UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Center For International Economics

Working Paper Series

Working Paper No. 2015-01

An iterative plug-in algorithm for realized kernels

Yuanhua Feng and Chen Zhou

January 2015

CIE Center For International Economics

An iterative plug-in algorithm for realized kernels

Yuanhua Feng[1] and Chen Zhou
University of Paderborn

Abstract

Realized kernels introduced by Barndorff-Nielsen et al. (2008) are consistent estimators of the daily integrated volatility in the presence of microstructure noise. A crucial problem by applying realized kernels is the selection of the bandwidth. This paper proposes an iterative plug-in algorithm to solve this problem under independent microstructure noise, which adapts the idea of Gasser et al. (1991) in nonparametric regression to the current context. It is shown that the selected bandwidth is consistent up to a bias factor due to the use of a biased formula of the asymptotically optimal bandwidth. The nice practical performance of the proposal is illustrated by application to data of a few German and French firms within a period of several years. Further analysis of the obtained realized kernels using a most recently proposed exponential SEMIFAR model is also discussed.

Keywords: Realized kernels, microstructure noise, bandwidth selection, iterative plug-in, slowly changing volatility trend, long memory

JEL Codes: C14, C58

[1]Corresponding author with email: yuanhua.feng@wiwi.upb.de.

1 Introduction

Estimation of the daily integrated volatility (IV) is an important topic in risk management, portfolio allocation and option pricing. Realized volatility (RV) introduced by Andersen et al. (2001a, b) is a model-free estimator of this quantity based on high-frequency financial data. The most simple definition of the RV, called RV_0, is the sum of the squared intraday returns. It is however found that high-frequency data often exhibit microstructure (MS) noise (Hasen and Lunde, 2006). Strong evidence for the existence of MS noise is illustrated in Figure 1 in the next section using numerical examples. Now, RV_0 is an inconsistent estimator of the IV (Zhang et al., 2005, Bandi and Russel, 2008). Different bias corrected estimators of the IV are introduced into the literature. For instance, Zhou (1996) proposed an improved estimator, called RV_Z, by including the cross-products between two consequent observations, which is unbiased under i.i.d. MS noise. Bandi and Russel (2006, 2008, 2011) and Oomen (2006) investigated the use of sparse equidistant high-frequency data and studied the choice of the optimal frequency to make a trade-off between the variance and bias of the proposed estimators. Zhang et al. (2005) and AÏt-Sahalia et al. (2011) proposed the use of a realized volatility estimator with two time scales to solve the bias problem caused by MS noise. Hansen and Lunde (2006) and Oomen (2005) proposed a simple kernel based estimator of the IV. Furthermore, Hansen et al. (2008) investigated correction of MS bias using moving average-based estimators.

Recently, Barndorff-Nielsen et al. (2008, 2009, 2011) introduced the realized kernels (RK), which are consistent estimators of the IV under given conditions. A crucial problem by applying realized kernels is the selection of the bandwidth, because an RK only works well, if the bandwidth is selected properly. This is illustrated in Figure 2 in the next section through the above mentioned numerical examples. Barndorff-Nielsen et al. (2009) proposed to select the bandwidth by plugging suitable estimates of two unknowns into a simplified (but biased) formula of the asymptotically optimal bandwidth of the RK. However, their proposal is very complex and not fully data-driven. And the selected bandwidth by this algorithm does not converge to the targeted bandwidth.

In this paper an iterative plug-in (IPI) bandwidth selector for realized kernels is developed by adapting the idea of Gasser et al. (1991) to the current context. So far as we know, this is the first IPI algorithm for RK. For simplicity, we also adopt the biased

targeted bandwidth proposed by Barndorff-Nielsen et al. (2009). The difference between RV_0 and the RK resulted in each iteration is used to estimate the variance of the MS noise. And RV_Z is used as an initial value of the RK so that the procedure is fully data-driven. It is shown that the proposed bandwidth selector is consistent in the sense that the relative error with respect to the targeted bandwidth tends to zero, as $n \to \infty$. Furthermore, the proposed bandwidth selection rule is very simple and the algorithm runs very fast, because only a few iterations are required. It is hence suitable to be applied to obtain data-driven RK in a long observation period. Theoretically, both of the resulted RK and the selected bandwidth become consistent from the third iteration, while their rate of convergence can still be improved in the fourth iteration. Thereafter, the resulted RK achieves its optimal rate of convergence of the order $O(n^{-1/5})$, which is also shared by the relative error in the selected bandwidth. The nice practical performance of the proposal is illustrated by application to data of a few German and French firms. These results show that in most of the cases the procedure converges within four iterations. And the distribution of the selected bandwidths is nearly normal. Empirical analysis showed that the resulted RK performs better than RV_0 and RV_Z. It seems that both of the bias and the standard deviation of RV_0 and RV_Z are clearly reduced by the data-driven RK. But this fact still needs to be confirmed through simulation. The performance of the proposed bandwidth selector on a few so-called 'challenging days' is discussed in detail.

Further analysis of the obtained results is of great interest. Andersen et al. (2001a, b, 2011) and Deo et al. (2006) find that the logarithmic RV may exhibit long memory. Choi et al. (2010) found that the observed long memory may be spuriously generated e.g. by a nonparametric trend or structural breaks. Hence, long memory, nonparametric trends and possible structural breaks should be studied simultaneously. We propose to analyze realized kernels use a piecewise version of the ESEMIFAR (exponential semiparametric fractional autoregressive, Beran et al., 2015). It is found that realized kernels exhibit long memory and a significant nonparametric trend at the same time. Estimation results for the two sub-periods before and after the 2008 financial crisis are clearly different.

The paper is organized as follows. Some necessary known results are summarized in Section 2. The data-driven bandwidth selector is proposed and studied in Section 3. Application to real data is reported in Section 4. In Section 5 modeling of realized kernels using the ESEMIFAR model is discussed. Final remarks in Section 6 conclude the paper.

2 Realized volatility and realized kernels

2.1 Effect of MS noise on realized volatility

Let $p^*(\tau)$ denote the logarithmic efficient asset price on a trading day, where $0 \leq \tau \leq T$, and 0 and T denote the opening and closing time. Assume that $p^*(\tau)$ are determined by the stochastic differential equation

$$dp^*(\tau) = \sigma(\tau)dW(\tau), \tag{1}$$

where $W(\tau)$ is a standard Brownian motion and $\sigma(\tau)$ is the spot volatility process. Furthermore, it is assumed that the $\sigma(\tau)$ and $W(\tau)$ processes are independent of each other. Estimation of the daily integrated volatility

$$\text{IV} = \int_0^T \sigma^2(\tau)d\tau \tag{2}$$

is of great interest. Realized volatility is introduced as a model free estimator of the IV based on high-frequency financial data. Let p_i be the logarithmic asset prices observed at time points $0 = \tau_0 < \tau_1 < \ldots < \tau_n < \tau_{n+1} = T$, where n is the (random) number of observations happened on that day. It is assumed that $\tau_i - \tau_{i-1} = O_p(n^{-1})$. The intraday returns are given by $r_i = p_i - p_{i-1}$. The most simple definition of the realized volatility is

$$\text{RV}_0 = \sum r_i^2, \tag{3}$$

which is a consistent estimator of the IV, if there is no MS noise such that $p_i = p_i^*$, where p_i^* stands for $p^*(\tau_i)$. In the presence of MS noise we have however

$$p_i = p_i^* + u_i, \tag{4}$$

where u_i represents a stationary noise process with mean zero and $\text{var}(u_i) = \omega^2$. It is assumed that u_i is independent of p_i^*. In this paper we will focus on the case with i.i.d. u_i. Let $r_i^* = p_i^* - p_{i-1}^*$ be the efficient returns. The corresponding noise contaminated observed returns are given by

$$r_i = p_i - p_{i-1} = r_i^* + e_i, \tag{5}$$

where $e_i = u_i - u_{i-1}$ is the noise in r_i. The observed returns are correlated to each other, while r_i^* are uncorrelated. Under the i.i.d. assumption on u_i, e_i follow an MA(1) model.

It can be shown that, the ACF of r_i at lag 1 is $\rho_r(1) = -\omega^2/(2\omega^2 + \sigma_i^2) \to -0.5$, as $n \to \infty$, where $\sigma_i^2 = \text{var}(r_i^*)$. If ω^2/σ_i^2 is large, RV_0 is clearly overestimated.

Empirical evidence of MS noise can be found by displaying the ACF of high-frequency returns. Figure 1 shows the correlograms of high-frequency returns on four selected trading days, one from each of the following German and French companies, Air France (AF), Allianz (ALV), BMW and Peugeot (PSA), respectively. From Figure 1 we see that $\rho_r(1)$ is always significantly negative, a clear evidence for the existence of MS noise. The independence assumption on the noise is about true in the first three cases. Furthermore, we see the noise on the first selected day is very strong with $\hat{\rho}(1) < -0.4$ as can be seen from Figure 1(a). Figures 1(b) and (c) show that the noise on the second and third selected days is at a middle and a relatively low level, respectively. The fourth example in Figure 1(d) is chosen to show that strong and dependent noise could also happen. For this example not only $\hat{\rho}_r(1)$ but also those at lags 2 and 3 are significantly non-zero with $\hat{\rho}_r(2) > 0$. But the sum of $\hat{\rho}_r(1)$ to $\hat{\rho}_r(3)$ is clearly negative.

In the presence of MS noise, RV_0 can be rewritten as

$$\text{RV}_0 = \sum_{i=1}^{n} (r_i^*)^2 + 2\sum_{i=1}^{n} r_i^* e_i + \sum_{i=1}^{n} e_i^2. \tag{6}$$

It is well known that the bias of RV_0 is $B(\text{RV}_0) = 2n\omega^2$ and the asymptotic variance of RV_0 is $\text{var}\,(\text{RV}_0) \approx 4nE(u_i^4)$, as $n \to \infty$. Different approaches are introduced into the literature to improve the performance of RV_0. Under the i.i.d. assumption on u_i, Zhou (1996) proposed to correct the bias in RV_0 by introducing the cross-products of lag 1 into RV_0. In this paper his proposal is slightly modified as follows:

$$\text{RV}_\text{Z} = \sum_{i=2}^{n-1} (r_i^2 + r_i r_{i+1} + r_i r_{i-1}). \tag{7}$$

Under independent MS noise RV_Z is unbiased and its variance is approximately $8n\omega^4$, as $n \to \infty$. That is this estimator is still inconsistent.

2.2 Realized kernels

To overcome the above mentioned problems of well known estimators of the IV, Barndorff-Nielsen et al. (2008, 2009, 2011) introduced the realized kernels, which are consistent

estimators of the IV in the presence of MS noise under regularity conditions. A RK is defined by

$$RK = \sum_{h=-H}^{H} k\left(\frac{h}{H+1}\right)\gamma_h, \quad \gamma_h = \sum_{j=|h|+1}^{n} r_j r_{j-|h|}, \tag{8}$$

where $k(u)$ is a kernel weight function, H is the bandwidth and γ_h is the h-th realized autocovariance. To ensure the non-negativity of the RK, it is assumed that $k(u)$ satisfies the Condition **K** in Barndorff-Nielsen et al. (2011). This implies in particular that the kernel is with a non-flat top such that $k''(0)$, the second derivative of $k(u)$ at the origin, is non-zero. A variety of kernel functions in this class may be found in Table 1 of Barndorff-Nielsen et al. (2011). The authors indicated that the use of the Parzen kernel is more preferable. For $u \geq 0$, the Parzen kernel is defined by

$$k(u) = \begin{cases} 1 - 6u^2 + 6u^3, & 0 \leq u \leq 1/2, \\ 2(1-u)^3, & 1/2 < u \leq 1, \\ 0 & u > 1. \end{cases} \tag{9}$$

This kernel will be used in the numerical part of this paper.

Asymptotic properties of the RK are studied by Barndorff-Nielsen et al. (2008, 2011). See also Ikeda (2013). Assume that the bandwidth H is of the order $H = O(n^\alpha)$ with $0 < \alpha < 1$, asymptotic bias and variance of an RK are given by

$$B(RK) \approx [k''(0)]^2 \omega^2 \frac{n}{H^2} \tag{10}$$

and

$$\text{var}\,(RK) \approx 4T k_\bullet^{0.0} \int_0^T \sigma^4(\tau) d\tau \frac{H}{n} + C_1 \frac{n}{H^3} + C_2 \frac{1}{H}, \tag{11}$$

where $k_\bullet^{0.0} = \int_0^\infty k^2(u)du$, and C_1 and C_2 are two constants. The quantity $\int_0^T \sigma^4(\tau)d\tau$ is called the daily integrated quarticity. These results indicate that variance of an RK is asymptotically negligible, if $\alpha > 1/3$, and both of its asymptotic variance and bias are negligible, if $\alpha > 1/2$. The asymptotic variance is dominated by the second term on the right-hand-side of (11), if $1/3 < \alpha < 1/2$, and by the first term, if $\alpha > 1/2$. The asymptotically optimal bandwidth (Barndorff-Nielsen et al., 2009, 2011), which minimizes the dominating part of the MSE (mean squared error) of an RK, is given by

$$H_{\text{A}} = c_0 \xi^{4/5} n^{3/5} \quad \text{with} \tag{12}$$

$$c_0 = \left\{ \frac{k''(0)^2}{k_\bullet^{0.0}} \right\}^{1/5} \quad \text{and} \quad \xi^2 = \frac{\omega^2}{\sqrt{T \int_0^T \sigma(\tau)^4 d\tau}}.$$

For the Parzen kernel we have $c_0 = 3.5134$. We see the optimal bandwidth for an RK with a non-flat top kernel is of the order $O(n^{3/5})$. If a bandwidth of this order is employed, the resulted RK will achieve its optimal convergence rate of the order $O(n^{-1/5})$.

The above theoretical results show that realized kernels work well, only if the bandwidth is chosen properly. To show this, the dependence of the RK on the bandwidth H is displayed in Figure 2 for the four selected examples, where the vertical line in each panel highlights the bandwidth selected by the procedure proposed in the next section with $\hat{H} = 55$, 20, 13 and 24, respectively. Figure 2 shows that an RK is very sensitive to the change of the used bandwidth, if H is small. In a common case, like those in Figures 2(a), (b) and (d), the change in H usually does not have a clear effect on the resulted RK, if the used bandwidth is large. However, Figure 2(c) indicates that sometimes both of a too large or a too small bandwidth can lead to a clearly wrong estimation result. Detailed discussion on the selected bandwidths will be given in Section 4.

3 Bandwidth selection for realized kernels

Examples in Figure 2 show that the selection of the bandwidth is a crucial problem for the application of the RK. In the current context the number of observations on a trading day is very large and one usually would also like to estimate the RK for a number of firms within a long observation period. Hence, we aim at the development of a quick bandwidth selector for the RK with nice theoretical and practical performance. An plug-in bandwidth selector can be obtained by inserting estimates of ω^2 and $\int_0^T \sigma^4(\tau)d\tau$ into H_{A}. However, the estimation of $\int_0^T \sigma^4(\tau)d\tau$ is not yet well solved in the literature. Barndorff-Nielsen et al. (2009) proposed a plug-in bandwidth selector based on the following formula:

$$H_{\mathrm{B}} = c_0 \xi_B^{4/5} n^{3/5} \tag{13}$$

with ξ^2 in H_A being replaced by $\xi_B^2 = \omega^2/\mathrm{IV}$, which is of the same order as H_{A} but with a biased factor in the constant. The reason is that $T \int_0^T \sigma^4(\tau)d\tau$ can be well approximated through IV^2, if $\sigma(\tau)$ does not vary too much. This biased version of the optimal bandwidth

will also be employed in the current paper. Now, assume that $\widehat{\mathrm{IV}}$ is an at least unbiased estimator of IV, it is easy to show that

$$\hat{\omega}^2 = \frac{\mathrm{RV}_0 - \widehat{\mathrm{IV}}}{2n} \tag{14}$$

is an consistent estimator of ω^2. Furthermore, if $\widehat{\mathrm{IV}}$ is a consistent estimator, H_{B} can be estimated consistently by replacing ξ^2_{B} with

$$\hat{\xi}^2_{\mathrm{B}} = \hat{\omega}^2/\widehat{\mathrm{IV}}. \tag{15}$$

The difference between H_{A} and H_{B} is an constant factor $H_{\mathrm{B}}/H_{\mathrm{A}} = (T\int_0^T \sigma_u^4 du/\mathrm{IV}^2)^{1/5}$, which is usually slightly bigger than one. In the following a consistent bandwidth selector of H_{B} is proposed by adapting the IPI idea of Gasser et al. (1991) to the current context with RV_0 and RV_{Z} as the initial values. The proposed algorithm reads as:

Step 1. In the first iteration let $\widehat{\mathrm{IV}}_1 = \mathrm{RV}_{\mathrm{Z}}$. Calculate $\hat{\omega}^2_1$ and $\hat{\xi}^2_1$ following (14) and (15). Insert the latter into (13) to obtain $\hat{H}_1$. Put $j = 2$.

Step 2. In the jth iteration with $j > 1$, calculate $\widehat{\mathrm{IV}}_j$ with $\hat{H}_{j-1}$. Then calculate $\hat{\omega}^2_j$ and $\hat{\xi}^2_j$, and obtain $\hat{H}_j$ similar to Step 1.

Step 3. Increase j by 1 and carry out Step 2 repeatedly. The procedure will be ended, if convergence is achieved or some stopping criterion is fulfilled, or a maximal number of iterations J is carried out. Put $\hat{H} = \hat{H}_j$.

We will see that $\hat{H}_1$ is an inconsistent bandwidth selector. But after a few iterations, $\hat{H}_j$ will become a consistent estimate of H_{B}. The detailed behavior of $\hat{H}_j$ in each iteration and the theoretical properties of the finally selected bandwidth are discussed in the following theorem and its proof.

Theorem 1. *Assume that Conditions* **K**, **SH**, **D** *and* **U** *in Theorem 2 of Barndorff-Nielsen (2011) hold. Assume further that u_i are i.i.d. and that the end-effect as indicated in that paper is treated suitably, so that it does not affect the asymptotic performance of $\widehat{\mathrm{IV}}_j$ in the proposed procedure. Then we have*

i) $\hat{H}_j$ selected by the proposed procedure with $j \geq 3$ is a consistent estimator of H_{B} in the sense that $(\hat{H}_j - H_{\mathrm{B}})/H_B = o_p(1)$.

ii) For $j \geq 4$ the selected bandwidth is consistent with a relative convergence rate of the order $n^{-1/5}$, i.e. $(\hat{H}_j - H_{\mathrm{B}})/H_{\mathrm{B}} = O_p(n^{-1/5})$.

A sketched proof of Theorem 1. *The following proof is carried out conditioning on given number of observations n on a trading day.*

i) In the first iteration, we have $\widehat{\mathrm{IV}}_1 = \mathrm{IV}_{\mathrm{Z}} = \mathrm{IV} + O_p(n^{1/2})$. Furthermore, it can be shown that $\hat{\omega}_1^2$ is $\sqrt{n}$-consistent such that $\hat{\omega}_1^2 = \omega^2[1 + O_p(n^{-1/2})]$ and $\hat{\xi}_1^2 = O_p(n^{-1/2})$. This results in an estimate $\hat{H}_1 = O_p(n^{\alpha_1})$ with $\alpha_1 = 2/5 > 1/3$. Following the asymptotic results summarized in (10) and (11), the use of $\hat{H}_1$ in the second iteration will lead to an estimate with an asymptotically negligible variance and a random bias term of the order $O\left(\frac{n}{H_1^2}\right) = O_p(n^{1/5})$. That is we have $\widehat{\mathrm{IV}}_2 = \mathrm{IV} + O_p(n^{1/5}) + o_p(1)$. Now, it can be shown that $\hat{\omega}_2^2 = \omega^2[1 + O_p(n^{-1/2}) + O_p(n^{-4/5})]$ with an additional term caused by the bias in $\widehat{\mathrm{IV}}_2$, which is still $\sqrt{n}$-consistent. Furthermore, we have $\hat{\xi}_2^2 = O_p(n^{-1/5})$. Insert these results into the proposed algorithm we obtain $\hat{H}_2 = O_p(n^{\alpha_2})$ with $\alpha_2 = 13/25 > 1/2$. The estimates $\widehat{\mathrm{IV}}_3$, $\hat{\omega}_3^2$ and $\hat{\xi}_3^2$ in the third iteration obtained with $\hat{H}_2$ are all consistent. Hence, $\hat{H}_3$ is a consistent estimate of H_{B} in the relative sense.

ii) Note that the error of $\hat{H}_j$ is dominated by that of $\widehat{\mathrm{IV}}_j$. Using Taylor expansion of a random function it can be shown that the rate of convergence of H_j is the same as that of $\hat{\xi}_j^2$. From the fourth iteration onwards, $\hat{H}_j$ achieves its optimal rate of convergence of the order $O_p(n^{-1/5})$ in a relative sense with $(\hat{H}_j - H_{\mathrm{B}})/H_{\mathrm{B}} = O_p(n^{-1/5})$ for any $j \geq 4$. $\diamond$

The proof above shows that, theoretically, at least three steps are required to achieve a consistent selector of H_{B}. Asymptotically, the performance of $\hat{H}_4$ might be slightly better than that of $\hat{H}_3$, because $\widehat{\mathrm{IV}}_4$ is obtained with a consistent bandwidth selector. The proposed data-driven algorithm and the above theoretical results can be easily adapted to the case, if an unbiased estimate of $\int \sigma^4(\tau)d\tau$ is used. Furthermore, note that the bandwidth for an RK is an integer. Small changes in the involved quantities often do not have any effect on the finally selected bandwidth. Hence, the proposed algorithm converges very quickly. This is also confirmed by the application in the next section.

An R code is developed for practical implementation of the proposed bandwidth selector. The procedure will be stopped, if $\hat{H}_j = \hat{H}_{j-1}$ is achieved. It is also found that sometimes the selected bandwidths take two consequent integers alternatively. If this happens, the procedure will also be ended. Now, the larger of the two selected bandwidths

will be used. Both cases will be considered as regular cases (Reg. Case). Furthermore, the following three special cases can also happen. The first special case (Sp. Case 1) is that with $\mathrm{RV_Z} < 0$. This indicates that the MS noise should be very strong (and maybe correlated). Now, $\hat{H}_1$ can not be calculated according to the proposed algorithm, because we have $\widehat{\mathrm{IV}}_1 < 0$. In this case we will manually set $\hat{\xi}_1^2 = 100/(2n)$, which will lead to a big starting bandwidth $\hat{H}_1$. Nevertheless the procedure runs very well and, after a few iterations, H_j will converge to the selected optimal bandwidth, which is independent of $\hat{H}_1$. The second special case (Sp. Case 2) is that with $\mathrm{RV_Z} > \mathrm{RV_0}$, which indicates probably that there is no MS noise on that day. Now, the procedure cannot be carried out, because we have $\hat{\omega}_1^2 < 0$. In this case, we will set $\hat{H} = 0$ and simply use $\mathrm{RV_0}$. The last special case (Sp. Case 3) will happen, when RK becomes bigger than $\mathrm{RV_0}$ in some iteration with $j > 1$. This means again that there is no strong MS noise in the observed prices on that day. And now the proposed algorithm cannot be carried out further. Hence, we will put $\hat{H} = \hat{H}_j$ as the selected optimal bandwidth.

Note that the optimal bandwidth for an RK under independent and dependent MS noise is of the same order of magnitude. The proposal bandwidth selector can hence be applied to the case with dependent MS noise. The example in Figure 2(d) also indicates that the proposal works in this case. But now, the selected bandwidth is only sub-optimal, because it is only of a correct order but with a clearly biased constant.

4 Application

The proposed algorithm is applied to the datasets of AF, ALV, BMW and PSA from 2. Jan. 2006 to 30. Jun. 2012, downloaded from the "Thomson Reuters" Corporation. The total number of trading days for the two German Stocks is 1655 and that for the two French Stocks is 1664. The numbers of days in the four cases with different behavior of the algorithm as described in the last section are listed in Table 1. We see the proposed algorithm converged for more than 99% of those datasets. The three special cases only happened with a very small chance. Sp. Case 1 occurred only once by PSA. Sp. Case 2 and Sp. Case 3 for AF and ALV also occurred rarely. For BMW and PSA, Sp. Case 3 occurred on 11 and 8 days, respectively. Now, the ratio of Sp. Case 3 is still clearly smaller than 1%. Therefore, the three special cases can be considered as some rare extreme events.

Trading days on which Sp. Cases 2 or 3 happened will be called challenging days. Now, the proposed algorithm does not work well. This will be discussed in Section 4.2 in detail.

4.1 Summary of the general findings

Figure 3 shows the histograms of the selected optimal bandwidths for the four companies. It seems that the proposed bandwidth selector is nearly asymptotically normally distributed. For finite samples the distribution is sometimes slightly skewed to the right with very fewer extremely large selected bandwidths. The selected bandwidths are usually between 5 and 25. The largest selected bandwidth is 55 by AF on 18. Jul. 2007 as indicated in Figure 3(a). As defined before, the selected bandwidth is 0, if Sp. Case 2 happened. From Figure 2 we can see that the use of the selected bandwidth leads to an estimate, which is at a very low level, but not the lowest value of all possible RK. This feature is as expected and shows that the proposed bandwidth selector works very well in practice. This nice property is particularly highlighted by Figure 2(c). Note that the bandwidth selected by the procedure of Barndorff-Nielsen et al. (2009) is usually very large. This is not only caused by the use of different algorithms, but also by the different features of the used datasets. It is of great interest to carry out a comparative study between the two proposals theoretically and through simulation. This is however beyond the aim of the current paper and will be discussed elsewhere. The histograms of the numbers of iterations for all companies are displayed in Figure 4. In the R code a maximal number of iteration $J = 15$ is used. For the datasets under consideration this limit is never achieved. The maximal number of iterations occurred is 11 by PSA on one day. The maximal numbers of iterations by AF, ALV and BMW are 6, 5 and 6, respectively. And the most possible number of iterations for all of the four companies is 3. Furthermore, in most of the cases the proposed algorithm converges within four iterations. This confirmed the results of Theorem 1.

The estimated RV_0, RV_Z and RK are summarized in Table 2, where the t statistic for the differences between RV_0 and RV_Z, and those between RV_Z and RK are also given in the second and third rows, respectively. These t values are calculated under the assumption that those differences in a given case are i.i.d. We see that for each of the four companies the mean of RV_0 is much larger than those of RV_Z and RK. The mean of RV_Z is also

bigger than that of RK. The differences between those mean values are always very highly significant. The difference between the means of RV_0 and RV_Z indicates the part of the bias caused by the MS noise, which can be discovered by $\hat{\rho}(1)$ of the returns. And the difference between the means of RV_Z and RK indicates additional bias caused by possible dependent MS noise, which can not be reflected by $\hat{\rho}(1)$ of the returns. The differences between the standard deviations of those estimators are similar to that mentioned above. It can also be shown that those differences are always significant. Details to this end are omitted to save space. In summary, the use of the proposed data-driven RK will lead to a clear reduction of the bias and the variation, comparing with the two well known estimators RV_0 and RV_Z. These empirical findings show that the proposed approach works well in practice. However, the practical performance of the proposed data-driven algorithm for RK still need to be confirmed by means of a simulation study.

The results of RV_0, RV_Z and RK for AF after logarithmic transformation are shown in Figure 5. From this figure we can see that, in addition to the differences among the estimates obtained by these different approaches, they also exhibit quite similar common patterns. In particular all of these series seem to have a non-stationary trend component and possible structural breaks caused by two financial crises, i.e. the global financial crisis in 2008 and the European debt crisis in 2011, respectively. The results of the data-driven RK for ALV, BMW and PSA, again in log-scale, are displayed in Figure 6. We see, these series also share similar patterns as those displayed in 5(c). Moreover, an interesting empirical finding is that, in addition to the common general tendency of those RK series, they seem also correlated to each other strongly. This feature is helpful for further modeling and forecasting of RK.

4.2 Detailed analysis of two challenging cases

Now, we will discuss briefly, why the proposed data-driven algorithm does not work well in Sp. Cases 2 and 3. The dataset of AF on the 15. Sept. 2011 was chosen as an example of Sp. Case 2 and that of BMW on the 21. May 2009 was chosen as an example of Sp. Case 3. Figure 7 shows the ACF of the intraday returns on those two challenging days. As shown in Figure 1, usually, MS noise will cause a clearly significant negative ACF at lag 1. From Figure 7 we can see that this is not true in both examples selected here.

Figure 7 (a) shows that almost all of the estimated ACF in this case are insignificant. But the ACF at lag 1 happens to be slightly positive. This results in turn in the fact that $\mathrm{RV_Z}$ is slightly larger than $\mathrm{RV_0}$. In this case we proposed the use of $\hat{H} = 0$, because now the effect of the MS noise seems to be unclear. One problem can arise in the presence of dependent MS noise. Now, it can happen that although the ACF at lag 1 is positive, but some ACF at higher lags can be negative so that $\mathrm{RV_0}$ is still biased. This kind of effect of MS noise can however not be corrected by the proposed data-driven RK. The problem in Sp. Case 3 is different. As we can see, now the ACF at lag 1 is negative and hence the proposed bandwidth selection algorithm can be started. However, some other ACF are clearly positive so that the sum of the ACF is now positive. This indicates again the existence of possible dependent MS noise. This kind of noise could however cause a negative bias in $\mathrm{RV_0}$. Following our proposal, the resulted RK in this special cease is always slightly bigger than $\mathrm{RV_0}$. The effect of this kind of possibly dependent MS noise can also not be captured by the proposed algorithm. Both examples indicate that the proposed algorithm should still be improved and it is worthy to development a data-driven RK by taking possibly dependent microstrucutre noise into account.

5 Further analysis using the Semi-FI-Log-ACD

Further analysis of the obtained RK is of great interest. Ebens (1999) showed that the distribution of the logarithmic volatility is approximately normal. Anderson et al. (2003), Corsi (2009) and Koopman et al. (2005) showed that logarithmic realized volatility may exhibit high persistence. From Figures 5 and 6 we can see that the logarithmic RK may also exhibit a deterministic nonparametric trend. The well known SEMIFAR (Beran and Feng, 2002) is a nonparametric regression model with long-range dependence. Most recently, Beran et al. (2015) proposed to apply the SEMIFAR model to logarithmic transformation of nonnegative financial time series. Their proposal is hence called an ESEMIFAR model, which can be applied to RK. See also Feng and Zhou (2015) for discussion on forecasting based on this approach. Assume that Z_t, the log-transformed RK, follow a SEMIFAR model

$$(1-B)^d \phi(B)[Z_t - \mu(\tau_t)] = \varepsilon_t, \tag{16}$$

where B denotes the backshift operator, $\phi(B)$ is the AR-characteristic polynomial, ε_t are i.i.d. normally distributed random variables with $E(\varepsilon_t) = 0$ and $\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$; $d \in (-0.5, 0.5)$ and $\tau = t/n$ denotes the rescaled time. The existing data-driven algorithms of the SEMIFAR can be used to fit (16), where the AR model is selected by the BIC. A very nice property of this proposal is that, if $d > 0$, the long memory parameter in the original and the log-data is the same. See Beran et al. (2015) for more details.

In the following, the RK series of Air France is used as an example. Like the nonparametric trend, a financial crisis will also cause spurious long memory, if long memory is estimated without taking possible structural breaks into account. Hence, we will apply the ESEMIFAR model to the whole series as well as to the two sub-series from 2. Jan. 2006 to 30. Sept. 2008, and from 1. Oct. 2008 to 30. Apr. 2011. These sub-periods are defined manually. Discussion on the detection of structural breaks under the SEMIFAR model is beyond the purpose of this paper. The sub-series after May 2011 is very short and is hence not considered.

An ESEMIFAR model with a third order local polynomial is fitted to the whole series and to each of the two sub-series mentioned above. The fitted trends together with the data are displayed in Figure 8(a) to (c), respectively. The trend in Figure 8(a) indicates clear effect of the two financial crises on the market volatility. However, it seems that there is no more structural breaks in the two sub-series. And now the ESEMIFAR fits the data well. The selected bandwidths ($\hat{b}$), the estimated long memory parameters ($\hat{d}$) and the selected AR model, if applicable, are listed in Table 3, where the 95%-confidence intervals and the results of the significant test of the fitted trend are also given. From this table we can see that in both sub-periods realized kernels exhibit significant long memory and a significant non-parametric trend simultaneously. In the second sub-period, the short memory part of this model is also significant. Comparing the results for the whole series with those for the two sub-periods, we can see that possible structure breaks cause by the two financial crises exhibit at least the following effects on the estimated ESEMIFAR model: The possible structure breaks resulted in clear overestimation of the long memory parameter, which in turn caused the wrong conclusion, that the estimated trend were insignificant.

6 Final remarks

An IPI algorithm for realized kernels under independent MS noise was proposed. To our knowledge this is the first IPI algorithm in the current context. It is shown that this proposal has some nice theoretical properties, runs very quickly and works usually very well in practice. Possible problems which can happen on some challenging days are discussed in detail. It is also proposed to analyze the resulted RK using the most recently proposed ESEMIFAR model. We also tried to apply this model to different pieces of the whole series. There are still some open questions in this context. Firstly, it is better, if one can find more reasonable solutions to the problems on the challenging days. Secondly, it is worthy to extend the current proposal to cases with dependent MS noise. Thirdly, the proposed bandwidth selector can also be improved, if an unbiased estimator of the daily integrated quarticity can be developed. Furthermore, to apply the idea of the piecewise ESEMIFAR model properly, a suitable approach for detecting structural breaks under the SEMIFAR model should also be developed. Finally, the development of a multivariate semiparametric long memory time series approach for jointly modeling of different RK series is also of great interest.

Reference

AÏt-Sahalia, Y., Mykland, P. A. and Zhang, L. (2011). Ultra high frequency volatility estimation with dependent microstructure noise. *Journal of Econometrics*, 160, 160-175.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Ebens, H. (2001a). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61, 43-76.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2001b). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, 96, 42-55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P. (2003). Modelling and forecasting realized volatility. *Econometrica*, 71, 579-625.

Andersen, T. G., Bollerslev, T. and Meddahi, N. (2011). Realized volatility forecasting and market microstructure noise. *Journal of Econometrics*, 160, 220-234.

Bandi, F. M. and Russell, J. R. (2006). Separating market microstructure noise from volatility. *Journal of Financial Economics*, 79, 655-692.

Bandi, F.M. and Russell, J. R. (2008). Microstructure noise, realized variance, and optimal sampling. *Review of Economic Studies*, 75, 339-369.

Bandi, F. M. and Russell, J. R. (2011). Market microstructure noise, integrated variance estimators, and the accuracy of asymptotic approximations. *Journal of Econometrics*, 160, 145-159.

Barndorff-Nielsen, O. E., Hansen, P., Lunde, A. and Shephard, N. (2008). Designing realized kernels to measure ex-post variation of equity prices in the presence of noise. *Econometrica*, 76, 1481-1536.

Barndorff-Nielsen, O. E., Hansen, P., Lunde, A. and Shephard, N. (2009). Realised kernels in practice: trades and quotes. *Econometrics Journal*, 12, C1-C32.

Barndorff-Nielsen, O. E., Hansen, P., Lunde, A. and Shephard, N. (2011). Multivariate realized kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and nonsynchronous trading. *Journal of Econometrics*, 162, 149-169.

Beran, J. and Feng, Y. (2002). SEMIFAR models- a semiparametric approach to modelling trends, long-range dependence and nonstationarity. *Comptat. Statist. & Data Anal.*, 40, 393-419.

Beran, J., Feng, Y. and Ghosh, S. (2015). Modelling long-range dependence and trends in duration series: an approach based on EFARIMA and ESEMIFAR models. *Statistical Papers*, in press.

Choi, K., Yu, W. and Zivot, E. (2010). Long memory versus structural breaks in modeling and forecasting realized volatiliy. *Journal of International Money and Finance*, 29, 857-875.

Corsi, F. (2009). A simple approximate long memory model of realized volatility. *Journal of Financial Econometrics*. 7, 174-196.

Deo, R., Hurvich, C. and Lu, Y. (2006). Forecasting realized volatility using a long-memory stochastic volatility model. *Journal of Econometrics*, 131, 29-58.

Ebens, H. (1999). Realized stock volatility. Working Paper No. 420, Department of Economics, Johns Hopkins University.

Feng, Y. and Zhou, C. (2015). Forecasting financial market activity using a semiparametric fractionally integrated Log-ACD. *Intern. J. Forecasting*, in press.

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc., 86, 643-652.*

Hansen, P. R., Large, J., Lunde, A. (2008). Moving average-based estimators of integrated variance. *Econometric Reviews*, 27, 79-111.

Hansen, P. R., Lunde, A. (2006). Realized variance and market microstructure noise (with discussion). *Journal of Business and Economic Statistics*, 24, 127-218.

Ikeda, S. (2013). Two scale realized kernels: a univariate case. *Journal of Financial Econometrics*, in press.

Koopman, S. J., Jungbacker, B. and Hol, E. (2005). Forecasting daily variability of the S&P100 stock index using historical, realised and implied volatility measurements. *Journal of Empirical Finance*, 12, 445-475.

Oomen, R. C. A. (2005). Properties of bias-corrected realized variance under alternative sampling schemes. *Journal of Financial Econometrics*, 3, 555-577.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business and Economic Statistics*, 24, 219-237.

Zhang, L., Mykland, P. A. and AÏt-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high frequency data. *Journal of the American Statistical Association*, 100, 1394-1411.

Zhou, B. (1996). High-frequency data and volatility in foreign exchange rates. *Journal of Business and Economic Statistics*, 14, 45-52.

Table 1: Numbers of days in different cases for the four companies

Cases / Firms	**Reg. Case**	**Sp. Case 1**	**Sp. Case 2**	**Sp. Case 3**
AF	1662	0	1	1
ALV	1652	0	1	2
BMW	1643	0	1	11
PSA	1653	1	3	8

Table 2: Statistics of RV_0, RV_Z and RK; t between RV_0 & RV_Z, and RV_Z & RK

	AF			ALV			BMW			PSA		
	mean	s.d.	t	mean	s.d.	t	mean	s.d.	t	mean	s.d.	t
RV_0	12.74	14.60	—	10.18	24.03	—	7.71	11.35	—	12.11	14.02	—
RV_Z	8.05	8.84	26.79	6.82	16.25	17.27	5.54	8.26	26.09	8.28	9.39	27.20
RK	6.29	6.62	23.48	5.28	10.02	9.41	5.16	6.55	6.90	7.07	7.57	19.03

Table 3: Results of ESEMIFAR for realized kernels of Air France

Series	$\hat{h}$	$\hat{d}$ **& %-CI**	$\hat{p}$	$\hat{\phi}_1$ **& 95%-CI**	**trend**
Jan. 2006 - Jun. 2012	0.191	0.437 [0.399, 0.475]	0	—	insign.
Jan. 2006 - Sep. 2008	0.269	0.380 [0.322, 0.438]	0	—	sign.
Oct. 2008 - Apr. 2011	0.149	0.285 [0.175, 0.395]	1	0.152 [0.014, 0.290]	sign.

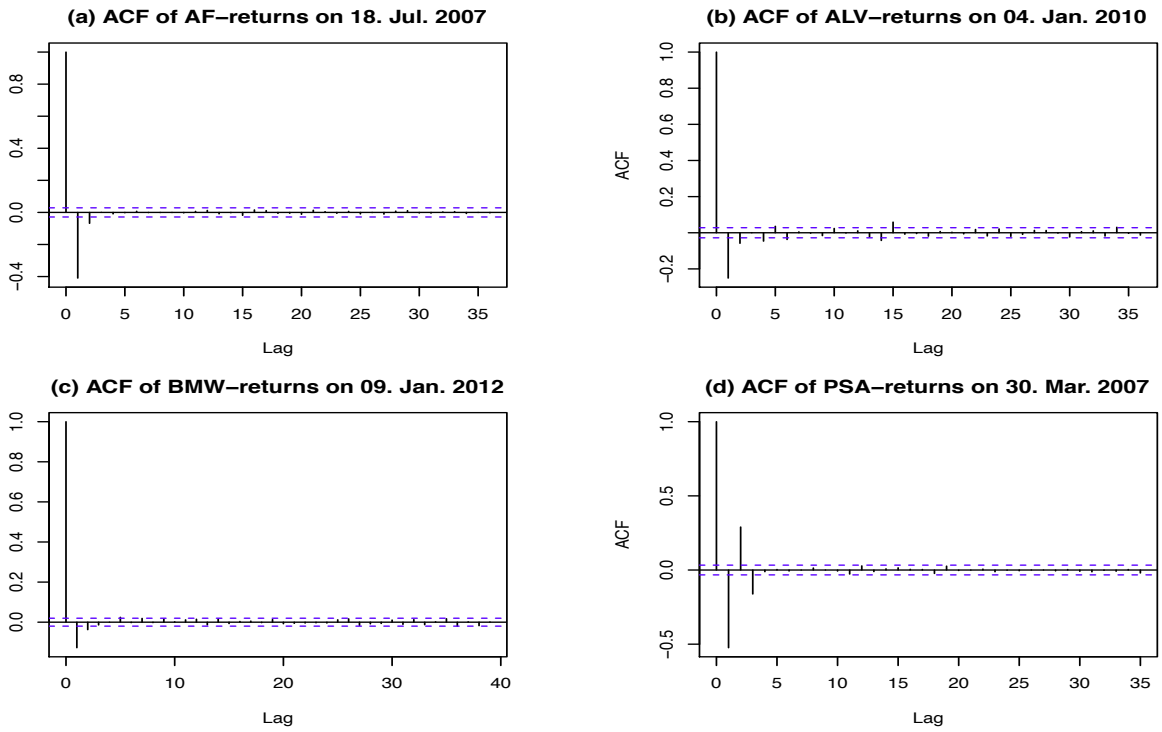

Figure 1: Examples of ACFs of high-frequency returns on four selected days

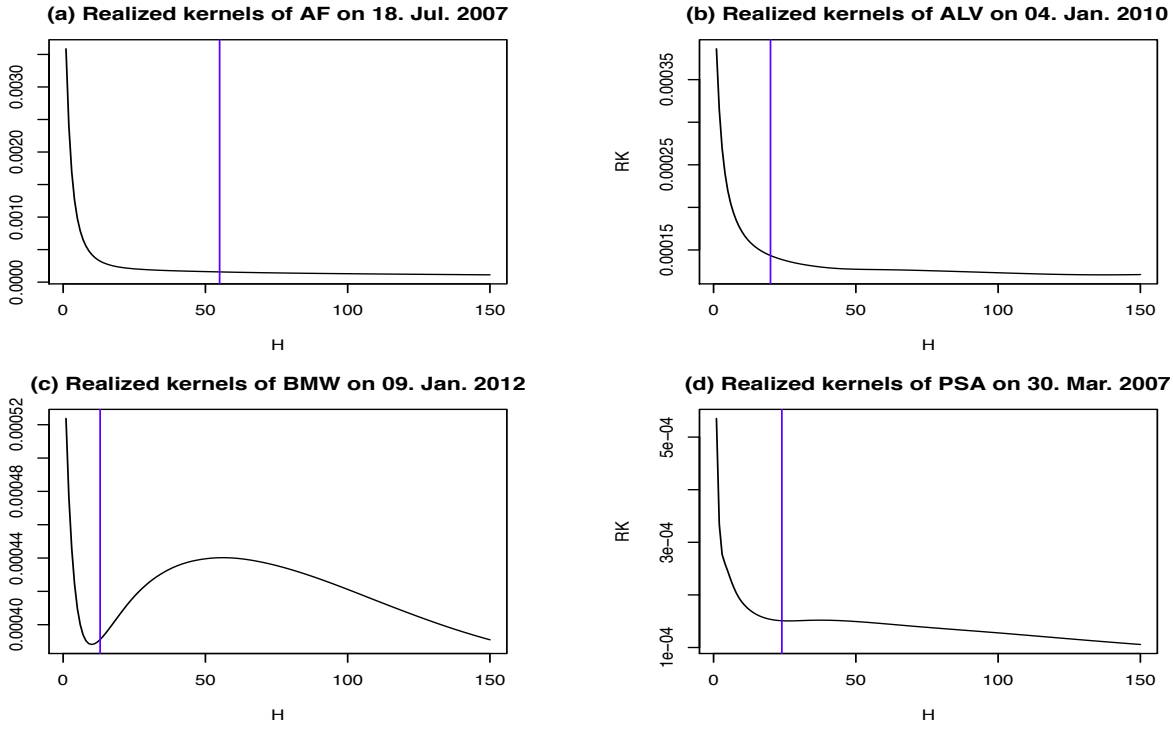

Figure 2: Realized kernels against H obtained on the four selected days

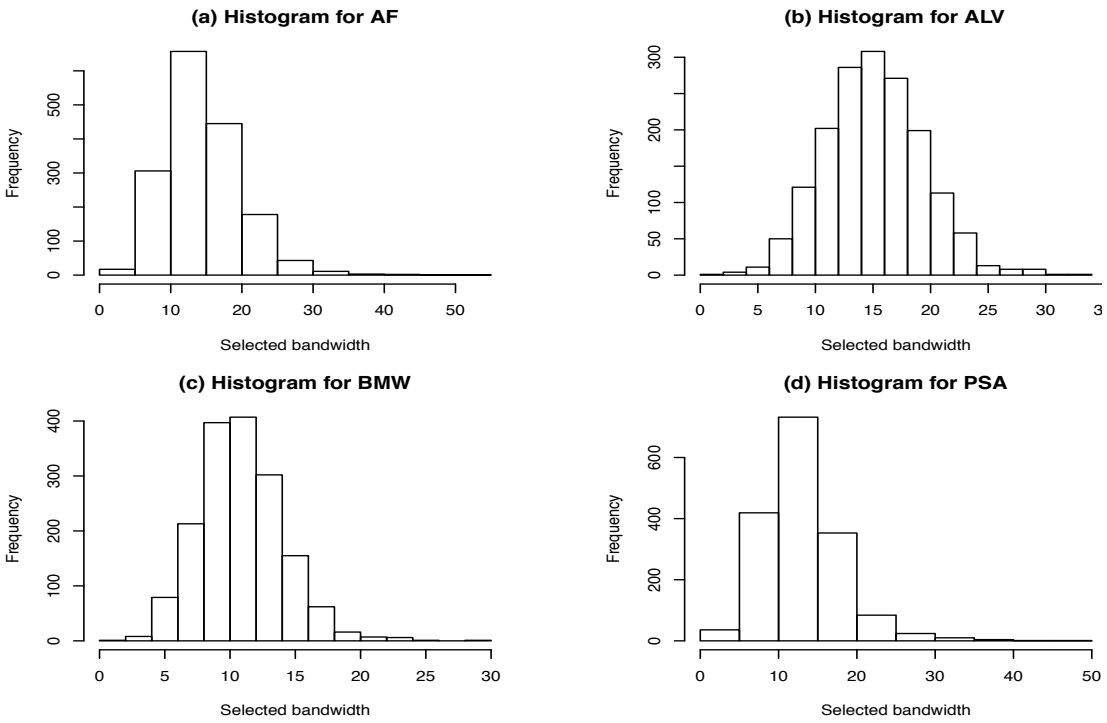

Figure 3: Histograms of selected bandwidth for all examples.

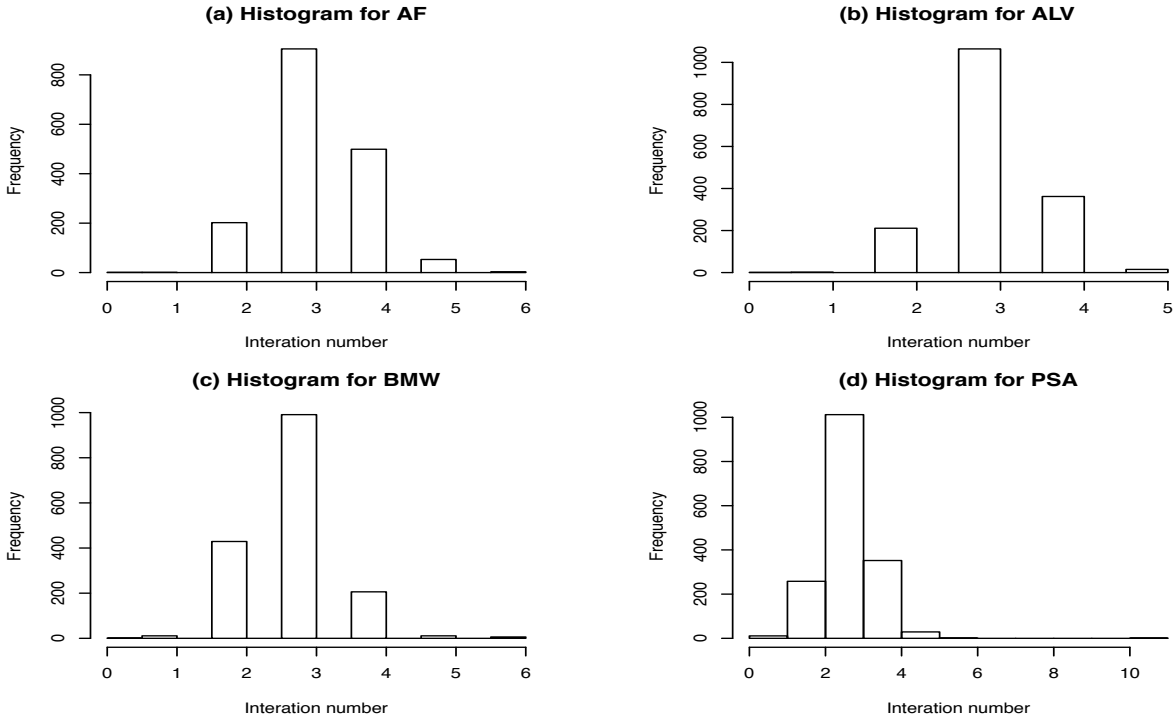

Figure 4: Histograms of the number of iterations for all examples.

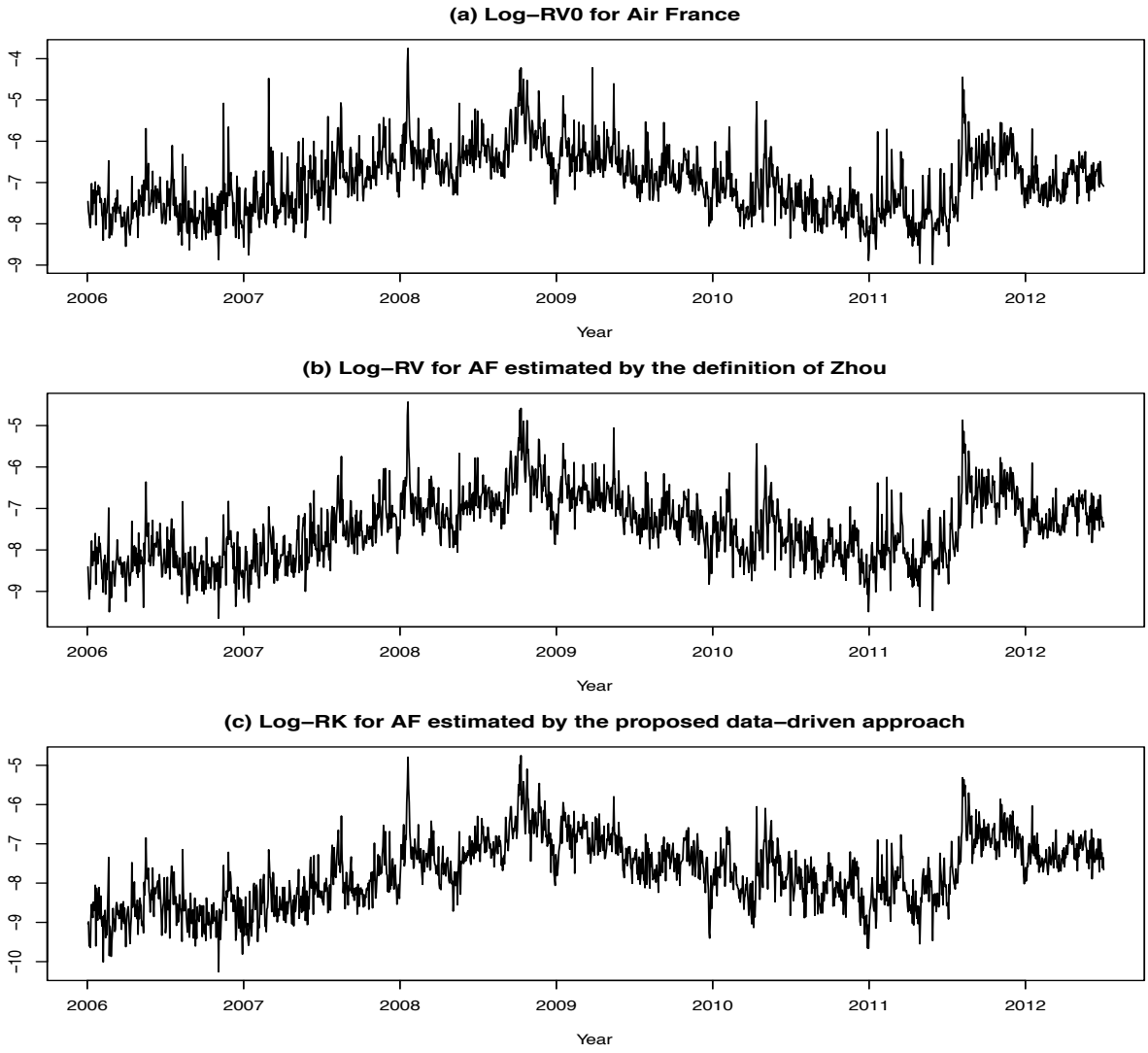

Figure 5: Logarithmic transformation of all realized volatility estimators for Air France

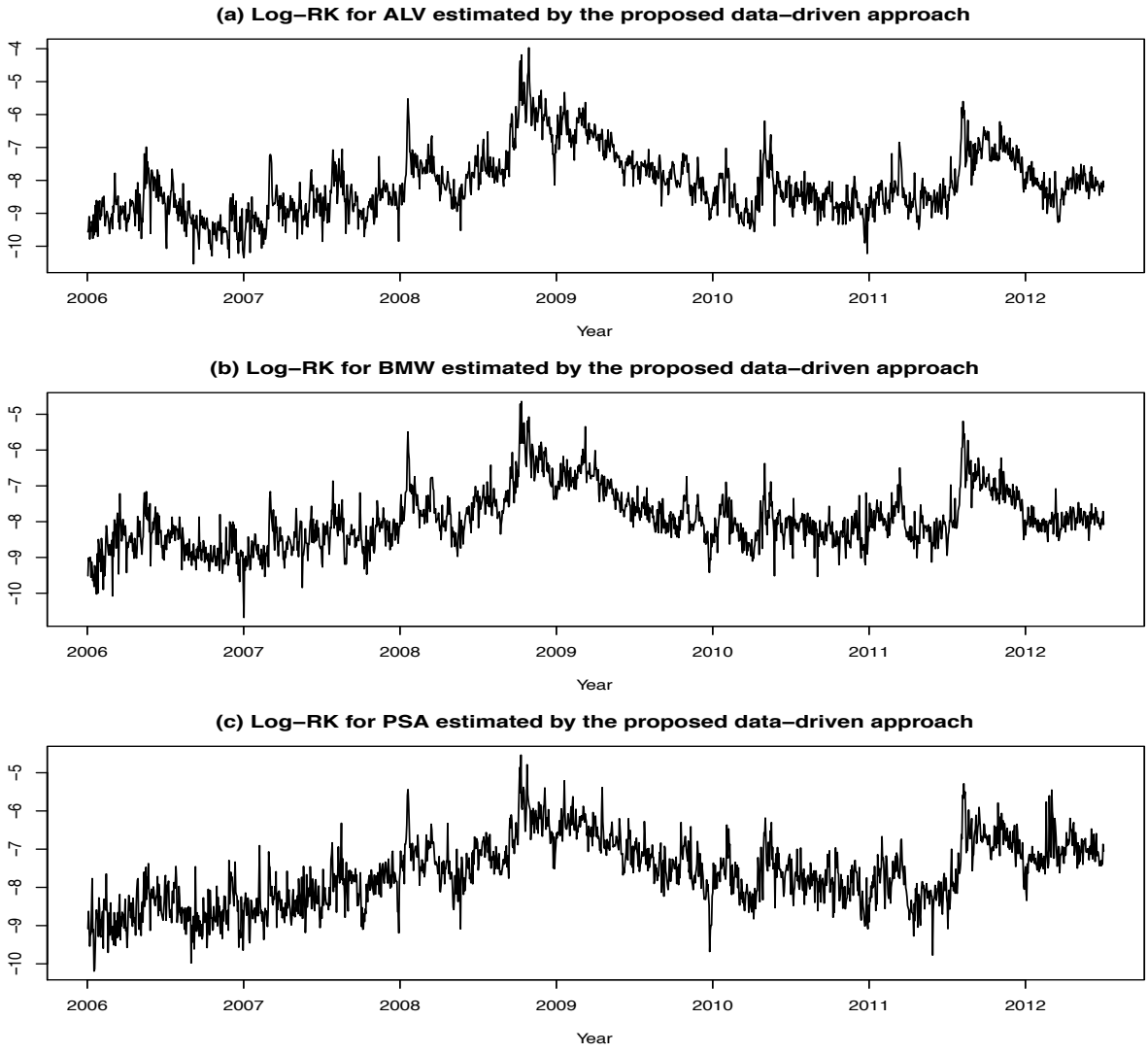

Figure 6: Logarithmic transformation of realized kernels for the other three companies

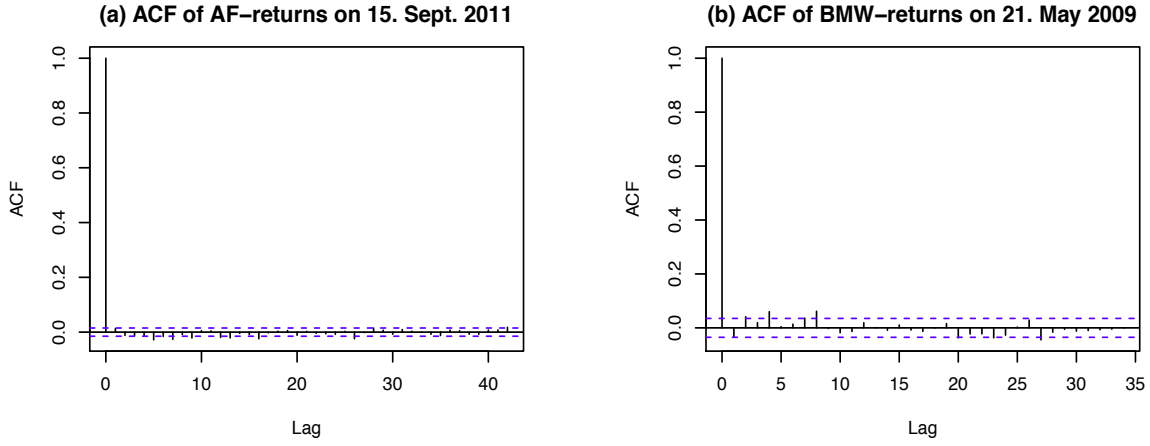

Figure 7: ACF of the high-frequency returns on the two challenging days

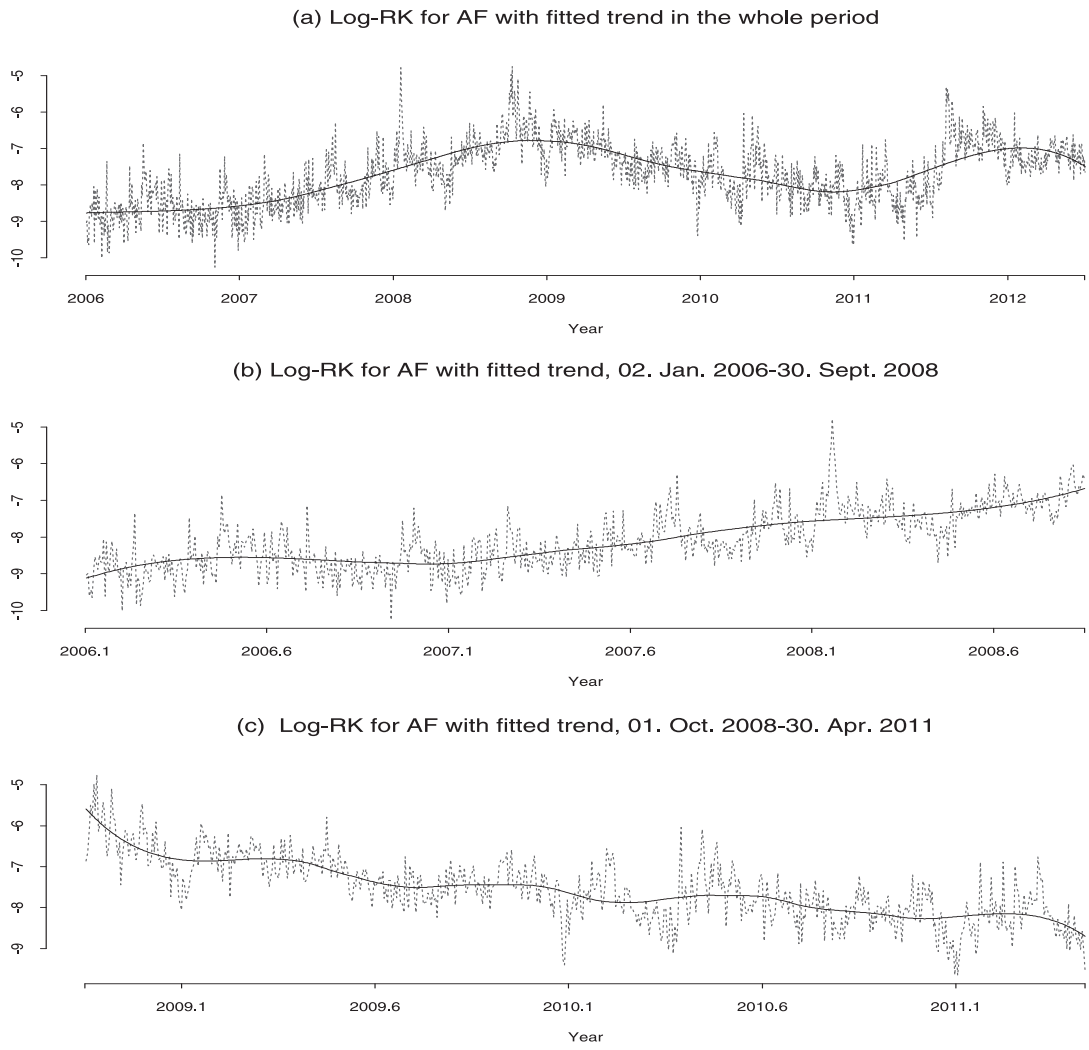

Figure 8: Estimated trend by ESEMIFAR together with the log-data

Recent discussion papers

2015-01	Yuanhua Feng Chen Zhou	An iterative plug-in algorithm for realized kernels
2014-13	Maximilian Müller Denis Schweizer Volker Seiler	Wealth Effects of Rare Earth Prices and China`s Rare Earth Elements Policy
2014-12	Sonja Brangewitz Behnud Djawadi René Fahr Claus-Jochen Haake	Quality Choices and Reputation Systems in Online Markets – An Experimental Study
2014-11	Daniel Kaimann Joe Cox	The Interaction of Signals: A Fuzzy set Analysis of the Video Game Industry
2014-10	Thomas Gries Ha Van Dung	Institutional environment, human capital, and firm growth: Evidence from Vietnam
2014-09	Ha Van Dung	Short-term precaution, insurance and saving mechanisms in rural Vietnam
2014-08	Thomas Gries Ha Van Dung	Household Savings and Productive Capital Formation in Rural Vietnam: Insurance vs. Social Network
2014-07	Achim Voß Mark Schopf	Lobbying over Exhaustible-Resource Extraction
2014-06	Eugen Dimant	The Nature of Corruption – An Interdisciplinary Perspective
2014-05	Natasa Bilkic Thomas Gries	Uncertainty and Conflict Decision
2014-04	Thomas Gries Natasa Bilkic	Investment under Threat of Disaster
2014-03	Natasa Bilkic Thomas Gries	Destructive Agents, Finance Firms and Systemic Risk
2014-02	Thomas Gries Margarete Redlin	Maritime Piracy: Socio-Economic, Political, and Institutional Determinants
2014-01	Ana Mauleon Nils Roehl Vincent Vannetelbosch	Constitutions and Social Networks
2013-16	Nils Roehl	Two-Stage Allocation Rules
2013-15	Zhichao Guo Yuanhua Feng Thomas Gries	Changes of China's agri-food exports to Germany caused by its accession to WTO and the 2008 financial crisis
2013-14	Eugen Dimant Tim Krieger Margarete Redlin	A Crook is a Crook … But is He Still a Crook Abroad? - On the Effect of Immigration on Destination-Country Corruption
2013-13	Republished as CIE Working Paper 2014-06	