_2, \ldots, r_{T+1})$ denotes the vector of subsequent realized returns. Finally, we just have to normalize equation (16) so that all model weights sum to one.

The parameter ϕ governs the informativeness of the researcher's prior information. The lower ϕ, the more weight is put on prior information. In the limit, if $\phi = 0$, $p(M_i|D_p)$ is equal for all models, i.e. the posterior probabilities are identical to the prior probabilities $p(M_i)$.

To provide a link with frequentist approaches and to get rid of the subjective aspects of the prior assumptions, we can increase ϕ to reduce the impact of the priors. In the limiting case, i.e. $\phi \to \infty$, the posterior model weights in equation (16) become proportional to $(SSE)^{-T/2}$. This result is also derived by Leamer (1978, p. 112) who is in search of a reasonable diffuse prior. Furthermore, it is easy to show that in this case the weights computed from equation (16) are identical to the weights that would be obtained from information-theoretical approaches that use AIC or BIC based on the following formula:[15]

$$p_{AIC}(i) = \frac{exp(0.5\Delta_{AIC,i})}{\sum_k exp(0.5\Delta_{AIC,k})}, \tag{17}$$

where $\Delta_{AIC,k} = AIC_k - maxAIC$. This subtraction is made merely for com-

13. For a discussion of these issues, see, for instance, Ley and Steel (2009).

14. I can ignore all terms here that are constant across models because they cancel out in equation (13).

15. Claeskens and Hjort (2008) give a good introduction into frequentist approaches of model selection and averaging and also motivates the formula given here.

putational reasons. In equation (17), we can replace AIC with BIC; since the model sizes are identical across models, the penalty term that normally differs between AIC and BIC does not matter.

To summarize, both a noninformative Bayesian approach as well as a frequentist approach yield identical results due to the simplicity of the setup (univariate linear regression for each model). Consequently, debates about which approach is superior are not relevant here and model weights, given the data, are easily computed. The better a proxy is able to explain subsequent realized returns, i.e. the lower the sum of squared errors is in the predictive regression, the more credible this proxy is in comparison to its competitors and the more weight a researcher should assign to it in the analysis of interest.

Due to the high level of noise inherent in realized returns, alternative proxies have been proposed. That is, the main motivation for these proxies is the replacement of realized returns. However, the previous analysis to infer the quality of these proxies has to rely again on the very same realized returns it wants to replace. In my opinion, this is a severe shortcoming of *any* alternative expected return proxy, a point that has mostly gone unnoticed in the literature.[16]

A main contribution of this study is that the introduction of model averaging techniques allows me to shed light on this issue. For the sake of simplicity, let's focus only on two alternative proxies with equal prior probability. In this case, the Bayes Factor and the likelihood ratio (again, they are identical due to the simplicity of the setup) can be interpreted as a summary of the evidence provided by the data in favor of one proxy, in comparison to another proxy.[17] In our case, the Bayes Factor is given by:

$$BF_{12} = \left(\frac{SSE_2}{SSE_1}\right)^{T/2} = \left(\frac{Var(w_{t,2}) + Var(u_{t+1})}{Var(w_{t,1}) + Var(u_{t+1})}\right)^{T/2}. \tag{18}$$

As argued above, the main motivation of any alternative proxy is the large variation in realized returns induced by u_{t+1}. Thus, equation (18) will be dominated by the term $Var(u_{t+1})$, which means that in small samples it will be notoriously hard to separate proxies with low measurement error from proxies with large measurement error. This means that the weights will not converge quickly to the best proxies in small samples. Only if sample size increases, even

16. I am only aware of Guay et al. (2011, p. 129), who give a qualitative assessment of this issue: "Like Easton and Monahan (2005) and a large literature in finance, we use realized returns as a metric to assess the cost of capital estimates and the effectiveness of our proposed remedies. Although our returns-based tests are consistent with a large asset-pricing literature, we acknowledge that realized returns are a noisy proxy for expected returns, and that this is, in fact, an important motivation behind implied cost of capital measures. However, despite the limitations, we are unaware of a superior benchmark to validate cost of capital measures that does not rely on realized returns."

17. For a detailed discussion of the Bayes Factor, see Kass and Raftery (1995).

SSE ratios close to one will eventually become large and reveal the superiority of one proxy over the other. Furthermore, it is often argued that u_{t+1} can be correlated with other variables in sample and therefore, inferences based on it can be misleading.[18] As a consequence, in small samples it might happen that inferior proxies get more weight.

3 Proxy selection and data

In general, the BMA approach can be applied to any empirical proxy of expected returns. However, the only proxy that has found widespread use in empirical applications so far is the ICC. Probably as a result of this success, it is also the only proxy with a multitude of alternative versions, which emphasizes the importance of incorporating model uncertainty into the empirical analysis. Hence, I focus on the ICC in the empirical part of my study.

The ICC is defined as the constant discount rate r_{ICC} that equates the current stock price of firm i, $P_{i,t}$, with the sum all future dividends, discounted to today:

$$P_{i,t} = \sum_{j=1}^{\infty} \frac{E_t[D_{i,t+j}]}{(1 + r_{i,ICC})^j}, \tag{19}$$

where $E_t[D_{i,t+j}]$ denotes the dividend in period $t + j$ that the investor expects at t. To get empirically traceable versions of equation (19), certain simplifying assumptions have to be made by the researcher, thereby introducing model uncertainty. Since market expectations about future dividends are just as unobservable as expected returns, proxies have to be used. The most common ones are analyst forecasts for *earnings* that I use here as well.[19] Since the dividend discount model of equation (19) can be transformed into a model that takes earnings as an input in several ways, this is a first differentiator between ICC methods: while derivatives of the residual income model rely on the clean-

18. Fama and French (2002), for instance, argue that the high realized returns in the U.S. stock market in the second part of the twentieth century were a result of a series of positive cash flow news that generated positive shocks u_{t+1} and not a result of high μ_t for this period. Additionally, Campello et al. (2008) also run the predictive regressions of their expected return proxy and do not find a large explanatory power of their proxy. Instead of attributing this result as a negative sign for the quality of their proxy, they blame the shock structure in the sample for this result and conclude that it is therefore worthwhile to explore alternative proxies. Yet another indication of the sensitivity of the specific sample is the evidence presented in Koijen and Van Nieuwerburgh (2011) that the regression coefficients in predictive regressions are instable over time.

19. Recently, Hou et al. (2012) proposed a pooled regression approach to obtain earnings forecasts from historical data. They claim that their forecasts are superior to analyst forecasts in terms of coverage, forecast bias, and earnings response coefficient. However, these proxies are only updated annually, while many studies require a higher updating frequency. Analyst forecasts are updated each month.

surplus relation to link dividends to earnings and book values, the abnormal growth in earnings model does not have to rely on book values.[20] Another distinguishing feature between ICC methods is their terminal value assumption that has to be made since earnings forecasts are only available for the next few years.

[Table 1 about here.]

In this paper, I focus on two derivatives of the abnormal growth in earnings model (OJ, MPEG), two derivatives of the residual income model (GLS, CT), and two direct implementations of equation (19) (PSS, CDZ) that have found widespread use in the literature. Table 1 presents the different methods and the specific assumptions made by them.

All methods rely on analyst forecasts, which are obtained monthly from IBES. IBES provides analyst forecasts for up to five years ahead, but mostly only the first two to three years are covered. Missing earnings forecast from period $t+3$ on are filled up by multiplying the forecast of the previous year and the long-term earnings growth rate provided by analysts: $eps_{i,t+j} = eps_{i,t+j-1} \times (1 + Ltg_t)$. IBES also provides the stock price, the shares outstanding, and a long-term growth rate Ltg. I match this data with data from Compustat that is publicly released. From Compustat, I obtain shareholder's equity (item SEQ) and common shares outstanding (item CSHO) to infer the book value per share.[21] Also, the payout ratios for the next three years are assumed to be constant and equal to the historical one, i.e. I divide common dividends (item DVC) by income before extraordinary item (IBCOM). For firms with a negative income, 6% of total assets is used as the denominator instead. With this data and further assumptions described in Table 1, I can solve numerically for the value of the ICC that sets the difference of the current price and discounted dividends to zero. I abort the root search as soon as the change in the ICC is less than 0.001% for one step. The ICC is computed at the IBES release date which is on the Thursday before the third Friday in each month. I match the ICC of this date with subsequent realized returns of the next calendar month.[22] The realized returns are the continuously compounded with-dividend monthly returns on the value-weighted portfolio of all NYSE, Amex, and NASDAQ stocks from the Center for Research in Security Prices (CRSP).[23]

20. Easton (2007) gives an introduction into the various ICC methods and their assumptions.

21. The reason I obtain shares outstanding both from Compustat and IBES is that the former is used to match it consistently to the historic shareholder's equity, while the latter is used to compute the monthly updated market capitalization. Calculating the market capitalization of a firm as the product of the IBES price and shares outstanding is common practice in the literature.

22. In untabulated results, I also compute monthly returns from the IBES release date to the day before the next IBES release date, which only has a marginal effect on the results.

23. Li et al. (2013) also entertain the ICC in predictive regressions. However, they use *excess*

I filter all observations that have at least one missing value of an ICC method, that have a negative book value and that have an IBES price below one dollar. With the remaining observations, I compute a monthly, value-weighted time-series for each ICC method from 1985 to 2011.

Note that the studies referred to in column 2 of Table 1 are the reference points for my procedures. This does not mean that I replicate their approaches exactly. For better comprehensibility, I reduce the assumptions made by the cited studies to a common denominator. For instance, empirical studies differ on how they fill up missing analyst forecasts. Some studies require at least three forecasts to be available, some studies use the long-term earnings growth rate to compute missing forecasts, and some studies compute this growth rate, if missing, from available earnings forecasts. Also, applied filters differ from study to study. Here, I apply the same rules of inferring missing data and filtering, as described above, consistently to all methods.

While the arbitrary decisions about issues discussed in the previous paragraph should have a negligible effect on the results, the column 4 in Table 1 shows assumptions that should have a far more pronounced impact. For instance, the MPEG and OJ method only rely on analyst forecasts for the next two periods, while the CT method takes the forecasts of all five years ahead into account. GLS, PSS, and CDZ use the first three forecasts. These methods also differ substantially in their terminal value assumptions: GLS linearly interpolates the three-year ahead ROE to the historical median industry ROE over the next nine years. In contrast, PSS and CDZ use an exponential rate of decline to mean-revert the year $t+3$ earnings growth rate g_{t+3} to a long-run growth rate in year $t+16$, g_{LT}. While the PSS method assumes that this growth rate is equal to the steady-state growth rate g, which is set to the historical average of the long-run nominal GDP growth rate, the CDZ method breaks this link between the long-run growth rate g_{LT} and the steady-state growth rate g. The former is set to the mean long-term analyst industry growth forecast, the latter is identical to g in the PSS case.

[Table 2 about here.]

Table 2 presents summary statistics for the ICC methods and realized returns. In line with previous research, all ICC methods have standard deviations that are an order of magnitude smaller than the standard deviation found in realized returns, which is considered to be one of the latter's main disadvantages. For example, Lee, Ng, et al. (2009) label realized returns an extremely noisy

ICCs, i.e. implied risk premiums, to predict *excess* realized returns to be consistent with prior literature. I do not follow this procedure to use the same variables throughout the paper. However, in untabulated analyses I check the robustness. Using excess returns leads to more equal posterior model probabilities, but the changes are minor.

proxy for expected returns and Table 2 confirms this view. The large noise in realized returns is also a driver of the low correlation between realized returns on the one hand and all ICC methods on the other hand. Nevertheless, all correlations are at least positive, a result that does not hold in the cross-section (see for example Easton and Monahan (2005)). Since all variables measure expected returns in theory, a positive correlation between those variables is confirmative, albeit weak, evidence that this is indeed the case. The correlation between the ICC methods is almost perfect, which contrasts evidence in the cross-section. This supports the view of Li et al. (2013) who claim that the aggregate ICC is less likely to be noisy because estimation errors present in firm-level ICCs are reduced through averaging. This result also implies that model uncertainty is far less an issue in the time-series than in the cross-section. Nevertheless, even in the time-series a researcher who wants to use the ICC in the empirical analysis will face quite some model uncertainty. First, while most of the correlation coefficients are over 90%, they are as low as 77%. Second, the mean across different ICC methods varies from 9.01% for method CT to 12.7% for method CDZ. Noteworthy differences are also present in the standard deviations of the various ICC methods. It becomes clear from Table 2 that a researcher would severely underestimate the true uncertainty in the statistical inference if he would base the results solely on one method. In this case he would focus on parameter uncertainty only, thereby completely ignoring model uncertainty.

I also want to emphasis that the high correlation between the ICC methods is no proof of the unbiasedness of these methods. It merely tells us that the proxies are not subject to large random measurement errors. It is still possible that all proxies are subject to the same sources of measurement error. In other words, a high correlation between proxies on the one hand and large measurement error on the other hand are not mutually exclusive. For the set of proxies chosen in this paper, there is at least reason to believe that these methods are all systematically biased: they are all based on potentially biased analyst forecasts and they are all based on a subset of the whole universe of U.S. stocks. It is just a reminder to the reader that, while the ICC is theoretically well founded, its empirical application needs a variety of ad-hoc assumptions, and it would be surprising if not at least one of these assumptions could induce a systematic bias. Furthermore, I have not implemented all variations of the ICC here. Therefore, the risk of misspecification cannot be ruled out.

4 Empirical results

4.1 Weights

[Table 3 about here.]

Table 3 shows the posterior model weights that are obtained from applying equation (16) with different shrinkage parameters ϕ. As has been argued in Section 2, a shrinkage parameter close to zero puts almost all weight on prior information and leaves little room for the data to change the researcher's view on his priors. Since the priors are equally weighted across models, so are the posteriors in the case of $\phi = 0.01$.

The more ϕ is increased, the more weight is put on the evidence in the data. And for this particular data set, the GLS method is performing best in predicting subsequent realized returns. In the limiting case in which the researcher discards all prior information ($\phi = \infty$), the posterior model weight of the GLS method is 39%. The ordering of the methods can also be inferred from the correlations between the ICC methods on the one hand and subsequent realized returns on the other (see Table 2). The higher the correlation, the higher the R^2, and the lower the sum of squared errors. And this is just the criterion to transform evidence in the data into model weights. Furthermore, it is interesting to see that the CDZ method gets almost no support from the data.

[Table 4 about here.]

How robust are these posterior model weights though? Table 4 gives the answer to this question. It shows the distribution of the posterior model weights for two different shrinkage parameters ($\phi = 1$ and $\phi = \infty$) and for 10,000 bootstrap runs. In each run, a random sample with replacement and the same size as the original data set (i.e. 324 months) is drawn and the posterior model weights are computed for this bootstrap sample.

Table 4 shows that a researcher faces considerable uncertainty about the performance of the various ICC methods. For instance, in the case of non-informative priors the 1% and 99% percentiles of the weights for the GLS method are 10% and 81%, respectively. This large variation is a result of the large noise inherent in realized returns. The true underlying expected return process is clouded by large, unsystematic shocks. Therefore, determining precisely the model weights with predictive regressions requires long samples that we do not have; and if we would have them, we would not need proxies in the first place for many research questions. This is an inherent circularity in any expected return proxy that tries to replace realized returns due to the latter's high noise.

To determine the quality of any such proxy, one needs the very same realized returns it ought to replace.

The large noise in realized returns and its consequences are well known in the finance literature. Goyal and Welch (2008) argue that apparent statistical significance of many predictors are exclusively due to years up to and especially on the years of the Oil Shock from 1973 to 1975. Fama and French (2002) find that the high realized returns of the second half of the 20th century are mostly driven by positive unexpected shocks, and not by high expected returns. And Campello et al. (2008) run the predictive regressions from above with their expected return proxy based on yield spreads and find no relation. They interpret this result as evidence that the shock structure in realized returns in their sample hindered the convergence to their expected return proxy, assumed to be correct, not as evidence that their proxy might be measured with error. From the perspective of the BMA approach advocated here, they have a very informative prior about the correctness of their proxy and therefore, discard any information in the data that casts doubt on this prior.

However, this leaves only one way to evaluate the performance of any proxy, that is, prior information. Of course, since in most empirical studies only one proxy class is under consideration, the implicit weight on this proxy is set to one. But this severely overstates, at least in my opinion, the confidence a researcher should have in his proxy. In the example of Campello et al. (2008), if their proxy is not able to explain subsequent realized returns, why should I choose their proxy instead of ICCs or proxies based on CDS spreads? In other words, a researcher who proposes a new proxy has to compare this proxy with existing ones. The only meaningful method of comparison that I know of are predictive regressions that are highly sensitive, so that a researcher might ignore these regressions altogether. But in this case, the researcher has to choose between two options. Either he considers evidence of all proxies simultaneously, which will weaken the power of statistical tests and therefore reduce one of the main advantages of alternative expected return proxies. Or he argues based on prior information why he deems his proxy more suitable than others and he has to quantify the superiority of his proxy.

This procedure is in sharp contrast to current practice. I am not aware of any study that compares different proxy classes such as expected returns based on the ICC, yield spreads, or CDS. Most papers ignore the evidence of predictive regressions completely and run ad-hoc robustness checks on their results. That is, they implicitly set the probability that their proxy is correct to one in the main empirical analysis and report their results, conditional on the assumption that their proxy is tracking expected returns perfectly. Afterwards, they repeat their analysis for variations of their proxy under consideration. Thus, the reader

has no reference for which of these variations is most supported by the data. It is also hard for him to combine the evidence from the battery of robustness tests to one coherent picture. And finally, the variations chosen in the robustness section are mostly chosen ad-hoc with no evident motivation. This is nicely illustrated in the ICC literature. Most asset pricing studies focus on the PSS approach and its derivatives, while most corporate finance studies implement abnormal growth in earnings models and residual income models. Newer studies mostly focus on one approach and change some input parameters, ignoring evidence based on other approaches. This procedure could be motivated by the fact that too many robustness checks will unnecessarily lengthen the paper and bore the reader, especially if the results are quite similar. However, as I show in subsequent examples, omitting such tests can result in quite dramatic misinterpretations.

The model averaging approach proposed in this paper is a solution to this problem. If a researcher is willing to make the extra effort to motivate this approach shortly, model averaging can take any number of expected return proxies into account without increasing the complexity of the analysis. Also, it automatically incorporates evidence about the quality of the proxies under consideration, if one is willing to take predictive regressions, despite their sensitivity to large shocks, into account. So if one proxy class gets no support in the data, it will not matter in the following empirical analysis. The weighting between the prior information and the data can easily be controlled by the researcher. Furthermore, this approach helps to protect a researcher of finding spurious relations between the variable of interest and expected returns by making sure that a researcher is not just selecting a proxy with a particular measurement error process that is related to the variable of interest. However, even this approach is not able to solve the problem of whether any proxy is tracking expected returns well. If no proxy does, the analysis will still be biased, even asymptotically. This is a severe shortcoming of any expected return proxy. Finally, the BMA approach subsumes current approaches, which also apply a model averaging approach implicitly by setting the probability of one proxy to one. So it is also possible to replicate current studies exactly with my approach. However, it requires the researcher to explicitly state his prior.

In the following, I present three empirical examples that show the impact of model uncertainty and apply the model averaging approach to deal with it.

4.2 The implied equity risk premium

In this section, I replicate the results of Claus and Thomas (2001) for an updated time period and for the six ICC methods introduced in Section 3. Claus and Thomas (2001) were one of the first studies to apply the ICC in empirical

research. They use the ICC to compute an implied risk premium, defined as the ICC minus the 10-year government bond yield, and find that the U.S. implied risk premium is only around 3% from 1985 to 1998. Due to a lack of alternative proxies back then, they only apply the CT approach.

I replicate their analysis for an updated time period from 1985 to 2011 and also incorporate model uncertainty into the analysis by considering six ICC methods simultaneously. In this example, I set $\phi = 0$, i.e. I consider each ICC method equally likely to track expected returns correctly. Hence, the posterior model weights are equal to the prior model weights; each ICC method gets the same weight. The reason why I do not consider the evidence from predictive regressions as relevant for this research question is because I assume that the level of the ICC, in which we are interested in here, is unrelated to the time-series process of the ICC. Only the latter is evaluated with predictive regressions, but since I assume that there is no relation to the former, these regressions do not help me in differentiating between the different methods. As a simple example, take two proxies, one that tracks expected returns perfectly, but is 10% too high every period, and one that is either 2% too high or 2% too low, with equal probability. While the former proxy is biased in levels, it will perfectly track the time-series of expected returns. The latter, on the other hand, will be unbiased, but not track expected returns reasonable well. In this application, we want to choose the latter, but the predictive regression would choose, at least asymptotically, the former. Hence, I ignore it.

[Figure 1 about here.]

Consequently, our inference for the implied equity risk premium should simultaneously consider the parameter uncertainty within each proxy and the uncertainty across proxies. Figure 1 does exactly that. For each proxy, 10,000 block bootstrap samples are generated with a block length of 24 months in which the mean of the implied risk premium is calculated. I use block bootstrapping here to preserve the autocorrelation structure of ICCs. The bootstrap samples for each proxy are then combined to one final sample. Based on the 10,000 bootstrapped means, I can compute the mean over all samples, which turns out to be 4.7%, or get the 2.5% and 97.5% percentile (2.7% and 7%, respectively). The plot illustrates that model uncertainty dominates parameter uncertainty considerably in this case. While the range of possible values for the implied risk premium mean, conditional on a specific proxy, is quite narrow – the largest 95% coverage region is 1.3% for the MPEG method; it is roughly three times as large when both parameter and model uncertainty are considered. Hence, it is of paramount importance to incorporate model uncertainty into the statistical inference.

Two additional points are worth repeating here. First, model uncertainty is not completely eliminated. For instance, all proxies are based on analyst forecasts and these forecasts are probably biased upwards. Of course, one could also adjust the model weights based on prior information. For example, the assumption made in the CDZ method that earnings grow with the analysts' long-term growth rate until year 15 is certainly a very aggressive growth assumption. If one deems this assumption to be unreasonable, the prior model weights of the method can be reduced accordingly. Second, this example still proves the usefulness of alternative proxies. The six ICC methods cover a wide range of earnings growth assumptions and yet, the results imply that the implied risk premium is positive and lies within a realistic range. Such a statement cannot be made for such a short period based on realized returns. Therefore, the increase in the variance, due to model uncertainty, is still considerably lower than the decrease, due to eliminating the large shocks that affect realized returns.

4.3 The intertemporal risk-return tradeoff

Although finance theory predicts a positive risk-return relation, empirical evidence based on realized returns does not conclusively find a positive sign. In simulations, Lundblad (2007) shows that even if there is a positive relation between the conditional variance and the conditional expected return, it takes very long samples to identify this relation with noisy realized returns.

Consequently, Pástor, Sinha, et al. (2008) replace realized returns with an ICC measure estimated with the PSS method. They find a positive relation between the conditional mean of market returns, approximated by their ICC, and the variance of market returns for the years 1981 to 2002. Empirically, they run the following regression specifications, which I replicate and extend with the model averaging approach:[24]

$$\widehat{\mu}_t = a + bVol_t + e_t \tag{20}$$

$$\Delta\widehat{\mu}_t = a + b\Delta Vol_t + e_t, \tag{21}$$

where $\widehat{\mu}_t$ is a proxy for expected excess returns and Vol_t is either the annualized variance or standard deviation of the daily value-weighted market returns from CRSP for this period. Since the IBES release date is typically a few days after the 15th of each month, I compute the conditional volatility based on returns ranging from the first trading day after the 15th of the previous month to

24. They also entertain a third specification in which they model both expected returns and the volatility as AR(1) processes and regress the former's residual on the latter's. To keep the analysis short, I omit this specification here.

the first trading day after the 15th of the current month.[25] The implied risk premiums are the difference between the ICC minus the 10-year government bond yield. $\Delta\widehat{\mu}_t$ and ΔVol_t are the first difference of the conditional market return mean and volatility proxies, respectively. Because the ICC is highly persistent, I follow Pástor, Sinha, et al. (2008) and use 12 Newey-West lags in regression (20). Since the first difference of ICCs does not show a persistent autocorrelation structure, they and I use one lag for specification (21).

[Table 5 about here.]

The rows labelled "PSS" in Table 5 repeat the analysis of Pástor, Sinha, et al. (2008) for a different time period (1985 to 2011 instead of 1981 to 2002). Despite the different time periods, the results are very similar. Like them, I find a positive risk-return tradeoff for both the levels and the first difference regressions and for equally and value-weighted implied risk premiums. These results are also robust: the 5th percentile based on Newey-West corrected standard errors is positive in all specifications.

In rows "MA1" and "MA2", I apply the model averaging approach to check whether these robust results could be overestimated due to the ignorance of model uncertainty. Since the density of the slope coefficient $\widehat{b}$, conditional on a specific ICC method, follows a t-distribution, the density across models is a weighted average of these conditional densities. This is simply a mixture t-distribution from which I sample 1,000,000 times. As a robustness check, I also implement a second model averaging approach, MA2, in which I generate 60,000 block-bootstrap samples with a block length of 24 months for equation (20) and 3 months for equation (21). In each sample, an ICC method is chosen randomly based on the posterior model weights. For this example, I decide to use a diffuse prior and set ϕ to ∞.

Table 5 shows that the consideration of model uncertainty has a negligible effect on the results. In all specifications, the mean of the sampled coefficients is very similar to the regression coefficient from the PSS case. Also, the 90% coverage region widens only marginally. There are now some cases in which 5% of the drawn coefficients are negative, but by and large almost all draws across the eight specifications are positive. This confirms the findings of Pástor, Sinha, et al. (2008) that there is a positive relation between the conditional market return and the conditional volatility.

[Figure 2 about here.]

25. Using conditional volatilities computed for the current calendar month yields very similar results.

Figure 2 gives the answer to the question why model uncertainty does not affect the results. It shows the histogram for 100,000 draws from the mixture t-distribution of the MA1 approach. In particular, the histogram plots draws from the case in which the first difference of the value-weighted implied risk premium is regressed on the first difference of the variance (lower left block in Table 5). It becomes clear from this figure that all methods lead to very similar results, with only slight variation in the mean and the variance of the slope coefficient's distribution. Not surprisingly, inferences based on a weighted average of similar distributions is similar to an inference based on any of these distributions.

In summary, this example shows that model averaging is an easy-to-use and flexible approach to incorporate model uncertainty. The results can be presented in a more concise way than based on separate evidence for each of the methods. It is also straightforward to extend this approach to more specifications of a specific proxy class or even across proxy classes. It also emphasizes that alternative expected return proxies have their merits over realized returns. In cases in which reasonable alterations of an expected return proxy lead to similar conclusions, model uncertainty has a negligible effect on the results. However, it is vital to check this, as the next example shows.

4.4 The importance of cash flow (CF) and discount rate (DR) news

In a recent study, Chen et al. (2013) entertain the ICC to determine whether stock prices move because of revisions in expected cash flows or discount rates. Other studies predominantly entertain a vector-autoregressive (VAR) approach to estimate the time-series of expected returns and back out cash flow news as the residual. Instead, Chen et al. (2013) use direct expected cash flow measures, namely analyst forecasts. They show that capital gain returns $Retx$ between $t+j$ and t can be separated into two parts. First, a cash flow part $CF_{j,k}$, which is the part that explains changes in stock prices due to changes in analyst forecasts between $t+j$ and t, holding the discount rate constant. Second, a discount rate part $DF_{j,k}$, which is the part that explains changes in stock prices due to changes in discount rates, holding the cash flows constant. As the subscript k indicates, both parts are dependent on the specific ICC method. In their paper, they estimate the discount rates with the CDZ method.

Concretely, recall from equation (19) and its derivatives that the stock price can be expressed as a function of the vector of future expected earnings $veps_k^t$ and an ICC proxy, $\widehat{\mu}_{t,k}$. Both are estimated at time t. $Retx$ over horizon j can

then be expressed as:

$$\begin{aligned} Retx_j &= \frac{P_{t+j} - P_t}{P_t} \\ &= \frac{f(veps_k^{t+j}, \widehat{\mu}_{t+j,k}) - f(veps_k^t, \widehat{\mu}_{t,k})}{P_t} \qquad (22) \\ &= CF_{j,k} + DR_{j,k}, \qquad (23) \end{aligned}$$

where

$$\begin{aligned} CF_{j,k} = \Bigg(&\frac{f(veps_k^{t+j}, \widehat{\mu}_{t+j,k}) - f(veps_k^t, \widehat{\mu}_{t+j,k})}{P_t} + \\ &\frac{f(veps_k^{t+j}, \widehat{\mu}_{t,k}) - f(veps_k^t, \widehat{\mu}_{t,k})}{P_t} \Bigg)/2 \qquad (24) \end{aligned}$$

and

$$\begin{aligned} DR_{j,k} = \Bigg(&\frac{veps^t, \widehat{\mu}_{t+j,k}) - f(veps^t, \widehat{\mu}_{t,k})}{P_t} + \\ &\frac{f(veps^{t+j}, \widehat{\mu}_{t+j,k}) - f(veps^{t+j}, \widehat{\mu}_{t,k})}{P_t} \Bigg)/2. \qquad (25) \end{aligned}$$

The slope coefficients obtained from regressing $CF_{j,k}$ and $DR_{j,k}$, respectively, on $Retx_j$ represent the portion of capital gain returns driven by CF news and DR news.

[Table 6 about here.]

Table 6 is a replication of Table 2 in Chen et al. (2013) for the aggregate market.[26] Although I use a different sample – for instance, I require that for every observation all ICCs are available –, the results are very similar. In both samples, the variation in capital gain returns is mostly explained by the DR news part for shorter horizons and by the CF news part for longer horizons. At a quarterly horizon, only 18%/16% of the return variation of the market portfolio is explained by CF news in my/their sample. This fraction increases to 70%/59% at a seven-year horizon. Also, the results are robust: at twelve quarters and beyond, the fraction of CF news is above 50%, even for the 5% percentile. In summary, these results imply that cash flow news is important in driving the stock price movements, based on evidence of the CDZ methods.

[Figure 3 about here.]

26. I winsorize $Retx$, DR, and CF for each horizon at the 1% and 99% breakpoints. This is the reason why the tautological relation in equation (22) is broken and the slope coefficients in Panel B do not add up to 1. However, the deviations are marginal. In line with Chen et al. (2013), I also use quarterly data, i.e. I only consider observations from March, June, September, and December of each year.

Do these results also hold if we incorporate model uncertainty into the analysis? The short answer is no and Figure 3 shows why. It plots the fraction of the variation in Retx that is explained by CF news over various horizons and for different ICC methods. As becomes apparent, the view that most of the variation in capital gain returns is driven by CF news is only supported by the CDZ method and, to a lesser extent, by the CT method. All other methods come to the conclusion that DR news, even for longer horizons, is more important. However, there is a very large variation across the different methods, which means that the return decomposition approach based on ICCs is very sensitive to the specific discount model.

The rationale behind this finding is that every ICC method equates the current stock price with a transformation of discounted expected dividends. Differences arise on how the second part is transformed. In this particular research question, the assumptions made here have a very large impact. Chen et al. (2013) assume that the earnings growth rate converges to the industry long-term growth rate provided by analysts over the next 15 years, although these growth rates are commonly interpreted to represent the next five years (see Claus and Thomas 2001) and are probably affected by analyst bias. Obviously, these growth assumptions are very sensitive to the current market environment. For example, during the dot-com bubble in 2001 the mean across the industry growth rates within my sample was as high as 25%. Assuming that investors expected earnings growth rates to converge to these growth rates for the next 15 years will obviously explain almost all of the capital gains that accrued over this period. Such an extreme assumption is not made by the other methods. For example, the PSS method assumes that the earnings growth rate in period 3 is the earnings growth rate provided by analysts and extrapolates this growth rate over 15 years to the historical average of the nominal GDP growth rate. This much more conservative assumption about expected earnings leaves a much larger part of capital gains unexplained and the ICC has to step in to fill the gap. The other extreme is the traditional GLS method, which anchors the current price on the very persistent current book value and the very persistent median ROE over the last decade to which ROEs for each firm are extrapolated from period 4 on. So the only parts left to explain price changes are the earnings forecasts of the first three years and the ICC. Hence, the latter has to account for a large part of the variation in returns, which actually results in a negative CF part for the GLS method in this sample.

This, of course, is an outcome of model uncertainty. We do not know if investors updated their long-term earnings growth assumptions or if they updated their expected returns. It is the question we want to answer. Each method emphasizes the two parts differently and hence, results conditional on only one

method ignore the uncertainty we have about these assumptions. Since we already know from Section 4.1 that the posterior model weights do not favor any ICC method unambiguously, it is obvious that the posterior distribution will be spread out. Nevertheless, it is instructive to apply the model averaging approach here.

First, I set ϕ to 1, which gives roughly equal weight to the evidence in the data and my prior beliefs that all ICC methods should be equally likely. The sampling from the parameter densities are done as in the previous example. That is, in the first case I sample from a mixture t-distribution, where each t-distribution's parameters are estimated from an OLS with Newey-West corrected standard errors. In the second case, I apply a bootstrap approach in which I choose in each run an ICC method randomly, based on the posterior model weights, and obtain the regression coefficient for the specific bootstrap sample.

[Table 7 about here.]

Table 7 presents the results. As expected, incorporating model uncertainty widens the coverage regions considerably. Only for shorter horizons can one be reasonably sure that returns are mostly driven by DR news. For longer horizons, the intervals become too large to draw any reasonable conclusions. The results also show that the two approaches of model averaging yield similar results here.

5 Conclusion and Outlook

In this paper, I incorporate model uncertainty into the statistical inference that is based on expected return proxies. In the theoretical part, I show how results can be biased if one ignores the uncertainty in the selection process of such proxies and propose a model averaging approach that incorporates it. In the empirical part, I apply this approach to three research questions that are based on the implied cost of capital approach.

My main findings are that ignoring model uncertainty can overestimate the confidence of the results considerably. Hence, many apparently significant results between expected return proxies and a variable of interest found in the literature could be due to ignoring both the performance evaluation of such a proxy and the uncertainty about this evaluation. Therefore, it should be an interesting endeavor to replicate previous studies with my model averaging approach. In particular, it would be interesting to apply this approach to studies that look at the cross-section of expected returns.

My approach serves also as a guideline on empirical research with expected return proxies. First, it shows that it is important to consider reasonable alter-

native specifications. This is in contrast with many studies in the ICC literature that focus only on one transformation of the dividend discount model and make minor adjustments to this model. If only proxies are considered that are virtually identical, it is obvious that the results across these proxies will not uncover model uncertainty. Second, a researcher must be explicit about any prior beliefs made about the quality of the proxies. In current studies, it is common practice to implicitly set the prior weight on one proxy to one and to ignore the evidence of other reasonable specifications, at least for the main part of the empirical analysis. This approach is unsatisfying for two reasons: First, it ignores evidence about the performance of each proxy to explain subsequent realized returns, which each proxy has to explain eventually. Second, it conceals the uncertainty inherent in the proxy selection process.

The model averaging approach inherits its weaknesses from the underlying proxies. If all proxies are systematically biased, the results based on evidence across these models will also be biased. If a specification is favored by the inclusion of many minor variations of this specification, the posterior results will also put too much emphasizes on this specification. If shocks on subsequent realized returns are correlated with a specific ICC method in sample, predictive regressions will unjustly favor it over other methods and results will be biased towards this method.

In summary, my results provide evidence that model uncertainty in alternative expected return proxies is the complement of parameter uncertainty in realized returns. My study is a first attempt to answer the question which of those is more worrisome for the applied researcher.

References

Avramov, Doron. 2002. Stock return predictability and model uncertainty. *Journal of Financial Economics* 64 (3): 423–458.

Binsbergen, Jules van, Wouter Hueskes, Ralph Koijen, and Evert Vrugt. 2013. Equity yields. *Journal of Financial Economics* Forthcoming.

Campello, Murillo, Long Chen, and Lu Zhang. 2008. Expected returns, yield spreads, and asset pricing tests. *Review of Financial Studies* 21 (3): 1297–1338.

Chen, Long, Zhi Da, and Xinlei Zhao. 2013. What drives stock price movements? *Review of Financial Studies* 26 (4): 841–876.

Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model selection and model averaging.* Vol. 330. Cambridge University Press Cambridge.

Claus, James, and Jacob Thomas. 2001. Equity premia as low as three percent? evidence from analysts' earnings forecasts for domestic and international stock markets. *The Journal of Finance* 56 (5): 1629–1666.

Cochrane, John H. 2011. Presidential address: discount rates. *The Journal of Finance* 66 (4): 1047–1108.

Cremers, Martijn. 2002. Stock return predictability: a bayesian model selection perspective. *Review of Financial Studies* 15 (4): 1223–1249.

Draper, David. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological):*45–97.

Easton, Peter. 2004. PE ratios, PEG ratios, and estimating the implied expected rate of return on equity capital. *The Accounting Review* 79 (1): 73–95.

———. 2007. Estimating the cost of capital implied by market prices and accounting data. *Foundations and Trends® in Accounting* 2 (4): 241–364.

Easton, Peter, and Steven Monahan. 2005. An evaluation of accounting-based measures of expected returns. *The Accounting Review* 80 (2): 501–538.

Elton, Edwin. 1999. Presidential address: expected return, realized return, and asset pricing tests. *The Journal of Finance* 54 (4): 1199–1220.

Fama, Eugene, and Kenneth French. 1997. Industry costs of equity. *Journal of Financial Economics* 43 (2): 153–193.

———. 2002. The equity premium. *The Journal of Finance* 57 (2): 637–659.

Fernandez, Carmen, Eduardo Ley, and Mark Steel. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16 (5): 563–576.

Friewald, Nils, Christian Wagner, and Josef Zechner. 2013. The cross-section of credit risk premia and equity returns. *The Journal of Finance* Forthcoming.

Gebhardt, William, Charles Lee, and Bhaskaran Swaminathan. 2001. Toward an implied cost of capital. *Journal of Accounting Research* 39 (1): 135–176.

Gode, Dan, and Partha Mohanram. 2003. Inferring the cost of capital using the Ohlson–Juettner model. *Review of Accounting Studies* 8 (4): 399–431.

Goyal, Amit, and Ivo Welch. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21 (4): 1455–1508.

Guay, Wayne, SP Kothari, and Susan Shu. 2011. Properties of implied cost of capital using analysts' forecasts. *Australian Journal of Management* 36 (2): 125–149.

Hail, Luzi, and Christian Leuz. 2009. Cost of capital effects and changes in growth expectations around us cross-listings. *Journal of Financial Economics* 93 (3): 428–454.

Hou, Kewei, Mathijs van Dijk, and Yinglei Zhang. 2012. The implied cost of capital: a new approach. *Journal of Accounting and Economics* 53 (3): 504–526.

Hughes, John, Jing Liu, and Jun Liu. 2009. On the relation between expected returns and implied cost of capital. *Review of Accounting Studies* 14 (2-3): 246–259.

Kass, Robert, and Adrian Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90 (430): 773–795.

Koijen, Ralph, and Stijn Van Nieuwerburgh. 2011. Predictability of returns and cash flows. *Annual Review of Financial Economics* 3:467–491.

Leamer, Edward. 1978. *Specification searches: ad hoc inference with nonexperimental data.* Wiley New York.

Lee, Charles, David Ng, and Bhaskaran Swaminathan. 2009. Testing international asset pricing models using implied costs of capital. *Journal of Financial and Quantitative Analysis* 44 (2): 307–335.

Lee, Charles, Eric So, and Charles Wang. 2011. Evaluating implied cost of capital estimates. *Working Paper.*

Ley, Eduardo, and Mark Steel. 2009. On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* 24 (4): 651–674.

Li, Yan, David Ng, and Bhaskaran Swaminathan. 2013. Predicting market returns using aggregate implied cost of capital. *Journal of Financial Economics* Forthcoming.

Lundblad, Christian. 2007. The risk return tradeoff in the long run: 1836–2003. *Journal of Financial Economics* 85 (1): 123–150.

Ohlson, James, and Beate Juettner-Nauroth. 2005. Expected EPS and EPS growth as determinants of value. *Review of Accounting Studies* 10 (2-3): 349–365.

Pástor, Ľuboš, Meenakshi Sinha, and Bhaskaran Swaminathan. 2008. Estimating the intertemporal risk–return tradeoff using the implied cost of capital. *The Journal of Finance* 63 (6): 2859–2897.

Pástor, Ľuboš, and Robert Stambaugh. 2009. Predictive systems: living with imperfect predictors. *The Journal of Finance* 64 (4): 1583–1628.

Ramnath, Sundaresh, Steve Rock, and Philip Shane. 2008. The financial analyst forecasting literature: a taxonomy with suggestions for further research. *International Journal of Forecasting* 24 (1): 34–75.

Sadka, Gil, and Ronnie Sadka. 2009. Predictability and the earnings–returns relation. *Journal of Financial Economics* 94 (1): 87–106.

Sala-I-Martin, Xavier, Gernot Doppelhofer, and Ronald Miller. 2004. Determinants of long-term growth: a bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94 (4): 813–835.

Wright, Jonathan. 2008. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics* 146 (2): 329–341.

Zellner, Arnold. 1986. On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: essays in honor of Bruno de Finetti,* ed. P. K. Goel and Arnold Zellner. North Holland, Amsterdam.

Table 1: Summary of empirical ICC proxies and data sources. This table shows the underlying valuation model for each of the implemented ICC proxies. For each model, the formula in column "P_t equals" is solved numerically for r. P_t is the share price of firm i from IBES (the firm index is suppressed throughout this table); if available, eps_{t+k} is the consensus mean earnings per share forecast for year $t+k$ from IBES. To be included in the sample, I require at least two analyst forecasts. Additional forecasts, if not available, are inferred by multiplying the forecast for the previous year with the analysts' long-term earnings growth rate: $eps_{t+k+1} = eps_{t+k} \times (1+Ltg_t)$. Expected dividends dps_{t+k} are computed as the product of eps_{t+k} and a payout ratio PO_t. For the first three years ahead, PO_t is inferred from Compustat data as common dividends (DVC) divided by income before extraordinary items (IBCOM) for profitable firms. For firms with a negative income, 6% of total assets is used as the denominator instead. bps_t is the equity book value per share at the end of period t, computed as shareholder's equity (SEQ) divided by common shares outstanding (CSHO) from Compustat.

Method	Based on	P_t equals	Implementation details
MPEG	Easton (2004)	$\frac{eps_{t+2}+r\times dps_{t+1}-eps_{t+1}}{r^2}$	
OJ	Ohlson and Juettner-Nauroth (2005), Gode and Mohanram (2003)	$\frac{eps_{t+1}}{r} + \frac{g_{st}\times eps_{t+1}-r\times(eps_{t+1}-dps_{t+1})}{r\times(r-g_{lt})}$	g_{st} is the average of the growth rate between eps_{t+1} and eps_{t+2} on the one hand and Ltg_t on the other hand. g_{lt} is the yield on a 10-year government bond minus 3%.
CT	Claus and Thomas (2001)	$bps_t + \sum_{j=1}^{5} \frac{eps_{t+j}-r\times bps_{t+j-1}}{(1+r)^j} + \frac{(eps_5-r\times bps_4)\times(1+g_{lt})}{(r-g_{lt})\times(1+r)^5}$	g_{lt} is the yield on a 10-year government bond minus 3%. Book values per share of year $t+j$ are recursively inferred from book values of year $t+j-1$, the earnings per share of year $t+j$ and a constant payout ratio PO_t: $bps_{t+j} = bps_{t+j-1} + eps_{t+j} \times (1-PO_t)$.
GLS	Gebhardt et al. (2001)	$bps_t + \sum_{j=1}^{11} \frac{(roe_{t+j}-r)\times bps_{t+j-1}}{(1+r)^j} + \frac{(eps_{12}-r)\times bps_{11}}{r\times(1+r)^{11}}$	$roe_{t+j} = eps_{t+j}/bps_{t+j-1}$ for $j = 1,2,3$. From $j=4$ on, roe_{t+j} are linearly interpolated to roe_{12}, which is defined as the median industry return on equity (ROE) (IBCOM/SEQ of previous year) over the last ten years for all profitable firms. I use the 48 Fama and French (1997) industry classification. bps_{t+j} is computed as in CT.
PSS	Pástor, Sinha, et al. (2008)	$\sum_{j=1}^{15} \frac{eps_{t+j}\times PO_{t+j}}{(1+r)^j} + \frac{eps_{t+16}}{r\times(1+r)^{15}}$	From $k=4$ on, earnings are computed as $eps_{t+j} = eps_{t+j-1} \times (1+g_{t+j})$. The earnings growth rate in year $t+3$, g_{t+3}, is set to Ltg_t and mean-reverted to its steady-state value by year $t+17$, g, with formula $g_{t+k} = g_{t+k-1} \times exp(log(g/g_{t+3})/14)$. The plowback rate, defined as $b_t = 1-PO_t$, is recursively computed as $b_{t+k} = b_{t+k-1}-(b_{t+2}-b)/14$. g is set to the average nominal GDP growth rate estimated using an expanding rolling window starting from 1947. The steady-state plowback rate is given as $b = g/r$.
CDZ	Chen et al. (2013)	As PSS	This method is identical to PSS except for the growth rate to which $g_{t+3} = Ltg_t$ mean-reverts. Here, g_{t+k} is computed as $g_{t+k} = g_{t+k-1} \times exp(log(g_{LT}/g_{t+3})/14)$, where g_{LT} is set to the industry-wide mean of long-term growth rates Ltg by the analysts. From $t+17$ on, in the period in which g_{LT} would be reached, it is instead assumed that earnings grow with g that are computed identical to PSS. I use the same industry classification as in the GLS case.

Table 2: Summary statistics. This table provides the mean, the standard deviation, and the correlations for the monthly time-series of the continuously compounded NYSE/Amex/Nasdaq value-weighted return (Ret) and the aggregate implied cost of capital for the six implemented methods (for computation details, refer to Table 1). All variables are reported in annualized percentages. The time period ranges from 1985 to 2011.

			Correlations						
Variable	Mean	SD	Ret	MPEG	OJ	CT	GLS	PSS	CDZ
Ret	9.63	16.34	100.00	9.41	7.84	6.59	11.67	9.13	5.21
MPEG	11.12	1.64	9.41	100.00	96.45	92.29	94.49	96.27	77.74
OJ	11.47	1.77	7.84	96.45	100.00	95.54	95.49	96.65	86.76
CT	9.01	1.29	6.59	92.29	95.54	100.00	89.91	97.30	87.37
GLS	9.91	1.89	11.67	94.49	95.49	89.91	100.00	95.48	77.02
PSS	9.91	1.48	9.13	96.27	96.65	97.30	95.48	100.00	85.04
CDZ	12.72	1.21	5.21	77.74	86.76	87.37	77.02	85.04	100.00

Table 3: Posterior model weights for different shrinkage parameters. This table shows the posterior model weights of the ICC methods for different shrinkage parameters ϕ. The weights are based on predictive regressions of the ICCs on subsequent continuously compounded monthly realized returns. The ICC methods are described in Table 1. The following priors are specified: Equal prior model probabilities $p(M_i)$ across ICC methods, an improper prior on σ^2, and the natural conjugate g-prior specification for β: $N(0, \phi\sigma^2(X_i'X_i)^{-1})$, where X_i is the $T \times 2$ matrix of a T vector of ones and the T vector $\widehat{\mu}_i$; the posterior model weights are computed via equation (16). Note that the case $\phi = \infty$ is identical to the AIC weighting shown in equation (17). The time period ranges from 1985 to 2011.

ϕ	MPEG	OJ	CT	GLS	PSS	CDZ
0.01	16.71	16.64	16.59	16.83	16.69	16.55
0.1	17.01	16.36	15.95	18.21	16.89	15.58
1	18.06	14.54	12.58	26.45	17.32	11.04
10	18.06	12.10	9.27	36.56	16.72	7.29
100	17.92	11.57	8.65	38.73	16.48	6.65
∞	17.90	11.50	8.58	39.00	16.44	6.58

Table 4: Bootstrapped posterior model weights for two shrinkage parameters. This table shows the distribution of posterior model weights of the ICC methods for two different shrinkage parameters ($\phi = 1$ and $\phi = \infty$) over 10,000 block-bootstrap samples with a block length of 24 months. For each bootstrap sample, the analysis outlined in Table 3 is run.

Percentile	MPEG	OJ	CT	GLS	PSS	CDZ
Posterior model weights for $\phi = 1$						
1%	8.01	6.11	4.73	14.56	7.97	1.61
5%	10.72	8.52	7.27	16.45	11.01	3.40
50%	16.78	14.10	12.69	24.32	16.69	12.00
95%	29.33	18.76	16.72	41.56	24.50	23.43
99%	39.68	21.86	19.48	50.83	28.54	31.56
Posterior model weights for $\phi = \infty$						
1%	2.23	1.18	0.76	10.24	2.27	0.08
5%	5.06	2.84	1.94	15.19	5.16	0.46
50%	15.42	10.72	8.58	32.24	15.29	7.96
95%	41.90	18.91	16.05	68.08	28.89	31.03
99%	63.79	24.22	20.62	81.37	36.97	50.37

Table 5: Implied premiums regressed on market volatility. This table replicates and extends the analysis in Pástor, Sinha, et al. (2008). It reports the slope coefficients as well as the 5% and 95% percentile of the coefficients' distributions from the regressions in equations (20) and (21). The independent variable is return volatility σ_t, estimated as the realized market volatility from daily returns of the first trading day after the 15th of month $t-1$ to the first trading day after the 15th of month t. I use the CRSP value-weighted index. The column labels σ_t^2 and σ_t denote the regressor. The sampling distributions of the slope coefficients are estimated in three ways. In the PSS case, it is from an OLS with the implied risk premium based on the PSS method and Newey-West corrected standard errors. In the first model averaging approach (MA1), I sample 1,000,000 times from a mixture t-distribution. Each t-distribution's parameters are determined from a regression of the implied risk premium for the six ICC methods under consideration with Newey-West corrected standard errors. The mixture probabilities are the posterior model weights with $\phi = \infty$. In the second model averaging approach (MA2), the statistics are based on 60,000 block-bootstrap samples with a length of 24 months for equation (20) and 3 months for equation (21). In each bootstrap run, an implied risk premium is chosen randomly based on the posterior model weights with $\phi = \infty$. The implied risk premiums are the difference between the ICCs and the 10-year government bond yield. In Panel A/B, I use equal/value-weighted premiums. The sample is monthly and ranges from 1985 to 2011.

	σ_t^2			σ_t		
	$\widehat{b}$	5th perc.	95th perc.	$\widehat{b}$	5th perc.	95th perc.
Panel A: Equal-weighted implied risk premiums						
Levels: $\widehat{\mu_t} = a + b\sigma_t^{(2)} + e_t$						
PSS	0.101	0.053	0.149	0.084	0.061	0.108
MA1	0.098	0.052	0.145	0.082	0.057	0.107
MA2	0.109	0.051	0.224	0.081	0.050	0.109
Differences: $\Delta\widehat{\mu_t} = a + b\Delta\sigma_t^{(2)} + e_t$						
PSS	0.024	0.010	0.039	0.024	0.015	0.034
MA1	0.023	0.007	0.039	0.023	0.012	0.034
MA2	0.024	0.001	0.046	0.023	0.010	0.033
Panel B: Value-weighted implied risk premiums						
Levels: $\widehat{\mu_t} = a + b\sigma_t^{(2)} + e_t$						
PSS	0.080	0.031	0.130	0.062	0.027	0.096
MA1	0.069	0.024	0.118	0.051	0.010	0.090
MA2	0.067	-0.004	0.132	0.046	-0.012	0.089
Differences: $\Delta\widehat{\mu_t} = a + b\Delta\sigma_t^{(2)} + e_t$						
PSS	0.016	0.002	0.031	0.017	0.008	0.027
MA1	0.015	-0.001	0.031	0.015	0.004	0.026
MA2	0.014	-0.008	0.033	0.015	0.003	0.025

Table 6: Return decomposition using CDZ method. This table replicates Table 2 in Chen et al. (2013). Panel A reports for the value-weighted market portfolio the mean as well as the variance of capital gain returns (Retx), cash flow (CF) news, and discount rate (DR) news, from one quarter up to 28 quarters. Panel B reports the portion of capital gain returns that can be explained by CF and DR news, respectively. These are determined by regressing CF news and DR news on aggregate Retx. The rows 5% and 95% report the confidence intervals around the coefficients and are based on Newey-West standard errors with the lag set to the number of overlapping quarters. The sample is quarterly and ranges from 1985 to 2011. All numbers are in percent.

	Horizons (Quarter)								
	1	2	4	8	12	16	20	24	28
Panel A: Summary statistics									
Mean(CF)	1.95	4.03	7.49	13.54	18.32	25.55	33.91	41.36	49.21
Mean(DR)	0.36	0.76	2.31	5.96	9.61	12.80	15.90	19.36	24.00
Mean(Retx)	2.30	4.74	9.69	19.40	27.98	38.61	50.26	61.55	73.92
Var(CF)	0.39	0.86	1.99	4.85	7.44	10.85	13.58	14.66	15.69
Var(DR)	0.76	1.27	2.37	3.35	3.64	4.30	4.50	4.03	5.59
Var(Retx)	0.60	1.27	2.36	5.47	9.58	15.55	19.82	21.74	25.31
Panel B: Decomposition									
5%	2.23	15.82	20.99	42.59	57.09	58.39	58.85	60.12	57.89
CF	18.00	33.08	41.83	63.44	69.87	70.85	72.72	74.16	70.11
95%	33.76	50.35	62.67	84.28	82.64	83.32	86.58	88.21	82.33
5%	65.00	47.21	35.03	13.11	14.52	13.62	9.72	7.57	13.50
DR	80.68	64.86	56.28	34.34	27.67	26.45	24.22	22.60	27.06
95%	96.36	82.50	77.54	55.56	40.82	39.27	38.71	37.64	40.63

Table 7: Return decomposition using the model averaging approach. This table updates Table 2 in Chen et al. (2013) by applying the model averaging approach proposed in this paper. The posterior model weights are based on $\phi = 1$. In Panel A, the slope coefficients are sampled from a mixture t-distribution where each t-distribution is scaled by the Newey-West standard errors with the lag set to the number of overlapping horizons and the slope coefficient added. The weighting across the t-distribution is based on the posterior model weights. 1,000,000 draws are taken. Panel B is based on 60,000 block-bootstrap samples with a length of 20 quarters, drawn with replacement. In each sample, the CF news based on an ICC method are chosen randomly, subject to the posterior model weights and the slope coefficient for the specific bootstrap sample and ICC methods returned. All Panels show the 5% percentile, the mean, and the 95% percentile of the generated samples. The sample is quarterly and ranges from 1985 to 2011. All numbers are in percent.

	Horizons (Quarter)								
	1	2	4	8	12	16	20	24	28
Panel A: Mixture t-distribution with Newey-West standard errors									
5%	-0.22	8.85	9.30	10.12	6.47	6.59	3.96	-6.10	-14.28
CF	15.22	25.39	27.69	37.01	36.98	37.05	36.01	34.30	28.94
95%	39.02	47.39	50.25	71.62	74.01	74.55	76.40	76.95	72.08
Panel B: Bootstrapped samples									
5%	2.66	11.01	7.65	11.75	7.33	6.15	3.63	-12.45	-19.99
CF	14.97	24.91	26.66	36.98	37.63	37.08	35.17	32.22	26.66
95%	35.11	43.57	49.17	70.63	75.29	75.84	76.80	77.31	71.44

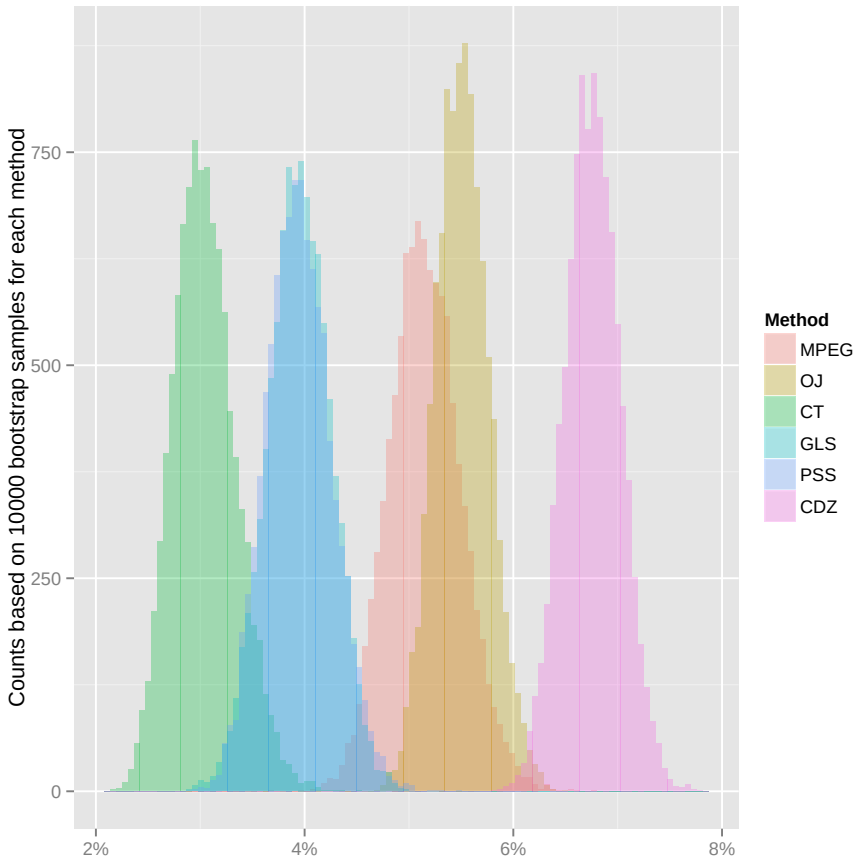

Figure 1: Histogram of bootstrapped means of implied risk premiums for six different ICC methods. This figure overlays the six histograms for the means of implied risk premiums computed from the six ICC methods. Each histogram consists of 10,000 means that are computed from block-bootstrapped samples with a block length of 24 months. The monthly sample ranges from 1985 to 2011.

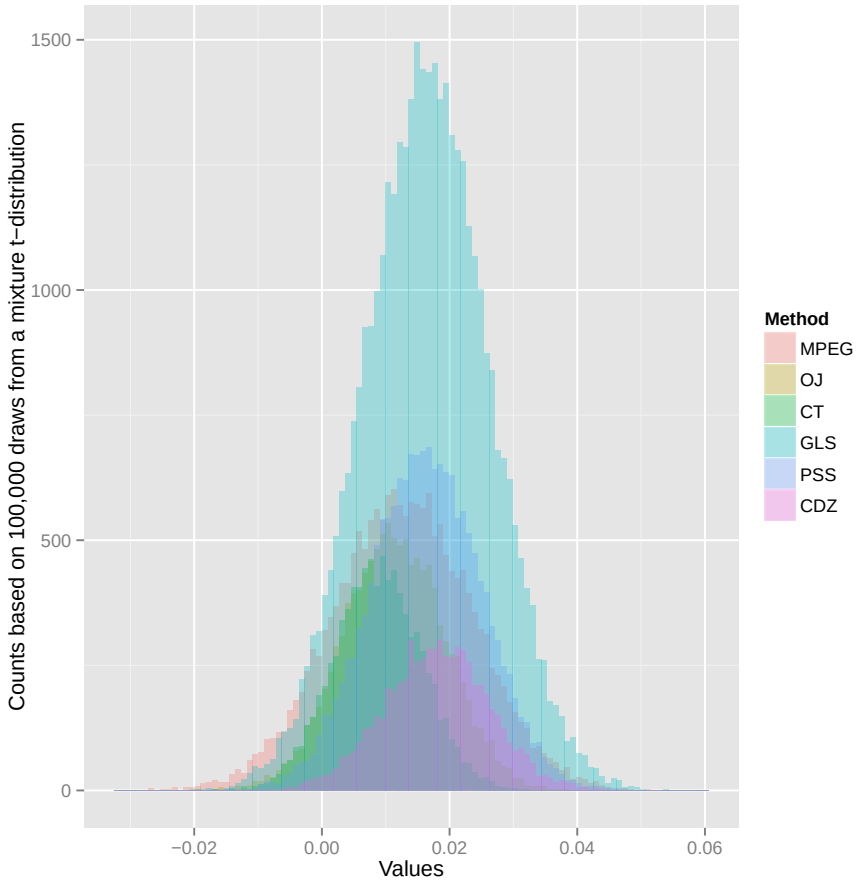

Figure 2: Mixture t-distribution of slope coefficients from regressing implied risk premiums on market volatility This plot shows 100,000 draws from a mixture t-distribution. Each t-distribution represents the sampling distribution of the slope coefficient from regressing the first differences of the value-weighted implied risk premium of the specific method on the first differences of the market volatility, which is measured here as the variance of daily stock returns. For more information, see the description in Table 5. This shows random samples from the MA1 model averaging approach. The monthly sample period begins in January 1985 and ends in December 2011.

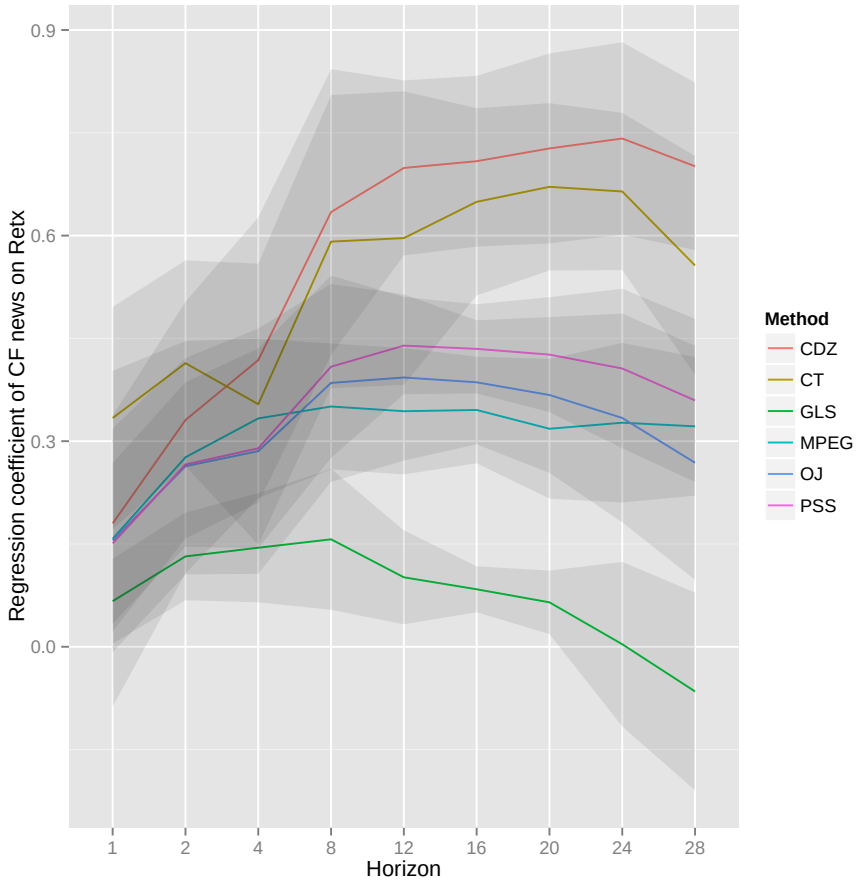

Figure 3: Fraction of Retx driven by CF news for different ICC methods and horizons. This figure shows the fraction of variation in Retx attributable to CF news for different ICC methods and horizons. The fraction is defined as the regression coefficient of CF news on Retx. The ICC methods are defined in Table 1. The shaded area around each line represents the 90% confidence bands around the coefficients. The bands are computed via Newey-West standard errors with the lag set to the number of overlapping quarters. The sample is quarterly and ranges from 1985 to 2011.