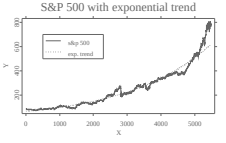
Semiparametric Diffusion Estimation and Application to a Stock Market Index

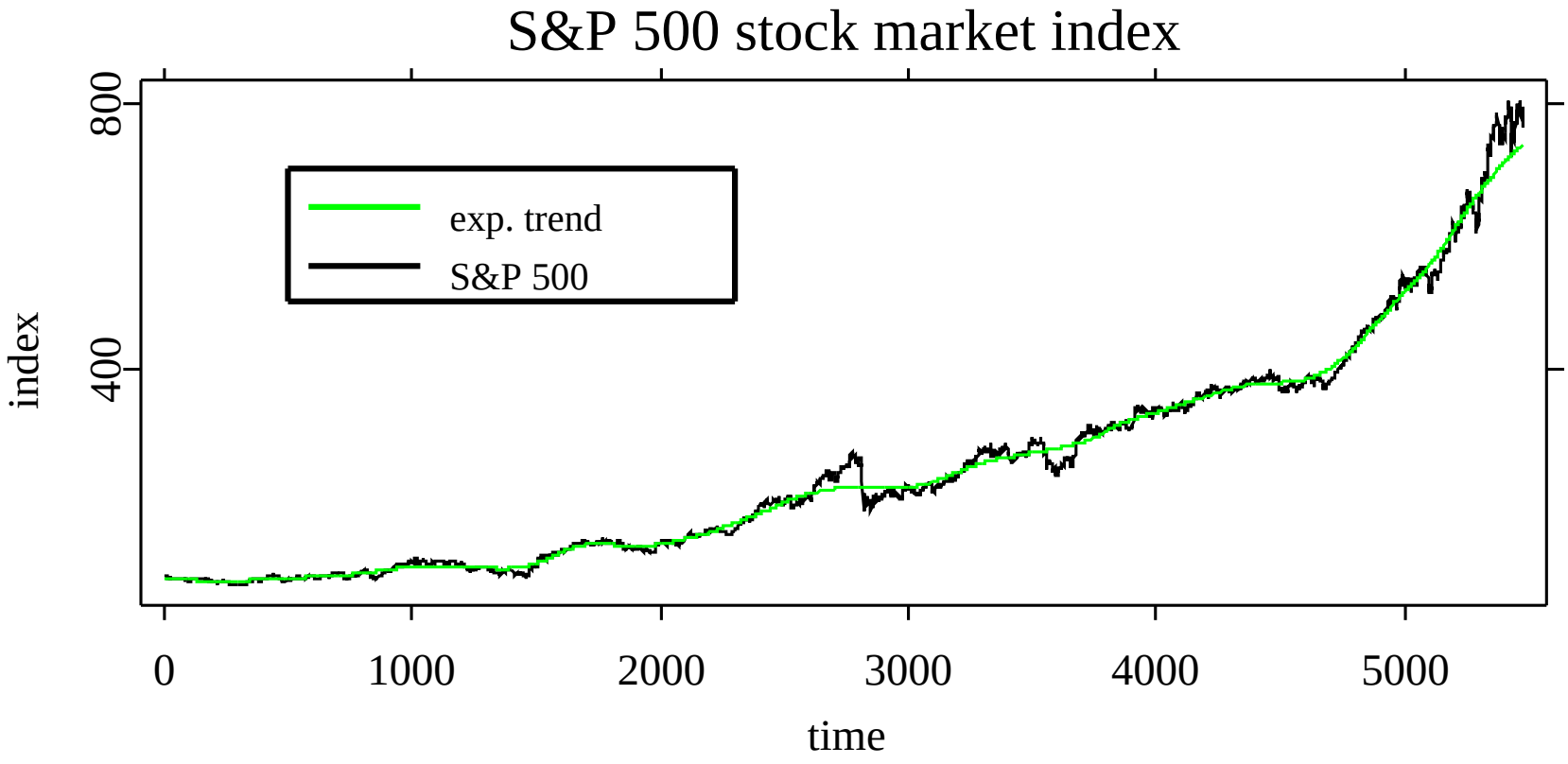
W. Härdle, T. Kleinow
Humboldt-Universität zu Berlin, Germany

A. Korestelev
Wayne State University, Detroit, USA

C. Logeay
Deutsches Institut für Wirtschaftsforschung, Berlin, Germany

E. Platen
University of Technology Sydney, Australia




S&P 500 stock market index
index
800
400
exp. trend
S&P 500
0
1000
2000
3000
4000
5000
time

Model

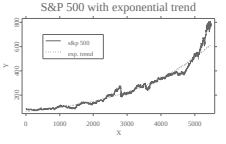
$$S(t) = S(0)Z(t)\exp\left(\int_0^t \eta(s)ds\right) \tag{1}$$

where $Z(t)$ is a stationary random process (diffusion)

$\eta(t)$ is a deterministic growth rate

Testing a parametric model for $Z(t)$ vs. a nonparametric alternative

Estimate the trend by kernel smoothing



Normalized Index

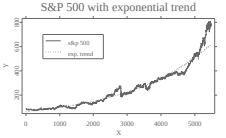
$$X(t) = \frac{S(t)}{\bar{S}(t)} \tag{2}$$

Here $\bar{S}(t)$ is an average index:

$$\bar{S}(t) = \exp\{(K_h * \ln S)(t)\} \tag{3}$$

where $(K_h * \ln S)(t)$ is a kernel smoother of $\ln S$ with a kernel K and a bandwidth h.

$$(K_h * \mu)(t) = \frac{\sum_{i=1}^{n} K_h(t - t_i)\mu(t_i)}{\sum_{i=1}^{n} K_h(t - t_i)} \tag{4}$$



Parametric models for $Z(t)$

square root process

$$dZ(t) = \beta\{1 - Z(t)\}dt + \gamma\sqrt{Z(t)}dW(t). \tag{5}$$

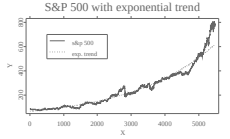
geometric Ornstein-Uhlenbeck process

$$\begin{aligned} Z(t) &= \exp\{U(t)\} \\ dU(t) &= -\beta U(t)dt + \gamma dW(t). \end{aligned} \tag{6}$$

Nonparametric alternative:

$$dZ(t) = m\{Z(t)\}dt + \sigma\{Z(t)\}dW(t). \tag{7}$$

Chen, Härdle, Kleinow (2001)

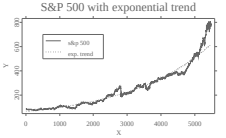


Normalized Index

$$
\begin{aligned}
X(t) &= \frac{S(t)}{\bar{S}(t)} = \frac{S(0)Z(t)\exp\left\{\int_0^t \eta(s)ds\right\}}{\exp\{(K_h * \ln S)(t)\}} \\
&= \frac{S(0)\exp\left\{\int_0^t \eta(s)ds\right\} Z(t)}{S(0)\exp\left\{\left(K_h * \int_0^{\cdot} \eta(s)ds\right)(t)\right\}\exp\{(K_h * \ln Z)(t)\}} \\
&= \exp\{L(t) - (K_h * L)(t)\} \qquad (8)
\end{aligned}
$$

where $L(t) = \ln Z(t)$ and

$$
\int_0^t \eta(s)\,ds - \left(K_h * \int \eta(s)\,ds\right)(t) \approx 0.
$$

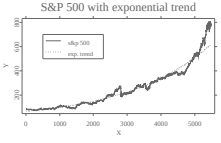


The logarithm of the normalized index $X(t)$ is a stationary process, but it is not the diffusion $\ln Z(t)$ itself.

$$\ln X(t) = L(t) - (K_h * L)(t) \tag{9}$$

Goal: Estimate the parameters β and γ of $Z(t)$ from the observations $X(t)$.

Puzzle: How to choose the bandwidth h for the nonparametric estimation of the growth rate $\eta(t)$?

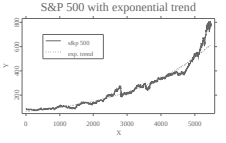


Stylized facts from the empirical data (S&P 500)

The diffusion coefficient γ is small for varying h, i.e. $\gamma \approx 0.01 \ll 1$.

h	200	250	300	350	400
$Var(X)$	0.0018303	0.0023465	0.0029246	0.0035622	0.0042183
$\hat{\gamma}$	0.0090593	0.0090703	0.0090849	0.0090991	0.0091103

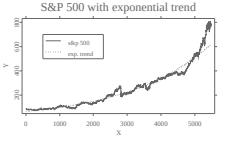
Table 1: Estimated values for γ and the estimated variance of X for different bandwidths h.



We concentrate on a geometric Ornstein-Uhlenbeck process, $\ln Z = U$.
The observed process $\ln X$ is then

$$\ln X = L - (K_h * L) \approx U - (K_h * U) \tag{10}$$

Goal: Choose h and estimate β and γ.



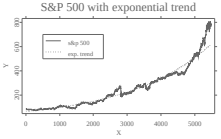
Proposition 1

$$Var[L - K_h * L] = \frac{\gamma^2}{2\beta}\left(1 - \frac{c}{\beta h} + O(h^{-2})\right) \tag{11}$$

and the autocorrelation of $\ln X = L - K_h * L$ is

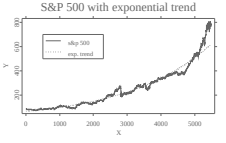
$$\begin{aligned}\rho_h^{(\ln X)}(\tau) &= Corr[(L - K_h * L)(\tau); (L - K_h * L)(0)] \\ &= \left(e^{-\beta\tau} - \frac{c_K}{\beta h} + O(\tau/h^2) \right) / \left(1 - \frac{c_K}{\beta h} + O(h^{-2}) \right)\end{aligned} \tag{12}$$

where c_K is a constant depending only on the Kernel K.



Algorithm to choose h

1. Estimate γ from the quadratic variation of $L - K_h * L$ (not sensitive to h)
2. Estimate β twice, once from $Var[L - K_h * L]$ and once from autocorrelation $\rho_h^{(\ln X)}(\tau)$.
3. Choose h that makes estimates equal.

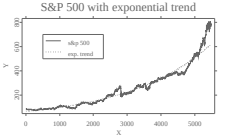


Estimators

$$
\begin{aligned}
\hat{\gamma}^2 &= T^{-1} \sum_{i=1}^{n} \big(\ln X(i\Delta t) - \ln X(i\Delta t - \Delta t) \big)^2 \\
&\approx T^{-1} \int_0^T d < \ln X > \approx \ T^{-1} \int_0^T d < L > \\
\hat{\beta}_1(h) &= \frac{\hat{\gamma}^2}{2Var[\ln X]} - \frac{c_K}{h} \\
\hat{\beta}_2(h) &= \Big| \frac{\partial^+}{\partial \tau} \rho_h^{(\ln X)}(\tau) \Big| \Big|_{\tau=0} - \frac{c_K}{h},
\end{aligned}
$$

The balance equation to choose h is

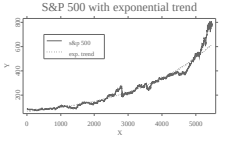
$$
\frac{\beta_1}{\beta_2} = 1 \tag{13}
$$



Empirical results

h	200	225	250	275	300
$\hat{\beta}_1(h)$	0.017913	0.01569	0.013826	0.012197	0.010776
$\hat{\beta}_2(h)$	0.01328	0.012098	0.011092	0.0098626	0.0086047
$\hat{\beta}_1(h)/\hat{\beta}_2(h)$	1.3489	1.2969	1.2465	1.2367	1.2523

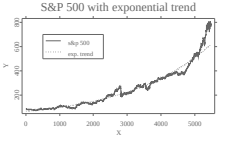
Table 2: Estimated values for β for different fixed bandwidths h.



Choice of the flexible bandwidth $h(t)$

- create M overlapping subintervals of length L with middle points m_i, $i = 1, \ldots, M$. We have chosen $m_{i+1} = m_i + 200$.
- choose a bandwidth for every subinterval and take this bandwidth as the optimal one for the middle point of the subinterval, $h_{opt}(m_i)$
- smooth the resulting function $h_{opt}(m_i)$
- interpolate $h_{opt}(m_i)$ to get $h_{opt}(t)$

This algorithm is applied to different lengths L (in our case: $2000, 2500, 3000, 3500, 4000$). We then used the "most stable" one.



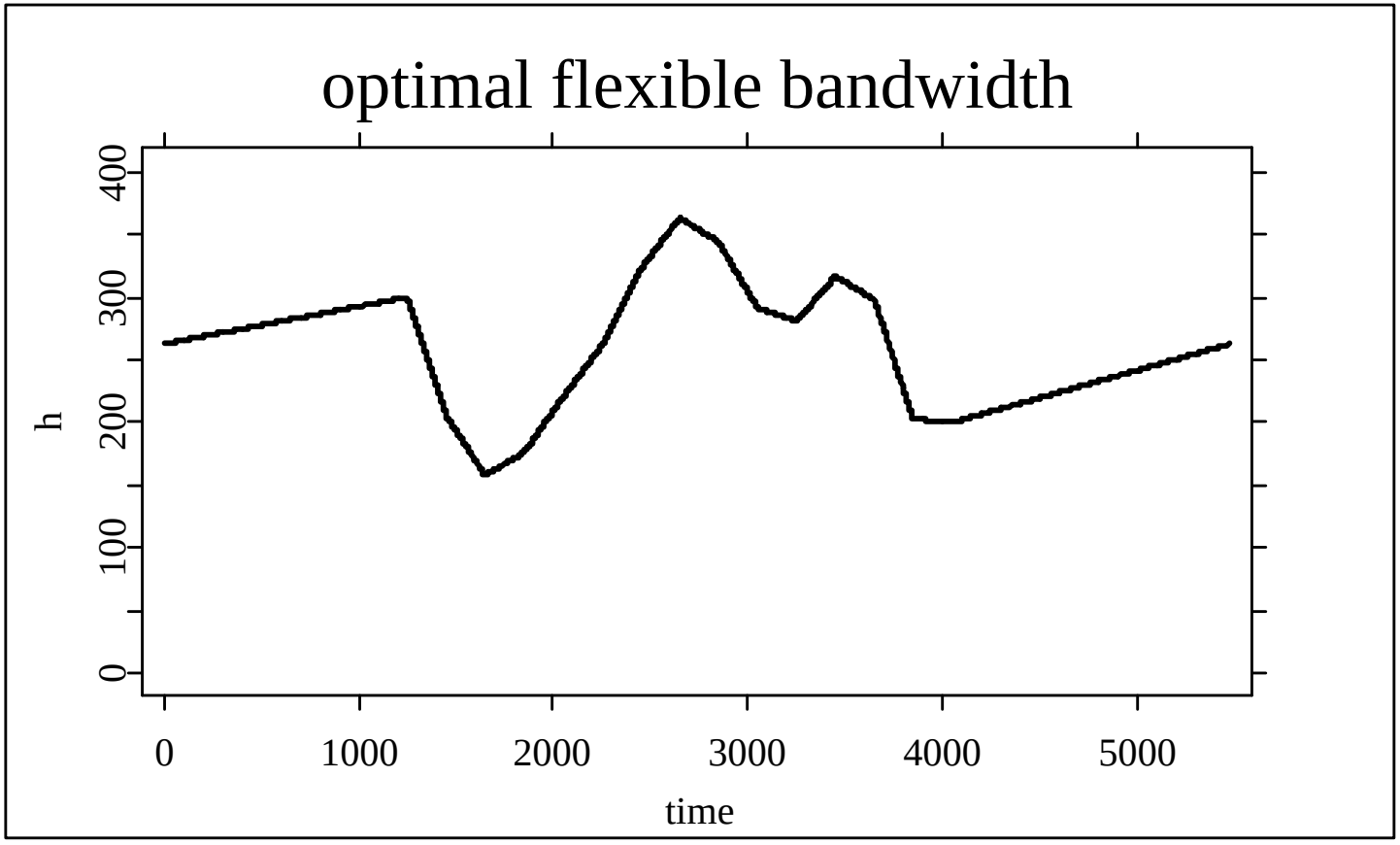
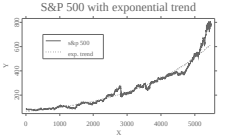
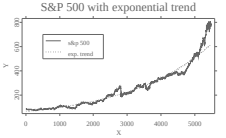


Figure 1: The optimal flexible bandwidth $h_{opt}(t)$.



Empirical results for the optimal flexible bandwidth

$$
\begin{aligned}
\hat{\beta}_1(h_{opt}) &= 0.010352 \\
\hat{\beta}_2(h_{opt}) &= 0.0089721 \\
\frac{\hat{\beta}_1(h_{opt})}{\hat{\beta}_2(h_{opt})} &= 1.1538 \\
\hat{\gamma}(h_{opt}) &= 0.0091033
\end{aligned}
$$



Restoration of Z

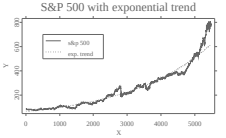
$$\ln X = L - K_h * L = \ln Z - K_h * \ln Z$$

$$\begin{aligned} \ln Z &= \ln X + K_h * \ln Z \\ \ln Z &= \ln X + K_h * \big(\ln X + K_h * \ln Z \big) \\ \ln Z &= \ln X + K_h * \ln X + K_h * \big(K_h * \ln X \big) \ldots \end{aligned} \tag{14}$$

By this approximation we have observations of Z

$$dZ(t) = m\{Z(t)\}dt + \sigma\{Z(t)\}dW(t). \tag{15}$$

Goal: estimate nonparametrically m and σ and test parametric vs. nonparametric



Parametric estimates from X vs. the estimates from the restored process Z

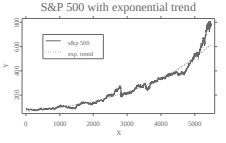
$$\hat{\beta}_1(h_{opt}) = 0.010352 \qquad \hat{\beta}_1 = 0.01003$$

$$\hat{\beta}_2(h_{opt}) = 0.0089721 \qquad \hat{\beta}_2 = 0.0093294$$

$$\frac{\hat{\beta}_1(h_{opt})}{\hat{\beta}_2(h_{opt})} = 1.1538 \qquad \frac{\hat{\beta}_1}{\hat{\beta}_2} = 1.0751$$

$$\hat{\gamma}(h_{opt}) = 0.0091033 \qquad \hat{\gamma} = 0.0092454$$

The estimated values from X and Z are in the same range and the ratio $\hat{\beta}_1/\hat{\beta}_2$ is close to one in both cases.



Nonparametric estimation

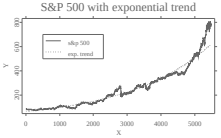
discrete observations

$$\begin{aligned} Y_i &= Z^{\Delta}(t_{i+1}) - Z^{\Delta}(t_i) \\ &= \Delta\, m\{Z^{\Delta}(t_i)\} + \sqrt{\Delta}\, \sigma\{Z^{\Delta}(t_i)\}\, \varepsilon_i \end{aligned} \tag{16}$$

with independent standard Gaussian random variables

$$\varepsilon_i = \frac{W(t_{i+1}) - W(t_i)}{\sqrt{\Delta}} \overset{iid}{\sim} \mathcal{N}(0,1).$$

Estimate m and σ by Kernel smoothing. Confidence bands are constructed via the wild bootstrap.



Testing

We apply Itô's formula to $Z(t) = \exp\{U(t)\}$

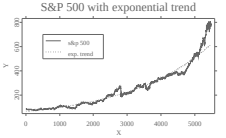
$$dZ(t) = Z(t)\left\{-\beta \ln Z(t) + \frac{1}{2}\gamma^2\right\}dt + \gamma Z(t) dW(t) \qquad (17)$$

Test

$$H_0(m) : m(z) = z\left\{-\beta \ln z + \frac{1}{2}\gamma^2\right\}$$

$$H_0(\sigma^2) : \sigma^2(z) = \gamma^2 z^2,$$

The alternatives are purely nonparametric functions $\hat{m}_{h_1}$ and $\hat{\sigma}^2_{h_1}$ where h_1 is the optimal bandwidth chosen with respect to the Silvermans rule of thumb. The Gaussian density function is used as the kernel. The conditional distribution of $\hat{m}_{h_1}$ and $\hat{\sigma}^2_{h_1}$ under the null hypothesis is constructed by bootstraping.



The bootstrap procedure

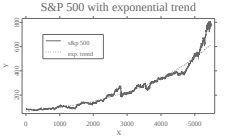
$$Y_i = \Delta m\{Z^{\Delta}(t_i)\} + \sqrt{\Delta}\sigma\{Z^{\Delta}(t_i)\}\,\varepsilon_i \tag{18}$$

- estimate $\hat{m}_{h_1}$ and $\hat{\sigma}^2_{h_1}$
- choose a larger bandwidth $g > h_1$ $(g = 2h_1)$
- estimate again $\hat{m}_g$ and $\hat{\sigma}^2_g(.)$ and calculate the residuals

$$\hat{\varepsilon}_i = \frac{Y_i - \Delta\hat{m}_g\{Z^{\Delta}(t_i)\}}{\sqrt{\Delta}\hat{\sigma}_g\{Z^{\Delta}(t_i)\}}$$

- replicate N times $\hat{\varepsilon}$ using wild bootstrap to obtain $(\varepsilon_i^{*,n})$ and simulate N new series $Z_i^{*,n}$ with $Z_1^{*,n} = Z^{\Delta}(t_i)$

$$Z_{i+1}^{*,n} - Z_i^{*,n} = \Delta\hat{m}_g(Z_i^{*,n}) + \sqrt{\Delta}\hat{\sigma}_g(Z_i^{*,n})\varepsilon_i^{*,n}$$



The bootstrap procedure (ctd.)

- reestimate $\hat{m}_{h_1}^{*,n}$ and $(\hat{\sigma}^2)_{h_1}^{*,n}$ for each $Z_i^{*,n}$
- build the statistics

$$T_m^* = \sup_z \frac{|\hat{m}_{h_1}^{*,n}(z) - \hat{m}_{h_1}(z)|}{\hat{\sigma}_{h_1}^{*,n}(z)}$$

and

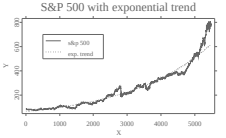
$$T_\sigma^* = \sup_z |(\hat{\sigma}^2)_{h_1}^{*,n}(z) - \hat{\sigma}_{h_1}^2(z)|$$

.

- construct confidence bands

$$KB(m(.)) = \left[\hat{m}_{h_1}(z) - \hat{\sigma}_{h_1}(z) t_{m,1-\alpha/2}\ ,\ \hat{m}_{h_1}(z) + \hat{\sigma}_{h_1}(z) t_{m,\alpha/2}\right]$$

$$KB(\sigma^2(.)) = \left[\hat{\sigma}_{h_1}^2(z) - t_{\sigma,1-\alpha/2}\ ,\ \hat{\sigma}_{h_1}^2(z) + t_{\sigma,\alpha/2}\right]$$



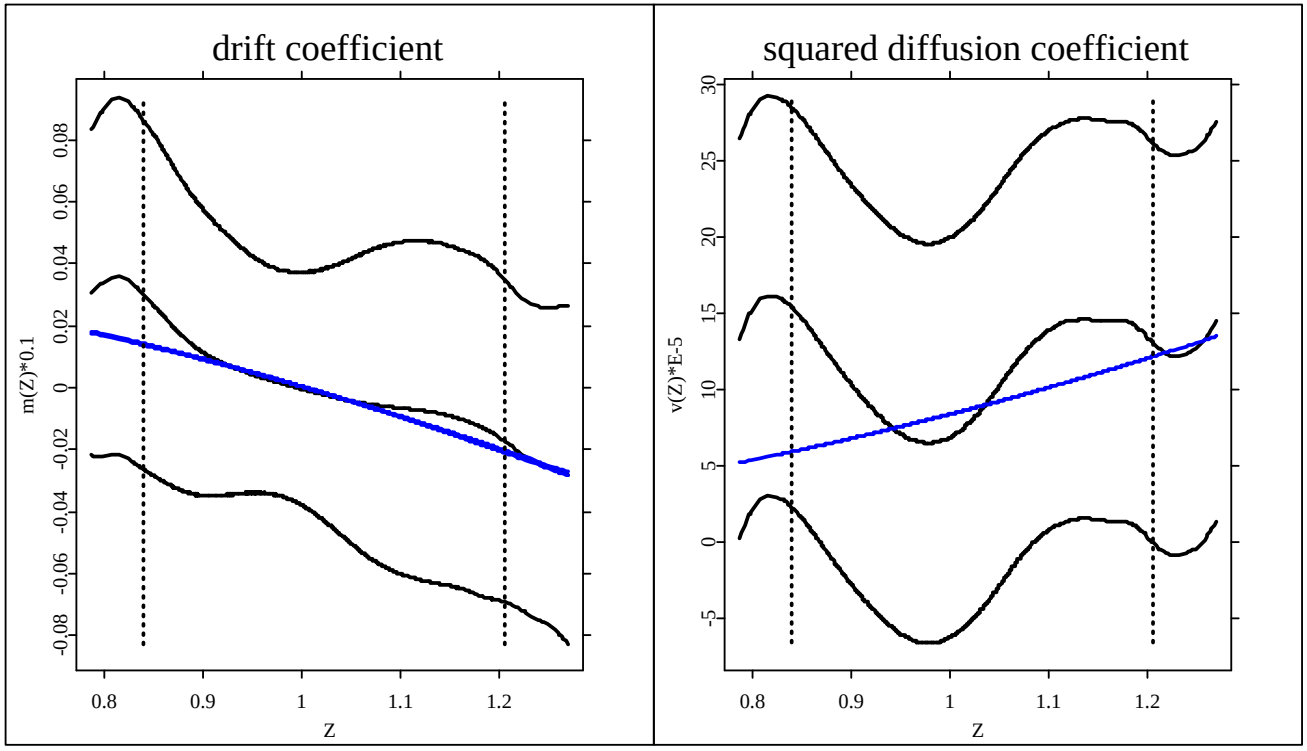
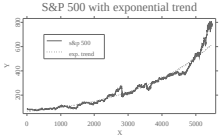


Figure 2: Nonparametric and parametric estimates of the drift $m(z)$ and squared diffusion coefficient $\sigma^2(z)$ with the 90% confidence bands. The vertical lines enclose the interval with 99% of the observerd data.



Conclusion

- An index is modeled as the product of an ergodic diffusion and a deterministic growth process
- The proposed methodology allows us to separate the estimation of the average growth of the index and that of the parameters of the ergodic diffusion.
- The empirical results show, that the null hypothesis of a geometric Ornstein-Uhlenbeck process is not rejected for the normalized S&P 500 data.