Smoothing long memory time series using R

Yuanhua Feng, Jan Beran and Sebastian Letmathe
Faculty of Business Administration and Economics, Paderborn University

November 20, 2020

Abstract

This paper provides first a brief summary of the SEMIFAR (semiparametric fractional autoregressiove) and ESEMIFAR (exponential SEMIFAR) models. Those models are extended slightly to include the moving average part. Under common distribution condition it is shown that the long memory parameter is not affected by the log-transformation. A simple data-driven algorithm is proposed, by which the selected bandwidth and the selected orders of the ARMA model are all consistent. An R package is developed for practical implementation. The application of the proposals are illustrated by different kind of time series.

Keywords: Nonparametric regression with long memory, SEMIFAR, ESEMIFAR, bandwidth selection, model selection, implementation in R,

JEL Codes: C14, C51

1 Introduction

Literature research and model research required.

In many areas of research data are observed spatially, depending on two separate dimensions in a lattice. In recent years one can observe more frequently some sort of apparent memory in the decay of spatial correlations to depend and change over its direction within the spatial process. For instance, long-memory in the sense of slowly decaying autocorrelations in (high frequency) financial data across trading time and trading day produces a random field on a lattice in both dimensions simultaneously. Beran, Feng, and Ghosh (2015) state that daily average trade duration data has often shown long memory with a clear non zero mode. Therefore a log-normal conditional distribution is suggested. The simplest approach to model long range dependence in a positive valued time series is to take the exponential of a linear long memory process such as FARIMA leading to stochastic volatility models. Due to the long range dependence there is an unobservable latent process which makes the estimation and interpretation of the fitted parameters very challenging.

The SEMIFAR and ESEMIFAR models introduced by Beran and Feng (2002c) and Beran, Feng, and Ghosh (2015) are designed for simultaneous modeling of stochastic trends, deterministic trends and stationary short- and long-memory components in a time series such that the trend generating mechanisms can be distinguished.

2 The SEMIFARIMA model

2.1 The SEMIFAR

A process Y_t is said to follow a SEMIFAR model, introduced by Beran (1999) if there exists an integer $m \in \{0, 1\}$ and a fraction $\delta \in (-0.5, 0.5)$ such that

$$\phi(B)(1-B)^{\delta}\{(1-B)^{m}Y_t - g(x_t)\} = \epsilon_t, \tag{1}$$

where $\phi(x) = 1 - \sum_{j=1}^{p} \phi x^j$ is a polynomial with all roots outside the unit circle, ϵ_t are iid normal with $E(\epsilon_t) = 0$, $\mathrm{var}(\epsilon_t) = \sigma_\epsilon^2$, $x_t = t/n$ with $t \in \mathbb{Z}$, B is the backshift operator and $g : [0, 1]$ is a nonparametric smooth trend function. The fractional differencing parameter δ was introduced by Granger and Joyeux (1980) and Hosking (1981) and is defined by

$$(1-B)^{\delta} = \sum_{k=0}^{\infty} b_k(\delta) B^k, \tag{2}$$

with

$$b_k(\delta) = (-1)^k \binom{\delta}{k} = (-1)^k \frac{\Gamma(\delta+1)}{\Gamma(k+1)\Gamma(\delta-k+1)}. \tag{3}$$

Considering the autocovariances $\gamma(k) = \text{cov}(Y_t, Y_{t+k})$, Y_t incorporates long memory if the spectral density given by

$$f(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} e^{ik\lambda} \gamma(k) \tag{4}$$

exhibits a pole at the origin of the frequency spectrum such that

$$f(\lambda) \sim c_f |\lambda|^{-2\delta}, \quad (\text{as } \lambda \to 0), \tag{5}$$

where $c_f > 0$ and "$\sim$" stands for the ratio of both sides converging one. Then, for $k \to \infty$ the autocovariances $\gamma(k)$ are proportional to $k^{2\delta-1}$ and hence yield an infinite sum. We can distinguish between three temporal dependency structures. The process $Z_t = \{(1-B)^m Y_t - g(x_t)\}$ has long memory for $\delta > 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) = \infty$, short memory for $\delta = 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) < \infty$ and is anitpersistent for $\delta < 0$ with $\sum_{k=-\infty}^{\infty} \gamma_U(k) = 0$ frequently reversing itself. Based on model (1) Beran and Feng (2002c) proposed an adapted version of a data-driven IPI (iterative plug-in) algorithm already introduced in Beran (1995) by replacing an estimate of the constant mean with a kernel estimate of g defined by

$$\hat{g}(x) = \frac{1}{nh} \sum_{t=1}^{n} K(\frac{x - x_t}{h})(1-B)^{\hat{m}} Y_t, \tag{6}$$

where $h > 0$ denotes the bandwidth, $x \in [0,1]$ and $K(\cdot)$ is a symmetric second order kernel with compact support (see e.g. Gasser and Müller 1979). Moreover, explicit expressions for the bias, variance, MISE and the optimal bandwidth which minimises the asymptotic MISE are stated in Theorem 1 in Beran and Feng (2002c). A comprehensive application of the SEMIFAR to finanical time series data was carried out by Beran and Ocker (2001). The authors found strong evidence of long memory in power transformed absolute return series in form of a stochastic- or deterministic trend and in some cases with both forms. Subsequently, these results indicate that conventional parametric short- and long memory models may not be suitable for modelling volatility of stock market indices. In Beran and Feng (2002a) two new IPI-algorithms are proposed which run much faster as they do not rely on a full search of the long memory parameter. Furthermore, an EIM- (exponential inflation method) bandwidth selector is defined. Beran and Ocker

(2001) and Beran and Feng (2002c) already suggested to use an EIM as it requires less iterations than the conventional multiplicative inflation method (MIM) used by Gasser et al. (1991), Herrmann et al. (1992) and Ray and Tsay (1997). Different choices for the inflation factor and asymptotic properties of the estimated bandwidths are derived by Beran and Feng (2002a). To control for the poor estimation quality of the kernel estimator at the boundaries, the authors introduced a small positive constant $\Sigma > 0$ such that as $n \to \infty$, $h \to 0$, $nh \to \infty$,

$$\text{MISE} = E\bigg\{ \int_{\Sigma}^{1-\Sigma} [\hat{g}(x) - g(x)]^2 dx \bigg\}. \tag{7}$$

In a following paper Beran and Feng (2002b) replaced the kernel estimator (6) with a local approximation of $g(x)$ given by the p order polynomial

$$g(x_t) \approx g(x) + g^{(1)}(x)(x_t - x) + \ldots + g^{(p)}(x)\frac{(x_t - x)^p}{p!} + R_p, \tag{8}$$

where R_p is a remainder term. Then the estimator $\hat{g}^{(\nu)}(x)$ with $(\nu \leq p)$ is obtained when the locally weighted sum of squared residuals is minimized such that

$$Q(x) = \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{p} \beta_j (x_t - x)^j \right\}^2 K\left(\frac{x_t - x}{h}\right) \Rightarrow min, \tag{9}$$

where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, h and K are defined as before.

2.2 The E-SEMIFARIMA

2.3 The E-SEMIFARIMA

A theorem about the memory property of the ESEMIFAR (direct after that model)

Theorem 1 *Assume that X_t follows an ESEMIFAR model defined above with $d > 0$ and a linear FARIMA error process ξ_t in the log-data, then $Z_t = X_t/v(\tau_t)$ is a stationary long memory process with the same memory parameter d.*

This result shows that the level of long memory of the stationary part of the process under consideration is not affected by the log- or exponential transformations. In the case when ξ_t is normal, those results are e.g. obtained in Dittmann and Granger (2002) and Beran et al. (2015). Theorem 2 extends those results to a common innovation distribution in the log-data. The proof of Theorem 2 is given in the appendix, where we will show that the process Z_t is a special case of the general framework defined in Surgailis and Viano (2002). Hence, their results apply to Z_t. Note however that the above result does not hold for $d = 0$. Detailed discussion on corresponding properties in case with $d = 0$ is beyond the aim of the current paper.

3 Data-driven estimation

3.1 The IPI-algorithm for estimating g

3.2 Data driven estimation of g' and g''

4 Implementation in R

The implementation of the SEMIFAR model in `R` is called via the `semifar.lpf` function. The script covers the kernel estimation of a nonparametric trend allowing for different error structures and maximum likelihood estimation of the parametric part of the model. The `R` function is designed for a fast and uncomplicated application of SEMIFAR and assumes some not necessarily fixed values a priori to be constant. For instance, the level of confidence used for η is by default set to 0.05 and is among other constants not included as input. Nevertheless the script can be made more flexible very easily by using more parameters in the function call. Besides the series there are currently seven variables to be set before the model is fitted. The minimal autoregressive order p, the maximal order p, the range of the MSE of the trend estimate, the degree of the polynomial approximating the nonparametric trend and its derivatives, the used kernel, the inflation factor inflating the previously estimated bandwidth and whether boundary protection is included or not.

Fitting a SEMIFAR process with the IPI algorithm required the purchase of the TIBCO

Spotfire software `SPLUS` and the additional `FinMetrics` package based on the `S` language. To this date, the `semifar.lpf` function has successfully been implemented into the language R. Major differences in computation are found in

1. the derivation of the ML parameter estimates for a fractionally-differenced ARIMA model, especially for the fractional filter δ
2. how the residuals of fractionally filtered data for an ARIMA model are obtained
3. the calculation of the spectral density and its update
4. a newly introduced inflation factor for bandwidth convergence

Concerning the first point, the difficulty in calculating δ comes with introducing antipersistent trend behavior allowing $\delta < 0$ and extending the interval of possible values for the `drange` to $(-0.5, 0.5)$. The main part of producing δ is given in (2) and implements the gamma function and the backshift operator in an infinte binomial sum over k subtracted from one such that $1 - \delta B - \frac{1}{2}\delta(1-\delta)B^2 - \frac{1}{6}\delta(1-\delta)(2-\delta)^2 B^3 - \ldots$ which gives $1 - \sum_{k=1}^{\infty} c_k(\delta) B^k$. For negative δ the signs reverse and the expression is < 0. However, the values in the interval `drange` in `fracdiff`{`fracdiff`} (over which the likelihood function is maximized as a function of δ) cannot exceed 0.5 since the binomial approaches 0 for $\delta = |\{-0.5, 0.5\}|$ and the series would be non stationary. Still, having δ set to $(-0.5, 0.5)$ is crucial for the identification of the trend mechanism of the SEMIFAR process, as the spectral density $f_X(\lambda) \sim \{0, c, \infty\}$ for δ in $\{(-0.5, 0), 0, (0, 0.5)\}$ in order to obtain the autocovariance behavior at the 0th frequency $f_X(0) = \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} \gamma_X(k) = \{0, c, \infty\}$ expressing the memory structure of the process. Therefore a decision criteria for setting `drange` to $(0, 0.5)$ or to $(-0.5, 0)$ is introduced. If either of the specified intervals for `drange` does not allow the computation of a correlation matrix of the parameter estimates, the alternative `drange` is chosen. In this way, the underlying temporal dependence structure of the data has to correlate positively with the fractional filter. Otherwise there must not be a correlation matrix.

```r
> m1=fracdiff(data,drange = c(0,0.5))
> m2=fracdiff(data,drange = c(-0.5,0))
> if(isTRUE(all.equal(m1$correlation.dpq,NULL))){result<-m2}
> if(isTRUE(all.equal(m2$correlation.dpq,NULL))){result<-m1}
```

By implementing two different persistence situations, `fracdiff` objects and a decision criterion, the fractional parameter is successfully estimated and equivalent to `arima.fracdiff` output in `S`.

The second change accounts for a different filtering technique used to obtain the residuals and variance of the fractionally filtered ARIMA process. The residuals account for the difference between an estimated and observed value. While in `S` a dynamic linear Kalman filter is employed to calculate the residuals, a much simpler autoregressive moving average filter is used in `R`. The Kalman filter predicts the future state of the series based on the current state vector times the change in time and gives the expected subsequent state. The previous predicted state is corrected using a current sensory measurement as innovation of the state estimate. However, the innovation is weighted relative to the prediction each time step. Its weight decreases if the current measurement matches the last predicted state and increases otherwise. Hence, the ratio of uncertainty in measurement and prediction set by the estimated ARMA parameters, filters for short memory and produces the residuals. An approach more straightforward is the successive application of standard recursive and convolution filtering methods for four different model order- and parameter estimates as filter input for the new `arma.filt` function. In this way all possible ARMA(p,q) contributions to short memory are addressed. Convolution filtering uses the product of different functions as weighted average and thus creates an individually weighted mean for each observation (therefore moving average) in line with

$$I(x,y) = \sum_{s=-i}^{i} \sum_{t=-j}^{j} w(s,t)(x-s, y-t). \tag{10}$$

Hence there are at least $max\left\{p,q\right\}$ `NA`s for each filter value. The recursive filtering method describes the re-use of its output as input which is also done within the Kalman filter. The filtered values will exceed their input values for positive w_j and will be lower for $w_j < 0$. The recursive filter method is given by

$$y_t = \sum_{j=0}^{J} y_{t-j} w_j, \tag{11}$$

where J is the set of filter criteria. Combining both filter mechanics ensures the consistent elimination of the short memory ARMA component moving the input series. Due to the recursive nature of the AR filter, the MA filter obviously has to be applied before the recursive filter. It is important to use the (fractionally) pre-filtered data since otherwise the filter will rule out movement which does not reflect short memory exclusively.

The third difference comes with the calculation of the spectral density update. `S` and `R` produce identical output, apart from the density spectrum which is no longer estimated in decibels such that the transformation `f=10**f$spec/10` is redundant and only `f=f$spec` remains. Moreover the frequency of the input `xfreq` can be fixed to one since we are considering an univariate time series.

The fourth change also concerns the update cycle by allowing different factors α used for inflating the previous bandwidth estimate in (**??**). The inflation method significantly changes the rate of convergence (Beran and Feng, 2002a). The optimal rate of convergence h_M expresses a trade-off between a bias and variance ratio producing an overall fit of the trend which is sufficiently stable while being sufficiently smooth. The `semifar.lpf` function offers the choice for three different inflation factors called `var`, `naive` and `opt`. If α is chosen as $\alpha_{opt} = (5-2\delta)/(7-2\delta)$ and δ is for simplicity assumed to be zero, than we have respectively

$$\begin{aligned}\hat{h} &= h_M \left\{1 + O(n^{2(2\delta-1)/(7-2\delta)}) + O_p(n^{2(2\delta-1)/(7-2\delta)}) + O_p(n^{(-1/2)})\right\} \\ &= h_M \left\{1 + O(n^{-2/7}) + O_p(n^{-2/7}) + O_p(n^{-1/2})\right\}\end{aligned} \tag{12}$$

where the first term gives the order of the bias under this particular inflation factor and the second term corresponds to the order of the variance. The third term is asymptotically negligible and for all choices of α the same. The overall rate of convergence under this inflation factor is $O(n^{-2/5})$. The lower bound of the variance is given if $O_p(n^{-1/2})$ since δ is in $(-0.5, 0.5)$ and the rearranged exponent contains $\delta - 1/2$ which is zero if the maximal value of δ is inserted. If the model includes persistent behavior and δ is non zero the rate of convergence is slower for $\delta > 0$ and respectively for $\delta < 0$ the rate of convergence is faster. Depending on α, either the variance term or the bias is of a smaller order and changes smoothing or stabilizes the trend function more. In general the best overall rate of convergence is achieved by applying an inflation factor α that minimizes the bias and variance of the SEMIFAR model.

5 Application to different kinds of time series

5.1 Application of the SEMIFARIMA

5.2 Application of the ESEMIFARIMA

5.3 Application to high-frequency financial data

6 The Semi-FI-Log-GARCH model

7 The Semi-FI-Log-ACD model* (nachfragen)

8 Concluding remarks

References

Beran, Jan and Yuanhua Feng (2002a). “Iterative plug-in algorithms for SEMIFAR models—definition, convergence, and asymptotic properties”. In: *Journal of Computational and Graphical Statistics* 11.3, pp. 690–713.

– (2002b). “Local polynomial fitting with long-memory, short-memory and antipersistent errors”. In: *Annals of the Institute of Statistical Mathematics* 54.2, pp. 291–311.

– (2002c). “SEMIFAR models—a semiparametric approach to modelling trends, long-range dependence and nonstationarity”. In: *Computational Statistics & Data Analysis* 40.2, pp. 393–419.

Beran, Jan, Yuanhua Feng, and Sucharita Ghosh (2015). “Modelling long-range dependence and trends in duration series: an approach based on EFARIMA and ESEMIFAR models”. In: *Statistical Papers* 56.2, pp. 431–451.

Beran, Jan and Dirk Ocker (2001). “Volatility of stock-market indexes—an analysis based on SEMIFAR models”. In: *Journal of Business & Economic Statistics* 19.1, pp. 103–116.

Granger, Clive WJ and Roselyne Joyeux (1980). “An introduction to long-memory time series models and fractional differencing”. In: *Journal of time series analysis* 1.1, pp. 15–29.

Hosking, Jonathan RM (1981). “Fractional differencing”. In: *Biometrika* 68.1, pp. 165–176.

Appendix