IWQW

Institut für Wirtschaftspolitik und Quantitative Wirtschaftsforschung

Diskussionspapier
Discussion Papers

No. 05/2013

Outliers & Predicting Time Series: A comparative study

Vlad Ardelean
University of Erlangen-Nuremberg

Thomas Pleier
University of Erlangen-Nuremberg

ISSN 1867-6707

Outliers & Predicting Time Series: A comparative study

June 4, 2013

Vlad Ardelean
University of Erlangen-Nuremberg
Vlad.Ardelean@fau.de
Thomas Pleier[1]
University of Erlangen-Nuremberg
Thomas.Pleier@fau.de

Keywords and phrases: Parametric prediction; Nonparametric prediction; Support Vector Regression; Outliers;

Abstract

Nonparametric prediction of time series is a viable alternative to parametric prediction, since parametric prediction relies on the correct specification of the process, its order and the distribution of the innovations. Often these are not known and have to be estimated from the data. Another source of nuisance can be the occurrence of outliers. By using nonparametric methods we circumvent both problems, the specification of the processes and the occurrence of outliers.

In this article we compare the prediction power for parametric prediction, semiparametric prediction and nonparamatric methods such as support vector machines and pattern recognition. To measure the prediction power we use the MSE. Furthermore we test if the increase in prediction power is statistically significant.

[1]Correspondence author: Thomas Pleier, Department of Statistics and Econometrics, University of Erlangen-Nuremberg, Lange Gasse 20, D-90403 Nuremberg, E-Mail: Thomas.Pleier@fau.de

1 Introduction

In the analysis of time series the underlying process is often assumed to be parametric. Especially for financial and macro economical time series two general processes and their various extensions have found widespread approval: Autoregressive Moving Avarage (ARMA) processes and generalized Autoregressive Heteroscedasticty (GARCH) processes. In order to estimate the unknown parameters from available data, the order of the process and the distribution of the innovations have to be known beforehand. Otherwise they can be estimated from the data. But the aim is not only to identify the underlying process, but also to predict future values. For example, the time varying volatility of returns can be forecasted when the parameters of the process are known. Such a forecast can be used as an input for pricing options, other derivatives, trading and hedging strategies. Furthermore, the risk of an asset can be measured by the predicted volatility.

An implicit assumption in the analysis of parametric time series is, that there are no aberrant observations, so called outliers. Outliers are observations that seem not to be consistent with the assumed model. When these observations are included to estimate the model parameters, the resulting estimates are biased.

Past shocks that markets have been affected by (i.e. East Asian crisis, Dot-com bubble, subprime mortgage crisis) question the assumption that no outlier is present. Any forecast is questionable, when outlying observations are present. By using non-parametric methods such as Support Vector Machines and Pattern Recognition we circumvent both problems, the specification of the processes and the occurrence of outliers.

The article is structured as follows, in the second chapter we describe some parametric models, their properties and estimation. Chapters three and four contain the non-parametric methods we decided to gather in our simulations. A simulation study that compares the methods according to the MSE criterion and some real live data are contained in chapter five.

2 Time Series

A common way to model the conditional mean is class of ARMA processes is, while the class of GARCH models is widely used to model time-varying volatility.

2.1 ARMA processes

A stochastic process $(X_t)_{\mathbb{Z}}$ is said to be ARMA(p,q) process, if:

$$X_t = \overbrace{\mu + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \nu_{t-j}}^{\mu_t} + \nu_t \; t \in \mathbb{Z} \tag{1}$$

where $\mathbf{\Lambda} = (\mu, \theta_1, \ldots, \theta_q, \phi_1 \ldots, \phi_q) \in \mathbb{R}^{1+p+q}$ and all ν_t follow some specified distribution with $E_F(\nu_t^2) < \infty$. A more rigorous treatment of GARCH(p,q) models and their properties can be found for example in Hamilton (1994).

2.1.1 GARCH processes

GARCH(p,q) processes are standard to model the volatility of financial asset returns as it captures many of the stylized facts for financial assets, see Cont (2001). A more rigorous treatment of GARCH(p,q) models and their properties can be found for example in Berkes et al. (2003) or Lindner (2009).
Following Bollerslev (1986), a stochastic process $X_{t\mathbb{Z}}$ is a GARCH(1,1) process if:

$$\begin{aligned} &X_t|\mathcal{F}_{t-1} = \sigma_t \nu_t. \\ &\sigma_t^2 = (\sigma_t(\theta))^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2,\ t \in \mathbb{Z} \end{aligned} \tag{2}$$

with $\mathbf{\Lambda} = (\alpha_0, \alpha_1, \beta_1),\ \alpha_0 > 0,\ \alpha_1 \geq 0,$ and $\beta_1 \geq 0$, where $\mathcal{F}_t$ denotes the information set of the process up to time t and $\nu_t \overset{iid}{\sim} F$, where F is some distribution function with density f, $E_F(\nu_t) = 0$ and $E_F(\nu_t^2) = 1$.
The simple GARCH(1,1) is sufficient in most applications, see Hansen and Lunde (2005)
For a squared GARCH(1,1) there exists an ARMA(1,1) representation, cf. Bollerslev (1986):

$$\begin{aligned} &X_t^2 = \alpha_0 + (\alpha_1 + \beta_1) X_{t-1}^2 - \beta_1 \varepsilon_{t-1} + \varepsilon_t \\ &\varepsilon_t = X_t^2 - \sigma_t^2. \end{aligned} \tag{3}$$

Note that the ε_t form a centered conditionally heteroscedastic series:

$$\begin{aligned} E(\varepsilon_t|\mathcal{F}_{t-1}) &= \sigma_t^2 \left(E(\nu_t^2) - 1\right) = 0 \\ E(\varepsilon_t^2|\mathcal{F}_{t-1}) &= \sigma_t^4 E((\nu_t^2 - 1)^2) = \sigma_t^4 m_4, \end{aligned}$$

if $E(\nu_t^4)$ exists and $m_4 = E((\nu_t^2 - 1)^2)$.

2.2 Prediction for parametric models

When the order and the innovations are known, the parameters of the process can be estimated via the maximum likelihood method. Under general assumptions the (quasi) maximum likelihood estimator is consistent and asymptotically normal even when the true distribution F of the innovations is not known, see for example Theorem 2.2 in Francq and Zakoïan (2004).

When the parameters are estimated the best predictor (in the sense that the mean square error is minimized) is given by the conditional expectation conditioned on the past observations, see Hamilton (1994). Let $x_1, \cdots, x_T$ be the observed data and $\hat{\mathbf{\Lambda}}$ the

estimated parameters of the assumed process. The best prediction of future values is given by:

$$\hat{x}_{T+1} = E(X_{t+1}|x_1, \cdots, x_t) = \mu + \sum_{i=1}^{p} \hat{\phi}_i x_{t-i} + \sum_{j=1}^{q} \theta_j \hat{\nu}_{t-j}, \tag{4}$$

$$\hat{\sigma}^2_{T+1} = E(\sigma^2_{T+1}|\mathcal{F}_T) = \hat{\alpha}_0 + \sum_{i=1}^{p} \hat{\alpha}_i x^2_{T+1-i} + \sum_{i=1}^{q} \hat{\beta}_i \sigma^2_{T+1-i} \tag{5}$$

where $\hat{\nu}_{n-q}, \cdots, \hat{\nu}_n$ and $\hat{\sigma}^2_{n-q}, \ldots, \hat{\sigma}^2_n$ are calculated recursively from (1) and (2). As starting values we set $x_i = 0$ for $i = 1, \ldots, p$, $\hat{\nu}_j = 0$ for $j = 1, \ldots, q$ and $\hat{\sigma}^2_j = 0$ for $j = 1 \ldots, q$.

2.3 Semiparametric prediction

The assumptions on the innovations are rather strong (independence, martingale difference series among others). These assumptions can be relaxed, but then the optimal prediction does not coincide with the conditional expectation. Dabo-Niang et al. (2010) propose a semi-parametric approach. Under some general assumptions the optimal semi-parametric prediction is given by:

$$\hat{x}_{T+1} = E(X_{T+1}|x_1, \cdots, x_T)) + E(\nu_{T+1}|\nu_1, \cdots, \nu_T).$$

The extra term is the optimal (nonlinear) prediction of the innovation process at time $T+1$. In order to estimate $E(\nu_{T+1}|\nu_1, \cdots, \nu_T)$ a Nadaraya-Watson estimator is proposed.

3 Support Vector Machines

Support Vector Machines (SVM) were first introduced by Vapnik (1995). SVM can be understood as learning system that uses a hypothesis space of linear functions in a high dimensional feature space. For more information on SVM see for example Schölkopf and Smola (2002).

In order to use SVM in the regression context we sssume that $y \in \mathbb{R}$. One is interested predicting unobserved y values. With small modifications the SVM can be applied in the regression context. Possible loss functions include:

- ε-insensitive loss function:

$$L^{\varepsilon}(x, y, f) = |y - f(x)|_{\varepsilon} = \max(\varepsilon, |y - f(x)|) - \varepsilon.$$

- Given a $\sigma > 0$ then Huber's robust loss function is given by:

$$L^{\varepsilon}(x, y, f, \sigma) = \begin{cases} \frac{1}{2\sigma}(y - f(x))^2 & \text{if } |y - f(x)| \leq \sigma \\ |y - f(x)| - \frac{\sigma}{2} & \text{else} \end{cases}$$

The resulting optimization problem is now (in its dual form)

$$\max_{\alpha,\alpha^*} \quad -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i+\alpha_i^*),$$

under the constraints

$$\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0 \text{ and } \alpha_i,\alpha_i^*\in[0,\gamma].$$

Let $\Phi : x \to \mathbb{R}^k$ be any transformation of the data into a higher dimension, then k is a kernel, if the following holds:

$$k(x,x') = \langle\Phi(x),\Phi(x')\rangle.$$

The kernel can be seen as a transformation of the to data to higher dimension. In this higher dimension the data is now linearly separable.

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)k(x_i,x)+b$$

gives the prediction of the y value of a new object x.

4 Pattern Recognition

The whole chapter is based on the book by Devroye et al. (1996). The idea of pattern recognition is that patterns recognized in the past are likely to be seen again.

The first task is to fix the recognition rule. We use two rules based on the euclidean distance. Given the actual pattern at time T of length p

$$p_T = \left(x_{T-(p-1)},\ldots,x_T\right)$$

we want to look at dates in the past that exhibit a similar pattern.

The so called elementary prediction of x_{T+1} is obtained either by averaging the observations $x_{t_1+1},\ldots x_{t_k+1}$, where $t_1,\ldots t_k \in \{1,\ldots,T\}$ are the k dates of which the corresponding patterns have smallest euclidean distance from the current pattern p_T - the nearest neighbour based prediction strategy (NNBP), or by a kernel-based weighted average of the whole past observations - the kernel based prediction strategy (KBP).

Another recognition rule we use is the generalized linear prediction strategy (GLP) proposed by Biau et al. (2010). The elementary predictor to the lag p of this strategy is a linear combination of the past p observations. In summary the elementary predictions are of the following form

$$\widehat{x}_{T+1,\,p} = \begin{cases} \sum_{i=p}^{T-1} x_{i+1}\cdot \mathrm{w}(p_i) & \text{(KBP \& NNBP)} \\ \sum_{i=1}^{p} c_i x_{T+1-i} & \text{(GLP)} \end{cases}$$

where $w(p_i)$ gives the above described weights determined by the euclidean distance from p_t and $(c_i)_{i=1}^p$ is the vector of coefficients of a fitted linear model.

For different fixed pattern lengths p different elementary predictors are obtained. Biau et al. (2010) propose a performance oriented criterion for the final prediction strategy: combine the different predictors linearly with weights determined by their past performance. They derive consistency properties for their algorithms under weak assumptions on the underlying process. We use a slight modification of their algorithms that enables comparison among the different predictions considered in this paper.

5 Outliers

Fox (1972) introduced two types of outliers for time series, namely additive (type I) and innovational (type II) outliers.

Type I outliers only effect a single observation while type II outliers also affect the following observations.

Intervention analysis introduced by Box and Tiao (1975) can be embedded in the definition of Fox (1972) when more then one outlier can occur.

Let X_t be the underlying and unobserved process and Y_t be the observable process which contains outliers, than additive outliers can be modelled the following way:

$$Y_t = X_t + \nu \mathbb{1}_t(\tau),$$

where $\nu \in \mathbb{R}$ the size of the outlier, $\tau \in \mathbb{Z}$ the time of occurrence of the outlier and $\mathbb{1}_t(\tau)$ is the indicator function which is one if $\tau = t$ and zero if $\tau \neq t$. Outlying observations lead to several problems. The estimate of the order of the process is biased towards 0, cf. Maronna et al. (2006). For an ARMA process, the estimated parameters are biased towards 0, cf. Maronna et al. (2006). For a GARCH process, Ardelean (2009) investigates the effect of outliers on the estimated parameters in a simulation study and finds that the resulting estimates are biased upwards, especially the offset parameter

6 Numerical Results

With a simulation study we compare the prediction performance of the methods described in the previous sections:

- Maximum Likelihood prediction (ML)
- Maximum Likelihood prediction with a non-parametric regression in the MA part (SP)
- Support Vector Machine with a gaussian Kernel (SVM)

- Kernel based prediction (KBP)
- Nearest neighbour based prediction(NNBP)
- Generalized linear prediction (GLP)

The prediction strategy based on the likelihood approach is based on the information of the full sample. In comparison the prediction strategies described by Devroye et al. (1996) are constructed in a sequential manner, which suits well the task of time series prediction, as new information is available is updated on a regular basis. As we want to compare different strategies we study performance in a non-sequential way. Furthermore we split the sample in two parts. The first part is the insample (first 80%) while the remaining part is the outsample (last 20%). The procedures learn only the pattern on the training sample. For both samples the performance measure is calculated.

As a measure of prediction power we use the MSE of the 1-step-ahead-predictions. For the GARCH processes we predict x_t^2, which is often used as a proxy for σ_t^2 see for example Andersen and Bollerslev (1998). As the squared observation of a GARCH(1,1) process can be written as an ARMA(1,1) process, we can use the prediction methods introduced in section 2.2. In these simulations the MSE turns out to be much lower than in the ARMA-type simulations, which is due to the processes and not a hint for easier predictability. For a more convenient look we calculate the relative prediction performance in means of the MSE with the optimal constant prediction as reference.

Furthermore we use a binomial test to find out weather the sophisticated methods significantly improve the predictions. For example, using a semiparametric method, the Null is H_{non}: *The prediction of the parametric strategy cannot be improved by the semi-parametric extension.* Looking at the proportion $\hat{p}$ of simulations in which we observe a lower prediction error in the semiparametric prediction we execute a standard test for $H_0 : \ p \leq 0.5$ versus $H_1 : p > 0.5$ (compare the *tea tasting lady*). This test provides information we can use to compare prediction performance of different strategies instead of only comparing average MSE which is a single value aggregating many possibly strongly deviating results.

We also use the test sample in order to see if any of the procedure suffers from so called 'overfitting', i.e. if a low insample cannot be traced in the outsample.

For the simulation study we use 3 different stationary data generating processes (DGP)

DGP	Process	Order	Parameter
DGP-1	AR	2	0.0, 0.7, -0.4
DGP-2	ARMA	1,1	0.0, 0.7, -0.2
DGP-3	GARCH	1,1	0.001, 0.05, 0.8

As innovations for the DGP's the standard normal and a t-distribution with 5 degrees of freedom are used. Sample size is $n = 500$.DGP-1 and DGP-2 are true ARMA processes (with order (2,0) respectively (1,1)). The squared instances of DGP-3 data follow a heteroscedastic ARMA-process. Alltogether we refer to these DGPs as ARMA-type processes.

In order to quantify the impact outliers have on the prediction performance we contaminate the samples with two outliers at $\tau_1 = 250$ and $\tau_2 = 300$. The size of the outliers is 5 (DGP1 and DGP2). For DGP3 we multiply 5 with the conditional variance at that time (relative) or with the unconditional variance (fixed).

Each setting is repeated 250 times.

6.1 Pure simulations of ARMA-type processes

Process	DGP-1 `N`		DGP-1 `t`		DGP-2 `N`		DGP-2 `t`	
MSE	IN	OUT	IN	OUT	IN	OUT	IN	OUT
ML AR(2)`N`	**0.989**	**1.001**	**0.989**	1.027	1.140	1.159	1.137	1.189
ML AR(2)`t`	0.990	1.002	0.991	**1.024**	1.141	1.160	1.138	1.186
ML ARMA(1,1) `N`	1.019	1.030	1.021	1.055	**0.990**	**0.999**	**0.989**	1.026
ML ARMA(1,1) `t`	1.019	1.031	1.022	1.052	0.991	1.001	0.991	**1.024**
SP AR(2)`N`	**0.991**	**0.996**	**0.995**	1.026	1.095	1.099	1.100	1.139
SP AR(2) `t`	0.991	0.996	0.997	**1.025**	1.095	1.099	1.102	1.139
SP ARMA(1,1) `N`	1.059	1.064	1.063	1.089	**0.989**	**0.993**	**0.994**	1.024
SP ARMA(1,1) `t`	1.060	1.064	1.064	1.088	0.989	0.993	0.995	**1.023**
SVM	0.843	1.512	0.912	1.705	1.256	3.489	1.523	4.119
KBP	**0.326**	1.269	**0.252**	1.349	**0.473**	2.714	**0.318**	3.050
NNBP	0.875	1.195	0.892	1.243	1.357	2.173	1.407	2.223
GLP	0.981	**1.004**	0.980	**1.031**	0.993	**1.019**	0.989	**1.040**

Table 1: MSE for the prediction performance for DGP-1 and DGP-2, total length $n = 500$.

Table 1 summarizes the results of the simulation using DGP1 and DGP2 with either normal or student `t`(5) distributed residuals. The variance of the residual distribution is rescaled to the value 1 so the average MSE ought to take values in a region of 1. It turns out that parametric ML estimates achieves the expected performance in our simulations. Note that correct specification in parameters and correct specification of the residuals' distribution lead to different improvements in performance. The parametric specification has more influence on the prediction results. The semiparametric estimates show rather similar performance. Looking at the nonparametric forecasts there are larger deviations both in in- and outsample which can be considered an indication for overfitting.

Comparing overall performance in in- and outsample this clue can be traced further. The loss in performance of the parametric models does not exceed 5% and that of the

semiparametric ones not even 4%. The loss in performance of the nonparametric models takes values from 2% to more than 800%. A tendency to overfitting now becomes obvious.

Process	DGP-1 N	DGP-1 t	DGP-2 N	DGP-2 t
AR(2)N	0.492 (0.624)	**0.504** (0.475)	0.980 (0.000)	**0.948** (0.000)
AR(2)t	**0.588** (0.003)	0.464 (0.885)	**0.984** (0.000)	0.916 (0.000)
ARMA(1,1) N	0.136 (1.000)	0.144 (1.000)	0.424 (0.993)	0.504 (0.475)
ARMA(1,1) t	0.156 (1.000)	0.128 (1.000)	0.604 (0.001)	0.468 (0.859)

Table 2: Proportion of simulations where semiparametric extension improved parametric prediction and corresponding p-value (H_0 : semiparametric extension does not improve prediction performance).

Process	DGP-1 N	DGP-1 t	DGP-2 N	DGP-2 t
SVM	0.096 (1.000)	0.104 (1.000)	0.000 (1.000)	0.000 (1.000)
KBP	0.056 (1.000)	0.068 (1.000)	0.000 (1.000)	0.000 (1.000)
NNBP	0.156 (1.000)	0.152 (1.000)	0.000 (1.000)	0.000 (1.000)
GLP	**0.812** (0.000)	**0.732** (0.000)	**0.988** (0.000)	**0.984** (0.000)

Table 3: Proportion of simulations where nonparametric strategies achieved lower MSE than falsely specified (in model specification and residual distribution) parametric prediction and corresponding p-value (H_0 : nonparametric prediction strategy does not achieve lower MSE).

In table 2 we see that in many of the simulations the semiparametric extension does not improve the parametric prediction. Even if the parametric model is falsely specified but flexible, there is no significant improvement. Table 3 shows that only the predictions of the GLP strategy are better than the falsely specified parametric prediction.

Process	DGP-3 N		DGP-3 t	
MSE	IN	OUT	IN	OUT
ML AR(2) N	0.986	**1.004**	0.977	**0.994**
ML AR(2) t	1.050	1.045	1.045	1.008
ML ARMA(1,1) N	**0.986**	1.027	**0.977**	1.061
ML ARMA(1,1) t	1.050	1.051	1.047	1.011
SP AR(2) N	0.995	0.999	0.978	0.972
SP AR(2) t	0.992	**0.982**	0.982	**0.950**
SP ARMA(1,1) N	**0.989**	1.021	**0.975**	1.040
SP ARMA(1,1) t	0.990	0.989	0.981	0.951
SVM	0.902	1.063	0.886	1.006
KBP	0.996	**0.999**	0.954	**0.998**
NNBP	**0.839**	1.038	**0.835**	1.050
GLP	1.067	1.099	1.013	1.045

Table 4: Relative MSE for the prediction performance with simulated DGP-3 data

Forecasting squared ARCH-type models is a hard task. Table 2 shows the performance of the different strategies relative to the performance of the best constant prediction. At first sight it seems almost impossible to outperform the constant prediction. The parametric predictions in this heteroskedastic setting are rather poor, in the semiparametric approach there are improvements and the nonparametric are somewhere in between. In contrast to table 1, the nonparametric strategies seem to suffer less from overfitting.

Process	DGP-3 `N`	DGP-3 `t`
AR(2)`N`	0.544 (0.092)	0.692 (0.000)
AR(2)`t`	**0.896** (0.000)	0.864 (0.000)
ARMA(1,1) `N`	0.544 (0.092)	0.676 (0.000)
ARMA(1,1) `t`	0.884 (0.000)	**0.880** (0.000)

Table 5: Proportion of simulations where semiparametric extension improved parametric prediction and corresponding p-value (H_0 : semiparametric extension doesn't improve prediction performance).

Process	DGP-3 `N`	DGP-3 `t`
SVM	0.144 (1.000)	0.396 (1.000)
KBP	**0.600** (0.001)	**0.584** (0.005)
NNBP	0.240 (1.000)	0.272 (1.000)
GLP	0.052 (1.000)	0.132 (1.000)

Table 6: Proportion of simulations where nonparametric strategies achieved lower MSE than AR(2) with normal distributed residuals and corresponding p-value (H_0 : nonparametric prediction strategy doesn't achieve lower MSE).

The improvement of semiparametric strategies is observed in significantly many simulations. Table 5 shows that in each setting the fraction of simulated series with improved prediction in the outsample exceeds 50%. Table 6 shows that the choice of using nonparametric strategies does not automatically lead to suitable predictions. In this setting the KBP strategy is the only competitive nonparametric approach.

6.2 Simulations of ARMA-type processes with outliers

Figure 1 shows the typical result for the different prediction strategies when two outlying observations are present ($\tau_1 = 250$ and $\tau_2 = 300$). The residuals at τ_1 and τ_2 are larger than the other residuals, indicating that all prediction methods (K-BP excluded) do not anticipate the outlying observations. Table 3 and 4 summarize the prediction performance when outliers are present. When calculating the MSE, the observations at time ($\tau_1 = 250$ and $\tau_2 = 300$) are excluded since we want to measure the impact of outliers on the prediction performance of 'typical' observations. For DGP-1 the outlier only have

influence on the insample performance. The best overall performance is achieved by the GLP method. For DGP-2, outliers also influence the insample as well as the outsample. The best overall performance is achieved by the semiparametric approach (SP ARMA(1,1)) followed by the GLP approach.

The prediction performance is similar regardless of the size of the outlier is relative or fixed. The SVM loses insample performance but gains outsample performance and has the best overall performance followed by semi-parametric model. Even though the NNBP has a very good insample performance the overfitting effect is clearly visible in the outsample.

Process	DGP-1 N		DGP-1 t		DGP-2 N		DGP-2 t	
MSE	IN	OUT	IN	OUT	IN	OUT	IN	OUT
ML AR(2) N	**1.092**	1.015	**1.065**	1.021	1.550	1.264	1.514	1.273
ML AR(2) t	1.094	**1.007**	1.066	**1.013**	1.689	1.159	1.669	1.176
ML ARMA(1,1) N	1.103	1.047	1.082	1.049	**1.520**	1.203	**1.485**	1.197
ML ARMA(1,1) t	1.108	1.038	1.086	1.041	1.712	**1.050**	1.696	**1.048**
SP AR(2) N	**1.085**	**0.997**	**1.063**	**1.011**	1.407	1.117	1.394	1.154
SP AR(2) t	1.020	1.006	1.072	1.014	1.635	1.105	1.624	1.137
SP ARMA(1,1)N	1.110	1.044	1.096	1.055	**1.375**	1.056	**1.366**	1.082
SP ARMA(1,1) t	1.144	1.066	1.118	1.068	1.654	**0.996**	1.656	**1.016**
SVM	0.856	1.463	0.921	1.606	1.407	3.316	1.605	3.903
KBP	**0.330**	1.278	**0.228**	1.364	**0.446**	2.7780	**0.279**	3.202
NNBP	0.926	1.190	0.931	1.228	1.493	2.126	1.523	2.235
GLP	1.075	**1.019**	1.050	**1.025**	1.509	**1.225**	1.462	**1.221**

Table 7: MSE for the prediction performance for DGP-1 and DGP-2 with two fixed outliers at $\tau_1 = 250$ and $\tau_2 = 300$, total length $n = 500$.

Process	DGP-1 N	DGP-1 t	DGP-2 N	DGP-2 t
AR(2)N	0.504 (0.475)	**0.624** (0.000)	**0.992** (0.000)	0.948 (0.000)
AR(2)t	0.524 (0.243)	0.480 (0.757)	0.972 (0.000)	0.908 (0.000)
ARMA(1,1) N	**0.528** (0.206)	0.440 (0.975)	**0.992** (0.000)	**0.956** (0.000)
ARMA(1,1) t	0.228 (1.000)	0.240 (1.000)	0.964 (0.000)	0.852 (0.000)

Table 8: Proportion of simulations where semiparametric extension improved parametric prediction and corresponding p-value (H_0 : semiparametric extension doesn't improve prediction performance).

Process	DGP-1 `N`	DGP-1 `t`	DGP-2 `N`	DGP-2 `t`
SVM	0.108 (1.000)	0.092 (1.000)	0.000 (1.000)	0.012 (1.000)
KBP	0.072 (1.000)	0.032 (1.000)	0.000 (1.000)	0.000 (1.000)
NNBP	0.212 (1.000)	0.108 (1.000)	0.000 (1.000)	0.004 (1.000)
GLP	**0.744** (0.000)	**0.768** (0.000)	**0.472** (0.829)	**0.732** (0.000)

Table 9: Proportion of simulations where nonparametric strategies achieved lower MSE than falsely specified (in model specification and residual distribution) parametric prediction and corresponding p-value (H_0 : nonparametric prediction strategy doesn't achieve lower MSE).

Comparing table 7 with table 1 we find prediction performance is generally weaker in the presence of outliers. Simulations with normally distributed residuals are affected stronger and the overfitting problem of nonparametric strategies seems to be reduced. In table 8 it can be seen that semiparametric predictions now lead to lower MSE more often and in table 9 the prediction performance is hardly affected by outliers.

The values in tables 10 and 13 are relative to the best constant prediction. This constant prediction is strongly affected by outliers, so the small values here does not show a prediction improvement *because* of the presence of outliers. This relative improvement shows that the considered models are more robust against (few) outliers than the constant prediction.

Tables 11 and 14 reveal the strength of semiparametric extension in a situation with heteroskedastic residuals and outliers in the insample. In every setting the proportion of time series with semiparametric improved prediction significantly exceeds 50%. Among the nonparametric strategies the KBP still is considerably well performing. The most interesting result in tables 12 and 15 is that SVM seems to suffer only very little from the presence of outliers and turn out to be very competitive in these simulations.

Process	DGP-3 N fixed		DGP-3 t fixed	
MSE	IN	OUT	IN	OUT
ML AR(2)N	0.995	**0.999**	0.992	**0.999**
ML AR(2) t	1.030	1.011	1.039	1.010
ML ARMA(1,1) N	**0.994**	1.064	**0.991**	1.097
ML ARMA(1,1) t	1.030	1.014	1.039	1.011
SP AR(2)N	0.992	0.974	0.988	0.974
SP AR(2) t	0.993	**0.973**	0.991	0.967
SP ARMA(1,1) N	**0.990**	1.014	**0.988**	1.051
SP ARMA(1,1) t	0.993	0.976	0.991	**0.966**
SVM	0.935	**0.979**	0.921	**0.984**
KBP	0.994	0.999	0.991	0.996
NNBP	**0.844**	1.144	**0.844**	1.126
GLP	1.067	1.123	1.046	1.067

Table 10: MSE for the prediction performance for DGP-3 with two fixed outlier at $\tau_1 = 250$ and $\tau_2 = 300$, total length $n = 500$.

Process	DGP-3 N fixed	DGP-3 t fixed
AR(2)N	0.680 (0.000)	0.768 (0.000)
AR(2)t	0.700 (0.000)	0.784 (0.000)
ARMA(1,1) N	0.680 (0.000)	0.736 (0.000)
ARMA(1,1) t	**0.704** (0.000)	**0.788** (0.000)

Table 11: Proportion of simulations where semiparametric extension improved parametric prediction and corresponding p-value (H_0 : semiparametric extension doesn't improve prediction performance).

Process	DGP-3 N fixed	DGP-3 t fixed
SVM	**0.528** (0.206)	**0.600** (0.001)
KBP	0.448 (0.956)	0.588 (0.003)
NNBP	0.064 (1.000)	0.100 (1.000)
GLP	0.104 (1.000)	0.136 (1.000)

Table 12: Proportion of simulations where nonparametric strategies achieved lower MSE than AR(2) with normal distributed residuals and corresponding p-value (H_0 : nonparametric prediction strategy doesn't achieve lower MSE).

Process	DGP-3 N relative		DGP-3 t relative	
MSE	IN	OUT	IN	OUT
ML AR(2) N	0.994	**1.000**	**0.983**	0.988
ML AR(2) t	1.030	1.011	1.043	**0.981**
ML ARMA(1,1) N	**0.993**	1.125	0.984	1.123
ML ARMA(1,1) t	1.030	1.011	1.045	0.985
SP AR(2) N	0.991	0.972	0.980	0.947
SP AR(2) t	0.993	0.970	0.985	**0.933**
SP ARMA(1,1) N	**0.988**	1.036	**0.978**	1.072
SP ARMA(1,1) t	0.992	**0.969**	0.984	0.934
SVM	0.933	**0.985**	0.893	**0.978**
KBP	0.994	0.999	0.956	0.981
NNBP	**0.844**	1.143	**0.838**	1.060
GLP	1.066	1.118	1.021	1.018

Table 13: MSE for the prediction performance for DGP-3 with two relative outlier at $\tau_1 = 250$ and $\tau_2 = 300$, total length $n = 500$.

Process	DGP-3 N relative	DGP-3 t relative
ML AR(2)N	0.728 (0.000)	0.796 (0.000)
ML AR(2)t	0.736 (0.000)	0.804 (0.000)
ML ARMA(1,1) N	**0.764** (0.000)	0.804 (0.000)
ML ARMA(1,1) t	0.740 (0.000)	**0.808** (0.000)

Table 14: Proportion of simulations where semiparametric extension improved parametric prediction and corresponding p-value (H_0 : semiparametric extension doesn't improve prediction performance).

Process	DGP-3 N relative	DGP-3 t relative
SVM	**0.532** (0.171)	0.556 (0.044)
KBP	0.516 (0.329)	**0.628** (0.000)
NNBP	0.100 (1.000)	0.232 (1.000)
GLP	0.084 (1.000)	0.264 (1.000)

Table 15: Proportion of simulations where nonparametric strategies achieved lower MSE than AR(2) with normal distributed residuals and corresponding p-value (H_0 : nonparametric prediction strategy doesn't achieve lower MSE).

6.3 Real Data

We analyse Germany's gross domestic product from 1970 until 2007. We use seasonally adjusted quarterly data available from Deutsche Bundesbank[2] , giving us 148 observations. The first 118 observations form the insample and the remaining are used as the outsample. Figure 1 shows the rescaled data, the empirical autocorrelation function and the empirical partial autocorrelation function.

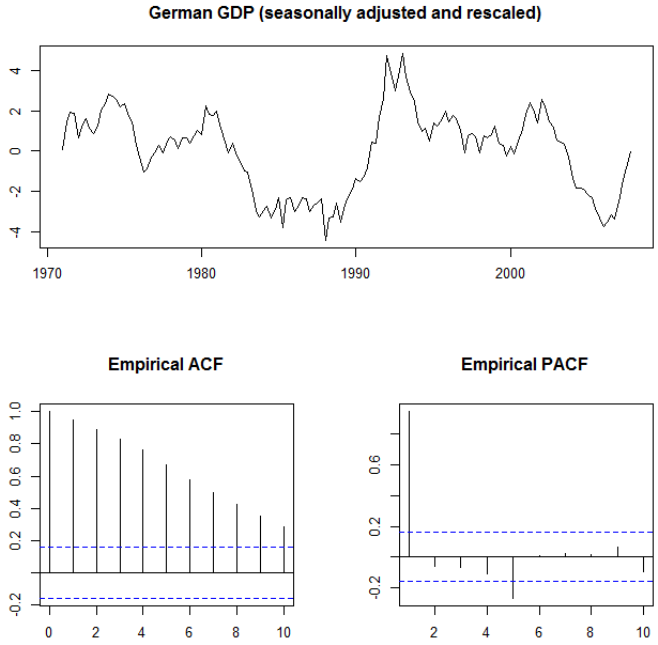

Figure 1: Germany's Gross Domestic Product

From what we learned of tables 1-3 and 7-9 the results are not astonishing: Semiparametric approaches achieve an improved prediction of the pure parametric ones and among the nonparametric predictions the GLP has lowest MSE in the outsample. The SVM, the KBP and the NNBP seemingly suffer from overfitting.

[2] `http://www.bundesbank.de/Navigation/DE/Statistiken/Zeitreihen_Datenbanken/Makrooekonomische_Zeitreihen/its_details_value_node.html?listId=www_s311_b40201&tsId=BBK01.JBB000!` , visited August 16th 2012.

MSE	IN	OUT
ML AR(2) `N`	**0.435**	**0.348**
ML AR(2) `t`	0.442	0.363
ML ARMA(1,1) `N`	**0.435**	0.349
ML ARMA(1,1) `t`	0.442	0.364
SP AR(2) `N`	0.406	0.270
SP AR(2) `t`	0.389	0.264
SP ARMA(1,1) `N`	0.393	0.269
SP ARMA(1,1) `t`	**0.378**	**0.262**
SVM	0.539	0.468
KBP	0.420	0.614
NNBP	0.426	0.595
GLP	**0.416**	**0.333**

Table 16: MSE of Germany's GDP prediction

As a GARCH-type setting we use DAX closing prices from November 16th, 2005 until December 28th, 2007, available at YAHOO! FINANZEN[3]. The squared log-returns are supposed to be similar to a squared GARCH-process. The first 428 observations are the insample, the successive 107 are predicted as outsample instances. Figure 2 shows the DAX log-returns and the corresponding empirical autocorrelation function (ACF) and the empirical partial autocorrelation function of the squared data.

[3] `http://de.finance.yahoo.com/q/hp?s=^GDAXI` , visited August 16th 2012.

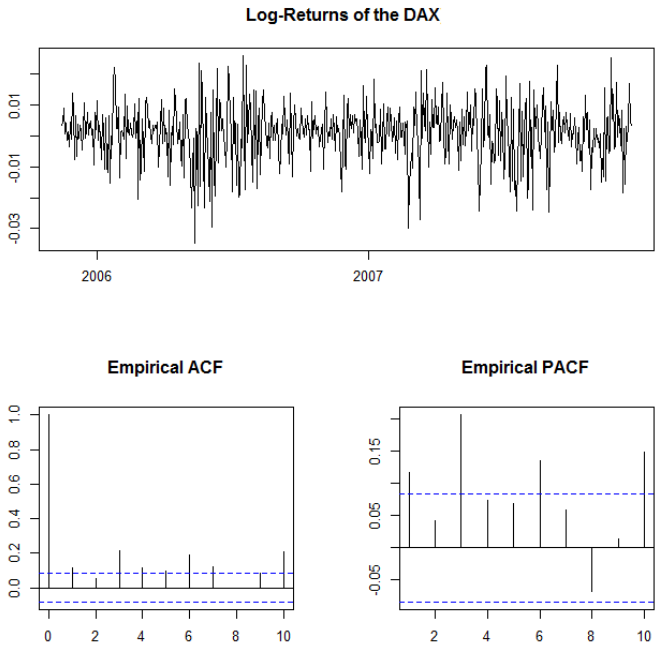

Figure 2: Log-Returns of pre-crisis DAX

We use the test by Engle (1982) to test for (G)ARCH effects in the residuals of a fitted ARMA-model on the squared log-returns. The p-value of the test is 0.0067. As mentioned above, a GARCH(1,1) suffices in most financial applications. Thus, we use a GARCH(1,1) as the parametric model. The MSE results of the prediction strategies show the expected performance in most parts. The outsample error is smaller than the insample error. The values given in the table are relative to the constant prediction strategy. There is more variation in the outsample than in the insample (compare Figure 2). As in times of high volatility it is easier to outperform this constant strategy, the special outsample performance is only due to the time series at hand.

Like in the simulations we can see that the improved performance of the semiparametric extension on the insample, can be traced in a improved outsample prediction. Interestingly all the nonparametric strategies show remarkably good results and no tendency of overfitting.

MSE	IN	OUT
ML AR(2) `N`	**1.008**	**0.867**
ML AR(2) `t`	1.648	1.649
ML ARMA(1,1) `N`	1.018	0.885
ML ARMA(1,1) `t`	1.301	1.247
SP AR(2) `N`	**0.979**	**0.841**
SP AR(2) `t`	1.006	0.884
SP ARMA(1,1) `N`	0.995	0.867
SP ARMA(1,1) `t`	1.002	0.881
SVM	**0.795**	**0.670**
KBP	1.018	0.887
NNBP	0.857	0.738
GLP	0.972	0.841

Table 17: MSE of predicting filtered DAX log-returns

7 Conclusion

We have compared prediction performance of parametric, semiparametric and nonparametric strategies in several ARMA-type settings. The results we find are the following:

- A prediction out of a model specification close to the true specification of the underlying process can hardly be beaten by any strategy.
- Semiparametric approaches lead to good performance even with only little deviation in model specification.
- The performance of nonparametric strategies depends mainly on the data at hand. Note that we used rather simple rules of thumb, when using more sophisticated techniques (i.e. cross validation) results may differ.
- Prediction strategies that use the linearity of the model perform well on homoskedastic time series simulations. Especially the GLP benefits in this situation.
- Simulations of heteroskedastic time series better are not predicted with the GLP. It turns out that now the KBP achieves better forecasts.

If outliers are added, there are some findings worth taking into care. Predicting power of the considered strategies on homoskedastic time series simulations is almost uneffected by outliers. Strategies using the linearity perform superior. Generally the specification of the model has more influence on the performance. Again the GLP which uses the linear structure is the most competitive among the nonparametric approaches. Note that the strategies suffer less from overfitting.

On the other hand, predicting heteroskedastic time series simulations turns out to be a hard task. As the constant predictor seems to be hardly beatable, the effort of applying time series theory seems to not always be worth it. Outliers strongly effect the estimation of the contant predictor, that's why the predictions benefit more of the time series approach. This benefit affects almost all considered strategies, so the ranking doesn't change, except for the KBP which almost doesn't benefit at all. In the presence of outliers (fixed or relative) the SVM are therefore the best choice among the nonparametric strategies when predicting heteroskedastic time series.

Two real datasets were used to look for similar results in non-artificial examples. It turned out that in the homoskedastic setting we were able to find the properties we expected from what we learned from the simulations. In the heteroskedastic setting the results were surprisingly extreme. It turned out that the nonparametric strategies achieved a better performance than expected. This may show that the deviation in model specification is much stronger in the real world than we implied in our simulations.

References

Andersen, T. G. and Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, 39(4):885–905.

Ardelean, V. (2009). The Impact of Outliers on Different Estimators for GARCH Processes: An empirical study. *IWQW Discussion Paper Series*.

Berkes, I., Horvát, K., and Kokoszka, P. (2003). GARCH processes: Structure and estimation. *Bernoulli*, 9:201–227.

Biau, G., Bleakley, K., Györfi, L., and Ottucsák, G. (2010). Nonparametric Sequential Prediction of Time Series. *Journal of Nonparametric Statistics*, 22(3):297–317.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31:307–322.

Box, G. E. P. and Tiao, G. C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, 70(349):70–79.

Cont, R. (2001). Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance*, 1:223–236.

Dabo-Niang, S., Francq, C., and Zakoïan, J.-M. (2010). Combining Nonparametric and Optimal Linear Time Series Predictions. *Journal of the American Statistical Association*, 105(492):1554–1565.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50:987–1007.

Fox, A. J. (1972). Outliers in Time Series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):350–363.

Francq, C. and Zakoïan, J. M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10:605–637.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, P. R. and Lunde, A. (2005). A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.

Lindner, A. M. (2009). Stationarity, Mixing, Distributional Properties and Moments of GARCH(p,q)-processes. In Mikosch, T., Kreiß, J.-P., Davis, R. A., and Andersen, T. G., editors, *Handbook of Financial Time Series*, pages 43–69. Springer Berlin Heidelberg.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley & Sons.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, London.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc, New York and NY and USA.

Diskussionspapiere 2013
Discussion Papers 2013

01/2013 **Wrede, Matthias**: Rational choice of itemized deductions

02/2013 **Wrede, Matthias**: Fair Inheritance Taxation in the Presence of Tax Planning

03/2013 **Tinkl, Fabian**: Quasi-maximum likelihood estimation in generalized polynomial autoregressive conditional heteroscedasticity models

04/2013 **Cygan-Rehm, Kamila**: Do Immigrants Follow Their Home Country's Fertility Norms?

Diskussionspapiere 2012
Discussion Papers 2012

01/2012 **Wrede, Matthias**: Wages, Rents, Unemployment, and the Quality of Life

02/2012 **Schild, Christopher-Johannes:** Trust and Innovation Activity in European Regions - A Geographic Instrumental Variables Approach

03/2012 **Fischer, Matthias:** A skew and leptokurtic distribution with polynomial tails and characterizing functions in closed form

04/2012 **Wrede, Matthias**: Heterogeneous Skills and Homogeneous Land: Segmentation and Agglomeration

05/2012 **Ardelean, Vlad**: Detecting Outliers in Time Series

Diskussionspapiere 2011
Discussion Papers 2011

01/2011 **Klein, Ingo, Fischer, Matthias and Pleier, Thomas:** Weighted Power Mean Copulas: Theory and Application

02/2011 **Kiss, David:** The Impact of Peer Ability and Heterogeneity on Student Achievement: Evidence from a Natural Experiment

03/2011 **Zibrowius, Michael:** Convergence or divergence? Immigrant wage assimilation patterns in Germany

04/2011 **Klein, Ingo and Christa, Florian:** Families of Copulas closed under the Construction of Generalized Linear Means

05/2011 **Schnitzlein, Daniel**: How important is the family? Evidence from sibling correlations in permanent earnings in the US, Germany and Denmark

06/2011 **Schnitzlein, Daniel**: How important is cultural background for the level of intergenerational mobility?

07/2011 **Steffen Mueller**: Teacher Experience and the Class Size Effect - Experimental Evidence

08/2011 **Klein, Ingo:** Van Zwet Ordering for Fechner Asymmetry

09/2011 **Tinkl, Fabian and Reichert Katja:** Dynamic copula-based Markov chains at work: Theory, testing and performance in modeling daily stock returns

10/2011 **Hirsch, Boris and Schnabel, Claus:** Let's Take Bargaining Models Seriously: The Decline in Union Power in Germany, 1992 – 2009

11/2011 **Lechmann, Daniel S.J. and Schnabel**, **Claus** : Are the self-employed really jacks-of-all-trades? Testing the assumptions and implications of Lazear's theory of entrepreneurship with German data

12/2011 **Wrede, Matthias**: Unemployment, Commuting, and Search Intensity

13/2011 **Klein, Ingo**: Van Zwet Ordering and the Ferreira-Steel Family of Skewed Distributions

Diskussionspapiere 2010
Discussion Papers 2010

01/2010 **Mosthaf, Alexander, Schnabel, Claus and Stephani, Jens:** Low-wage careers: Are there dead-end firms and dead-end jobs?

02/2010 **Schlüter, Stephan and Matt Davison:** Pricing an European Gas Storage Facility using a Continuous-Time Spot Price Model with GARCH Diffusion

03/2010 **Fischer, Matthias, Gao, Yang and Herrmann, Klaus:** Volatility Models with Innovations from New Maximum Entropy Densities at Work

04/2010 **Schlüter, Stephan and Deuschle, Carola:** Using Wavelets for Time Series Forecasting – Does it Pay Off?

05/2010 **Feicht, Robert and Stummer, Wolfgang:** Complete closed-form solution to a stochastic growth model and corresponding speed of economic recovery.

06/2010 **Hirsch, Boris and Schnabel, Claus:** Women Move Differently: Job Separations and Gender.

07/2010 **Gartner, Hermann, Schank, Thorsten and Schnabel, Claus:** Wage cyclicality under different regimes of industrial relations.

08/2010 **Tinkl, Fabian**: A note on Hadamard differentiability and differentiability in quadratic mean.

Diskussionspapiere 2009
Discussion Papers 2009

01/2009 **Addison, John T. and Claus Schnabel:** Worker Directors: A German Product that Didn't Export?

02/2009 **Uhde, André and Ulrich Heimeshoff:** Consolidation in banking and financial stability in Europe: Empirical evidence

03/2009 **Gu, Yiquan and Tobias Wenzel:** Product Variety, Price Elasticity of Demand and Fixed Cost in Spatial Models

04/2009 **Schlüter, Stephan:** A Two-Factor Model for Electricity Prices with Dynamic Volatility

05/2009 **Schlüter, Stephan and Fischer, Matthias**: A Tail Quantile Approximation Formula for the Student t and the Symmetric Generalized Hyperbolic Distribution

06/2009 **Ardelean, Vlad:** The impacts of outliers on different estimators for GARCH processes: an empirical study

07/2009 **Herrmann, Klaus:** Non-Extensitivity versus Informative Moments for Financial Models - A Unifying Framework and Empirical Results

08/2009 **Herr, Annika:** Product differentiation and welfare in a mixed duopoly with regulated prices: The case of a public and a private hospital

09/2009 **Dewenter, Ralf, Haucap, Justus and Wenzel, Tobias**: Indirect Network Effects with Two Salop Circles: The Example of the Music Industry

10/2009 **Stuehmeier, Torben and Wenzel, Tobias**: Getting Beer During Commercials: Adverse Effects of Ad-Avoidance

11/2009 **Klein, Ingo, Köck, Christian and Tinkl, Fabian**: Spatial-serial dependency in multivariate GARCH models and dynamic copulas: A simulation study

12/2009 **Schlüter, Stephan**: Constructing a Quasilinear Moving Average Using the Scaling Function

13/2009 **Blien, Uwe, Dauth, Wolfgang, Schank, Thorsten and Schnabel, Claus:** The institutional context of an "empirical law": The wage curve under different regimes of collective bargaining

14/2009 **Mosthaf, Alexander, Schank, Thorsten and Schnabel, Claus:** Low-wage employment versus unemployment: Which one provides better prospects for women?

Diskussionspapiere 2008
Discussion Papers 2008

01/2008 **Grimm, Veronika and Gregor Zoettl:** Strategic Capacity Choice under Uncertainty: The Impact of Market Structure on Investment and Welfare

02/2008 **Grimm, Veronika and Gregor Zoettl:** Production under Uncertainty: A Characterization of Welfare Enhancing and Optimal Price Caps

03/2008 **Engelmann, Dirk and Veronika Grimm:** Mechanisms for Efficient Voting with Private Information about Preferences

04/2008 **Schnabel, Claus and Joachim Wagner:** The Aging of the Unions in West Germany, 1980-2006

05/2008 **Wenzel, Tobias:** On the Incentives to Form Strategic Coalitions in ATM Markets

06/2008 **Herrmann, Klaus:** Models for Time-varying Moments Using Maximum Entropy Applied to a Generalized Measure of Volatility

07/2008 **Klein, Ingo and Michael Grottke:** On J.M. Keynes' "The Principal Averages and the Laws of Error which Lead to Them" - Refinement and Generalisation