Discovery of Stock Chart Patterns by Kernel Smoothing and Automatic Outlier Detection

Hoang Tran

Department of Statistics, Florida State University

Outline

Technical Analysis

- The study of price trends for profitable trading
- **Charting**: identify *geometric patterns* for trading signals
- "Voodoo finance"

Technical Analysis

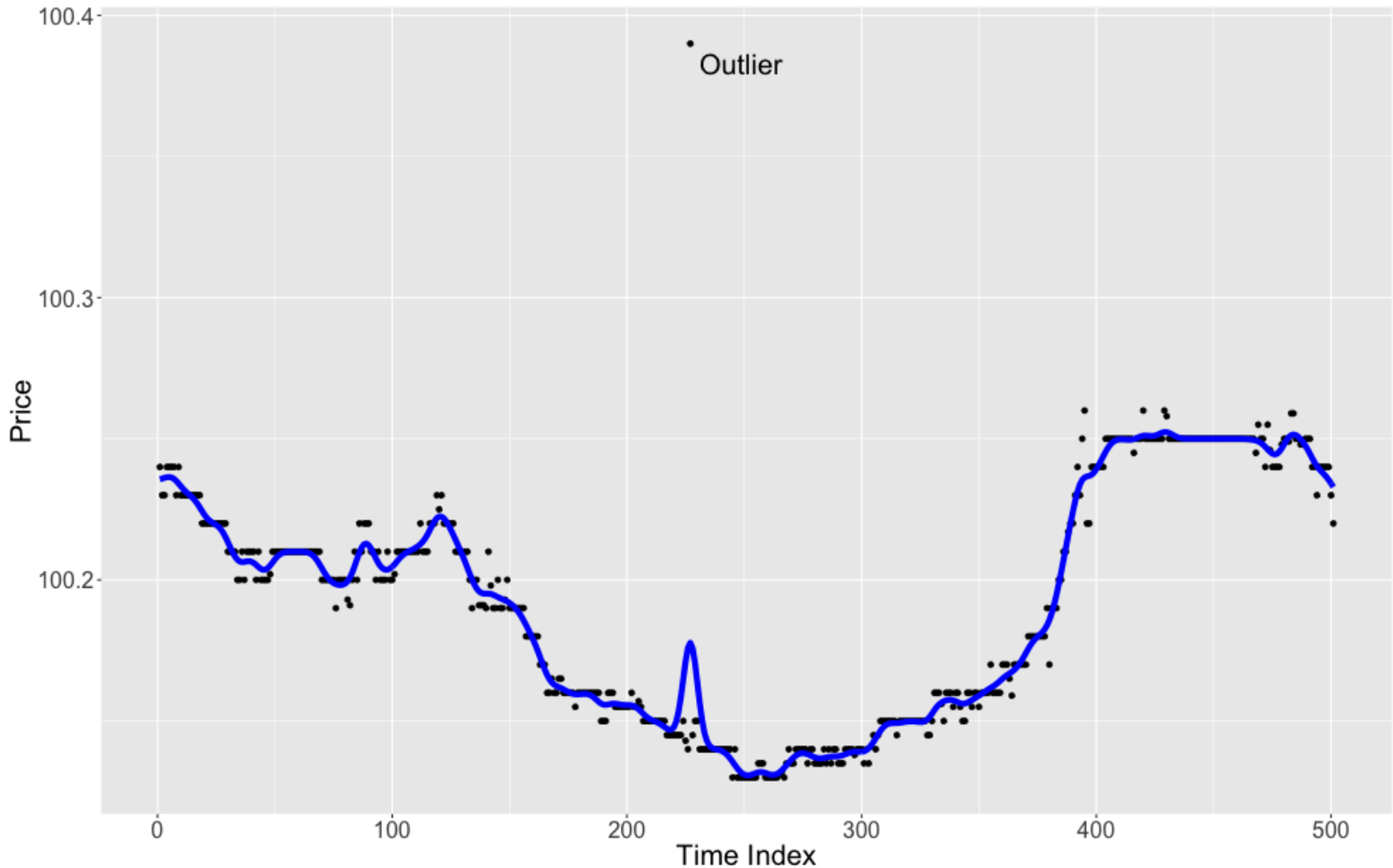
Figure: Cup and handle chart pattern [StockCharts.com, 2016]

Technical Analysis

- Difficult and/or time consuming to manually identify patterns
- A previous approach: use kernel smoothing to detect patterns [Lo et al., 2000]
- Outliers can be problematic

Chart Patterns and Outliers

Figure: McDonald's trading prices on Jan. 3rd, 2012 (10:18:18 to 10:20:00 am)

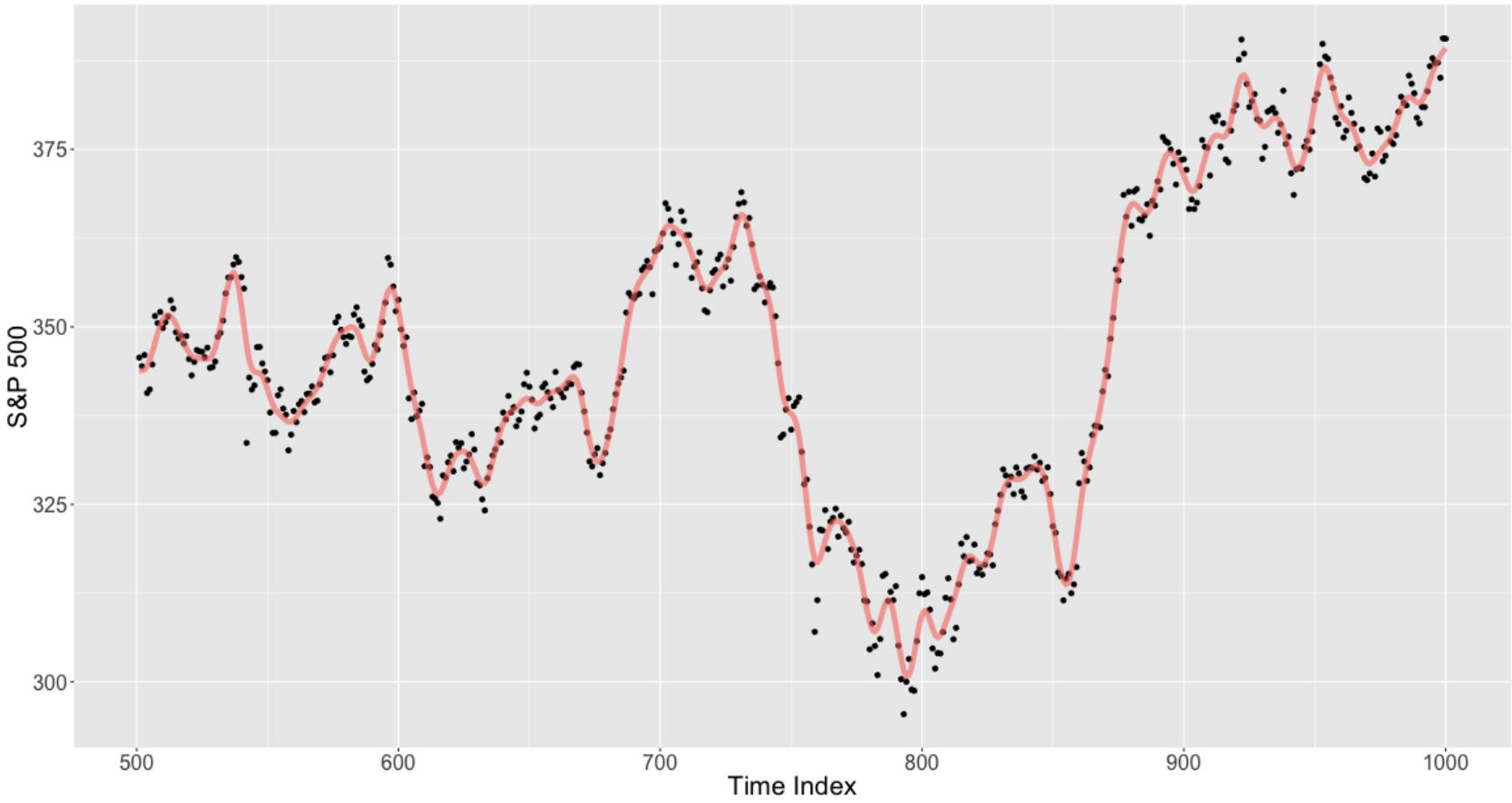


NOIS

- **N**on-Parametric **O**utlier **I**dentification and **S**moothing (NOIS)
- Pointwise confidence bands - **NOIS bundles**
- Applicable to both high frequency intra-day and inter-day data

Non-Linearity of Stock Price Data

Figure: S&P 500 closing prices with kernel smoothing



Non-Linearity of Stock Price Data

- Parametric procedures are unsuitable
- Non-parametric methods
  - Basis approaches (splines) - more **global**
  - Kernel smoothing - more **local**

Nadaraya-Watson estimator for kernel smoothing:

$$\hat{f}(x) = \frac{\sum_{t=1}^{n} K_h(x - x_t) y_t}{\sum_{t=1}^{n} K_h(x - x_t)}$$

Kernel Smoothing Bias Correction

Bias in areas of high curvature and at the boundaries

$$E[\hat{f}(x)] = f(x) + \frac{h^2}{2} f''(x) + o(h^2)$$

- Modify the estimating equation to reduce the effect of the bias [Lian, 2012]
- Bias correction after outlier detection and estimation

Bandwidth Selection

- Too *small* - curve will be very "jagged"
- Too *large* - curve will be too smooth
- Need a data-dependent tuning
  - LOOCV - small bandwidths for positively correlated data

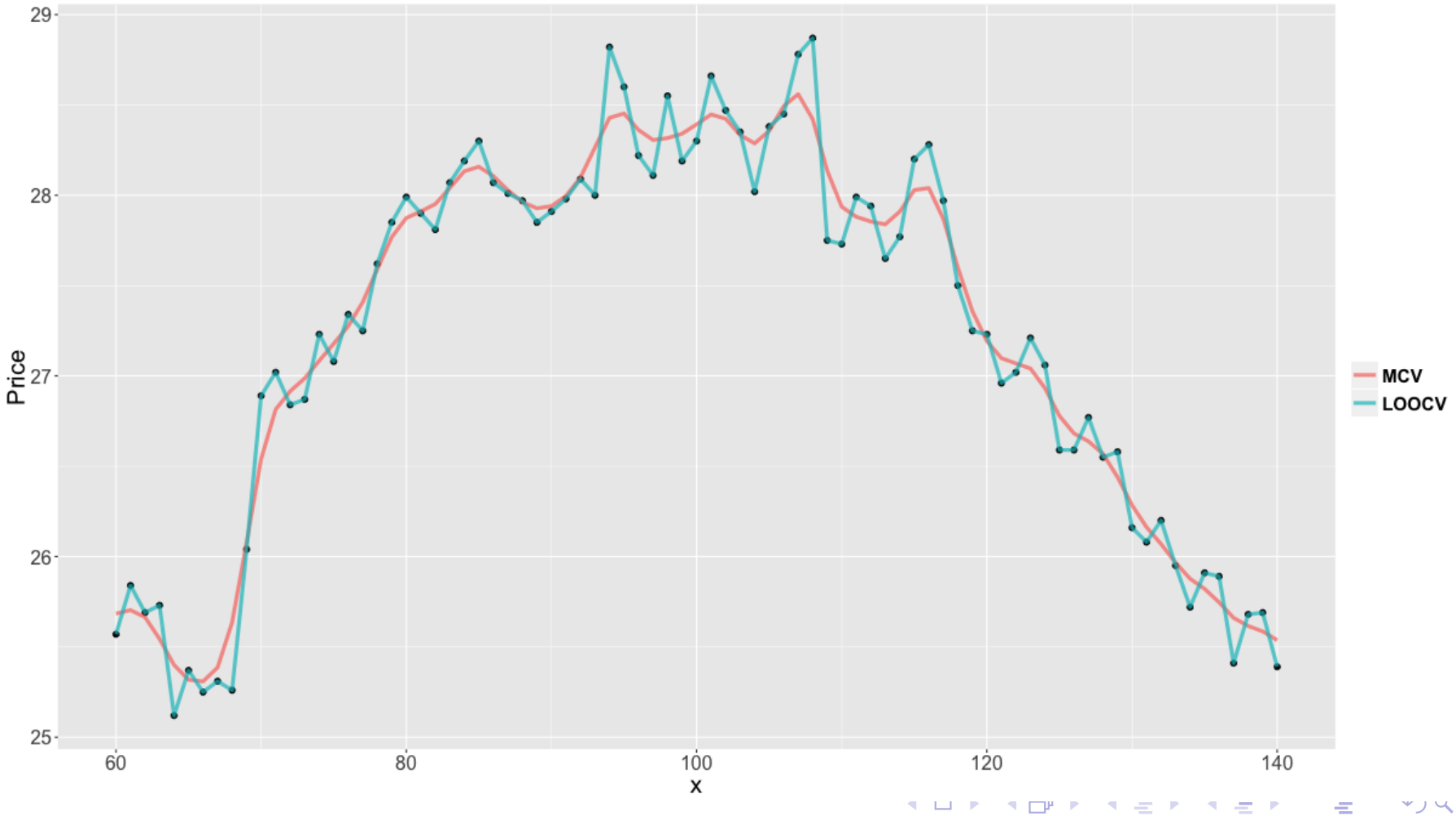
$$CV(h) = \frac{1}{n}\sum_{t=1}^{n}\left[y_t - \hat{f}_{h,t}(x_t)\right]^2$$

  - "Generalized" cross-validation - suitable for dependent data [Yao and Tong, 1998]

$$ECV_s(h) = \frac{1}{n-s}\sum_{t=s+1}^{n}\left[y_t - \hat{f}_{h,t}(x_t)\right]^2$$

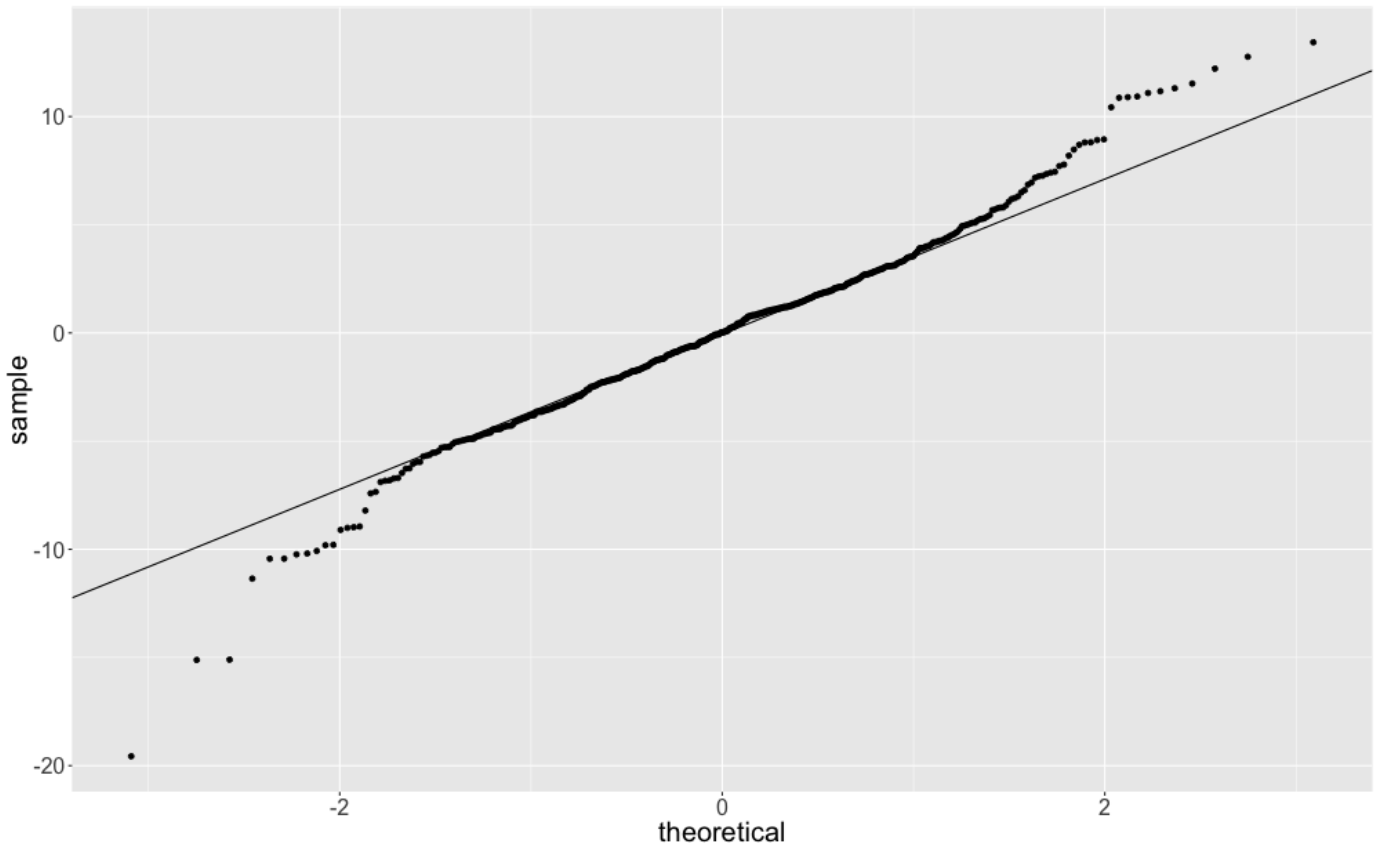
Bandwidth Selection

Figure: Comparison of cross-validation methods for daily prices of MSFT



Heavy-Tails and Outliers

Figure: Normal Q-Q plot of S&P 500 residuals



Kernel Smoothing for Stock Price Data

- We first consider:

$$y_t = f(x_t) + \delta_t \quad t = 1, \ldots, n \quad \delta_t \sim N(0, \sigma^2(x_t)) \qquad (1)$$

- The objective function for the Nadaraya-Watson estimator:

$$\arg\min_{f(x)} \sum_{t=1}^{n} K_h(x - x_t)[y_t - f(x)]^2 \qquad (2)$$

- Previous approaches - Huber's loss, L_1, convex functions

Outlier Detection and Robust Estimation

- New form with mean-shift parameter

$$y_t = f(x_t) + \gamma_t + \epsilon_t \quad t = 1, \ldots, n \quad \epsilon_t \sim N(0, \sigma^2(x_t)) \quad (3)$$

$\gamma_t \neq 0$ for outliers, $\gamma_t = 0$ otherwise

- $\boldsymbol{\gamma}$ should be sparse
- **Goal**: estimate $f(x_t)$ and $\boldsymbol{\gamma}$

Outlier Detection and Robust Estimation

- The constrained form:

$$\underset{f(x),\boldsymbol{\gamma}}{\arg\min} \sum_{t=1}^{n} K_{h,t}[y_t - \gamma_t - f(x)]^2 \quad \text{s.t.} \quad \sum_{t \in \mathcal{J}(x)} \mathbb{1}_{\gamma_t \neq 0} \leq q(x) \tag{4}$$

  where $K_{h,t} \equiv K_h(x - x_t)$, $\mathcal{J}(x) = \{\, t : K_h(x - x_t) \neq 0 \,\}$

- L_0 - non-convex, non-differentiable and discrete

Outlier Detection and Robust Estimation

$\Theta^{\#}$ is quantile thresholding:

$$\Theta^{\#}(\boldsymbol{x}; q) = \begin{cases} x_{(j)} & \text{if } j \leq q \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$x_{(1)}, \ldots, x_{(n)}$ are the order statistics: $|x_{(1)}| \geq \cdots \geq |x_{(n)}|$

Outlier Detection and Robust Estimation

Alternating minimization

- Let $w_{h,t} \equiv K_h(x_k - x_t) / \sum_{t=1}^n K_h(x_k - x_t)$
- At a target point x_k: start with $\boldsymbol{y}^{\text{adj}} = \boldsymbol{y} - \boldsymbol{\gamma}^{(j)}$ and $j = 1$
- $\hat{f}(x_k) = \sum_{t=1}^n w_{h,t} y_t^{\text{adj}}$
- $\boldsymbol{r}^{(j)} = \text{diag}\{\sqrt{\boldsymbol{w}_{h,t \in \mathcal{J}(x_k)}}\}(\boldsymbol{y} - \hat{f}(x_k))$
- $\boldsymbol{\gamma}^{(j+1)} = (\text{diag}\{\sqrt{\boldsymbol{w}_{h,t \in \mathcal{J}(x_k)}}\})^{-1} \Theta^{\#} \left(\boldsymbol{r}^{(j)}; q(x_k)\right)$
- Stop if $\|\boldsymbol{\gamma}^{(j+1)} - \boldsymbol{\gamma}^{(j)}\|_\infty$ is small

Outlier Detection and Robust Estimation

- A pooling procedure to accumulate the outlier estimates
- Collect each $\hat{\boldsymbol{\gamma}}$ column vector into the $n \times n$ matrix $\hat{\boldsymbol{\Omega}}$:

$$\hat{\boldsymbol{\Omega}} = \begin{bmatrix} \hat{\boldsymbol{\gamma}}_1{}^T \\ \vdots \\ \hat{\boldsymbol{\gamma}}_n{}^T \end{bmatrix}^T$$

- Apply $m(\boldsymbol{x}) = \text{sign}(\boldsymbol{x}) \odot \max(|\boldsymbol{x}|)$ to each row $\hat{\boldsymbol{\omega}}_i$ in $\hat{\boldsymbol{\Omega}}$, combine into an $n \times 1$ vector:

$$\boldsymbol{\nu} = \begin{bmatrix} m(\hat{\boldsymbol{\omega}}_1) \\ \vdots \\ m(\hat{\boldsymbol{\omega}}_n) \end{bmatrix}$$

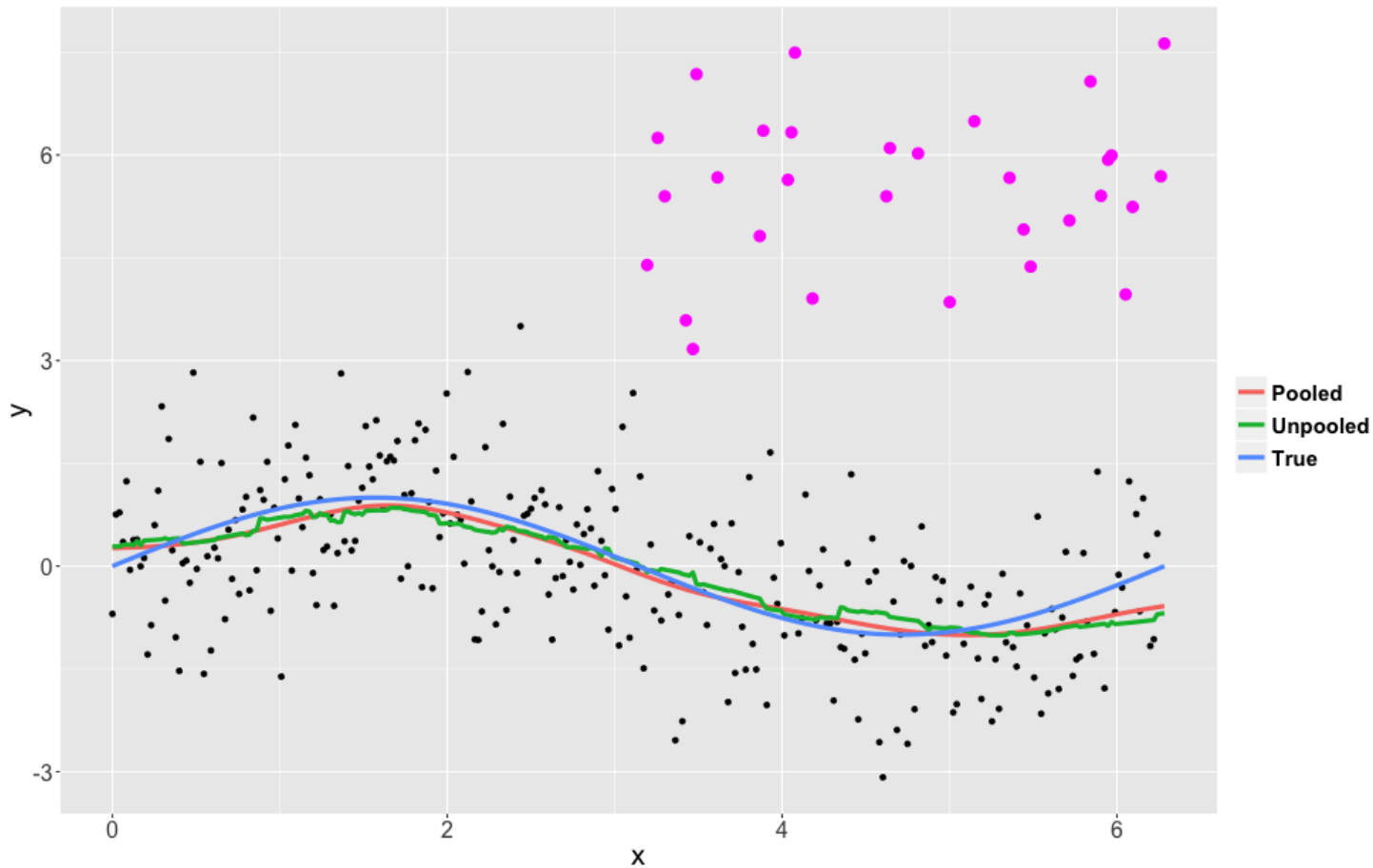
Outlier Detection and Robust Estimation

- Compute $\hat{\boldsymbol{\gamma}}^{\mathrm{P}} = \Theta^{\#}(\boldsymbol{\nu}; q^{\mathrm{p}})$
- The pooled clean data set $\{(x_t, y_t^{\mathrm{adj}} = y_t - \hat{\gamma}_t^{\mathrm{P}}), 1 \leq t \leq n\}$
- The pooled function estimate
  $\hat{f}(x_k)^{\mathrm{P}} = \sum_{t=1}^{n} K_h(x_k - x_t) y_t^{\mathrm{adj}} / \sum_{t=1}^{n} K_h(x_k - x_t)$

Non-Parametric Outlier Identification and Smoothing (NOIS)

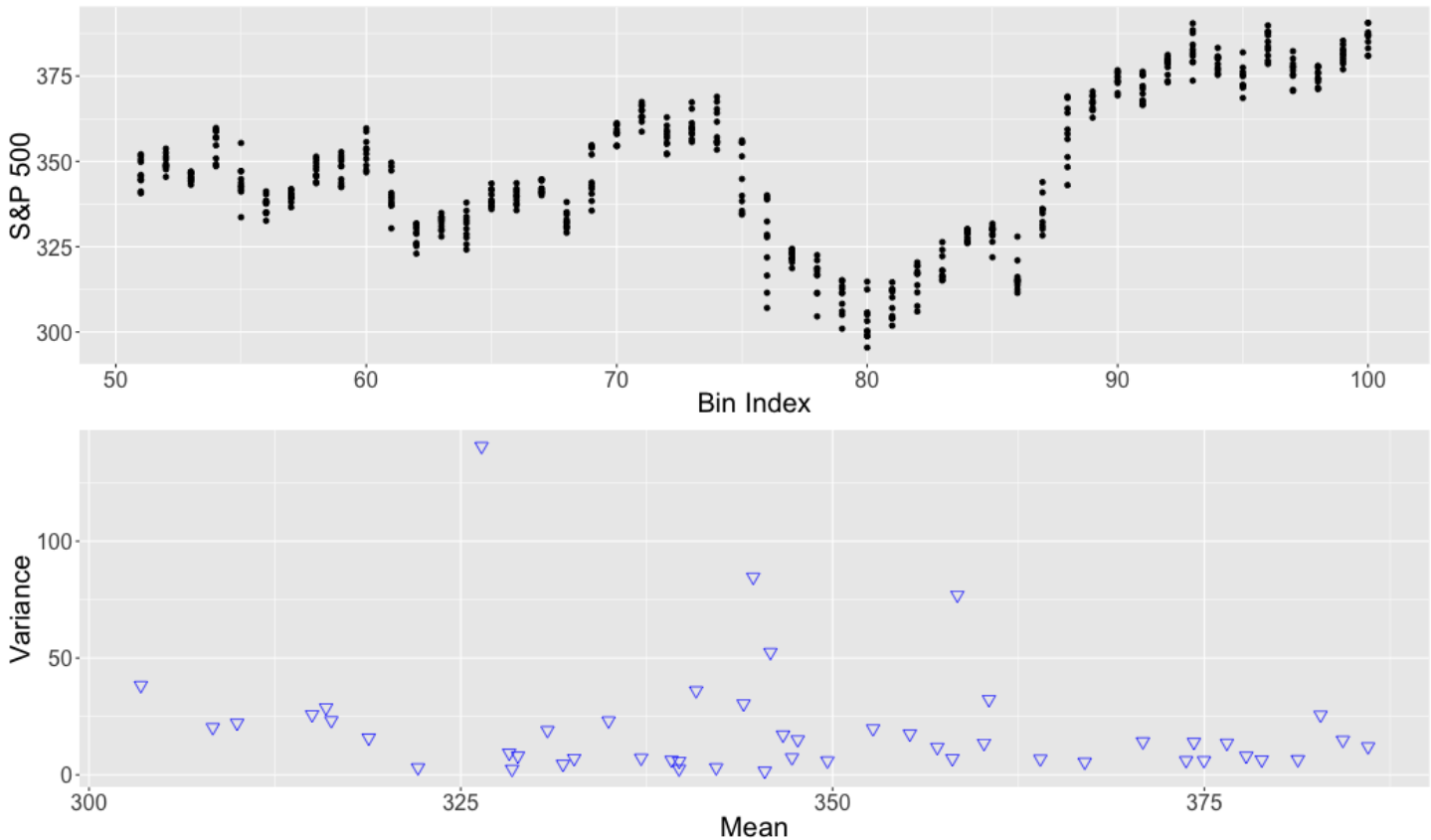
Outlier Detection and Robust Estimation

Figure: Pooled and unpooled estimates



Heteroskedasticity

Figure: S&P 500 closing prices and mean-variance plot



NOIS Bundles

- Account for heteroskedasticity and uncertainty
- Bundle widths can show pertinent features
- Naive approach - normal approximation
  - Asymptotic result - how large is n?
  - Symmetric
  - Requires $\hat{\sigma}^2$
- We use non-parametric methods

Empirical Likelihood

- Non-parametric LRT for testing:

$$H_0 \ : f(x) = \theta_0 \quad \text{vs} \quad H_1 \ : f(x) \neq \theta_0$$

- Assign probability weights p_t to the y_t^{adj}'s

The log EL **ratio** $lr_n(\theta_0)$:

$$\max \sum_{t=1}^{n} \log(np_t) \quad \text{s.t.} \quad \sum_{t=1}^{n} p_t K_{h,t}(y_t^{\text{adj}} - \theta_0) = 0, \ \sum p_t = 1, \ p_t \geq 0$$

We can use Lagrange multipliers for the optimization

EL Pointwise Confidence Bands

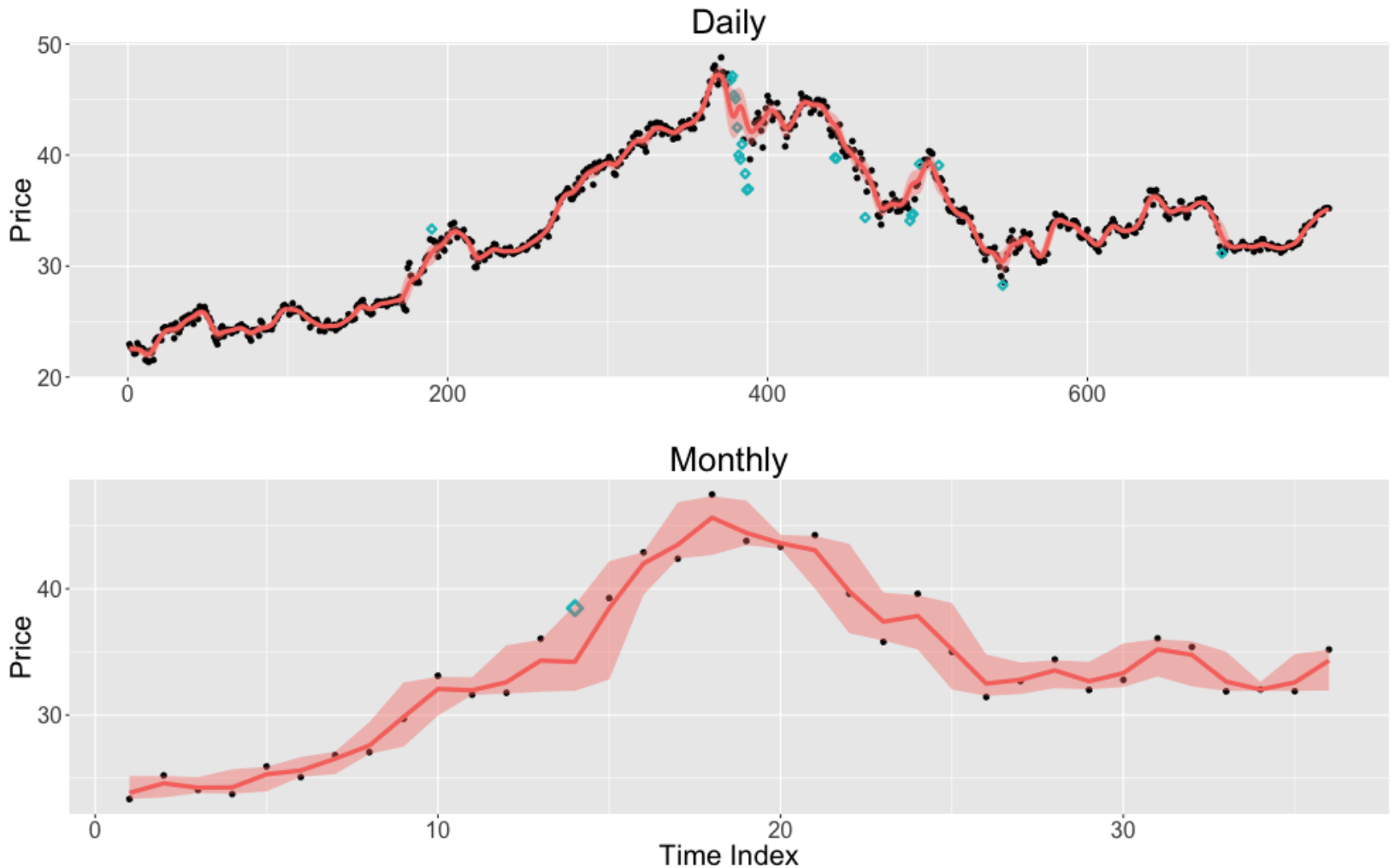
- Under H_0, $-2lr_n(\theta_0) \to \chi^2_1$ in distribution [Chen, 1996]
- A $1-\alpha$ confidence interval:

$$\left\{\theta : -2lr_n(\theta) \leq \chi^2_{1,1-\alpha}\right\}$$

- This is a root-finding problem (bisection, Brent's method, etc.)
- No explicit $\hat{\sigma}^2$!
- Parallel computation is possible

Pointwise Confidence Bands

Figure: Daily and monthly stock prices for LMT



Simulation Studies

- 200 points: $y_t = \sin(x_t) + e_t$ with x_t's evenly spaced on $[0, 2\pi]$ and $e_t \sim N(0, 1)$
- Outliers: randomly sample $O \in \{10, 20, 35\}$ of the y_t's where $x_t \in [\pi, 2\pi]$, shift them up by $U(10, 12)$
- 50 simulations

Simulation Studies

Outlier detection: fix $q^{\mathrm{P}} = 30$ and report

- **MSE**: the mean squared error
- **M**: the mean masking probability (fraction of undetected true outliers)
- **S**: the mean swamping probability (fraction of good points labeled as outliers)
- **JD**: the joint outlier detection rate (fraction of simulations with 0 masking)

Simulation Studies

Table: Outlier Detection Simulation Results

	$O = 10$	$O = 20$	$O = 35$
MSE	0.12	0.14	0.85
M (%)	0	0	18
S (%)	10	5.6	0.78
JD (%)	100	100	0

Simulation Studies

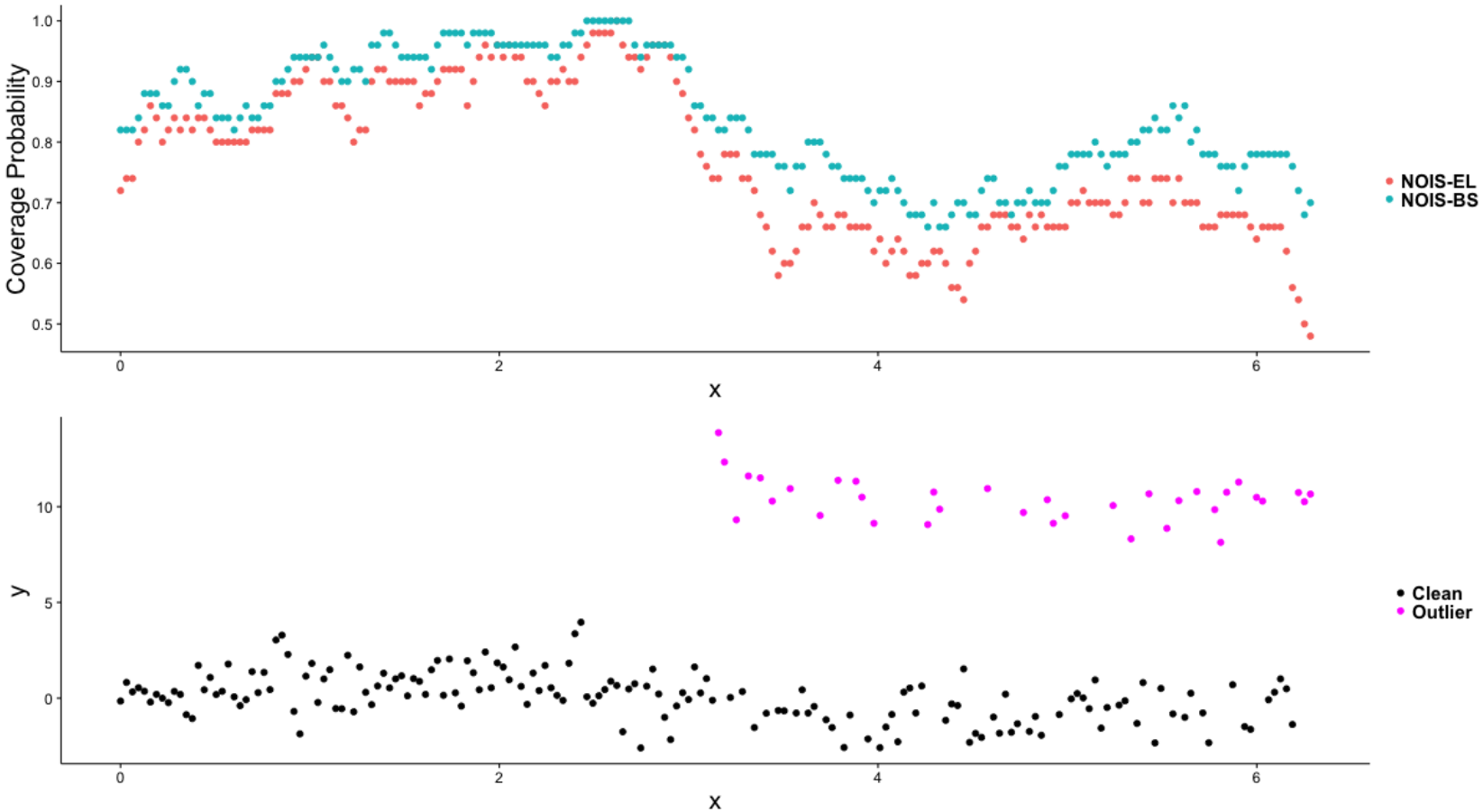
Pointwise confidence band coverage levels: let $q^{\mathrm{P}} = O$

Table: Coverage Probability Simulation Results ($\alpha = 0.05$)

	NOIS-EL	NOIS-BS	KS-EL	KS-BS
$O = 10$	0.87 (0.62)	0.90 (0.72)	0.42 (1.03)	0.56 (1.45)
$O = 20$	0.84 (0.57)	0.88 (0.70)	0.39 (1.33)	0.52 (1.65)
$O = 35$	0.77 (0.68)	0.84 (0.99)	0.37 (1.88)	0.49 (2.60)

Simulation Studies

Figure: Coverage probabilities for NOIS-EL and NOIS-BS ($O = 35$)

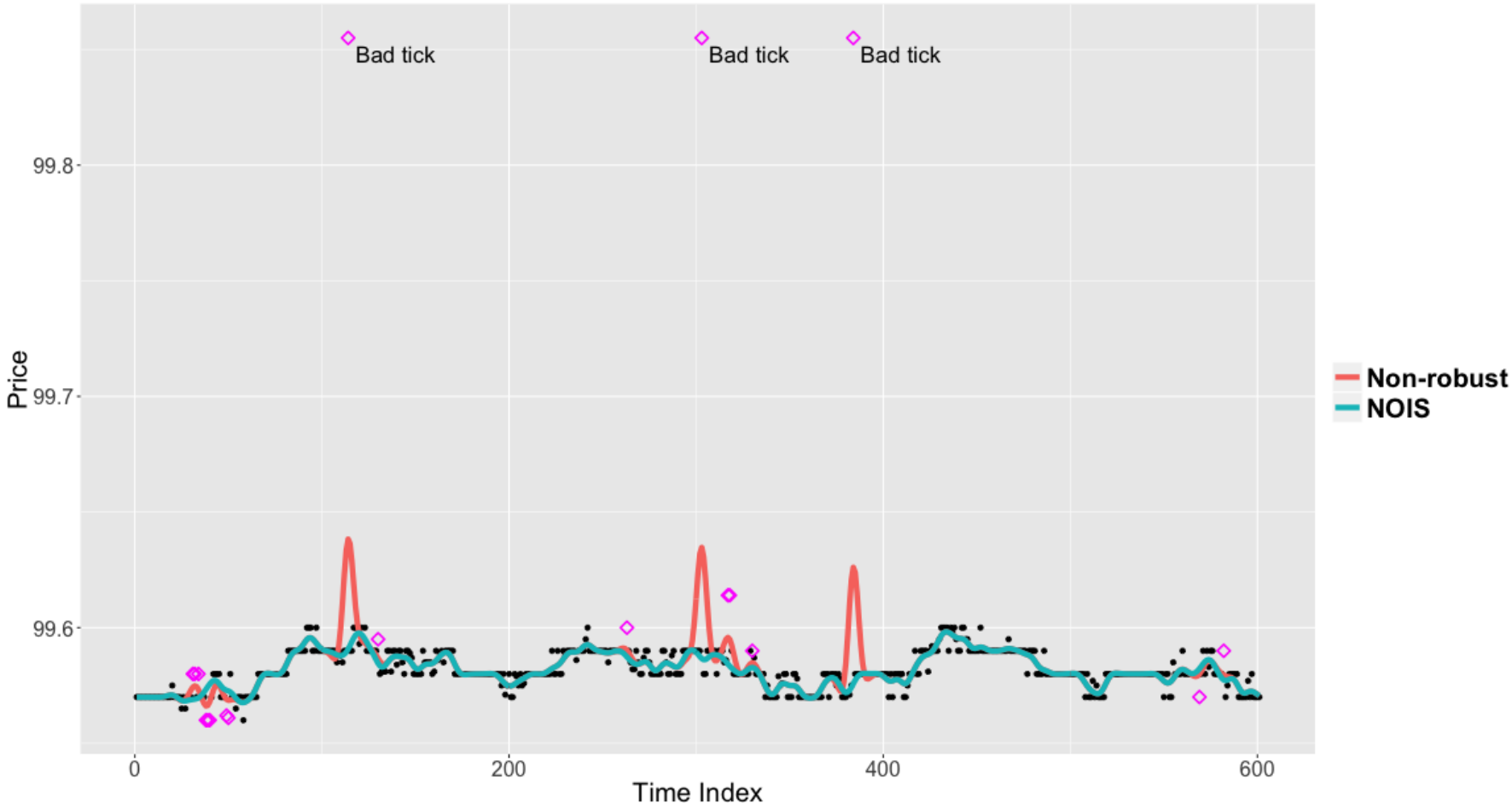


High-Frequency Data

- Intra-day data - as fine as microseconds (1 millionth of a second)
- As many as $2\% - 3\%$ of high-frequency prices are outliers [Falkenberry, 2002]
- "Bad ticks" - likely erroneous
- "Borderline cases" - subjective

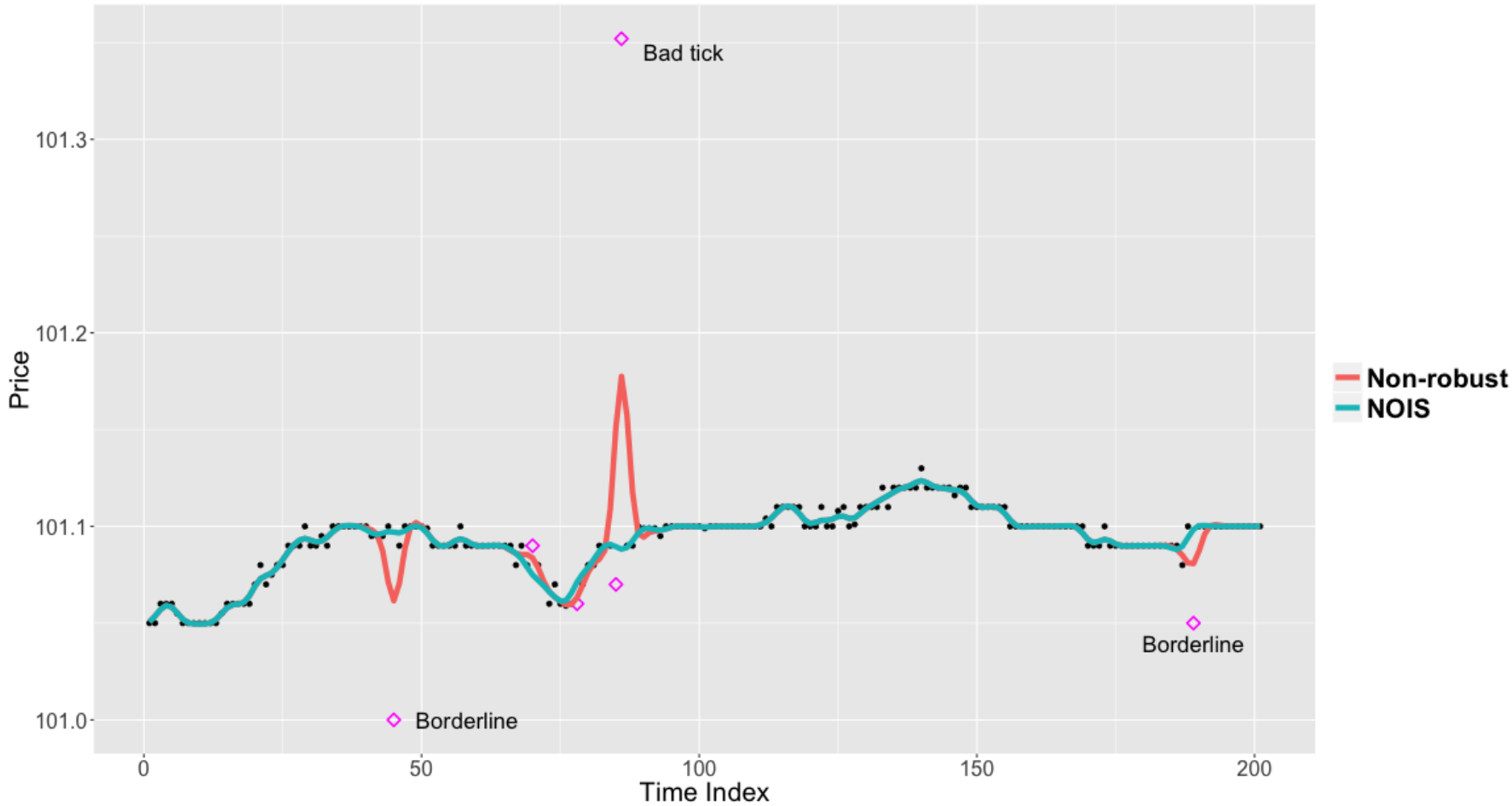
High-Frequency Data

Figure: NOIS and non-robust function estimates for MCD on Jan. 3rd, 2012 (3:00:37 to 3:08:55 pm)



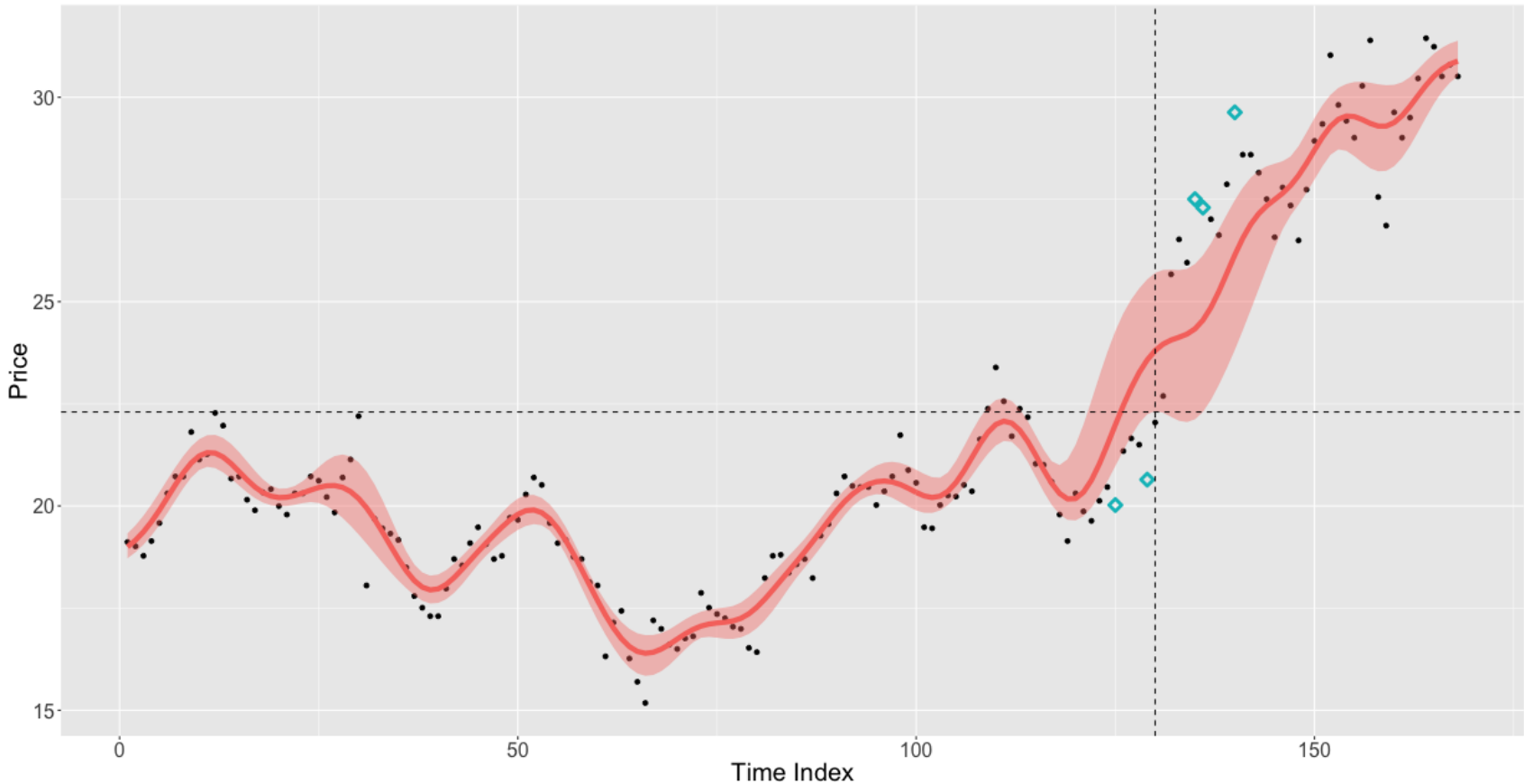
High-Frequency Data

Figure: NOIS and non-robust estimates for MCD on Jan. 3rd, 2012 (9:40:01 to 9:41:17 am)



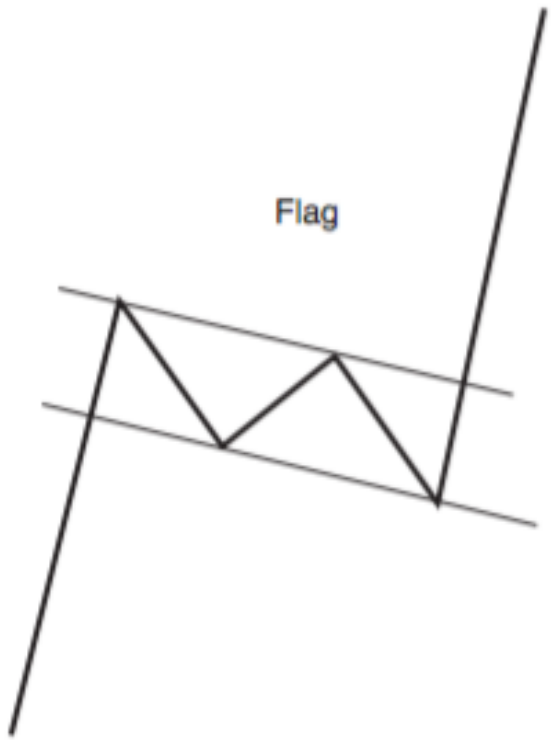
JBL Cup and Handle

Figure: NOIS bundle for JBL cup and handle pattern



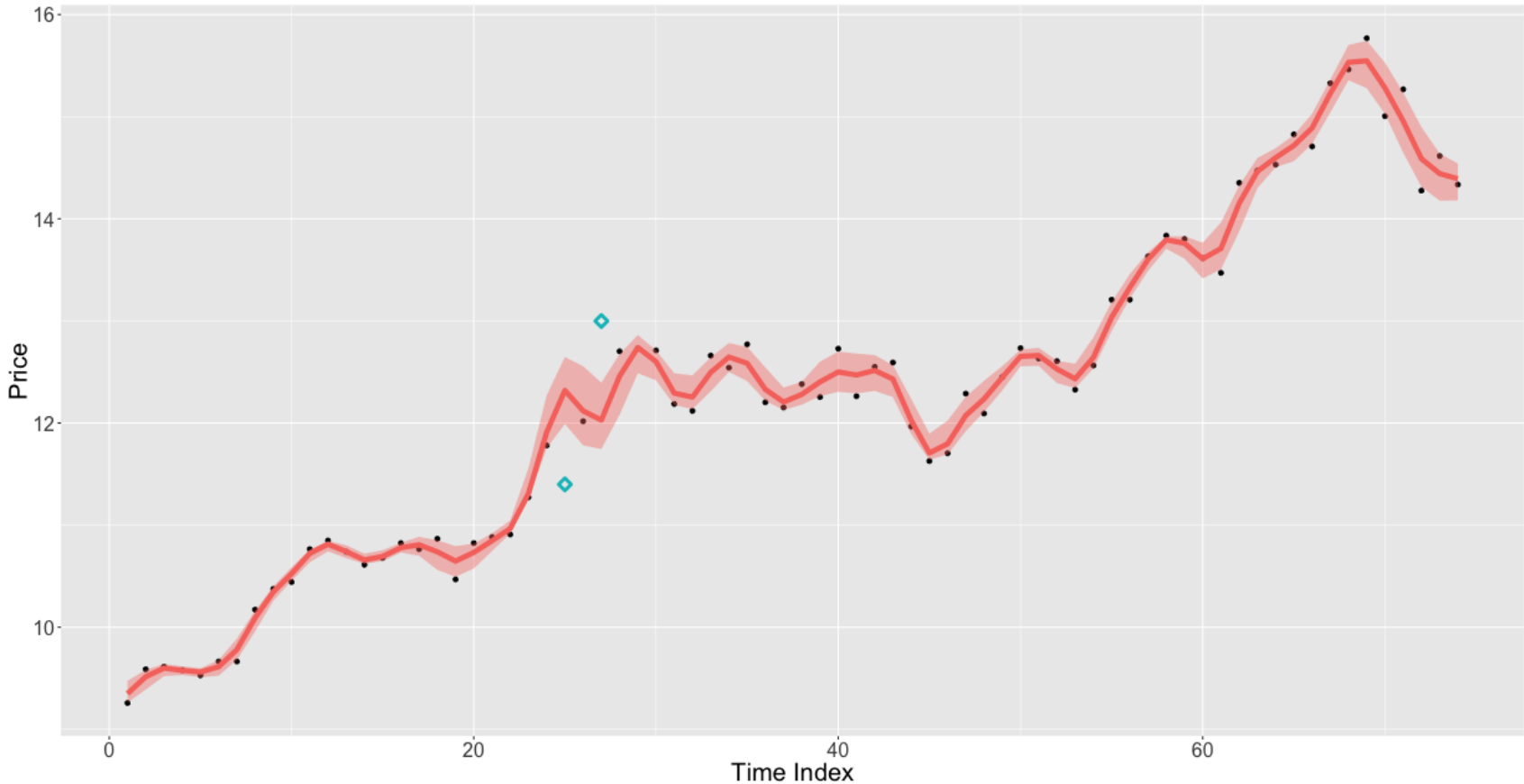
Flag

Figure: Flag chart pattern [Kirkpatrick II and Dahlquist, 2010]



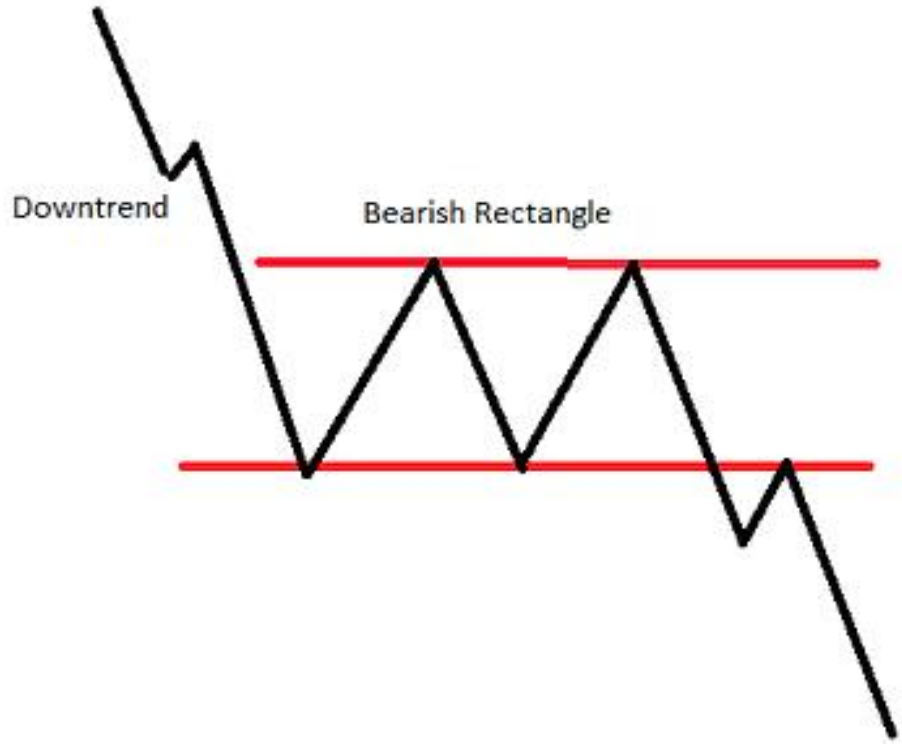
HPQ Flag Pattern

Figure: NOIS bundle for HPQ flag pattern



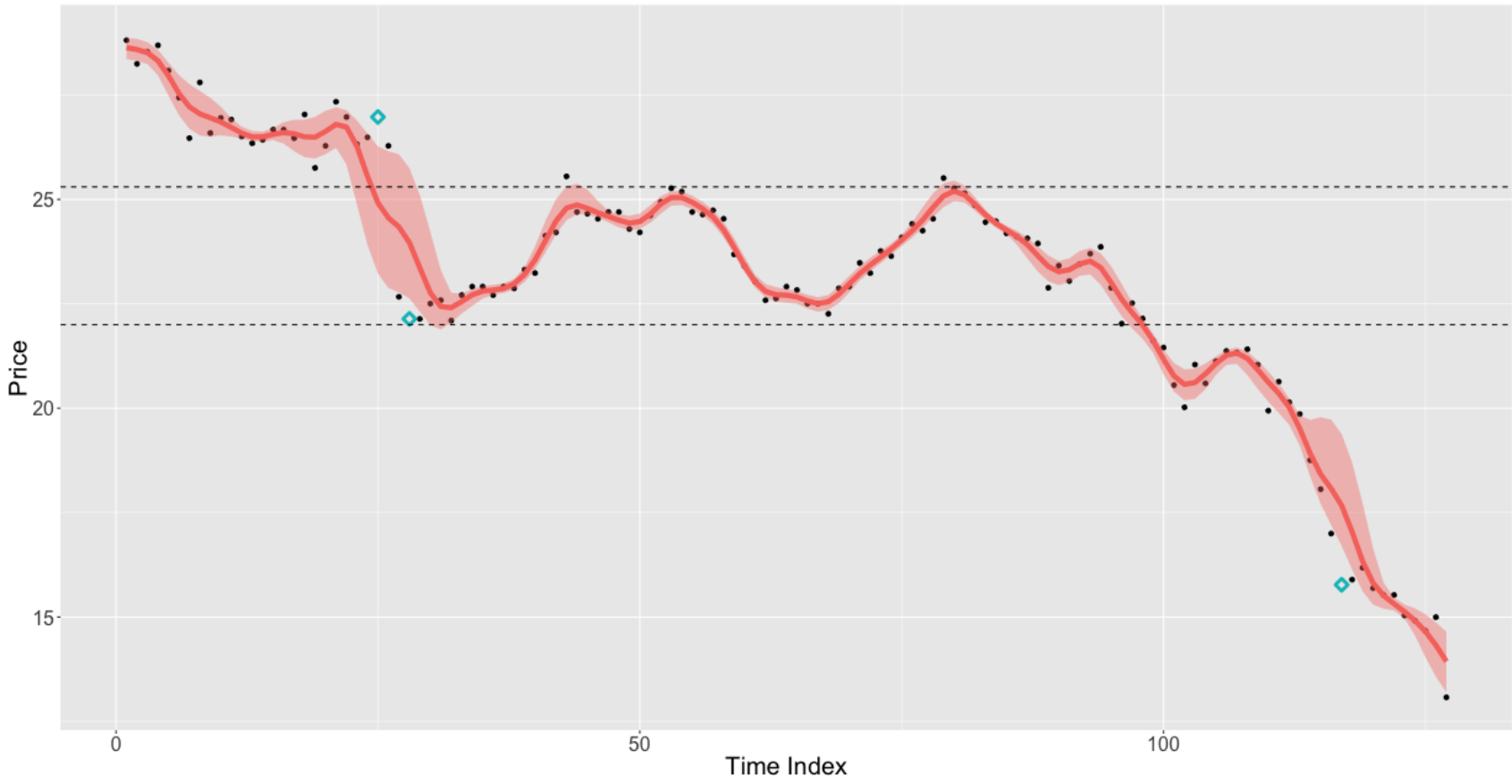
Rectangle

Figure: Rectangle chart pattern [Forex 10 Best, 2016]



LMT Rectangle

Figure: NOIS bundle for LMT rectangle pattern



Discussion

- Three challenges of financial data - non-linearity, outliers and heteroskedasticity
- NOIS can detect "bad ticks", but we might be overscrubbing the data
- NOIS bundles - new insights in technical analysis chart patterns
- Simulations - NOIS is successful at outlier detection but has undercoverage issues

Future Work

- Joint parameter tuning of q^{P} and h
- Undercoverage issues - maybe an average instead of $\max(\cdot)$
- Additional L_2 penalty for shrinkage
- Algorithm for **automatically** detecting chart patterns
- Incorporate volume of stocks
- Augmented Lagrangian for EL in big data

References I


Chen, S. X. (1996).
Empirical likelihood confidence intervals for nonparametric density estimation.
Biometrika, 83(2):329–341.


Falkenberry, T. N. (2002).
High frequency data filtering.
Tick Data, Technical.


Forex 10 Best (2016).
Bearish rectangle.
[Online; accessed September 28th, 2016].


Kirkpatrick II, C. D. and Dahlquist, J. (2010).
Technical analysis: the complete resource for financial market technicians.
FT press.

References II


Lian, H. (2012).
Empirical likelihood confidence intervals for nonparametric functional data analysis.
Journal of Statistical Planning and Inference, 142(7):1669–1677.


Lo, A. W., Mamaysky, H., and Wang, J. (2000).
Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation.
The journal of finance, 55(4):1705–1770.


StockCharts.com (2016).
Chart patterns.

References III


Yao, Q. and Tong, H. (1998).
Cross-validatory bandwidth selections for regression estimation based on dependent data.
Journal of Statistical Planning and Inference, 68(2):387–415.