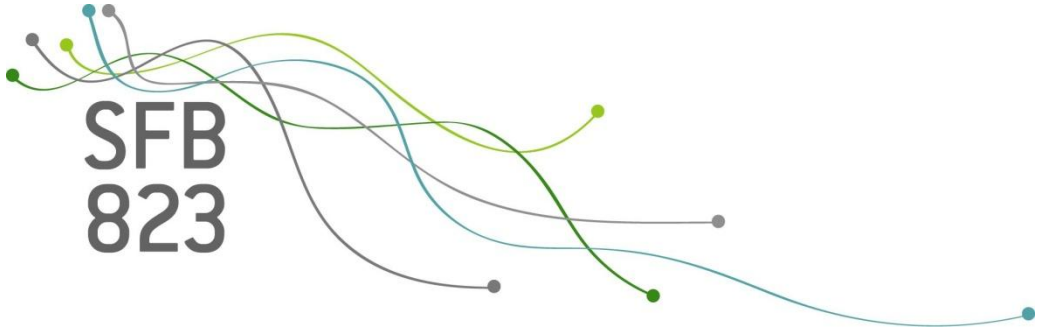
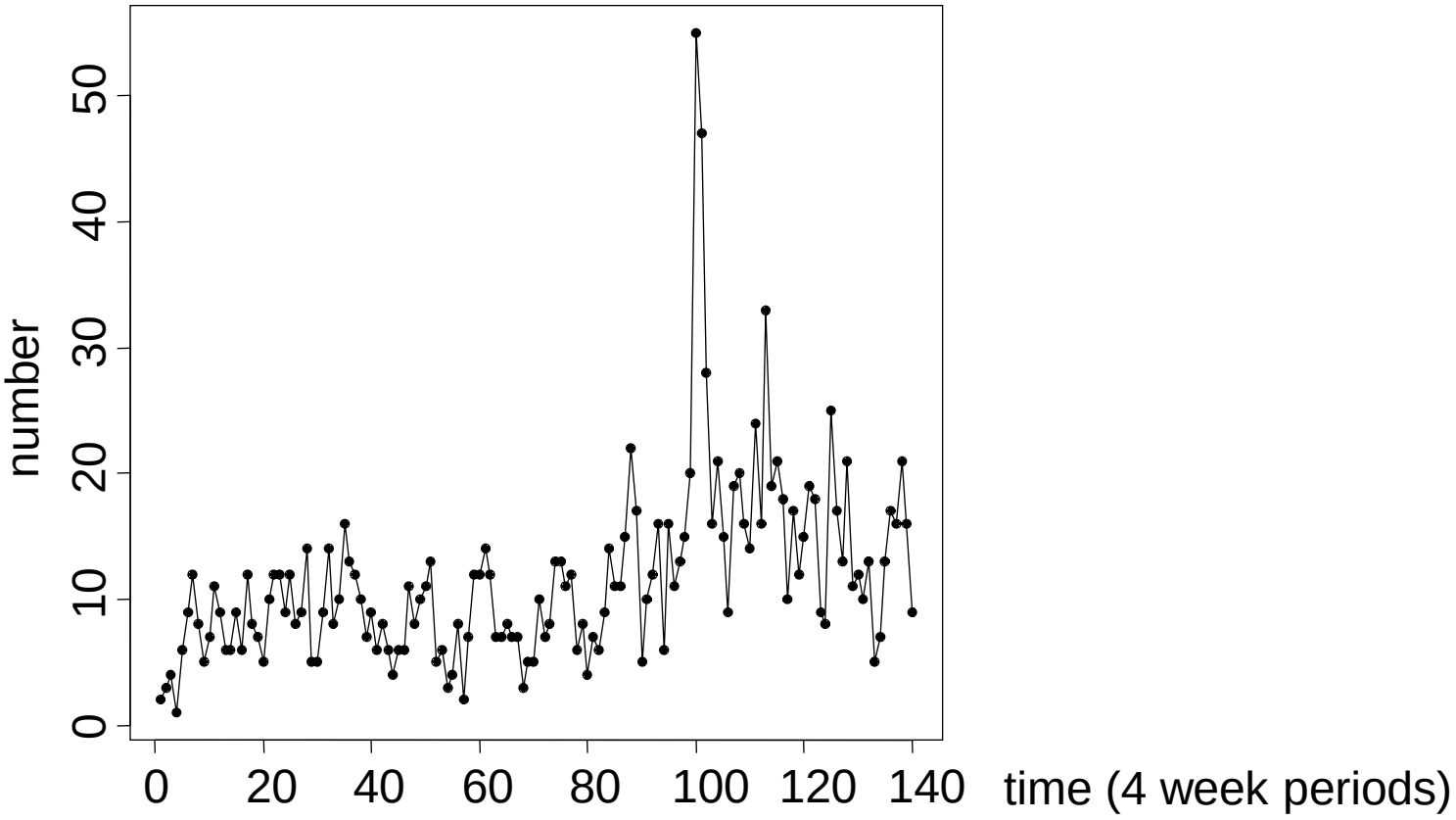
Outliers and Interventions in Count Time Series

Roland Fried, Tobias Liboschik
Department of Statistics, TU Dortmund

Konstantinos Fokianos
Dept. Mathematics and Statistics, University of Cyprus



Motivation: Number of Campylobacterosis Infections



INGARCH-model: $Y_t \mid (Y_s, s<t) \sim \text{Poi}(\lambda_t)$ *Ferland, Latour, Oraichi (2006)*

$$\lambda_t = \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-13}$$

Possible outliers at times 100, 101; level shift after time 80?

Time Series following GLM

GLM: $Y_t \big| (Y_s, s<t) \sim \text{Poi}(\lambda_t)$

$$\eta(\lambda_t) = \beta_0 + \sum_{i=1}^{p} \beta_i \eta(Y_{t-i} + c) + \sum_{j=1}^{q} \alpha_j \eta(\lambda_{t-j})$$

with p and q model orders, η link function, e.g.:

Loglinear generalized autoregression: η = log, c=1

Ergodicity for p=q=1 *Fokianos & Tjøstheim (2011)*

Linear generalized autoregression (INGARCH): η = id , c=0

Ergodicity for INGARCH(1,1) *Fokianos, Rahbek & Tjøstheim (2009)*
Neumann (2011)

Overview

1) **Intervention analysis for INGARCH models**

   - Intervention models: internal / external

   - Detection: Time & type known

   - Type unknown: classification

   - Time & type unknown: parametric bootstrap

2) **Bayesian modelling of additive outliers**

3) **Robust fitting of IN(G)ARCH models**

1) Interventions in the INGARCH(1,1)-Model

Clean process: $Y_t \mid (Y_s, s<t) \sim \text{Poi}(\lambda_t)$

$$\lambda_t = \beta_0 + \beta_1 Y_{t-1} + \alpha_1 \lambda_{t-1}$$

Additive internal / external effect of size ν at time τ:

$$Z_t \mid (Z_s, s<t) \sim \text{Poi}(\kappa_t), \qquad X_t = \delta^{t-\tau} I_{[\tau,\infty)}(t)$$

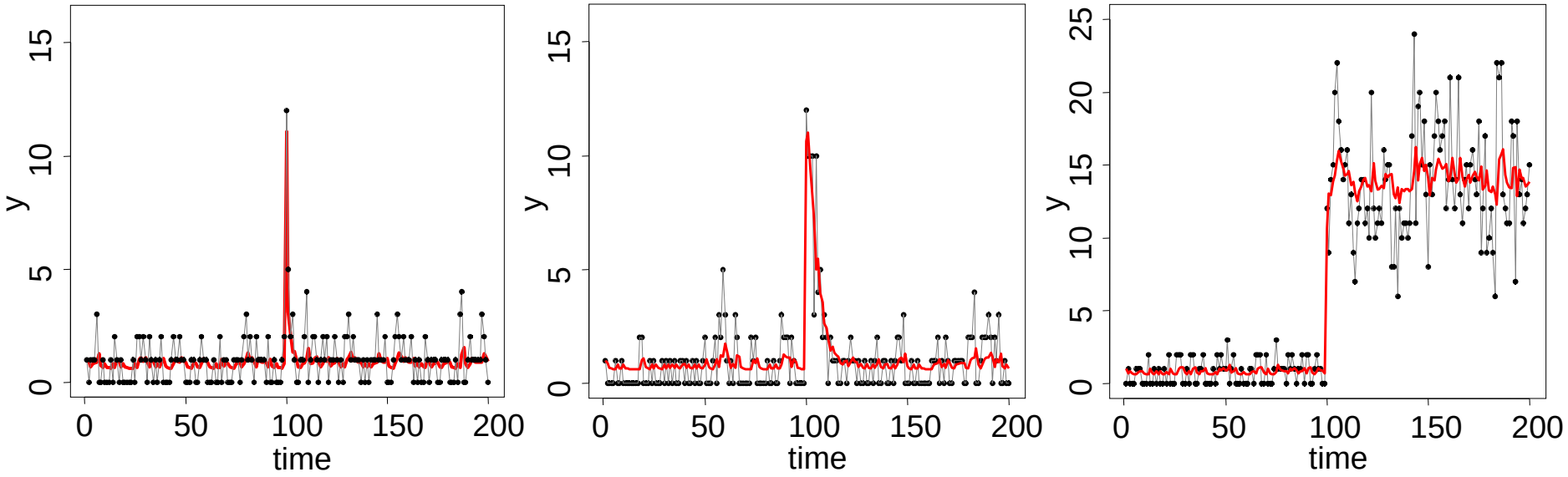
$$\kappa_t = \beta_0 + \beta_1 Z_{t-1} + \alpha_1 (\kappa_{t-1} - \nu X_{t-1}) + \nu X_t$$

Equivalently $Z_t = Y_t + C_t, \quad C_t \mid (C_s, s<t) \sim \text{Poi}(\mu_t)$

for $\nu>0$ and $t \geq \tau$: $\mu_t = \beta_1 C_{t-1} + \alpha_1 \mu_{t-1} + \nu \delta^{t-\tau}$

Different Types of Interventions

δ=0 spiky outlier (SO) δ=0.8 transient shift (TS) δ=1 level shift (LS)



underlying mean process (κ_t)

Goal: Detect and classify different outliers

Conditional Maximum Likelihood: Time & Type known

Score function

$$S_{n\tau}(\theta) = \sum_{t=1}^{n}\left(\frac{z_t}{\kappa_t(\theta)} - 1\right)\frac{\partial \kappa_t(\theta)}{\partial \theta}$$

$$\theta = (\beta_0, \beta_1, \alpha_1, \nu)$$

with the recursions:

$$\frac{\partial \kappa_t(\theta)}{\partial \beta_0} = 1 + \alpha_1 \frac{\partial \kappa_{t-1}(\theta)}{\partial \beta_0}$$

$$\frac{\partial \kappa_t(\theta)}{\partial \beta_1} = z_{t-1} + \alpha_1 \frac{\partial \kappa_{t-1}(\theta)}{\partial \beta_1}$$

$$\frac{\partial \kappa_t(\theta)}{\partial \alpha_1} = \kappa_{t-1}(\theta) + \alpha_1 \frac{\partial \kappa_{t-1}(\theta)}{\partial \alpha_1} \quad -\nu\delta^{t-1-t}I(t>\tau)$$

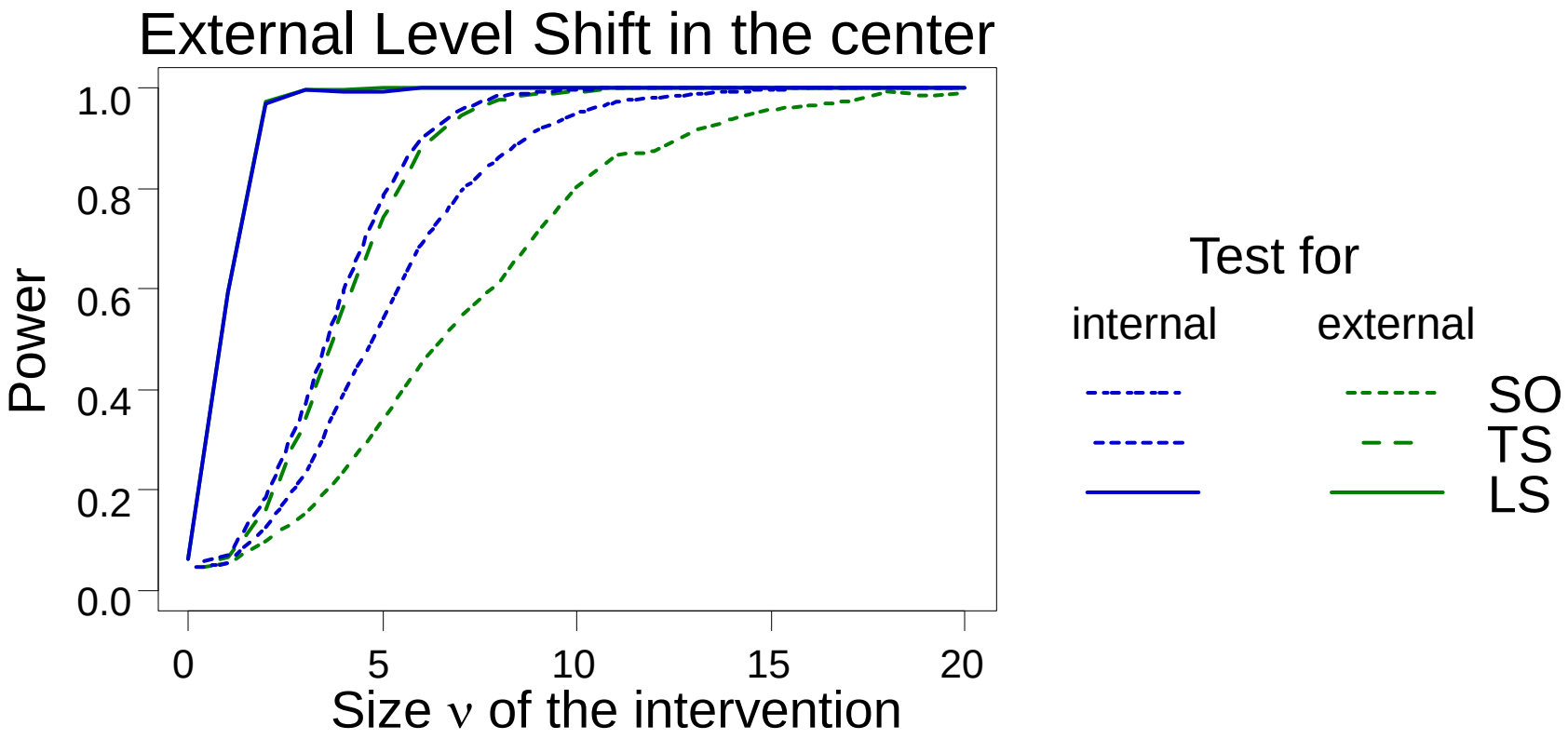
$$\frac{\partial \kappa_t(\theta)}{\partial \nu} = \delta^{t-\tau} I_{[\tau,\infty)}(t) + \alpha_1\left(\frac{\partial \kappa_{t-1}(\theta)}{\partial \nu} \quad -\delta^{t-1-t}I(t>\tau)\right)$$

Conditional information

$$G_{n\tau}(\theta) = \sum_{t=1}^{n} \mathrm{Cov}_{t-1}\left[\frac{\partial \ell_t(\theta)}{\partial \theta}\right] = \sum_{t=1}^{n} \frac{1}{\kappa_t(\theta)}\left(\frac{\partial \kappa_t(\theta)}{\partial \theta}\right)\left(\frac{\partial \kappa_t(\theta)}{\partial \theta}\right)'$$

Simultaneous score tests for outliers at all time points

Power in case of external Level Shift



Tests detect also other patterns:

Robust vs. Misspecification, but classification rules needed

Restrict to internal effects in the following

Classification for time τ=100 known, n=200

β_0=1, α_1=0.3, β_1=0.3

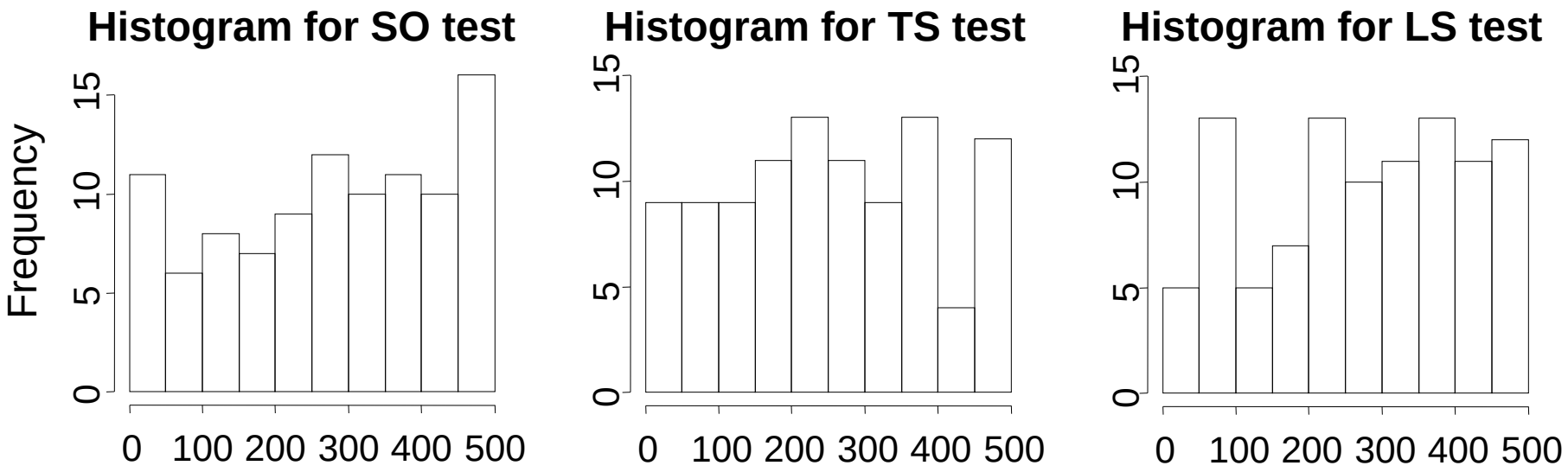
		LS	TS	SO
	ν= 0	4.0	2.5	1.0
LS	ν= 1	50.0	1.5	11.5
	ν= 2	96.0	0.0	3.0
TS	ν= 5	1.5	35.0	7.5
	ν=10	2.2	92.8	2.5
SO	ν= 5	2.0	1.0	18.0
	ν=10	3.0	0.2	68.0

Classify as LS if LS detected
classify as SO / TS according to larger test statistic otherwise

Time τ unknown: parametric bootstrap

1) Fit INGARCH model without outliers to the data
   Calculate maximum score test statistics for each type

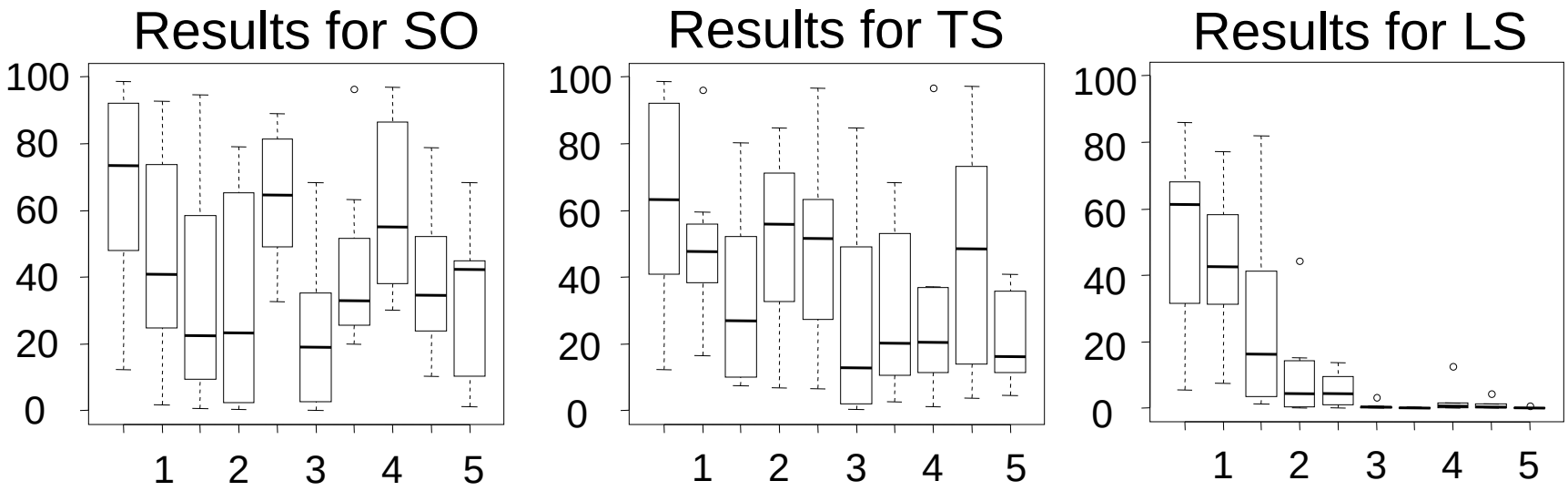
2) Generate b clean INGARCH series from the fitted model
   Calculate maximum score test statistics using true parameters



Number of bootstrap test statistics exceeding that of the original series (b=500)

Bootstrap p-values roughly uniform

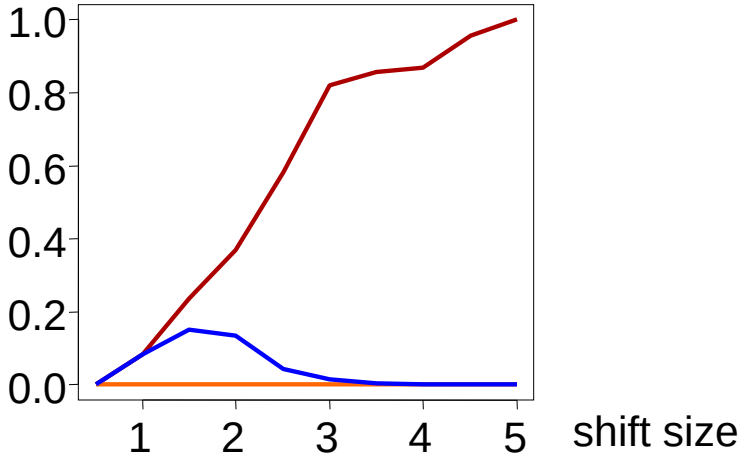
Bootstrap results for internal Level Shift



Boxplots: percentages of bootstrap test statistics exceeding that of original series in case of a level shift of increasing height ν=0.5, 1,…, 5

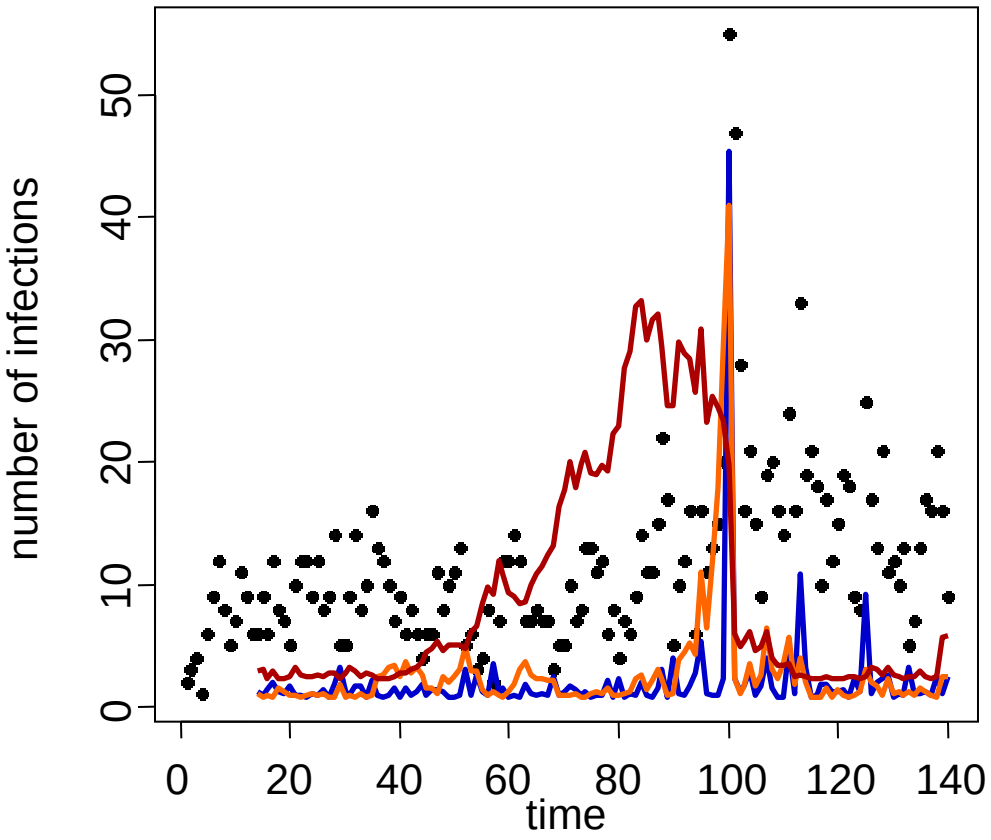
Resulting classification rate for

SO TS LS



Analysis of the Campylobacterosis Infections

Initial model: $\lambda_t = 2.439 + 0.196\lambda_{t-13} + 0.591 z_{t-1}$



SO test statistic

p-value 0 at τ=100

TS test statistic

p-value 0 at τ=100

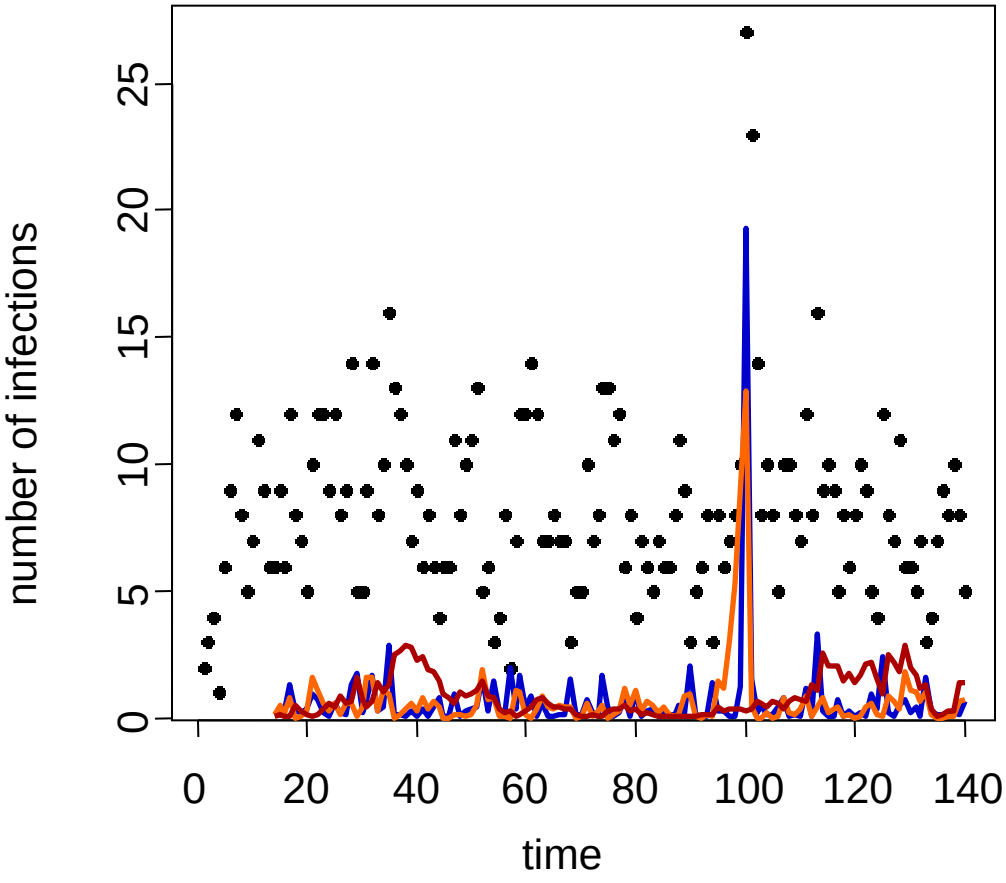
LS test statistic

p-value 0 at τ=84

LS of size 7.64 (external: 4.6) detected at τ_1=84

Analysis of the Campylobacterosis Infections (2)

New model: $\kappa_t = 3.681 + 0.150\kappa_{t-13} + 0.409 z_{t-1}$



SO test statistic

p-value 0.0 at τ=100

TS test statistic

p-value 0.0 at τ=100

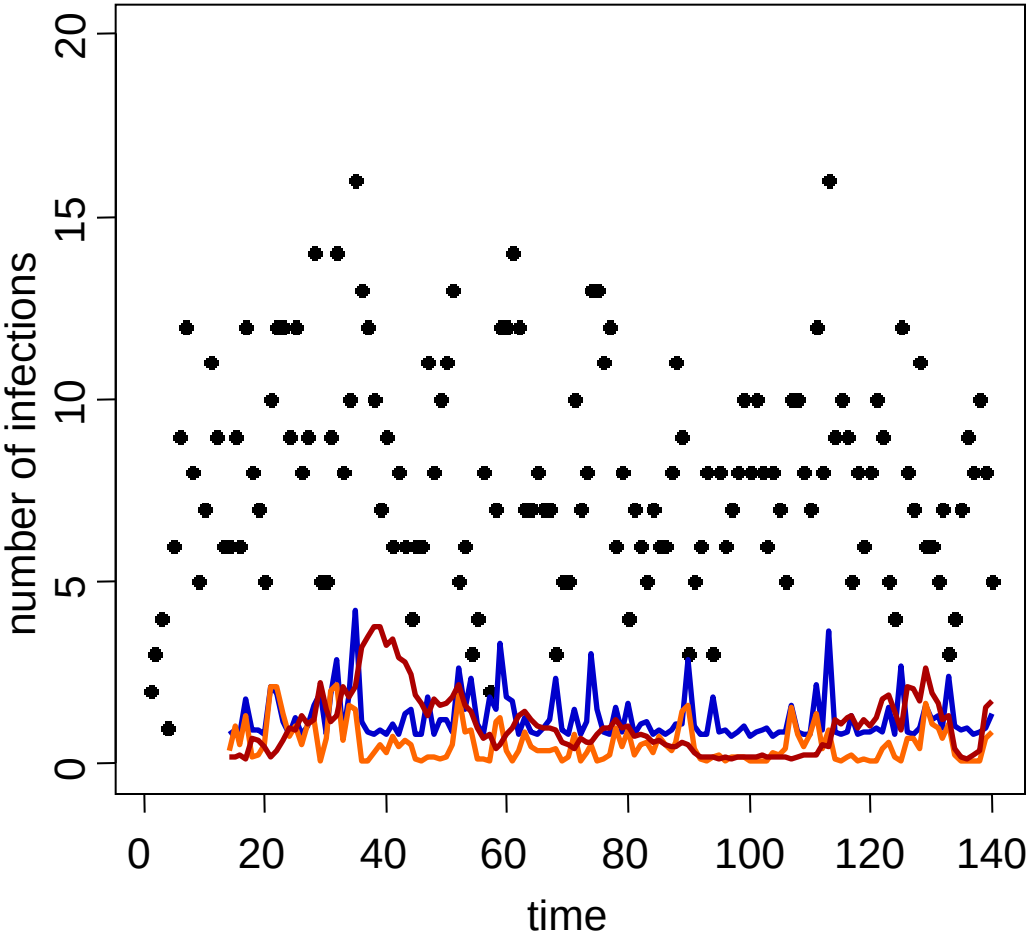
LS test statistic

p-value 0.9

SO of size 22.55 (external: 16.4) detected at τ_2=100

Analysis of the Campylobacterosis Infections (3)

New model: $\kappa_t = 2.300 + 0.387\kappa_{t-13} + 0.323 z_{t-1}$



SO test statistic

p-value of 0.65

TS test statistic

p-value of 0.90

LS test statistic

p-value of 0.83

Stop

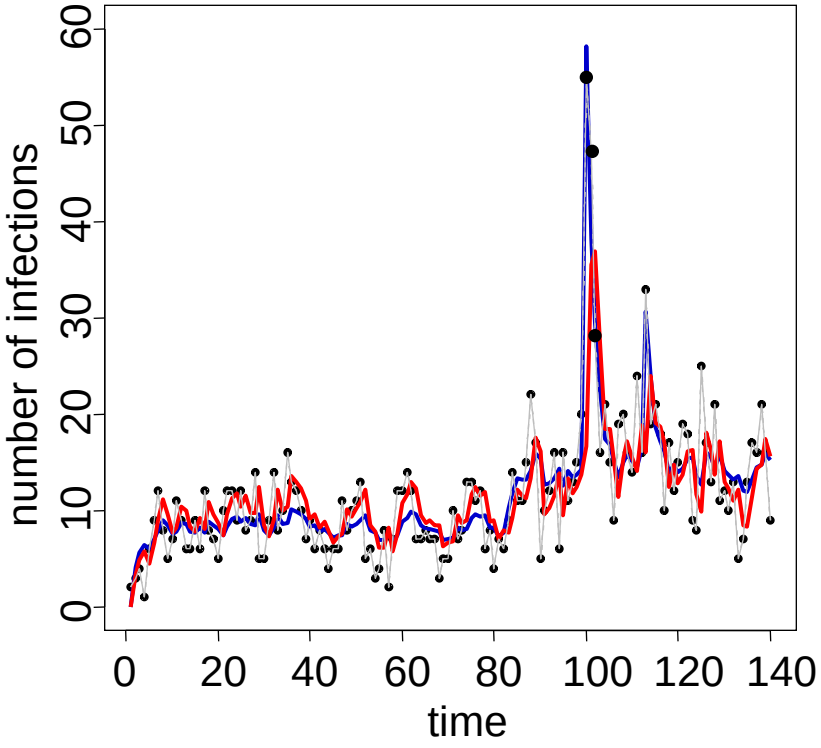
log-link: shift at time 84, spikes at times 100 and 101 detected.

Diagnostic Plots

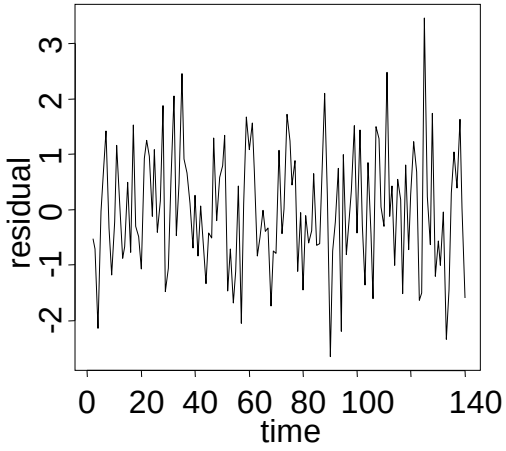
Joint estimation: $\hat{\kappa}_t = 3.64 + 0.342\kappa_{t-1} + 0.217 z_{t-1}$

$$+ 2.7 \cdot I(t \geq 83) + 42.7 \cdot I(t = 100) + 15.2 \cdot I(t = 113)$$

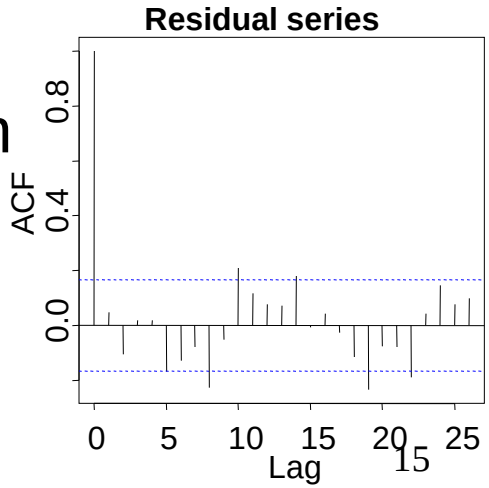
Time series and predictions $\hat{\kappa}_t$ & $\hat{\lambda}_t$



Pearson residuals



Auto-correlation function



2) Bayesian Modelling of Additive Outliers

Process contaminated by additive outlier (size ω, time τ):

$$Z_t = Y_t + C_t, \quad C_t = \delta_t \cdot \text{Poi}(\omega), \quad \delta_t \sim \text{Bern}(\pi_t),$$

$$\lambda_t = E(Y_t \mid Y_s, s < t) = \beta_0 + \alpha_1 \lambda_{t-1} + \beta_1 Y_{t-1}$$

$$(\alpha_1, \beta_1) \sim \text{Dirichlet}\,(a_1 = 1, a_2 = 1, a_3 = 1)$$

$$\pi_t \sim \text{Beta}(b_1, b_2)$$

$$\beta_0, \kappa_0, \omega \sim \text{Gamma}$$

Model Fitting by MCMC

Likelihood $$L(\theta) = \prod_{t=2}^{n} \frac{(\lambda_t + \delta_t \omega)^{z_t}}{z_t!} e^{-(\lambda_t + \delta_t \omega)}$$

Metropolis within Gibbs, using full conditionals:

$$\delta_t \,|\, \mathbf{z}, \theta_{-\delta_t} \sim Bern(A/(A+B))$$

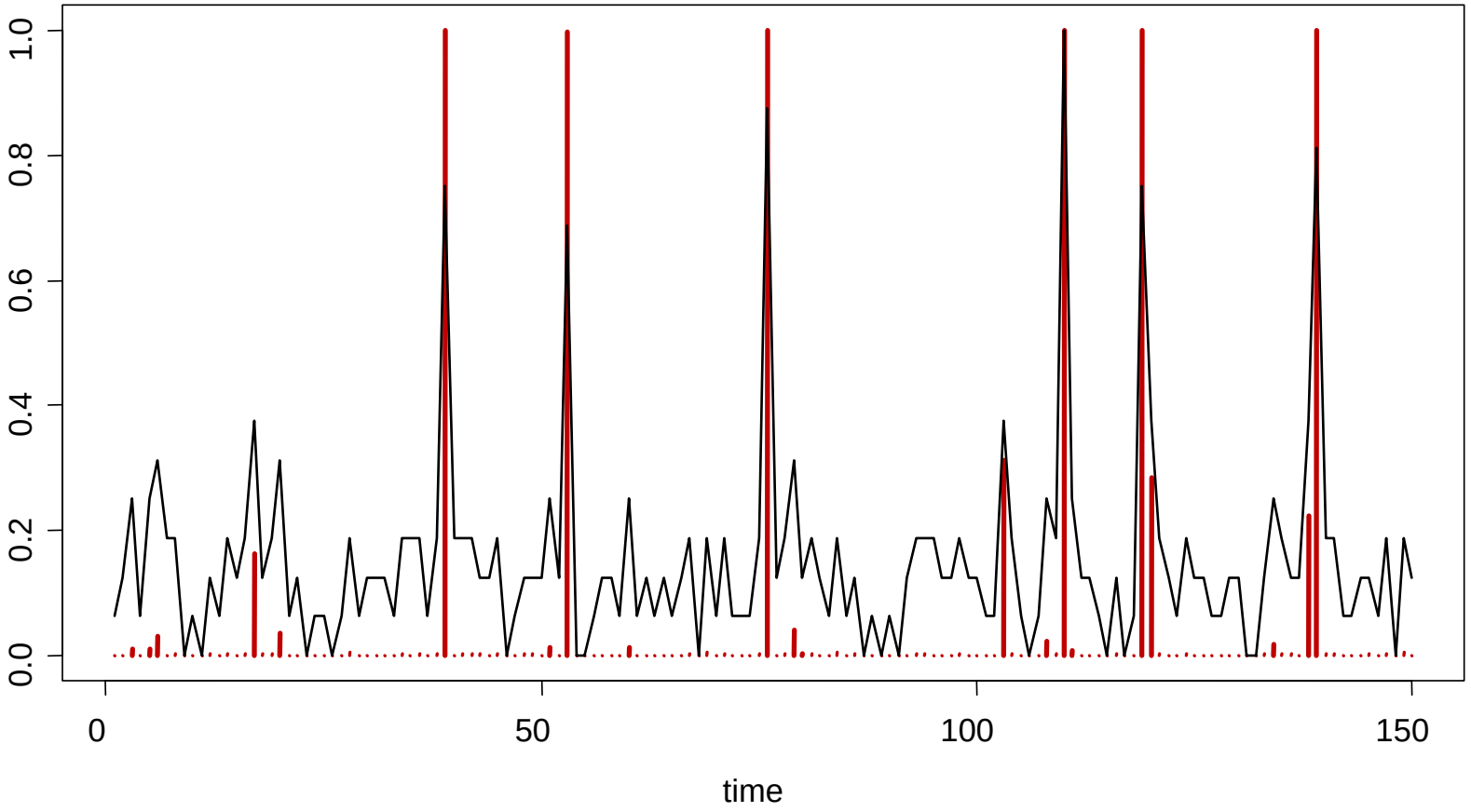
$$A = (\lambda_t + \omega)^{z_t} e^{-(\lambda_t + \omega)} \pi_t$$

$$B = \lambda_t^{z_t} e^{-\lambda_t} (1 - \pi_t)$$

$$\pi_t \,|\, \mathbf{z}, \theta_{-\pi_t} \sim Beta(a_{\pi_t} + \delta_t, b_{\pi_t} + 1 - \delta_t)$$

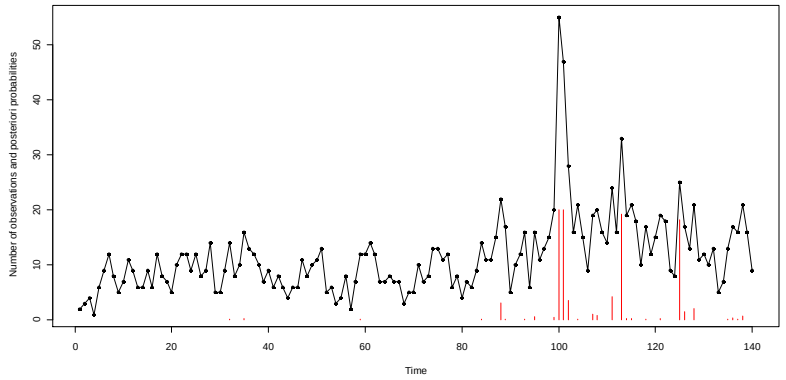
Simulated Data Example

INGARCH series with 6 additive outliers and
posterior probabilities of outlyingness

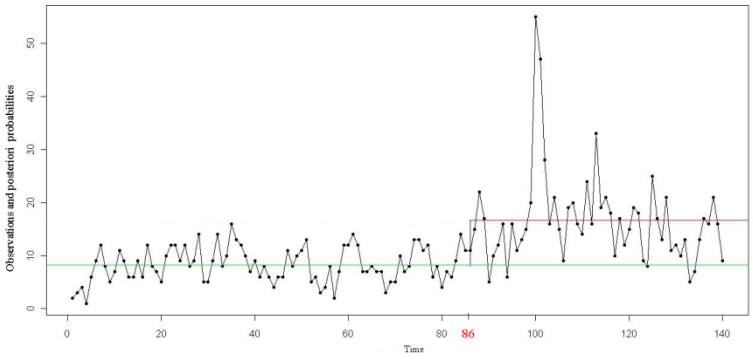


Analysis of the Campylobacterosis Data

Model with Additive Outliers



Model with Level Shift



$\hat{\beta}_0 = 1.71(0.72), \quad \hat{\alpha}_1 = 0.421(0.14)$

$\hat{\beta}_1 = 0.424(0.10), \; \hat{\omega} = 22.1(4.94)$

Outliers at 100, 101, 113, 125

Probability 1.0 1.0 .96 .91

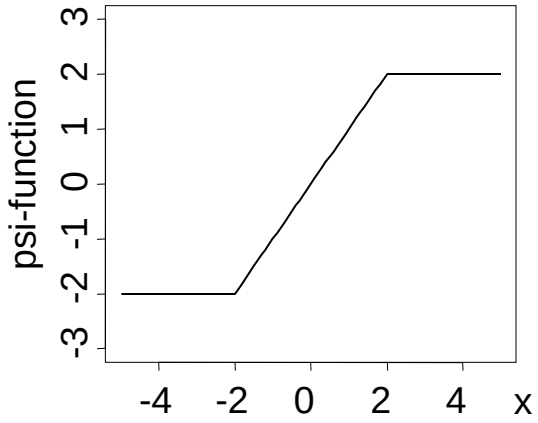
Shift at time 86

3) M-estimation of Location μ for i.i.d. Data

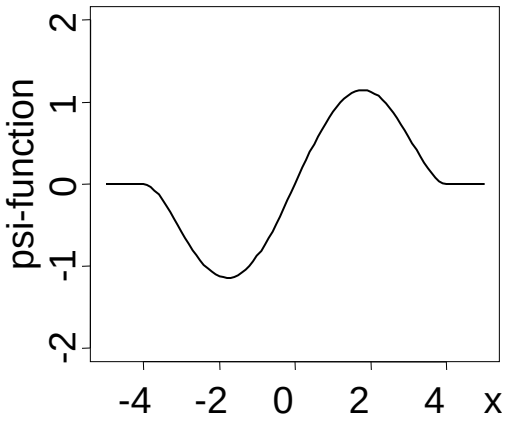
Minimize $\sum_{t=1}^{n} \rho\left(\frac{y_t - \mu}{\sigma}\right) \quad \Leftrightarrow \quad \sum_{t=1}^{n} \psi\left(\frac{y_t - \mu}{\sigma}\right) = 0$

e.g. $\rho = -\log f$ gives ML-estimator, $\psi = \rho'$

Huber ψ-function



Tukey ψ-function



$$\psi_k(x) = \begin{cases} x, & |x| < k \\ k \cdot \text{sign}(x), & |x| \geq k \end{cases}$$

$$\psi_k(x) = x\left(1 - \left(\frac{x}{k}\right)^2\right)^2 \cdot I(|x| < k)$$

Conditional likelihod estimation for INARCH

INARCH-model: $Y_t \mid (Y_s, s<t) \sim \text{Poi}(\mu_t), \quad \mu_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p}$

Conditioning on first p observations $y_1, \ldots, y_p$:

$$\sum_{t=p+1}^{n} \frac{y_t - \mu_t}{\sqrt{\mu_t}} \frac{1}{\sqrt{\mu_t}} \begin{pmatrix} 1 \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

M-estimation:

$$\psi\left(\frac{y_t - \mu_t}{\sqrt{\mu_t}}\right)$$

μ, σ^2 marginal mean & variance

$$\begin{pmatrix} 1 \\ \mu + \sigma\psi\left(\frac{y_{t-1} - \mu}{\sqrt{\sigma}}\right) \\ \vdots \\ \mu + \sigma\psi\left(\frac{y_{t-p} - \mu}{\sqrt{\sigma}}\right) \end{pmatrix}$$

Bias Correction for INARCH(p) Model

M-estimator with bias correction:

$$\sum_{t=p+1}^{n}\left(\psi\left(\frac{y_t-\mu_t}{\sqrt{\mu_t}}\right)\frac{1}{\sqrt{\mu_t}}\begin{pmatrix}1\\ \mu+\sigma\psi\left(\frac{y_{t-1}-\mu}{\sqrt{\mu}}\right)\\ \vdots\\ \mu+\sigma\psi\left(\frac{y_{t-p}-\mu}{\sqrt{\mu}}\right)\end{pmatrix}-\begin{pmatrix}a_0\\ \vdots\\ \vdots\\ a_p\end{pmatrix}\right)=\begin{pmatrix}0\\ \vdots\\ \vdots\\ 0\end{pmatrix}$$

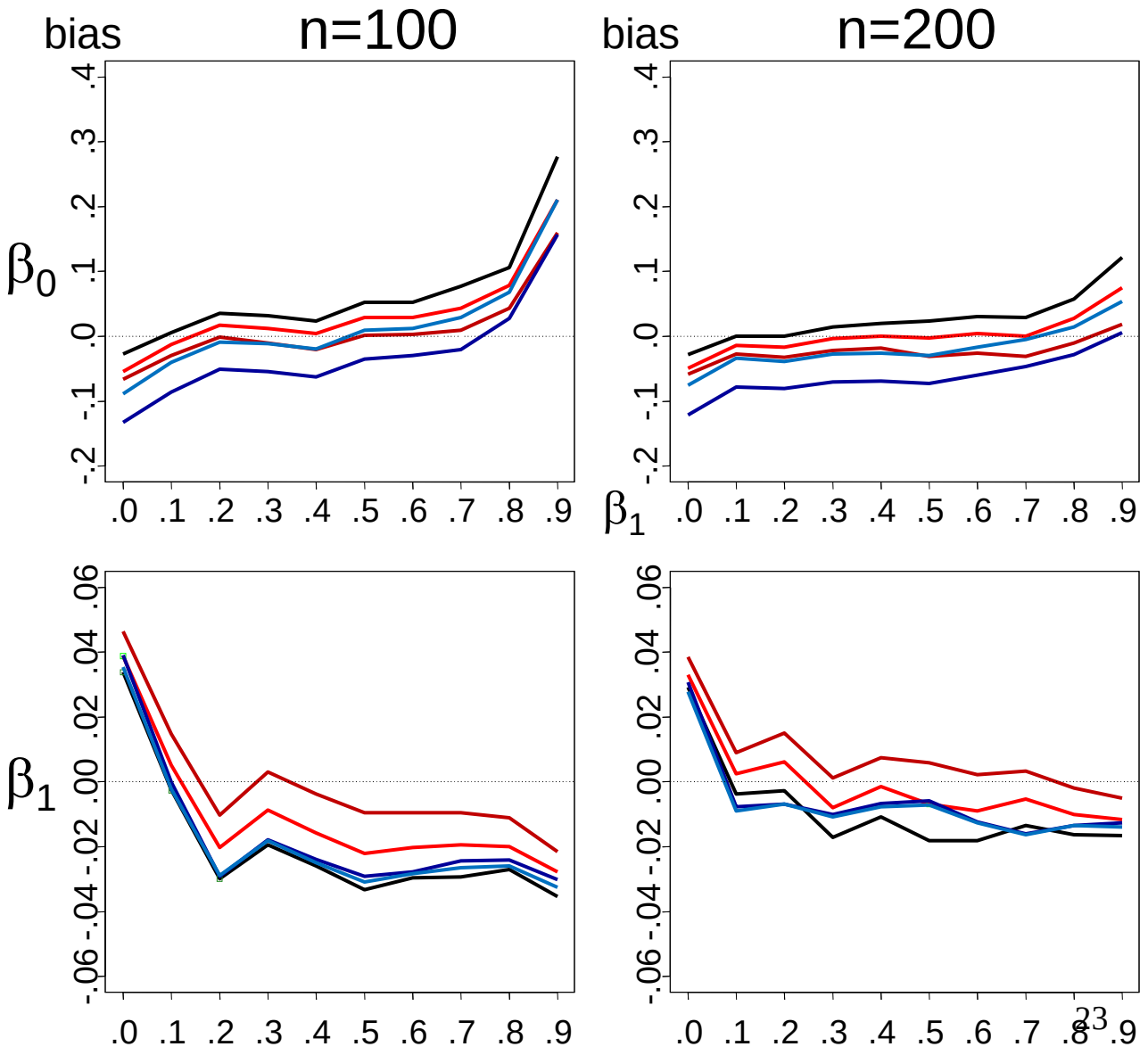
with $a_0,\ldots, a_p$ depending on $\beta_0, \ldots, \beta_p$ such that expectation of left hand side is 0.

Asymptotically normal *ElSaied (2012)*

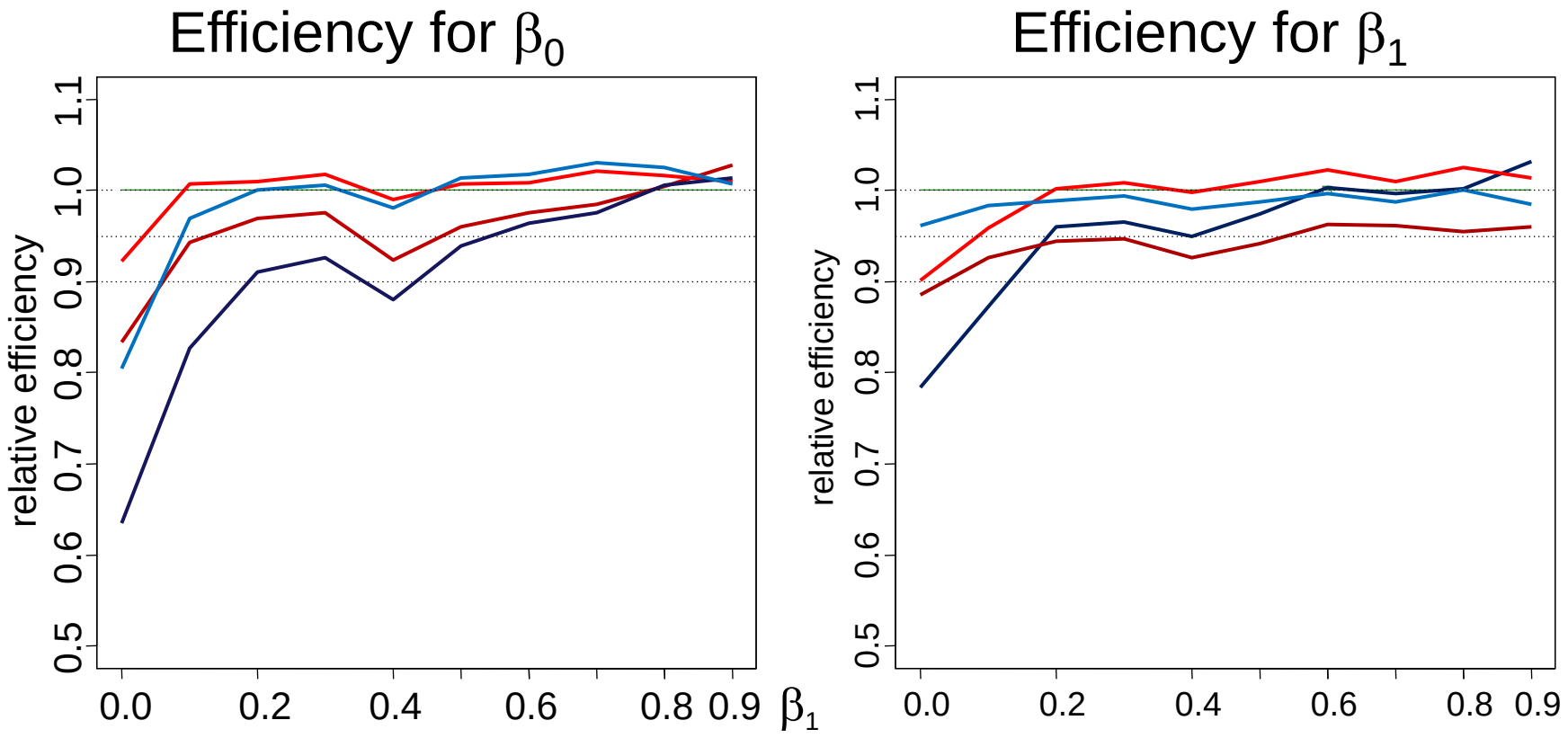
Bias for INARCH(1) as a function of β_1

Conditional ML
Tukey, k=7 (corrected)
Tukey, k=5 (corrected)

Bias correction better in large samples



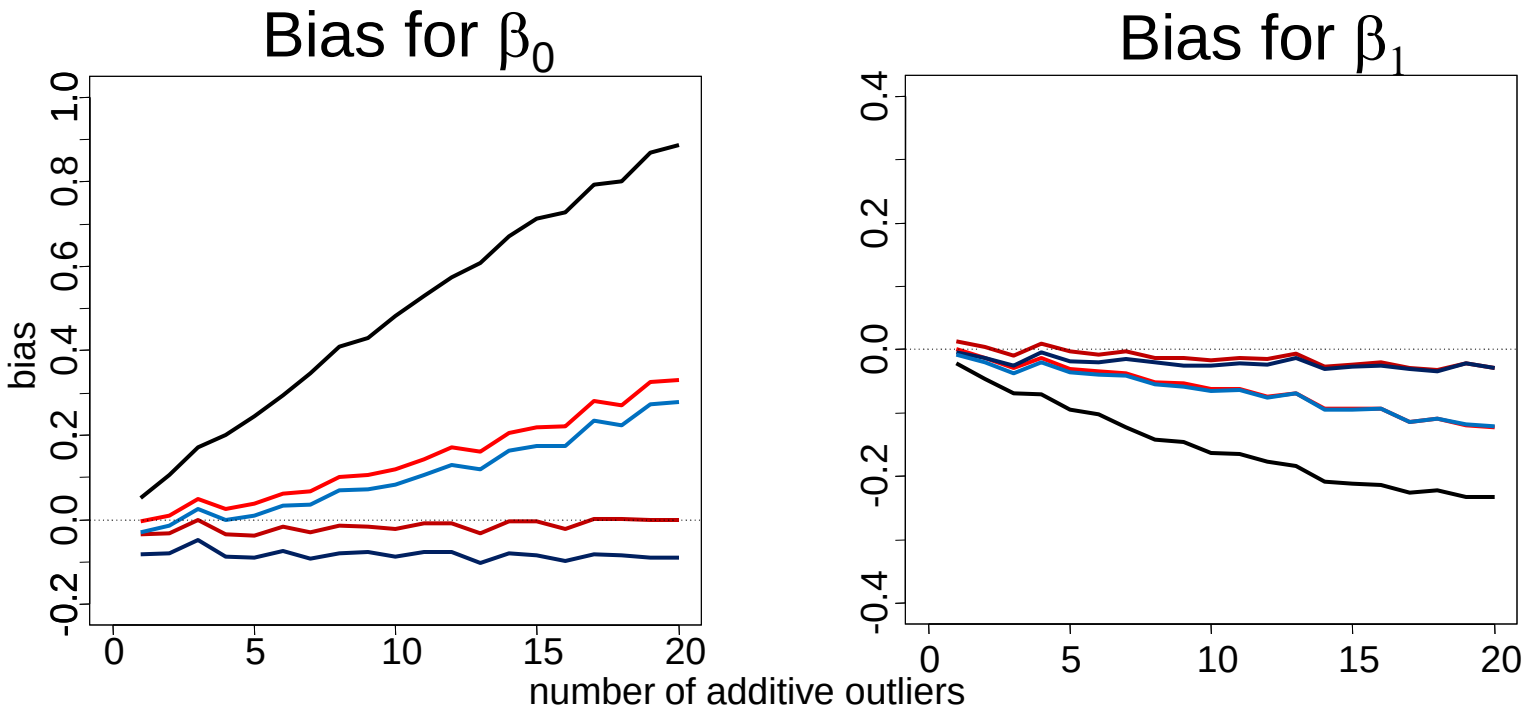
Efficiencies: INARCH(1), β_0=1,several β_1, n=100



Tukey, k=5 (corrected), Tukey, k=7(corrected)

Robustness: INARCH(1) with β_0=1, β_1=.4

Increasing number j of separated outliers



Conditional ML

Tukey, k=5 (corrected), Tukey, k=7 (corrected)

Conclusion

Modeling of external / internal intervention effects in INGARCH processes feasible via latent mean process

(Maximum) Score tests for simultaneous testing of all types of outliers at all time points, using parametric bootstrap

Misspecification not a problem, but correct classification

(Multiplicative) Log-linear model: same observations suspicious

Bayesian detection of additive outliers using MCMC

Robust generalized M-estimators

Surveillance: 1st approach using 1-step ahead predictions extend to h-step ahead prediction to distinguish spikes - shifts

References

Cadigan, N.G., Chen, J. (2001). Properties of Robust M-estimators for Poisson and Negative Binomial Data. *J. Statist. Comput. Simul.* 70, 273-288.

Davis, R.A., Dunsmuir, W.T.M., Street, S.B. (2003). Observation driven models for Poisson counts. *Biometrika* 90, 777-790.

ElSaied, H. (2012). Robust Modelling of Count Data. Applications in Medicine. PhD-thesis, Dept. of Statistics, TU Dortmund University, Germany.

Ferland, R.A., Latour, A., Oraichi, D. (2006). Integer-valued GARCH processes. *J. Time Series Analysis* 27, 923-942.

Fried, R., Agueusop, I., Bornkamp, B., Fokianos, K., Fruth, J., Ickstadt, K. (2012). Bayesian Modelling of Additive Outliers in INGARCH time series, submitted.

Fokianos, K., Fried, R. (2010). Interventions in INGARCH Processes. *J. Time Series Analysis 31, 210-225*.

Fokianos, K., Fried, R. (2012). Interventions in log-linear Poisson Autoregression. *Statistical Modelling 12, 299-322.*

Fokianos, K., Rahbek, A., Tjøstheim, D. (2009). Poisson Autoregression. *J. American Statistical Association 104, 1430-1439.*

Fokianos, K., Tjøstheim, D. (2011). Loglinear Poisson Autoregression. *J. Multivariate Analysis* 102, 563-578.

Simpson, D.G., Carroll, R.J., Ruppert, D. (1987). M-Estimation for Discrete Data: Asymptotic Distribution Theory & Implications. *Annals of Statistics 15, 657-669*.

Interventions in a loglinear model

Clean process: $Y_t \mid (Y_s, s<t) \sim \text{Poi}(\lambda_t)$

$$\log(\lambda_t) = \beta_0 + \sum_{i=1}^{p} \alpha_i \log(\lambda_{t-i}) + \sum_{j=1}^{q} \beta_j \log(Y_{t-j}+1)$$

Ergodicity for p=q=1 *Fokianos, Rahbek & Tjøstheim (2009)*

Process contaminated by outlier of size ν at time τ:

$$Z_t \mid (Z_s, s<t) \sim \text{Poi}(\kappa_t)$$

$$\log(\kappa_t) = \beta_0 + \sum_{i=1}^{p} \alpha_i \log(\kappa_{t-i}) + \sum_{j=1}^{q} \beta_j \log(Z_{t-j}+1) + \nu\delta^{t-\tau} I(t \geq \tau)$$

Equivalently $Z_t = Y_t + C_t, \quad C_t \mid (C_s, s<t) \sim \text{Poi}(\lambda_t(\exp(\mu_t)-1))$

for $\nu>0$ and $t \geq \tau$: $\mu_t = \sum_{i=1}^{p} \alpha_i \mu_{t-i} + \sum_{j=1}^{q} \beta_j \log\left(1 + \frac{C_{t-j}}{Y_{t-j}+1}\right) + \nu\delta^{t-\tau}$

Outliers in the INGARCH(p,q)-Model

Clean process: $Y_t \mid (Y_s, s<t) \sim \text{Poi}(\lambda_t)$

$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \alpha_i \lambda_{t-i} + \sum_{j=1}^{q} \beta_j Y_{t-j}$$

Ergodicity for INGARCH(1,1): *Fokianos, Rahbek & Tjøstheim (2009)*

Process contaminated by outlier of size ν at time τ:

$$Z_t \mid (Z_s, s<t) \sim \text{Poi}(\kappa_t)$$

$$\kappa_t = \beta_0 + \sum_{i=1}^{p} \alpha_i \kappa_{t-i} + \sum_{j=1}^{q} \beta_j Z_{t-j} + \nu \delta^{t-\tau} I(t \geq \tau)$$

Equivalently $Z_t = Y_t + C_t, \quad C_t \sim \text{Poi}(\kappa_t - \lambda_t)$

for $\nu>0$ and $t \geq \tau$: $\kappa_t - \lambda_t = \sum_{i=1}^{p} \alpha_i (\kappa_{t-i} - \lambda_{t-i}) + \sum_{j=1}^{q} \beta_j C_{t-j} + \nu \delta^{t-\tau}$