City Research Online

City, University of London Institutional Repository

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: http://openaccess.city.ac.uk/15053/

Link to published version: http://dx.doi.org/10.1016/j.insmatheco.2016.04.007

City Research Online: http://openaccess.city.ac.uk/   publications@city.ac.uk

Nonparametric Long-Term Prediction of Stock Returns With Generated Bond Yields

Michael Scholz[a,*], Stefan Sperlich[b], Jens Perch Nielsen[c]

[a]*Department of Economics, University of Graz, Universitätsstraße 15/F4, 8010 Graz, Austria*
[b]*Département des sciences économiques and & Research Center for Statistics, Université de Genéve, Bd du Pont d'Arve 40, 1211 Genévre 4, Switzerland*
[c]*Faculty of Actuarial Science and Insurance, Cass Business School, 106 Bunhill Row, London, EC1Y8TZ, UK*

Abstract

Recent empirical approaches in forecasting equity returns or premiums found that dynamic interactions among the stock and bond are relevant for long term pension products. Automatic procedures to upgrade or downgrade risk exposure could potentially improve long term performance for such products. The risk and return of bonds is more easy to predict than the risk and return of stocks. This and the well known stock-bond correlation motivates the inclusion of the current bond yield in a model for the prediction of excess stock returns. Here, we take the actuarial long term view using yearly data, and focus on nonlinear relationships between a set of covariates. We employ fully nonparametric models and apply for estimation a local-linear kernel smoother. Since the current bond yield is not known, it is predicted in a prior step. The structure imposed this way in the final estimation process helps to circumvent the curse of dimensionality and reduces bias in the estimation of excess stock returns. Our validated stock prediction results show that predicted bond returns improve stock prediction significantly.

Keywords: Prediction; Stock returns; Bond yield; Cross validation; Generated regressors

JEL: C14; C53; C58; G17; G22

---

*Corresponding author at: Department of Economics, University of Graz, Universitätsstraße 15/F4, 8010 Graz, Austria. Tel.: +43 316 380 7112; fax: +43 316 380 697112.

Email addresses: michael.scholz@uni-graz.at (Michael Scholz), stefan.sperlich@unige.ch (Stefan Sperlich), Jens.Nielsen.1@city.ac.uk (Jens Perch Nielsen)

1. Introduction and Motivation

For a long time predicting asset returns has been a main objective in the empirical finance literature. It started with predictive regressions of independent variables on stock market returns. Typically, valuation ratios are used that primarily characterise the stock, for example the dividend price ratio, the dividend yield, the earnings price ratio or the book-to-market ratio. Other variables related to the interest rate like treasury-bill rates and the long-term bond yield, or macroeconomic indicators like inflation and the consumption wealth ratio, are often incorporated to improve prediction. For a detailed overview we refer to the examples and discussion in Rapach et al. (2005) or Campbell and Thompson (2008).

In this paper, we take the actuarial long term view using yearly data, and focus on nonlinear relationships between a set of covariates. There are not many historical years in our records and data sparsity is of great importance in our approach. One could also use data of higher frequency as weekly or daily data, but one has to remember that the logistics of prediction is then very different. In our approach using yearly data bias might be of big importance while variance becomes less of an issue. In other words, the usual variance-bias trade-off depends on the horizon. An adequate model for monthly data might perform worse for yearly data and vice versa. The reason for the use of yearly data is our interest in actuarial models of long term savings and their possible econometric improvement (see e.g. Bikker et al. (2012), Guillen et al. (2013a), Guillen et al. (2013b), Owadally et al. (2013), Guillen et al. (2014), or Gerrard et al. (2014)). Our favored methodology of validating the fully nonparametric models that we employ for the long term yearly data also originates from the actuarial literature (see Nielsen and Sperlich (2003)).

The apparent predictability found by many authors was controversially discussed. As Lettau and Nieuwerburgh (2008) note, correct inference is problematic due to the high persistence of financial ratios, which have poor out-of sample forecasting power that moreover shows significant instability over time. Therefore, the question of whether empirical models are really able to forecast the equity premium more

accurately than the simple historical mean was intensively debated in the finance literature. Recently, Goyal and Welch (2008) fail to provide benefits of predictive variables compared to the historical mean. In contrast, Rapach et al. (2010) recommend a combination of individual forecasts. Their method includes the information provided from different variables and reduces this way the forecast volatility. Elliott et al. (2013) suggest a new method to combine linear forecasts based on subset regressions and show improved performance over the classical linear prediction methods. More recently, Scholz et al. (2015) propose a simple bootstrap test about the true functional form to evidence that the null of no predictability of returns can be rejected when using information such as earnings.

A direct comparison of stocks and bonds, mostly used by practitioners, makes the so-called FED model. It relates yields on stocks, as ratios of dividends or earnings to stock prices, to yields on bonds. Asness (2003) shows the empirical descriptive power of the model, but notes also that it fails in predicting stock returns. One of his criticisms is the comparison of real numbers to nominal ones. Actually, most studies discuss separately the predictability in stock and bond markets. However, Shiller and Beltratti (1992) analyse the relation between stock prices and changes in long-term bond yields. Fama and French (1993) find that stock returns have shared variation due to the stock-market factors, and they are linked to bond returns through shared variation in the bond-market. Engsted and Tanggaard (2001) pose the interesting question of whether expected returns on stocks and bonds are driven by the same information, and to what extent they move together. In their empirical setting, they find that excess stock and bond returns are positively correlated. Aslanidis and Christiansen (2014) adopt quantile regressions to scrutinize the realized stock-bond correlation and the link to the macroeconomy. Tsai and Wu (2014) analyse the bond and stock market responses to changes in dividends. Lee et al. (2013) find dynamic interactions among the stock, bond, and insurance markets. For additional literature on the relation between stock and bond returns (especially co-movements, joint distributions, or correlations), see, for example, Lim et al. (1998), Ilmanen (2003), Guidolin and Timmermann (2006), Connolly et al. (2010),

Baele et al. (2010), or Bekaert et al. (2010).

One overall idea of the this paper is to exploit the interrelationship of present values of stock returns and bond returns. They are after all both discounted cash flows. Our underlying assumption implies that expected returns are associated with variables related to longer-term aspects of business conditions, as mentioned in Campbell (1987). Consequently, we include in a nonparametric prediction model of excess stock returns the bond yield of the same year. This way, the bond captures a most important part of the stock return, namely the part related to the change in long-term interest rate. Nonlinear forecasting methods are a growing area of empirical research, see for example Guidolin and Timmermann (2006), McMillan (2007), or Guidolin et al. (2009). Nielsen and Sperlich (2003) find a significant improvement in the prediction power of excess stock returns due to the use of nonlinear smoothing techniques. Based on their findings, we focus on nonlinear relationships between a set of covariates and the bond yield of the same year. We apply for estimation a local-linear kernel smoother which nests the linear model without bias. For the purpose of bandwidth selection and to measure the quality of prediction we use a cross-validation measure of performance. It is a generalized version of the validated R^2 of Nielsen and Sperlich (2003) and allows for a direct comparison of the proposed model with the historical mean.

An obvious problem is that the current bond yield is unknown. Thus, we have to predict it in a first step. Here, we also employ fully nonparametric models and use a local-linear kernel smoother. This raises the question why it is necessary to use a two-step procedure. One could directly include the variables used for the bond prediction when forecasting stock returns. The problem is that such a model would suffer from the curse of dimensionality and complexity in several aspects: The dimension of the covariates, possible over-fitting, and the interpretability. In nonparametrics it is well known that the import of structure is an appropriate way to circumvent these problems[1]. Furthermore, Park et al. (1997)

---

[1]An other possibility could be the optimal choice of regressors, see Vieu (1994).

showed that an appropriate transformation of the predictors can significantly improve nonparametric prediction. In our approach, we utilize the additional knowledge about the structure that is inherent in the economic process that generates the data. We find that the inclusion of the generated variable shows notable improvement in the prediction of excess stock returns. Note that one does not achieve computational efficiency, but rather estimation efficiency from adding information. To our knowledge we are the first including nonparametrically generated regressors for nonparametric prediction of time series data. Therefore we also have to develop the theoretical justification for the use of constructed variables in nonparametric regression when the data are dependent.

For the empirical part we use annual Danish stock and bond market data (also used in Lund and Engsted (1996), Engsted and Tanggaard (2001), or Nielsen and Sperlich (2003)). We find that the inclusion of predicted bond yields greatly improves the prediction quality of stock returns in terms of the validated R^2. With our best prediction model for one-year stock returns we not only beat the simple historical mean but we also observe a large increase in validated R^2 from 5.9% to 28.3%. To underline our findings, we also include in our empirical analysis the prediction of the ratio of stock returns and dividend yields getting similar results.

The paper proceeds as follows. Section 2 describes the prediction framework and the measure of validation. The mathematical justification is introduced in Section 3. Section 4 presents our findings from an empirical and a small simulation study. Section 5 concludes. Finally, the Appendix contains proofs of our theoretical results.

2. The prediction framework

In the financial and actuarial literature traditional approaches like the classic R^2, the adjusted R^2, goodness-of-fit or testing methods are mainly used to measure in-sample forecasting power. More recently, out-of-sample statistics and tests are discussed, see for example Inoue and Kilian (2004), Clark and West

(2006), Goyal and Welch (2008), or Campbell and Thompson (2008). In our study, we use a generalized version of the *validated* R^2 (R_V^2) of Nielsen and Sperlich (2003) based on leave-k-out cross-validation. It measures how well a model predicts in the future compared to the historical mean. The classical R^2 is often used, easy to calculate and has a straight forward interpretation. But it can hardly be used for prediction nor for comparison issues as it always prefers the most complex model. See also Valkanov (2003) or Dell'Aquila and Ronchetti (2006) for more relevant arguments for disregarding the classical R^2 measure when selecting a model. For comparison often the adjusted R^2 is applied, which penalises complexity via a degree of freedom adjustment. It is well known that this correction does not work in our case, see for example Sperlich et al. (1999).

The idea of the R_V^2 is to replace total variation and not explained variation by their leave-k-out cross-validated analogs. Note that cross-validation (cv) is a quite common in the nonparametric time series context, see Györfi et al. (1990). More formally, consider the two models

$$Y_t = \mu + \varepsilon_t \qquad \text{and} \qquad Y_t = g(X_t) + \zeta_t,$$

where μ is estimated by the sample mean $\bar{Y}$ and the unknown function g by local linear kernel regression. The explanatory variables X_t in the fully nonparametric model can be multidimensional vectors containing information from the past – accounting this way for a possible long term dependence. We suppress a subscription for the chosen smoothing parameter h, since we always apply the bandwidth h that maximises the R_V^2. It is defined as

$$R_V^2 = 1 - \frac{\sum_t \{Y_t - \hat{g}_{-t}\}^2}{\sum_t \{Y_t - \bar{Y}_{-t}\}^2}, \tag{1}$$

where leave-k-out cross-validated values $\hat{g}_{-t}$ and $\bar{Y}_{-t}$ are used, i.e. the function g and the mean $\bar{Y}$ both are computed removing k observations around the t-th point in time. Here we use $k = 1$, the classical leave-one-out cross-validation estimator. We have also checked the robustness of the results for different values of k. We found no significant differences between the validated R_V^2 values for $k \le 7$. Note that

cross-validation with $k = 1$ is not always appropriate in a time series context, as it is a cross-validation estimator of prediction errors. It is well known that cross-validation for time series often requires to leave out more than one data point and possibly requires some extra correction when the leave-out fraction of data is non-negligible (see, for example, Burman et al. (1994)). Burman and Nolan (1992) give simulation evidence and show that under proper dependence conditions even leave-one-out cross-validation is appropriate. Burman and Nolan (1992) further point out that for stationary processes the test and training set need not to be independent for cross-validation to work. Bandi et al. (2016) show in recent work that the cross-validated bandwidth continues to be optimal with respect to the average squared error even when the data generating process is a β-recurrent Markov chain (covering stationary as well as nonstationary processes).

A possible generalisation of our approach would be the use of the Do-validation principle to obtain more robust validated R^2-values. See Mammen et al. (2011) and Gámiz-Pérez et al. (2013) for some recent improvement of the cross-validation approach. Since maximizing the R_V^2 is equivalent to minimize the cross-validation criterion, we also use the R_V^2 to find the optimal prediction bandwidth for the kernel function used (Quartic kernel). Please keep in mind that the choice of the kernel function is automatically corrected by the (adaptive) bandwidth choice, even in small data sets. As the R_V^2 measures how well a given model and estimation principle predicts compared to the historical mean, an $R_V^2 < 0$ indicates that one predicts worse than the historical mean. Note further that cross validation (and thus the R_V^2) punishes not just underfitting but also overfitting (pretending a functional relationship that is not real, see Györfi et al. (1990)). This would result in $R_V^2 < 0$.

In the following, we study excess stock returns defined as

$$S_t = \log\{(P_t + D_t)/P_{t-1}\} - r_{t-1},$$

where D_t denotes the (nominal) dividends paid during year t, P_t the (nominal) stock price at the end of

year t, and r_t the short-term interest rate, which is $r_t = \log(1 + R_t/100)$ using the discount rate R_t. In our article, we concentrate on forecasts over the one-year horizon, but also longer periods can easily be included with $Y_t = \sum_{i=0}^{T-1} S_{t+i}$, the excess stock return at time t over the next T years. However, this would pose greater statistical challenges. Based on the motivations from the introduction, we include the same years bond yield as a single regressor or together with further lagged covariates in the model equation, i.e. we consider the model

$$Y_t = g(\hat{b}_t, v_{t-1}) + \varepsilon_t, \tag{2}$$

with the unknown function g, the constructed bond yield $\hat{b}_t$, a vector of further regressors v_{t-1} and error terms ε_t, i.e. mean zero variables given the past. Note again that $\hat{b}_t$ could be a more general multidimensional regressor which contains elements from the past and accounts for long term dependence. However, in our practical implementation we consider $\hat{b}_t$ to be the constructed bond yield. In addition, we do not assume any explicit distribution on the asset returns. An explicit understanding of this distribution, see for example Eling (2014), could perhaps enhance efficiency of our estimation. Remember, the problem which occurs is that the current bond yield is unknown. Therefore, we must predict them in a prior step, i.e. we construct the bond yield with the fully nonparametric model

$$b_t = p(w_{t-1}) + \zeta_t, \tag{3}$$

where p is an unknown function, w_{t-1} is a vector of explanatory variables as for example, lagged interest rates or bond yields, and ζ_t an error. In our empirical study we find evidence for the use of a first-order auto-regression in model (3), justified by a test for long-run memory that is robust to short-range dependence, see Lo (1991). Both, model (2) and (3), are estimated with a local linear kernel smoother using cross validation. For the choice of the bandwidth, we basically have two possibilities. Either we treat each model separately, determining first the best (in terms of R_V^2) bond model and using this in the second step, or we choose the bandwidth in both steps according to the best R_V^2 for the stock return

8

prediction.

As discussed, not only economic intuition motivates the inclusion of the constructed bond yields but also statistical arguments. In Theorem 3.7 we will develop the mathematical justification for the use of constructed variables in the case of dependent data. In other words, when we estimate nonparametrically stock returns using a generated regressor, we asymptotically obtain the same function as if we had observed the real bond yield. Basically, the bias of the final estimate is enlarged by an additive factor which is proportional to the bias of the predicted variable from the first step. A similar relationship holds for the variance which is increased by an additive term proportional to the variance of the constructed regressor. This relates the bond to the stock prediction. For simplification ignore for a moment v_{t-1} in (2), and call the function containing the real bonds $\tilde{g}$. A closer look to the prediction error ε_t gives

$$Y_t - g(\hat{b}_t) = [Y_t - \tilde{g}(b_t)] + [\tilde{g}(b_t) - \tilde{g}(\hat{b}_t)] + [\tilde{g}(\hat{b}_t) - g(\hat{b}_t)] \tag{4}$$

$$\simeq \tilde{\varepsilon}_t + \tilde{g}'(\xi)(b_t - \hat{b}_t). \tag{5}$$

The last term in (4) vanishes as we will see in Theorem 3.7 and the second term can be easily approximated. The gain in our two-step procedure comes now from the fact that the bond in the second term in (5) is quite predictable. We confirm this fact in the empirical part 4.2 (see Table 2). In the same vain, Lin et al. (2014) find that bond returns are more predictable then stock returns. Another idea would be the following: first, estimate g with the available bond data b_{t-1}, and second, evaluate $\hat{g}$ at the constructed $\hat{b}_t$. Since, however, this procedure did not improve the stock forecasts, we skip it from further considerations.

One could directly use the variables in the vector w_{t-1} as regressors in model (2). But the model would suffer from complexity and dimensionality in several aspects: The dimension of the covariates as well as their interplay. In the nonparametric literature, typically two strategies are proposed to circumvent these problems–either semiparametric modeling or additivity, both to import structure. Nielsen and Sperlich

9

(2003) showed that additive models fail to improve the prediction of stock returns due to a non-ignorable interaction between the predictors. We improve these results by providing additional structure which is inherit by the underlying data generating process. We think of the same years bond yield as an important factor which captures some of the relevant features for the expected stock returns. Then, the inclusion of bond yields when predicting stock returns nonparametrically acts as a kind of complexity and dimension reduction due to the import of more structure.

To see if it is possible to further improve the predictive power in our setting, we will also analyse the model (2) with a different dependent variable. We consider the ratio between current stock returns and dividend yield, i.e. $Y_t^* = Y_t/d_t$ (see Section 4).

3. Mathematical justification

We prove the consistency of a function estimate which makes use of constructed variables and derive its asymptotic properties. For the prediction in the time series context, we follow the steps from Ferraty et al. (2001) and combine them with Sperlich (2009)[2]. Let us consider a sample of real random variables $\{(X_i, Y_i), i = 1, \ldots, n\}$ which are not necessarily independent and want to estimate the unknown function $m(x) = E(Y|X = x)$, $x \in \mathbb{R}$, that should always exist. Note that for time series $\{(Z_i), i \in \mathbb{N}\}$ a k-step ahead forecast is included in a natural way setting $Y_i = Z_{i+k}$ and $X_i = Z_i$. We concentrate only on the case of an auto-regression function of order one. Since we face constructed realisations for X, we assume a predictor[3] with an additive bias and a stochastic error:

$$\hat{x} = x + b(x) + u(x) \tag{6}$$

---

[2]For more technical details of the proofs we refer to the appendix.

[3]We don't specify a particular one but we will need some assumptions on it. The used sample of size N, for instance, consists of some instruments $\boldsymbol{Z} \in \mathbb{R}^\delta$. In the following, we have N=n, since we use the same series in both the prediction and final step.

uniformly, where $u(x) := u \cdot \sigma_u(x)$. The independent random variables u are normalised versions of $u(x)$. The conditional variance at point x is $\sigma_u(x)$. This is rather general as it holds for almost all common predictors. For technical reasons, we further assume finite higher moments for u. Then, for example the Nadaraya-Watson estimator is

$$\hat{m}_{NW}(x) = \frac{\hat{q}(x)}{\hat{f}(x)}, \tag{7}$$

$$\text{with} \quad \hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\hat{X}_i - x}{h}\right) \quad \text{and} \quad \hat{q}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{\hat{X}_i - x}{h}\right)$$

where K denotes some kernel function with bandwidth h.

To measure the strength of dependence in the time series, we limit us to the *strong-* or α-*mixing*[4] defined in Doukhan (1994) or Fan and Yao (2003) as $\lim_{n \to \infty} \alpha(n) = 0$, for the *mixing coefficient*

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_n^\infty} |P(A)P(B) - P(AB)|,$$

where $\mathcal{F}_i^j$ is the σ-algebra generated by $\{X_k, i \le k \le j\}$. We further assume that the sequence $\{(X_i, Y_i), i = 1, \ldots, n\}$ is *algebraic* α-mixing, i.e. that for some real constants $a, c > 0$ we have $\alpha(n) \le cn^{-a}$. To get the asymptotic properties in the context of strong mixing, we make use of an exponential inequality of the Fuk-Nagaev type, cf. Rio (2000).

Lemma 3.1. *For an algebraic α-mixing sequence of random variables $\{(Z_i), i \in \mathbb{N}\}$, with $s_n^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} |\text{cov}(Z_i, Z_j)|$ and $||Z_i||_\infty < \infty$ for all i, holds for some $\varepsilon > 0$ and $r > 1$*

$$P\left(\left|\sum_{i=1}^{n} Z_i\right| > 4\varepsilon\right) \le 4\left(1 + \frac{\varepsilon^2}{rs_n^2}\right)^{-\frac{r}{2}} + 2ncr^{-1}\left(\frac{2r}{\varepsilon}\right)^{a+1}.$$

Furthermore, we need the Billingsley inequality from Bosq (1998) to bound from above the covariance of two elements of a strong-mixing time series.

Lemma 3.2. *For an α-mixing sequence of random variables $\{(Z_i), i \in \mathbb{N}\}$, with $||Z_i||_\infty < \infty$ for all $i \ne j$, holds $|\text{cov}(Z_i, Z_j)| \le 4||Z_i||_\infty ||Z_j||_\infty \alpha(|i - j|)$.*

To prove the asymptotic behaviour of the kernel regression estimator (7), we make some common

---

[4]The weakest of the usually defined mixing conditions.

assumptions. As noted above, we analyse an algebraic α-mixing sequence of real random variables $\{(X_i, Y_i), i = 1, \ldots, n\}$. We suppose that for all $i \neq j$ the joint density f_{ij} for the pair (X_i, X_j) exists and that $|Y| < C < \infty$ almost surely. Also for the unobservables X_i, we assume a density function f_X which is bounded and has a continuous second derivative. At the fix $x \in \mathbb{R}$ we suppose $f_X(x) > 0$. Let the kernel K be integrable, bounded, with compact support and continuous second derivative. It fulfils $\int K(s)ds = 1$ and $\int sK(s)ds = \int K'(s)ds = \int K''(s)ds = 0$. For both, the deterministic and the stochastic part of the predicted realisations $\hat{x}$ in (6), we assume that $b(x)$ and $b'(x)$ are at least of order $O(h_0^2)$ uniformly, and $\sigma_u^2(x)$ of order $O((nh_0^\delta)^{-1})$. Here, h_0 is a smoothing parameter tending to zero when the sample size n goes to infinity and δ refers to the dimension of the used instruments in the prediction step. Let further be $b(\cdot)$ and $\sigma_u(\cdot)$ Lipschitz-continuous. To simplify our calculations, we further suppose that $h_0^2 h^{-1}$ and $(nh_0^\delta h)^{-1}$ go to zero, and use the usual assumption that nh and nh_0^δ go to infinity as $n \to \infty$.

Before we state the main result of the section, we collect some important facts. First, we define the following variables for $l \in \{0, 1\}$

$$Z_i = Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right) - E\left[Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right)\right] \tag{8}$$

and analyse the asymptotic behaviour of

$$s_n^{2*} = \sum_{i=1}^n \sum_{j \neq i} |\text{cov}(Z_i, Z_j)|.$$

Proposition 3.3. *Under the above assumptions holds $s_n^{2*} = o(nh) + O(n^2 \alpha(\tilde{\Delta}))$, where $\tilde{\Delta}$ has the same order like the slowest from $\left\{\frac{1}{h \log n}, \frac{nh_0^\delta h}{\log n}, \frac{(nh_0^\delta h)^2 h}{\log n}\right\}$.*

When $(nh_0^\delta)^{-1} = O(h^2)$ the above proposition reduces to $s_n^{2*} = o(nh) + O(n^2 \alpha((h \log n)^{-1}))$ as in the case without any prior prediction, i.e. $b(x) = \sigma_u(x) = 0$.

Proposition 3.4. *Under the given assumptions and if exists an $\varepsilon > 0$ such that*

$$\Delta^{a-1} = O(n^{-1-\varepsilon}), \tag{9}$$

with Δ from $\{h, (nh_0^\delta h)^{-1}, ((nh_0^\delta)^2 h^3)^{-1}\}$, it holds that $s_n^{2*} = o(n\Delta)$.

Proposition 3.5. *Under the assumptions of Proposition 3.4 we have* $\text{var}(Z_1) = O(\Delta)$.

With Δ from Proposition 3.4 we can direct conclude

$$s_n^2 = \sum_{i=1}^n \sum_{j=1}^n |\text{cov}(Z_i, Z_j)| = n \cdot \text{var}(Z_1) + s_n^{2*} = O(n\Delta). \tag{10}$$

Before we specify the result about the convergence of estimator (7) we need

Proposition 3.6. *Under the assumptions and Δ from Proposition 3.4, verifying*

$$c_1 n^{\frac{3-a}{a+1}+\theta} \le \Delta \le c_2 n^{\frac{1}{1-a}-\theta}, \tag{11}$$

with existing c_1, c_2, $\theta > 0$, it holds for ν and $\varepsilon > 0$ with $\psi = g$ or $\psi = f$

$$P\left(\left| E\hat{\psi}(x) - \hat{\psi}(x) \right| > \varepsilon \sqrt{\frac{\log n}{nh^2} \Delta} \right) = O(n^{-1-\nu}). \tag{12}$$

Now we can state the main theorem. For continuous (around x) functions m and f we get the quasi complete convergence[5] of Nadaraya-Watson estimators with constructed regressors.

Theorem 3.7. *Under the above assumptions and* (11)*, it holds quasi completely that* $|\hat{m}_{NW}(x) - m(x)| \longrightarrow 0$.

The extension to the local linear estimator is almost straightforward. For $j = 0, 1, 2$

$$s_j(x) = \sum_{i=1}^n K\left(\frac{\hat{X}_i - x}{h}\right)(\hat{X}_i - x)^j \quad \text{and} \quad t_j(x) = \sum_{i=1}^n K\left(\frac{\hat{X}_i - x}{h}\right)(\hat{X}_i - x)^j Y_i,$$

we can define $\hat{m}_{LL}(x) := (t_0(x)s_2(x) - t_1(x)s_1(x))/(s_0(x)s_2(x) - s_1^2(x))$ what leads to

$$\hat{m}_{LL}(x) = \sum_{i=1}^n C\left(\frac{\hat{X}_i - x}{h}\right) Y_i \Bigg/ \sum_{i=1}^n C\left(\frac{\hat{X}_i - x}{h}\right), \tag{13}$$

with $C\left(\frac{\hat{X}_i - x}{h}\right) = \sum_{j \ne i} K\left(\frac{\hat{X}_i - x}{h}\right)(\hat{X}_j - \hat{X}_i)K\left(\frac{\hat{X}_j - x}{h}\right)(\hat{X}_j - x)$ as a discretisation of $C(u) = \int K(u - v)vK(u)u\,du$[6]. Since equation (13) is of the same form like (7) and the kernel C fulfils the same conditions

---

[5]For a sequence of real random variables X_n exists a real random variable X such that for all $\varepsilon > 0$ holds $\sum_{i=1}^\infty P(|X_i - X| > \varepsilon) < \infty$, cf. Serfling (1980).

[6]Note that C is a bimodal kernel. Since it puts more weight to points close to x, except if they are too close, than to points far from x, it is a natural choice in the case of strong mixing data, see Kim et al. (2009).

Table 1: Danish stock and bond market data (1923-1996)

	Min	Max	Mean	Sd
CSE Stock Price Index	64.78	3177.88	511.07	662.99
Dividend Accruing to Index	2.99	44.76	15.11	11.81
Excess Stock Returns	-42.44	72.10	2.10	17.19
Bond Yield	-13.70	60.30	8.61	12.07
Dividend Yield	0.01	0.08	0.04	0.01
Long-term Interest Rate	3.80	19.45	8.24	4.29
Short-term Interest Rate	2.50	17.86	6.96	3.46

as K, the application of Theorem 3.7 yields

Corollary 3.8. *Under the assumptions of Theorem 3.7, it holds quasi completely that* $|\hat{m}_{LL}(x) - m(x)| \longrightarrow 0$.

For mean square convergence, asymptotic normality and higher order polynomials, one could directly extend the work of Masry and Fan (1997) to the case of predicted regressors.

4. Empirical evidence and simulation studies

We interpret our method presented as a two stage regression approach. Based on the idea that the bond of the same year captures an important part of the stock return we search in the first step the optimal prediction model for the bond. Afterwards, as we have seen in Theorem 3.7, we can consistently predict stock returns using the predicted bond yields.

4.1. Data description

Consider the annual Danish stock and bond market data for the period 1923 – 1996 from Lund and Engsted (1996). In the appendix of their work, a detailed description of the data can be found. We use a stock index based on a value weighted portfolio of individual stocks chosen to obtain maximum coverage of the market index of the Copenhagen Stock Exchange (CSE). Notice that the CSE was open during the second world war. When constructing the data, corrections were made for stock splits and new equity issues below market prices. Table 1 presents summary statistics of the available variables. In

14

Table 2: R_V^2-values (in percent) for bond model (3)

w_{t-1}	d	S	L	r	b	d,L	d,r	S,L	S,r	L,b	r,b	S,L,b	S,r,b	L,r,b
par.	3.2	11.6	24.0	22.3	-1.3	21.9	19.4	31.9	33.1	29.2	35.2	31.9	37.4	30.9
nonpar.	5.3	16.3	24.0	26.8	-1.2	23.2	19.4	31.9	33.1	29.2	35.5	31.9	37.4	31.8

the following, we use the dividend price ratio, d, the stock return, S, the long-term interest rate, L, the short-term interest rate, r, and the bond yields, b, as explanatory variables.

4.2. The prior step: A simple bond yield predictor

We speak of a simple predictor as in the literature quite complex models can be found for this problem. Our main target, however, are the stock returns where bond yield prediction is just an auxiliary step in order to reduce complexity and dimension. Therefore, the model and bandwidth selection for (3) has to be based on the objective of maximising the R_V^2 of the stock return problem (2). Recognising that the model that maximises the R_V^2 for bond prediction is not necessarily the one that maximises the R_V^2 for stock returns, it becomes clear that it is worth to consider nonparametric alternatives for (3), even if parametric models seem to do a very good job for bond yield prediction alone. This is the reason why we need Theorem 3.7; for parametric predictors $\hat{x}$ the consistency of (7) follows trivially.

If we just look at the bond yield prediction, then we get mostly positive R_V^2 for the models listed in Table 2. We observe that only in few cases a local linear predictor does a better job than a linear model as far as we look at the R_V^2 for bond yields. The interesting numbers, however, we will see only when looking at the R_V^2 for stock returns in Table 3, next section.

Figure 1 shows the estimated functions $\hat{p}$ for the bond yield prediction step with a single covariate w_{t-1} from the set $\{S, r, d, L\}$ using a linear model (triangles) and a fully nonparametric model (diamonds). For some of the models a clear nonlinear behavior can be observed. Figure 2 displays the estimation results of the combination of the variables r, b that gives the largest validated R^2 value for 2-dim. models in Table 2, again for the linear model (triangles) and the fully nonparametric model (diamonds). Note that we set one variable at a certain level (25%, 50%, 75% quantile) and plot the relationship of bond

15
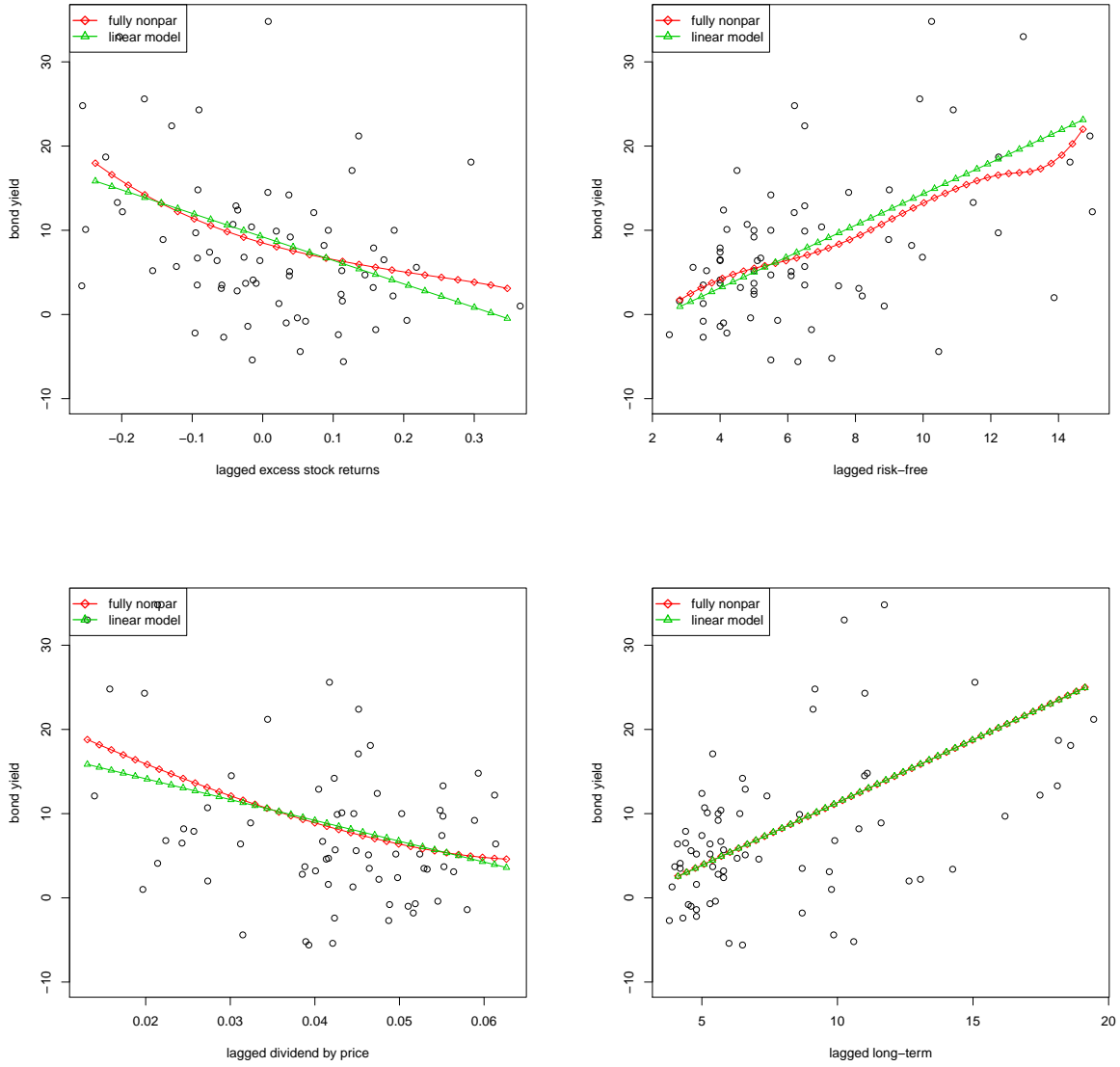
Figure 1: Estimated function $\hat{p}$ for bond yield prediction with different covariates w_{t-1}. Upper panel: excess stock returns, risk-free rate; Lower panel: dividend by price, long-term interest rate; Estimated with: linear model (triangles), fully nonparametric model (diamonds)

yields with the remaining variable. For example, on the left-hand side of Fig. 2, we set the lagged bond yield at values of 2.0, 5.7, and 12.2. The linear model and the fully nonparametric model behave very similar (what is not surprising, since both have more or less the same validated R^2 value). Only at the boundaries a clear difference of both models is visible. Note again that we are interested in stock return prediction and that the predicted bond used in the final step not necessarily has to be the best possible one.

16
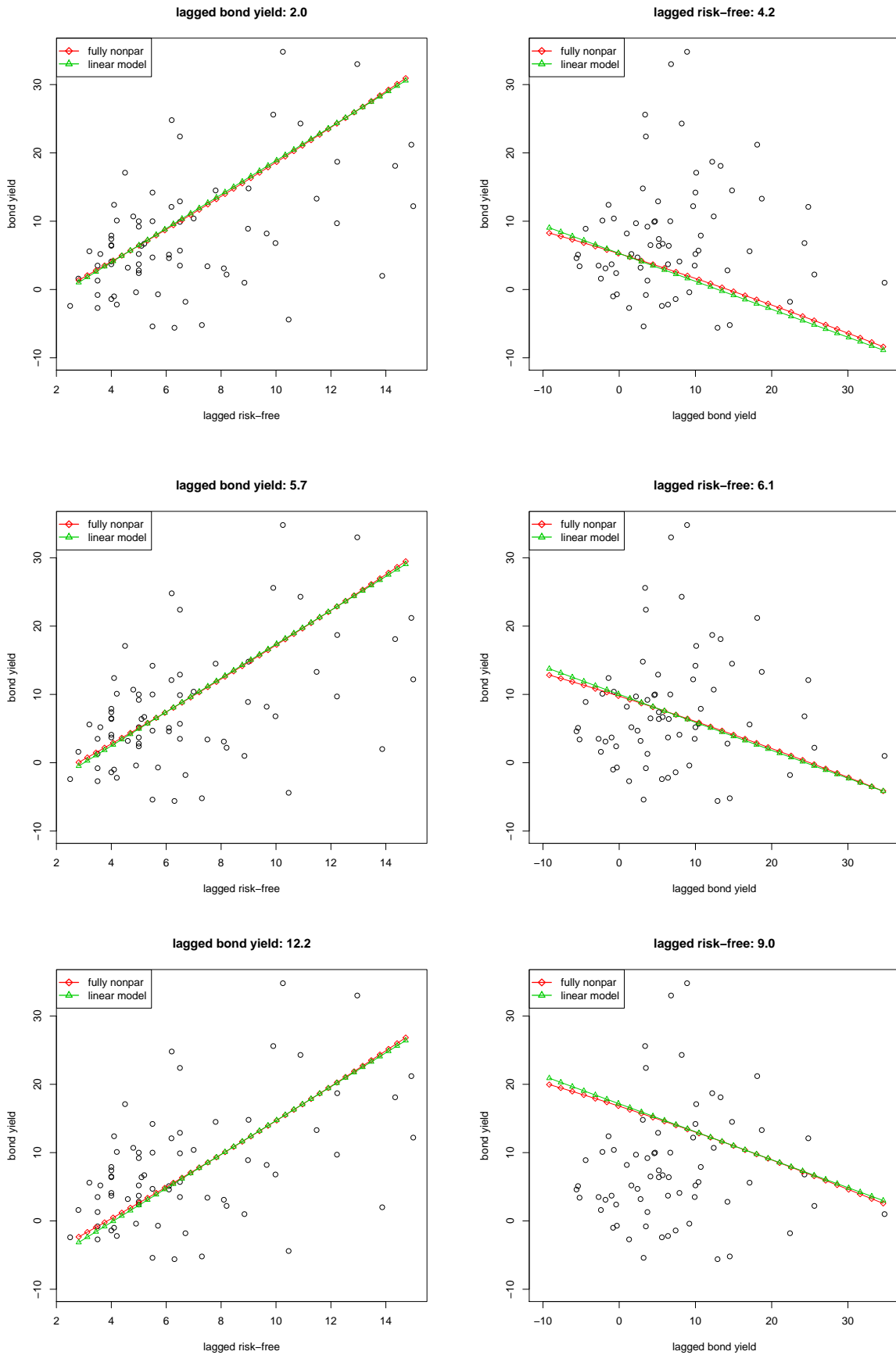
Figure 2: Estimated function $\hat{p}$ for bond yield prediction with lagged bond yield and risk-free rate as covariates w_{t-1}; Left: at 25%, 50%, 75% quantile level for lagged bond yield, Right: at 25%, 50%, 75% quantile level for lagged risk-free rate; Estimated with: linear model (triangles), fully nonparametric model (diamonds)

4.3. Stock prediction

Now we examine the predictive relationship of excess stock returns Y_t and a set of financial variables v_{t-1} using different models. Results of this exercise are summarised in Table 3. First, for the sake of illustration, we develop our strategy step by step and start with the estimation of the model $Y_t = g(v_{t-1}) + \varepsilon_t$ with a fully nonparametric kernel based method as well as the parametric counterpart (not including the constructed bond yield $\hat{b}_t$). Part (a) of Table 3 reports the results and shows that all parametric models produce negative validated R_V^2 values. It means that with a linear regression approach we cannot better forecast one-year stock returns than the simple mean. A more sophisticated technique is needed. In fact, our so far best nonparametric model[7] uses actual lagged bond yields, b_{t-1}, and gives an R_V^2 of 5.9%. But even better results are possible when we include the generated bond yield $\hat{b}_t$ in our analysis.

Second, we follow our procedure proposed in Section 2 and generate the current bond yield with model (3). Then we include this constructed variable as a regressor in the final step, the model for excess stock returns as stated in equation (2). Let us do this first without any further regressor v_{t-1}. As discussed before, we have to choose the model and bandwidths along the largest R_V^2 value for predicting stock returns[8]. How much the predictive power has increased by this method can be seen when comparing part (a) with part (b) of Table 3. The best model in (b) uses as single regressor lagged bond yields in the first step and only the predicted bond as covariate in the second step (R_V^2 of 10.6%). Even for the parametric counterpart our strategy helps to improve prediction power since we can observe positive R_V^2 for some models. As one can clearly see, the nonparametric version produces better results, recall our discussion in the previous section.

---

[7]Nielsen and Sperlich (2003) report in an analog setting a R_V^2 value of 5.5 for a fully nonparametric two-dimensional model with dividend-price ratio, d_{t-1}, and lagged excess stock returns, S_{t-1}, as explanatory variables, but don't use bond yields in their analysis.

[8]If one chooses for example the bandwidth that predicts best bond yields in the prior step, then the values in Table 3 will shrink or remain the same. The best prediction model will still be $g(\hat{b}_t)$ but with an R_V^2 value of 9.0% instead of 10.6%.

Table 3: R_V^2-values (in percent) for stock model (2)

v_{t-1}	d	S	L	r	b	d,S	d,L	d,r	d,b	S,L	S,r	S,b	L,r	L,b	r,b
(a) $Y_t = g(v_{t-1}) + \varepsilon_t$															
nonpar	-1.4	1.8	-4.2	-3.6	5.9	5.5	-6.0	-7.4	3.1	-3.5	-7.1	4.6	-9.4	0.8	0.5
par	-1.3	-1.8	-4.2	-5.7	-4.0	-3.5	-5.8	-7.2	-6.2	-6.8	-7.9	-6.6	-9.3	-7.5	-8.6
(b) $Y_t = g(\hat{b}_t) + \varepsilon_t$ with $b_t = p(v_{t-1}) + \zeta_t$															
nonpar	10.0	1.3	-3.5	1.4	10.6	-1.5	-3.8	2.9	10.1	-1.1	-3.1	3.9	-0.6	-0.9	-3.6
par	10.0	-2.6	-4.0	-4.2	-11.0	-4.2	-3.8	-4.1	-3.5	-3.7	-3.9	3.7	-3.7	-3.3	-3.5
(c) $Y_t = g(\hat{b}_t, v_{t-1}) + \varepsilon_t$ with $b_t = p(v_{t-1}) + \zeta_t$															
nonpar	16.4	5.1	9.1	16.3	8.9	-1.6	28.3	21.6	10.2	1.6	13.5	-1.3	15.8	15.6	20.3
par	16.4	-23.7	3.3	-0.7	5.1	-25.2	12.0	17.3	10.2	-1.5	7.9	-13.4	-6.1	7.7	12.3

v_{t-1}	d,S,L	d,S,r	d,S,b	d,L,r	d,L,b	d,r,b	S,L,r	S,L,b	L,r,b	S,r,b
(a) $Y_t = g(v_{t-1}) + \varepsilon_t$										
nonpar	-2.9	-6.7	3.3	-11.2	-3.8	1.0	-11.0	0.3	-4.4	-1.6
par	-8.6	-9.8	-8.8	-10.9	-9.9	-11.2	-12.5	-11.1	-13.0	-11.9
(b) $Y_t = g(\hat{b}_t) + \varepsilon_t$ with $b_t = p(v_{t-1}) + \zeta_t$										
nonpar	1.3	-3.5	8.9	-2.8	-1.0	-3.7	-2.1	1.8	-3.6	1.6
par	-3.3	-3.8	-4.7	-3.4	-3.0	-3.4	-3.9	-3.6	-3.4	-3.5
(c) $Y_t = g(\hat{b}_t, v_{t-1}) + \varepsilon_t$ with $b_t = p(v_{t-1}) + \zeta_t$										
nonpar	10.8	14.2	0.8	17.5	16.6	20.4	10.0	1.6	15.9	7.6
par	-1.2	5.7	-15.5	-2.5	11.4	18.0	-1.7	-3.6	-8.0	7.6

Performance of nonparametric and parametric estimators for excess stock return predictions on different covariates: (a) Without constructed bond $\hat{b}_t$; (b) With $\hat{b}_t$ as unique regressor, constructed with variables v_{t-1}; (c) With $\hat{b}_t$ and the same variables v_{t-1} as in the first bond prediction step. Bandwidth choice in the final step.

Third, we construct the current bond as before but accompany this regressor in model (2) by any combination of lagged variables from the predictor set $\{d, S, L, r, b\}$ as our vector v_{t-1}. Then, the two largest R_V^2 were achieved by $\hat{g}(\hat{b}_t, d_{t-1}, S_{t-1}, L_{t-1})$ where $\hat{b}_t = \hat{p}(d_{t-1})$ (yielding $R_V^2 = 30.3\%$) or $\hat{b}_t = \hat{p}(d_{t-1}, L_{t-1})$ (yielding $R_V^2 = 28.9\%$), respectively. Note that for an increasing set of regressor variables the corresponding multidimensional bandwidth grid on which we looked for the best predicting one had to be reduced for numerical reasons. Consequently, lower dimensional models have the tendency to be slightly favoured in our study. The full set of results for the 25 times 25 combinations of $\{d, S, L, r, b\}$ is not shown for the sake of presentation, but available on request. We include in part (c) of Table 3

only the 'diagonal' of those results since the predictive power is among the best of all possible models. In other words, in part (c) of Table 3 holds $w_{t-1} = v_{t-1}$, exactly the same regressors used for the bond construction in step one accompany $\hat{b}_t$ again in the second step. For completeness, also the results of the parametric counterpart are included in part (c) of Table 3. We see that our new prediction procedure improves the predictive power for stock returns. We find again convincing evidence that the two nonparametric steps are better than the parametric counterpart. For the best model in Table 3 – we have $w_{t-1} = v_{t-1} = (d_{t-1}, L_{t-1})$ – we find a large increase in the R_V^2 value from 5.9% to 28.3%, an about factor five improvement compared to the best model without constructed bonds. This finding again indicates that the bond captures a quite important part of the stock return which is related to the change in long-term interest rate.

Figure 3 shows the estimated functions $\hat{g}$ for the excess stock return predictions based on a single covariate from the set $\{S, r, d, b\}$ using a fully nonparametric model (diamonds), a fully nonparametric model with the constructed bond as single regressor in the second step (crosses), and a fully nonparametric model based on the predicted bond together with the regressor of the first step (pluses). Again, for some of the models a clear nonlinear pattern can be observed. Figure 4 displays the estimation results of the combination of the variables d, L that gives the largest validated R^2 value in Table 3, again for three different models used in Figure 3. Note that we set one variable at a certain level (25%, 50%, 75% quantile) and plot the relationship of excess stock returns with the remaining variable. For example, on the left-hand side of Fig. 4, we set the lagged long-term interest rate at values of 5.1, 6.4, and 10.6. It seems that the model which uses only the predicted bond as a covariate is too inflexible in its functional form and needs the additional information which is still inherent in the covariates of the first step when it comes to stock return prediction in step 2. This underlines the findings of Table 3 where the largest validated R^2 values can be found in part (c) for the nonparametric estimators of the models based on the predicted bond together with the same covariates of the first step.
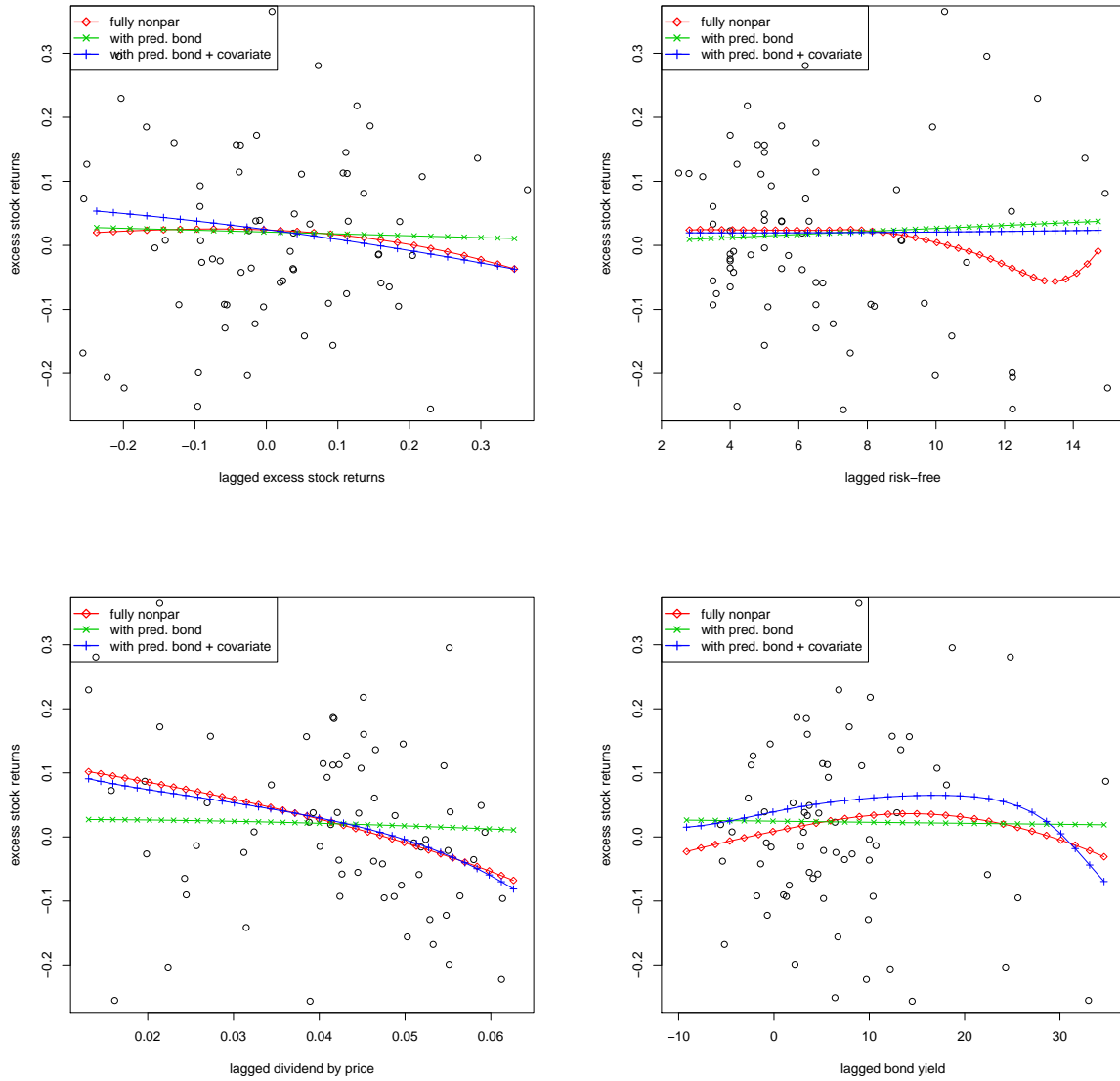
Figure 3: Estimated function $\hat{g}$ for excess stock return prediction with different covariates. Upper panel: excess stock returns, risk-free rate; Lower panel: dividend by price, bond yield; Estimated with: fully nonparametric model (diamonds), predicted bond as single regressor in the second step (crosses), predicted bond together with the regressor of the first step (pluses)

21

22
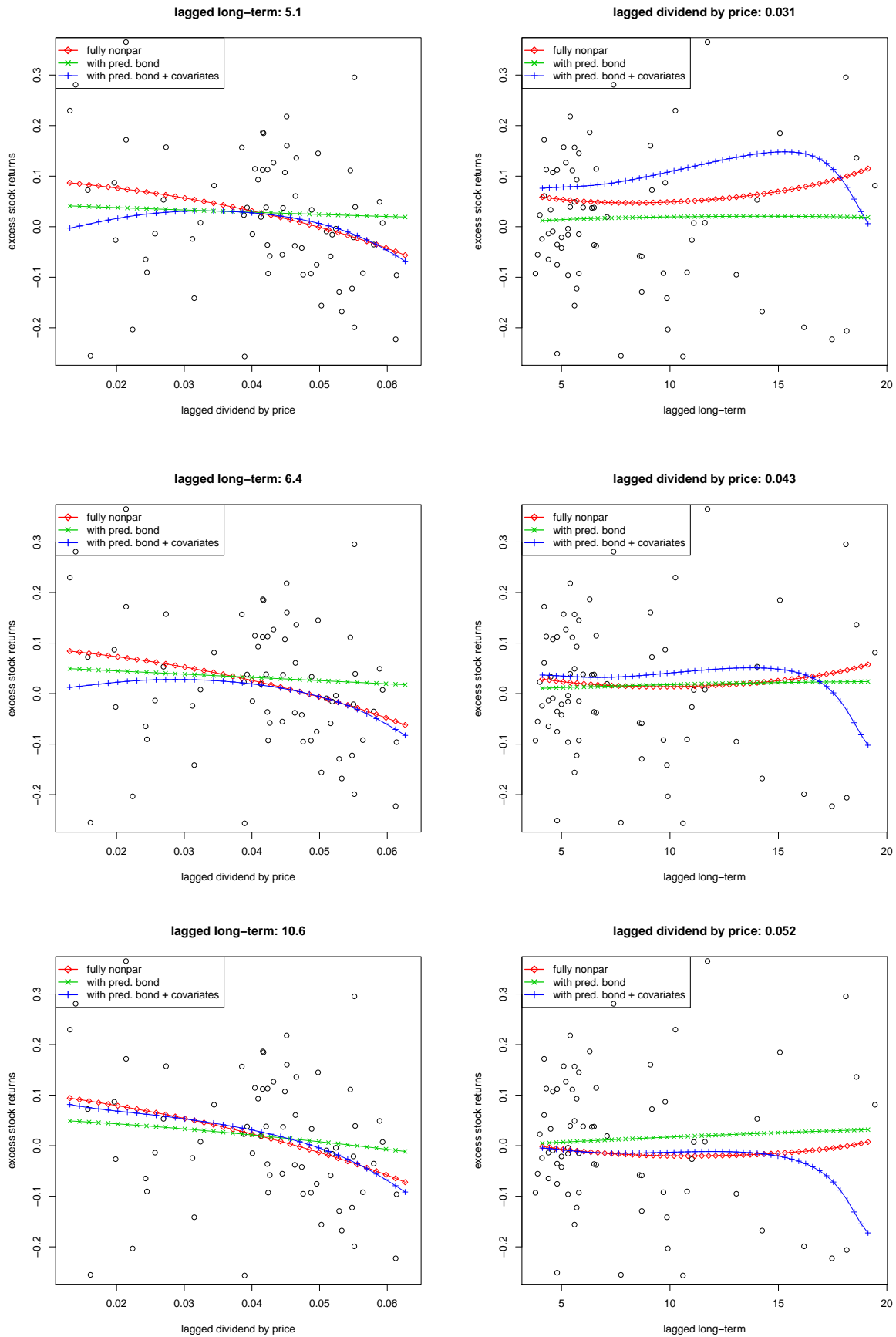
Figure 4: Estimated function $\hat{g}$ for excess stock return prediction with lagged dividend by price and long-term interest rate as covariates; Left: at 25%, 50%, 75% quantile level for lagged long-term interest rate, Right: at 25%, 50%, 75% quantile level for lagged dividend by price; Estimated with: fully nonparametric model (diamonds), predicted bond as single regressor in the second step (crosses), predicted bond together with dividend by price and long-term interest rate in second step (pluses)

In order to get a better view of the potential of our proposed method, we analyze the out-of-sample mean-squared error (oos-mse) for a one-step ahead prediction with an expanding estimation sample for the new method in comparison to the oos-mse of the corresponding fully nonparametric and linear models as well as the historical mean. For an illustration we use dividend by price and long-term interest rate since we observed for this combination of covariates the largest R^2_V in Table 3. Figure 5 shows the predicted annual excess stock returns of the different models in comparison to the realized annual excess stock returns of the CSE. We observe the smallest oos-mse for the new method (0.044), followed by the historical mean (0.049), the linear model (0.058), and the fully nonparametric model (0.059).

The last part of our empirical study concentrates on the change of the dependent variable. Up to now, we used the excess stock return but for the following we divide this value by the dividend yield, i.e. we use $Y_t^* = Y_t/d_t$. Table 4 summarises our findings for Y^*.

Table 4: R^2_V-values (in percent) for model (2) with $Y_t^* = Y_t/d_t$ as dependent variable.

w_{t-1}	d, L	d, r	L, r	L, b	r, b	d, L, r	d, r, b	L, r, b
par.	-8.1	-10.4	-11.2	-4.5	-6.8	-15.0	-11.2	-12.7
nonpar.	-8.3	-10.5	-11.3	1.3	11.4	-15.2	2.4	12.3
final with $v = w$	39.9	35.4	41.4	31.3	35.8	39.5	42.6	43.5

First line: fully parametric model, second line: nonparametric estimation without constructed bond, third line: nonparametric with constructed bond and bandwidth selection in the final step.

The first line refers again to the parametric version of model (2) and the second line to the fully nonparametric method, both without constructed bonds. Almost all of the parametric models have negative R^2_V values and also only a small number of nonparametric models beat the simple mean. In contrast, when we include the constructed bond in the nonparametric prediction, a large increase of the validated R^2_V can be observed. For example, the model which uses long- and short-term interest rate, and lagged bond yields for both the bond generation and following stock prediction, has a R^2_V value (43.5%) that is over three and a half times larger than the value of the best model without constructed bonds (12.3%).
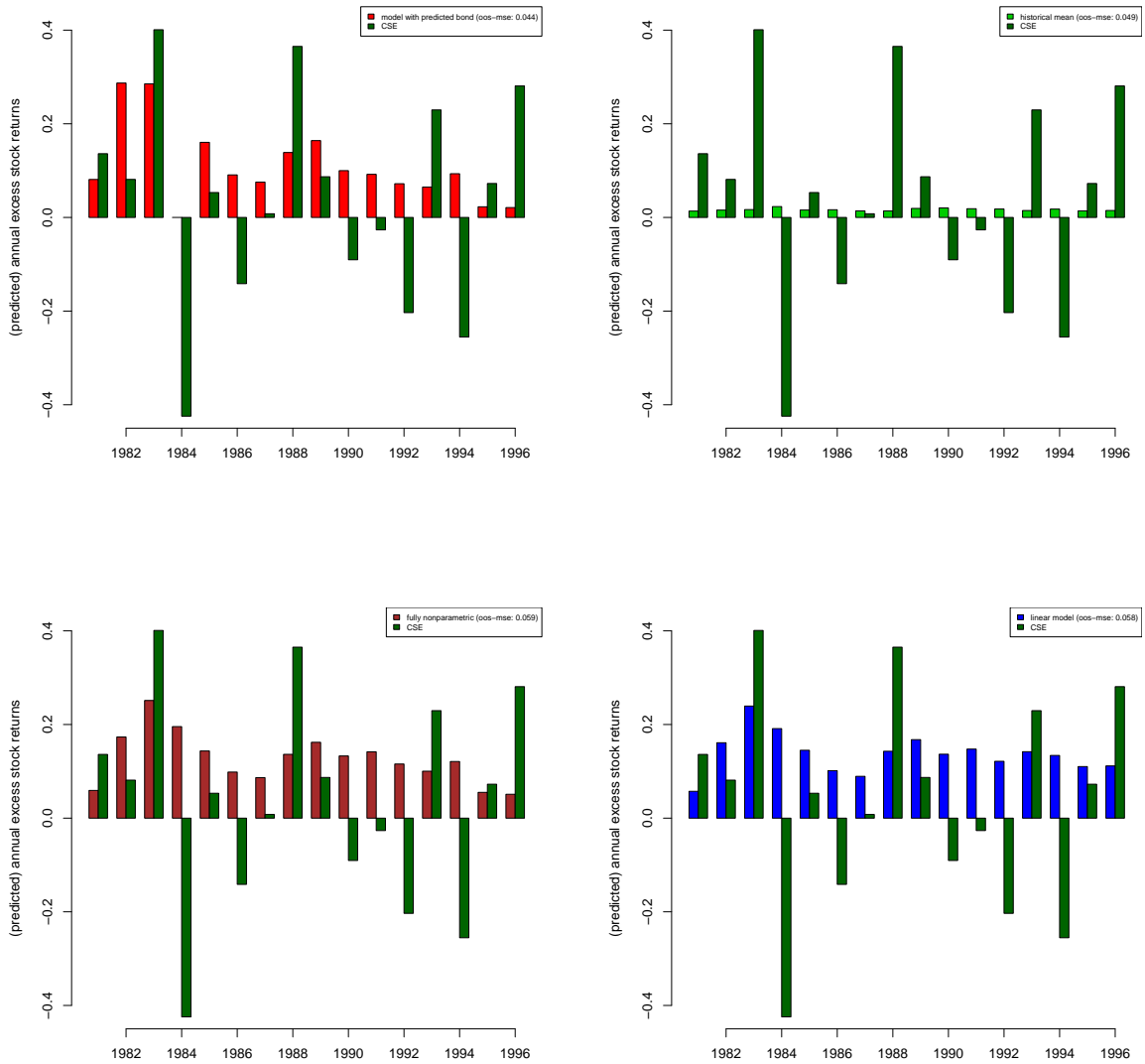
Figure 5: Predicted annual excess stock returns of different models using dividend by price and long-term interest rate as covariates in comparison to the realized annual excess stock returns of the CSE (upper left: new method, upper right: historical mean, lower left: fully nonparametric model, lower right: linear model)

Table 5: R_V^2-values (in percent) for dimension reduction.

σ	0.3	0.6	0.9	0.3	0.6	0.9
s_1	57.3 (10.4)	23.7 (12.7)	10.5 (11.5)	49.2 (6.2)	19.2 (5.6)	9.2 (4.4)
s_2	88.1 (3.5)	66.4 (8.7)	47.0 (11.8)	85.2 (2.9)	59.6 (5.9)	39.7 (6.7)
2 step method	89.1 (3.3)	68.1 (8.5)	47.9 (11.7)	86.7 (2.4)	62.2 (5.7)	42.0 (6.6)
fully nonpar.	45.8 (11.7)	35.0 (12.9)	24.3 (13.5)	47.6 (6.6)	34.2 (6.5)	23.0 (6.1)

Note: For simplicity's sake we use $\sigma_{s_1} = \sigma_{s_2} = \sigma_m$ and refer to it as σ. Averaged values over 500 simulation runs with standard errors in brackets. Left panel: n=50, right panel: n=200.

4.4. Simulation studies

A simulation study gives us the possibility to highlight the potential of our method. We first show the effects of a dimension reduction and afterwards of a pronounced curvature.

Let us consider a four dimensional function that is separable into two terms: $m(x_1, \ldots, x_4) = \tilde{m}(s_1, s_2)$ with $s_1 = s_1(x_1, x_2)$ and $s_2 = s_2(x_3, x_4)$. We simulated data from the following models: $S_1 = x_1 + x_1 x_2 + \varepsilon_{\sigma_{s_1}}$, $S_2 = \exp(x_3 + x_4) + \varepsilon_{\sigma_{s_2}}$, and $Y = m(x_1, \ldots, x_4) + \varepsilon_{\sigma_m} = \tilde{m}(s_1, s_2) + \varepsilon_{\sigma_m} = s_1 + s_2 + \varepsilon_{\sigma_m}$. For each explanatory the support is $[0, 1]$. An autoregressive design with $\phi = 0.75, 0.2, 0.02$ for $x_1, \ldots, x_3$ was used; also a normal for x_4. Different parameter values σ for the zero mean normal error distributions were investigated as well as different sample sizes n. The kernel used was the Gaussian. For computational reasons the bandwidths are chosen separately in each step of the simulation[9]. In step one we predict s_1 and s_2, used in step 2 to estimate function $\tilde{m}$.

Lines three and four of Table 5 present the results for the two-step approach for $\tilde{m}$ and the fully nonparametric method estimating m in terms of R_V^2 values, averaged over 500 runs. The proposed two-step procedure succeed in improving on the fully nonparametric estimator in all cases by far. The effect of the dimension reduction is of course more pronounced for the smaller sample size and results in an almost factor 2 improvement.

---

[9]Note that this is again suboptimal, i.e. a bandwidth choice along the final objective would give even better results for our method than those presented in Table 5.

**Table 6:** R_V^2-values (in percent) for pronounced curvature.

σ_m	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
s	95.7	95.6.	95.7	85.8	85.5	85.9	69.8	70.8	70.2
	(2.1)	(2.3)	(2.0)	(4.5)	(4.4)	(3.7)	(8.1)	(7.0)	(7.1)
2 step method	86.9	51.0	26.9	82.4	48.4	25.0	73.7	43.7	23.5
	(5.1)	(10.8)	(11.6)	(6.9)	(10.8)	(11.6)	(11.4)	(11.6)	(12.1)
fully nonpar.	72.6	38.3	16.1	72.6	39.3	16.2	72.4	39.2	17.0
	(8.7)	(12.5)	(12.4)	(9.1)	(11.8)	(12.4)	(10.3)	(12.2)	(12.1)

Note: Averaged values over 500 simulation runs with standard errors in brackets. Sample size: n=50. Left panel: $\sigma_s = 0.1$, middle panel: $\sigma_s = 0.3$, right panel: $\sigma_s = 0.5$.

For the second part we consider the function composition $m(x) = \tilde{m} \circ s(x)$, where the inner function s has a pronounced curvature. We simulated data from the following models: $S = \sin(4\pi(x - 1/8)) + \cos(4/3 \cdot \pi(x - 1/2)) + 1.6 + \varepsilon_{\sigma_s}$ and $Y = m(x) + \varepsilon_{\sigma_m} = \sin \circ s(x) + \varepsilon_{\sigma_m}$, i.e. $\tilde{m}(x) = \sin(x)$. Note that s is one of the example functions used in Park et al. (1997). A uniform design was used with the support $[0, 1]$. Different parameter values σ for the zero mean normal error distributions were investigated for a sample size of $n = 50$. The kernel used was the Gaussian and the bandwidths are chosen separately in each step for the two-step part. Again we are aware of the suboptimality, i.e. we could even do better with respect to the R_V^2 but at the cost of computing time. Table 6 reports the results. We find that already in this simple example the proposed two-step approach can help to obtain clearly better results, i.e. much larger R_V^2-values in all cases. Figure 6 shows the used inner function (left) and estimates of m and $\tilde{m}$ (right). We see that our method can better estimate problematic regions, in particular by bias reduction.

5. Concluding remarks and outlook

Motivated by economic theory and statistical arguments, we include the same years bond yield in the fully nonparametric prediction approach for excess stock returns. Since the current bond yield is unknown, we propose to construct it in a prior step using again nonparametric techniques. The bandwidths should be chosen in such a way that they maximise the R_V^2 of the final step. The empirical study demonstrates that this two-step approach can improve the stock return prediction enormously. We moreover prove the
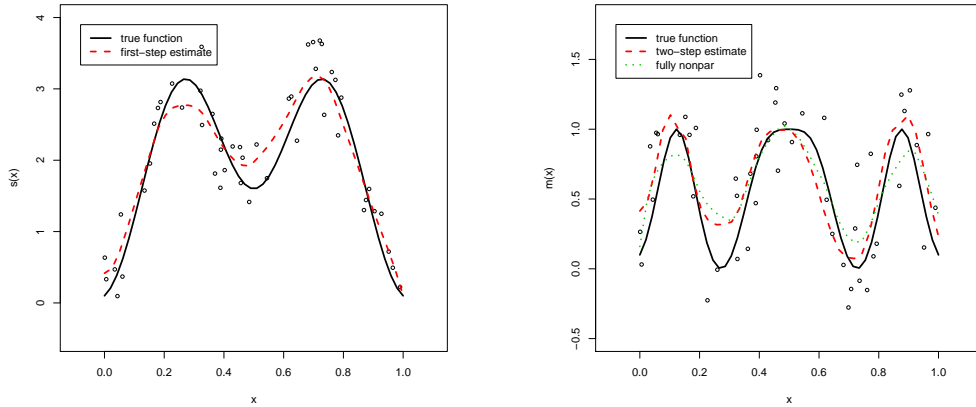
Figure 6: Left: Simulated Data (dots) with $\sigma_s = 0.3$ from inner function s (solid), first-step estimate (dashed); Right: Simulated Data (dots) with $\sigma_m = 0.3$ from function composition m (solid), two-step estimate (dashed), usual nonparametric estimate (dotted)

consistency of our method and derive the asymptotic behaviour of our final predictor. We illustrate the improvement due to our method using annual Danish stock and bond market data which were studied in detail in former articles by different authors. Our results confirm our motivation of including the same years bond yield, namely that it captures the most important part of the stock return, that one related to the change in long-term interest rate. This actually holds not only for stock returns but also for transformed variables, as for example returns divided by dividend yields.

The statistically insights are the following. It is clear that we face a regression model that exhibits high complexity and dimensionality. An obvious remedy would be the imposing of structure. Since it has been shown that additive separability is inappropriate because of unknown interactions, we make use of financial theory to exploit the inherit structure of stock returns. Alternatively, one could interpret the first stage as an optimal nonparametric transformation that maps, for example, the long-term interest rate to the current bond yield, $L_{t-1} \to \hat{b}_t$. The subsequent nonparametric smoother of the transformed variable is than characterised by less bias. Here, we present a practical example in the spirit of the somewhat theoretical method proposed by Park et al. (1997) which improves nonparametric regression

with transformation techniques. Although we extend their method in several aspects, their paper provides some statistical intuition for the success of our approach. Our simulations additionally underpin the key idea of complexity and dimension reduction.

Appendix A. Proofs

Proof of Proposition 3.3. Since the variables Z_i in (8) are centered, we calculate for $i \neq j$

$$|EZ_iZ_j| = \left| EY_i^l Y_j^l K\left(\frac{\hat{X}_i - x}{h}\right) K\left(\frac{\hat{X}_j - x}{h}\right) \right.$$
$$\left. - EY_i^l K\left(\frac{\hat{X}_i - x}{h}\right) EY_j^l K\left(\frac{\hat{X}_j - x}{h}\right) \right|. \tag{A.1}$$

First, we analyse the second term in the last equation and use the assumption that all Y_i are bounded.

$$EY_i^l K\left(\frac{\hat{X}_i - x}{h}\right) \leq C \int \int K\left(\frac{u - x + b(u) + v\sigma(u)}{h}\right) f(u,v)dudv.$$

A simple Taylor-expansion of the kernel leads to

$$C \int \int \left\{ K\left(\frac{u - x}{h}\right) + K'\left(\frac{u - x}{h}\right)\left(\frac{b(u) + v\sigma(u)}{h}\right) + \right.$$
$$\left. K''\left(\frac{u - x}{h} + \kappa \frac{b(u) + v\sigma(u)}{h}\right)\frac{(b(u) + v\sigma(u))^2}{2h^2} \right\} f(u,v)dudv,$$

where $\kappa \in (0,1)$. With the common substitution $s = (u - x)h^{-1}$ we get

$$EY_i^l K\left(\frac{\hat{X}_i - x}{h}\right) = O(h + (nh_0^\delta h)^{-1}).$$

Analog steps lead to

$$E\left[Y_i^l Y_j^l K\left(\frac{\hat{X}_i - x}{h}\right) K\left(\frac{\hat{X}_j - x}{h}\right)\right] = O(h^2 + (nh_0^\delta)^{-1} + (nh_0^\delta h)^{-2}),$$

and thus the covariance $|\text{cov}(Z_i, Z_j)|$ for $i \neq j$ is of the same rate.

On the other hand, we can directly make use of Lemma 3.2 because all Y_i and K are bounded so that $||Z_i||_\infty < \infty$. It follows

$$|\text{cov}(Z_i, Z_j)| \leq C\alpha(|i - j|).$$

The idea is now to combine both results. When the indices of the two variables Z_i and Z_j are close[10] to each other we use the first one, and when they are far from each other the second one. To control this,

---

[10]Since we use time series data, this means that the two events are close in time.

we introduce a sequence of integers a_n and obtain

$$s_n^{2*} = \sum_{i=1}^{n} \sum_{j \neq i} |\text{cov}(Z_i, Z_j)|$$

$$\leq C \Big[\sum \sum_{0 < |i-j| \leq a_n} \{h^2 + (nh_0^\delta)^{-1} + (nh_0^\delta h)^{-2}\} + \sum \sum_{|i-j| > a_n} \alpha(|i-j|) \Big].$$

Since $i \neq j$, the largest possible term for a_n is $0 < |i - j| \leq a_n \cong n - 1$ and the smallest $|i - j| > a_n \cong 1$. Furthermore, the maximum number of elements in s_n^{2*} is $n^2 - n$, and we obtain for Δ that is of the same order like the slowest term in $O(h^2 + (nh_0^\delta)^{-1} + (nh_0^\delta h)^{-2})$

$$\frac{\sum \sum_{0 < |i-j| \leq a_n} \Delta}{\Delta n a_n} \leq \frac{n^2 - n}{n(n-1)} \quad \Longleftrightarrow \quad \sum \sum_{0 < |i-j| \leq a_n} \Delta = O(\Delta n a_n)$$

and

$$\frac{\sum \sum_{|i-j| > a_n} \alpha(|i-j|)}{n^2 \alpha(a_n)} \leq \frac{n^2 - n}{n^2} \quad \Longleftrightarrow \quad \sum \sum_{|i-j| > a_n} \alpha(|i-j|) = O(n^2 \alpha(a_n)).$$

This means that

$$s_n^{2*} = O(\{h^2 + (nh_0^\delta)^{-1} + (nh_0^\delta h)^{-2}\} n a_n + n^2 \alpha(a_n)).$$

Choosing a_n from $\{\frac{1}{h \log n}, \frac{n h_0^\delta h}{\log n}, \frac{(nh_0^\delta)^2 h^3}{\log n}\}$ proves the Proposition. $\qquad \square$

Proof of Proposition 3.4. Using the algebraic mixing condition and Proposition 3.3

$$s_n^{2*} = o(nh) + O(n^2 \tilde{\Delta}^{-a}),$$

with $\tilde{\Delta}$ from

$$\left\{ \frac{1}{h \log n}, \frac{n h_0^\delta h}{\log n}, \frac{(nh_0^\delta h)^2 h}{\log n} \right\}.$$

Using the assumption (9) and noting that $\frac{(\log n)^a}{n^\varepsilon} \to 0$ for $n \to \infty$ closes the proof. $\qquad \square$

Proof of Proposition 3.5. We use again that Y_1 is bounded so that remains to analyse

$$E\Big[K\Big(\frac{\hat{X}_1 - x}{h}\Big)\Big]^2 = \int \int K\Big(\frac{u - x + b(u) + v\sigma(u)}{h}\Big)^2 f(u, v) du dv.$$

With a Taylor-expansion and analog steps like in the proof of Proposition 3.3 we get

$$= \int \int \Big\{ K\Big(\frac{u - x}{h}\Big) + K'\Big(\frac{u - x}{h}\Big)\Big(\frac{b(u) + v\sigma(u)}{h}\Big) +$$

$$K''\Big(\frac{u - x}{h} + \kappa \frac{b(u) + v\sigma(u)}{h}\Big) \frac{(b(u) + v\sigma(u))^2}{2h^2} \Big\}^2 f(u, v) du dv,$$

where $\kappa \in (0,1)$, and find that

$$E\Big[K\Big(\frac{\hat{X}_1 - x}{h}\Big)\Big]^2 = O(h + (nh_0^\delta h)^{-1} + (nh_0^\delta)^2 h^3)^{-1})$$

what proves the Proposition. $\qquad\square$

Proof of Proposition 3.6. Using $l = 0$ for $\psi = f$ and $l = 1$ for $\psi = q$, respectively, we directly get with (8)

$$|E\hat{\psi}(x) - \hat{\psi}(x)| = \left| E\left(\frac{1}{nh}\sum_{i=1}^n Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right)\right) - \frac{1}{nh}\sum_{i=1}^n Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right) \right|$$

$$= \frac{1}{nh}\left| \sum_{i=1}^n \left\{Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right) - E\left(Y_i^l K\left(\frac{\hat{X}_i - x}{h}\right)\right)\right\} \right| = \frac{1}{nh}\left| \sum_{i=1}^n Z_i \right|.$$

Therefore, applying Lemma 3.1 we obtain

$$P\left(|E\hat{\psi}(x) - \hat{\psi}(x)| > \delta\right) = P\left(\left| \sum_{i=1}^n Z_i \right| > nh\delta\right)$$

$$\leq 4\left(1 + \frac{\delta^2 n^2 h^2}{16 r s_n^2}\right)^{-\frac{r}{2}} + 2ncr^{-1}\left(\frac{8r}{nh\delta}\right)^{a+1}.$$

Since we have seen in (10) that $s_n^2 = O(n\Delta)$, with Δ from Proposition 3.4, with $\delta = \varepsilon\sqrt{\frac{\log n}{nh^2}\Delta}$ we get

$$P\left(|E\hat{\psi}(x) - \hat{\psi}(x)| > \varepsilon\sqrt{\frac{\log n}{nh^2}\Delta}\right)$$

$$\leq 4\left(1 + \frac{\varepsilon^2 \log n}{16r}\right)^{-\frac{r}{2}} + 2ncr^{-1}\left(\frac{8r}{\varepsilon}\right)^{a+1}(n\Delta \log n)^{-\frac{a+1}{2}}. \qquad (A.2)$$

Now, we can choose $r > 1$ such that $\log n = o(r)$, and use the limit definition

$$\exp(x) = \lim_{z \to \infty}\left(1 + \frac{x}{z}\right)^z,$$

with $z = -r/2$. For the first term of the right hand side of (A.2), we obtain for $z \to \infty$

$$\left(1 + \frac{\varepsilon^2 \log n}{16r}\right)^{-\frac{r}{2}} = \left(1 - \frac{\varepsilon^2 \log n}{32z}\right)^z \longrightarrow \exp\left(-\frac{\varepsilon^2 \log n}{32}\right) = n^{-\frac{\varepsilon^2}{32}}.$$

Noting that $C(\log n)^{-(a+1)/2} \leq C$ for $n > 2$ and a constant C, (A.2) can be expressed as

$$P\left(|E\hat{\psi}(x) - \hat{\psi}(x)| > \varepsilon\sqrt{\frac{\log n}{nh^2}\Delta}\right) \qquad \leq Cn^{-\frac{\varepsilon^2}{32}} + C\varepsilon^{-(a+1)}n^{1-\frac{a+1}{2}}r^a \Delta^{-\frac{a+1}{2}}.$$

With $r = n^b$ for $b > 0$, i.e. $\log n = o(r)$, and the left hand side of the assumption (11),

$$n^{1+ab-\frac{a+1}{2}}\Delta^{-\frac{a+1}{2}} \le n^{1+ab-\frac{a+1}{2}-\frac{3-a}{2}-\theta\frac{a+1}{2}} = n^{-1-\theta\frac{a+1}{2}+ab} = n^{-1-\nu}.$$

Thus, we obtain for a sufficiently small b that

$$P\Big(|E\hat{\psi}(x) - \hat{\psi}(x)| > \varepsilon\sqrt{\frac{\log n}{nh^2}}\,\Delta\Big) \le Cn^{-\frac{\varepsilon^2}{32}} + C\varepsilon^{-(a+1)}n^{-1-\nu}.$$

Finally, for a sufficiently large ε, we get that exist $\nu, \varepsilon > 0$ such that

$$P\Big(|E\hat{\psi}(x) - \hat{\psi}(x)| > \varepsilon\sqrt{\frac{\log n}{nh^2}}\,\Delta\Big) \le Cn^{-1-\nu},$$

what proves the assertion. $\qquad\qquad\square$

Proof of Theorem 3.7. From Proposition 3.6 follows directly

$$E\hat{q}(x) - \hat{q}(x) \longrightarrow 0, \quad \text{and} \quad E\hat{f}(x) - \hat{f}(x) \longrightarrow 0, \qquad\qquad (A.3)$$

both quasi completely. With the first part of the proof of Proposition 3.3 we obtain[11]

$$E\hat{f}(x) = \frac{1}{h}EK\left(\frac{\hat{X}-x}{h}\right) = f(x) + B_f(x) + o(h_0^2 + h),$$

with $B_f(x) = h^2/2f''(x)\mu_2(K) + \{b(x)f'(x) + b'(x)f(x)\}\mu_1(K')$, and thus

$$E\hat{f}(x) - f(x) \longrightarrow 0. \qquad\qquad (A.4)$$

The analog can be shown for $E\hat{q}(x)$. With

$$E\hat{q}(x) = \frac{1}{h}EYK\left(\frac{\hat{X}-x}{h}\right),$$

and taking the conditional expectation for $X = x$, we get

$$E\hat{q}(x) = \frac{1}{h}\int\int m(u)K\left(\frac{u-x+b(u)+v\sigma(u)}{h}\right)f(u,v)dudv.$$

Repeating the same steps as in the first part of the proof of Proposition 3.3, using $q = m \cdot f$ as well as that the function q is continuous over the compact support of the kernel K, i.e. that $q(x+hs) \longrightarrow q(x)$ uniformly in s, we obtain

$$E\hat{q}(x) - q(x) \longrightarrow 0. \qquad\qquad (A.5)$$

Furthermore, from (A.4) and (12) follows the quasi complete convergence of $\hat{f}(x)$ to $f(x)$, i.e. for all

---

[11] A similar result can be found in Theorem 2.1 (i) in Sperlich (2009).

$\varepsilon > 0$, it holds that

$$\sum_{n=1}^{\infty} P(|\hat{f}(x) - f(x)| > \varepsilon) < \infty.$$

Since $f(x) > 0$, we can define $\delta = \varepsilon = f(x)/2$ and get for $\delta > 0$

$$\sum_{n=1}^{\infty} P(\hat{f}(x) \leq \delta) < \infty. \tag{A.6}$$

Note, that with (7) and $q = f \cdot m$ we can state

$$\hat{m}_{NW}(x) - m(x) = \frac{\hat{q}(x) - q(x)}{\hat{f}(x)} + (f(x) - \hat{f}(x)) \frac{m(x)}{\hat{f}(x)}, \tag{A.7}$$

and thus with (A.3) − (A.7) follows the assertion. □

References

Aslanidis, N., Christiansen, Ch. (2014). Quantiles of the realized stockbond correlation and links to the macroeconomy. *Journal of Empirical Finance*, 28, 321–331.

Asness, C. (2003). Fight the Fed Model: The Relationship Between Future Returns and Stock and Bond Market Yields. *Journal of Portfolio Management*, 30(1), 11–24.

Bandi, F., Corradi, V., and Wilhel, D., 2016, Possibly Nonstationary Cross-Validation, Available at SSRN: http://ssrn.com/abstract=2748510

Baele, L., Bekaert, G., Inghelbrecht, K. (2010). The Determinants of Stock and Bond Return Comovements. *The Review of Financial Studies*, 23(6), 2374–2428.

Bekaert, G., Engstrom, E., Grenadier, R. (2010). Stock and bond returns with Moody Investors. *Journal of Empirical Finance*, 17, 867–894.

Bikker, J., Broeders, D., Hollanders, D., and Ponds, E. (2012). Pension Funds Asset Allocation and Participant Age: A Test of the Life-Cycle Model, *Journal of Risk and Insurance* 79(3), 595–618.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*, 2nd ed. Springer, NY.

Burman, P., Chow, E., and Nolan, D., 1994, A Cross-Validatory Method for Dependent Data, *Biometrika* 81, 351–358.

Burman P, Nolan D. 1992. Data-Dependent Estimation of Prediction Functions. *Journal of Time Series Analysis* **13**: 189-207.

Campbell, J. Y. (1987). Stock Returns and the Term Structure. *Journal of Financial Economics*, 18, 373–399.

Campbell, J. Y., Thompson, S. B. (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies*, 21(4), 1455–1508.

Clark, T., West, K. (2006). Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale

Difference Hypothesis. *Journal of Econometrics*, 135, 155–186.

Connolly, R., Stivers, C., Sun, L. (2010). Stock Market Uncertainty and the Stock-Bond Return Relation. *Journal of Financial and Quantitative Analysis*, 40(1), 161–194.

Dell'Aquila, R., Ronchetti, E. (2006). Stock and Bond Return Predictability: The Discrimination Power of Model Selection Criteria. *Computational Statistics and Data Analysis*, 50, 1478–1495.

Doukhan, P. (1994). *Mixing.* Springer, New-York.

Eling, M. (2014). Fitting asset returns to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*, 59, 45–56.

Elliott, G., Gargano A., and Timmermann, A., 2013, Complete Subset Regressions, *Journal of Econometrics* 177, 357–373.

Engsted, T., Tanggaard, C. (2001). The Danish Stock and Bond Markets: Comovement, Return Predictability and Variance Decomposition. *Journal of Empirical Finance*, 8, 243–271.

Fama, E. F., and French, K. R. (1993). Common risk factors in the returns on stock and bonds. *Journal of Financial Economics* 33 3–56.

Fan, J., Yao, Q. (2003). *Nonlinear Time Series.* Springer, New-York.

Ferraty, F., Núñez-Antón, V., Vieu, P. (2001). *Regresión No Paramétrica: Desde la Dimensión Uno Hasta la Dimensión Infinita.* Servicio Editorial da la Universidad del Pais Vasco, Bilbao.

Gámiz-Pérez, M. L., Martínez-Miranda, M. D., Nielsen, J. P. (2013). Bandwidth selection in marker dependent kernel hazard estimation. *Computational Statistics and Data Analysis*, 68, 155-169.

Gerrard, R., Guillen, M., Nielsen, J. P., and Perez-Marin, A. (2014). Long-run savings and investment strategy optimization. *Scientific World Journal* Article ID 510531.

Goyal, A., Welch, I. (2008). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4), 1455–1508.

Guidolin, M., Hyde, St., McMillan, D., Ono, S. (2009). Non-linear Predictability in Stock and Bond Returns: When and Where is it Exploitable? *International Journal of Forecasting*, 25(2), 373–399.

Guidolin, M., Timmermann, A. (2006). An Econometric Model of Nonlinear Dynamics in the Joint Distribution of Stock and Bond Returns. *Journal of Applied Econometrics*, 21(1), 1–22.

Guillen, M., Jarner, S., Nielsen, J. P., and Perez-Marin, A. (2014). Risk-Adjusted Impact of Administrative Costs on the Distribution of Terminal Wealth for Long-Term Investment. *Scientific World Journal* Article ID 521074.

Guillen, M., Konicz, A., Nielsen, J. P., and Perez-Marin, A. (2013a). Do not pay for a Danish interest guarantee. The law of the triple blow. *Annals of Actuarial Science* 7(2), 192–209.

Guillen, M., Perez-Marin, A., Nielsen, J. P., and Petersen, K. (2013b). Performance measurement of pension strategies: a case study of Danish life cycle products, *Scandinavian Actuarial Journal* 1, 49–68.

Györfi, L., Härdle, W., Sarda, P., Vieu, Ph. (1990). *Nonparametric Curve Estimation from Time Series*(Lecture Notes in Statistics). Springer, Heidelberg.

Ilmanen, A. (2003). Stock-bond correlations. *Journal of Fixed Income.* 13, 55–66.

Inoue, A., Kilian, L. (2004). InSample or OutofSample Tests of Predictability? Which One Should We Use? *Econometric Reviews*, 23(4), 371–402.

Kim, T. Y., Park, B. U., Moon, M. S., Kim, K. (2009). Using Bimodal Kernel for Inference in Nonparametric Regression with Correlated Errors. *Journal of Multivariate Analysis*, 100, 1487–1497.

Lee, Ch., Huang, W., Yin, Ch. (2013). The dynamic interactions among the stock, bond and insurance markets. *The North American Journal of Economics and Finance*, 26, 28–52.

Lettau, M., Nieuwerburgh, S. V. (2008). Reconciling the Return Predictability Evidence. *The Review of Financial Studies*, 21(4), 1607–1652.

Lim, E., Gallo, J., Swanson, P. (1998). The Relationship between International Bond Markets and International Stock Markets. *International Review of Financial Analysis*, 7(2), 181–190.

Lin, H., Wang, J., Wu, Ch. (2014). Predictions of corporate bond excess returns. *Journal of Financial Markets*, 21, 123–152.

Lo, A. W. (1991). Long-Term Memory in Stock Market Prices. *Econometrica*, 59(5), 1279–1313.

Lund, J., Engsted, T. (1996). GMM and Present Value Tests of the C_CAPM: Evidence from Danish, German, Swedish, and UK Stock Markets. *Journal of International Money and Finance*, 15, 497–521.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P., Sperlich, S. (2011). Do-validation for Kernel Density Estimation. *Journal of the American Statistical Association*, 106(494), p. 651–660.

Masry, E., Fan, J. (1997). Local Polynomial Estimation of Regression Functions for Mixing Processes. *Scandinavian Journal of Statistics*, 24, 165–179.

McMillan, D. G. (2007). Non-linear Forecasting of Stock Returns: Does Volume Help? *International Journal of Forecasting*, 23(1), 115–126.

Nielsen, J. P., Sperlich, S. (2003). Prediction of Stock Returns: A New Way to Look at it. *Astin Bulletin*, 33, 399–417.

Owadally, I., Haberman, S., and Gómez Hernández, D. (2013). A Savings Plan With Targeted Contributions, *Journal of Risk and Insurance* 80(4), 975–1000.

Park, B. U., Kim, W. C., Ruppert, D., Jones, M. C., Signorini, D. F., Kohn, R. (1997). Simple Transformation Techniques for Improved Non-parametric Regression. *Scandinavian Journal of Statistics*, 24(2), 145–163.

Rapach, D. E., Strauss, J. K., Zhou, G. (2010). Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *The Review of Financial Studies*, 23(2), 821–862.

Rapach, D. E., Wohar, M. E., Rangvid, J. (2005). Macro Variables and International Stock Return Pre-

dictability. *International Journal of Forecasting*, 21(1), 137–166.

Rio, E. (2000). Théorie Asymptotique pour des Processus Aléatoires Faiblement Dépendants. In *SMAI, Mathématiques et Applications* 31. Springer, Berlin.

Scholz, M., Nielsen, J. P., Sperlich, S. (2015), Nonparametric Prediction of Stock Returns Based on Yearly Data: The Long-Term View, *Insurance: Mathematics and Economics*, 65, 143–155.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics* . Wiley, NY.

Shiller, R. J., Beltratti, A. E. (1992). Stock Prices and Bond Yields. *Journal of Monetary Economics*, 30, 25–46.

Sperlich, S. (2009). A Note on Non-parametric Estimation with Predicted Variables. *The Econometrics Journal*, 12, 382–395.

Sperlich, S., Linton, O., Härdle, W. (1999). Integration and Backfitting Methods in Additive Models – Finite Sample Properties and Comparison. *Test*, 8, 419–458.

Tsai, H., Wu, Y. (2014). Bond and Stock Market Response to Unexpected Dividend Changes. *Journal of Empirical Finance*, to appear.

Valkanov, R. (2003). Long-horizon Regressions: Theoretical Results and Applications. *Journal of Financial Economics*, 68, 201–232.

Vieu, Ph. (1994). Choice of Regressors in Nonparametric Estimation. *Computational Statistics and Data Analysis*, 17, 575–594.