Diagnosing Overdispersion in Count Data Time Series

Christian H. Weiß
Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

Sebastian Schweer
Institute of Applied Mathematics,
University of Heidelberg

This talk is based on the articles

Schweer, S., Weiß, C.H. (2014).

Compound Poisson INAR(1) processes:

stochastic properties and testing for overdispersion.

Comp. Stat. Data Anal. **77**, 267–284.

Weiß, C.H., Schweer, S. (2014).

Detecting overdispersion in INARCH(1) processes.

Work in progress.

Further details and references are provided by this article.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

INAR(1) Model for Time Series of Counts

Motivation & Properties

Popular for **real-valued** stationary processes:

ARMA(p,q) model. Let $(\epsilon_t)_{\mathbb{Z}}$ white noise, then

$$X_t \;=\; \alpha_1 \cdot X_{t-1} + \ldots + \alpha_\mathrm{p} \cdot X_{t-\mathrm{p}} + \epsilon_t + \beta_1 \cdot \epsilon_{t-1} + \ldots + \beta_\mathrm{q} \cdot \epsilon_{t-\mathrm{q}},$$

where $\alpha_1, \ldots, \alpha_\mathrm{p}, \beta_1, \ldots, \beta_\mathrm{q} \in \mathbb{R}$ suitably chosen.

Autocorrelation via **Yule-Walker equations**:

$$\rho_X(k) \;=\; \sum_{j=1}^{\mathrm{p}} \; \alpha_j \cdot \rho_X(|k-j|) \;+\; \frac{\sigma_\epsilon^2}{\sigma_X^2} \cdot \sum_{i=0}^{\mathrm{q}-k} \; \beta_{i+k} \cdot a_i \;+\; \delta_{k0} \cdot \frac{\sigma_\epsilon^2}{\sigma_X^2}.$$

Example: AR(1) model $X_t \;=\; \alpha \cdot X_{t-1} + \epsilon_t$ with $\rho_X(k) = \alpha^k$.

Not applicable to count data processes: generally, $\alpha \cdot X \notin \mathbb{N}_0$.

Several approaches in literature of

how to avoid the "multiplication problem".

In first part of this talk, we consider models based on

binomial thinning operator (Steutel & van Harn, 1979):

$$\alpha \circ X := \sum_{i=1}^{X} Y_i, \qquad \text{where } Y_i \text{ are i.i.d. Bin}(1, \alpha),$$

i. e., $\alpha \circ X \sim \text{Bin}(X, \alpha)$ and has range $\{0, \ldots, X\}$.

($\approx$ number of "survivors" from population of size X)

INAR(1) Processes

Let $(\epsilon_t)_{\mathbb{Z}}$ be i.i.d. with range $\mathbb{N}_0 = \{0, 1, \ldots\}$, denote $\mathbb{E}[\epsilon_t] = \mu_\epsilon$, $\mathrm{Var}[\epsilon_t] = \sigma_\epsilon^2$. Let $\alpha \in (0;1)$.

$(X_t)_{\mathbb{Z}}$ referred to as **INAR(1) process** if

$$X_t \;=\; \alpha \circ X_{t-1} \;+\; \epsilon_t,$$

plus appropr. independence assumptions. (McKenzie, 1985)

Properties: (also see lecture)

Homogeneous Markov chain with

$$\mathbb{P}(X_t = k \mid X_{t-1} = l) \;=\; \sum_{j=0}^{\min\{k,l\}} \binom{l}{j} \alpha^j (1-\alpha)^{l-j} \cdot \mathbb{P}(\epsilon_t = k - j).$$

INAR(1) Processes

If INAR(1) process stationary (see below), then

$$\text{pgf}_X(z) \;=\; \text{pgf}_X(1-\alpha+\alpha z)\cdot \text{pgf}_\epsilon(z).$$

In particular, if $\mu_\epsilon, \sigma_\epsilon < \infty$, we have

$$\mu_X \;=\; \frac{\mu_\epsilon}{1-\alpha}, \qquad \sigma_X^2 \;=\; \mu_X \cdot \frac{\frac{\sigma_\epsilon^2}{\mu_\epsilon}+\alpha}{1+\alpha}.$$

(Note: X_t equidispersed iff ϵ_t equidispersed.)

Autocorrelation function: $\rho_X(k) = \alpha^k$, i. e., AR(1)-type.

For further properties and references, see Weiß (2008).

Poisson INAR(1) Processes

Most popular instance of INAR(1) family:

Poisson INAR(1) model, $X_t \;=\; \alpha \circ X_{t-1} \;+\; \epsilon_t.$

Here, innovations $(\epsilon_t)_{\mathbb{Z}}$ i.i.d. Poi(λ), such that $\mu_\epsilon = \sigma_\epsilon^2 = \lambda$.

Stationary marginal distribution:

also Poisson distribution, Poi($\frac{\lambda}{1-\alpha}$), because:

additivity of Poisson distribution, and

Poisson distribution **invariant to binomial thinning**:

If $X \sim$ Poi(μ), then $\alpha \circ X \sim$ Poi($\alpha \cdot \mu$).

Poisson INAR(1) Processes

Poisson INAR(1) has **equidispersed** marginals.

Modifications to get **overdispersed** marginals:

- modify thinning operation (Weiß, 2008), or
- modify distribution of innovations, e. g.,
  as in Jung et al. (2005), Pedeli & Karlis (2011):
  negative binomially (NB) distributed ϵ_t's.

In the following, we more generally consider
compound Poisson (CP) distributed innovations.

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Compound Poisson INAR(1) Processes

Definition & Properties

Compound Poisson Distribution

$Y_1, Y_2, \ldots$ i.i.d. with range $\mathbb{N} = \{1, 2, \ldots\}$,
denote pgf as $H(z)$ (**compounding distribution**).

Let $N \sim \text{Poi}(\lambda)$, independent of $Y_1, Y_2, \ldots$

$\epsilon := Y_1 + \ldots + Y_N$ **compound Poisson distributed**:
$\epsilon \sim \text{CP}(\lambda, H)$.

Then $\text{pgf}_\epsilon(z) = \exp\big(\lambda(H(z) - 1)\big)$. (Feller, 1968)

More precisely:
$\epsilon \sim \text{CP}_\nu(\lambda, H)$ if $H(z) = h_1 z + \ldots + h_\nu z^\nu$ with $h_\nu > 0$,
$\epsilon \sim \text{CP}_\infty(\lambda, H)$ if Y_i have infinite range.

Some examples:

- $\mathsf{CP}_1(\lambda, H) = \mathsf{Poi}(\lambda)$.
- CP_ν with $\nu < \infty$ and $h_x = 1/\nu$ for $x = 1, \ldots, \nu$:

  Poisson distribution of order ν, abbr. $\mathsf{Poi}_\nu(\lambda)$.
- $\mathsf{NB}(n, \pi)$-distribution: CP_∞ with $\lambda := -n \ln \pi$ and

$$H(z) = \frac{\ln\big(1-(1-\pi)z\big)}{\ln \pi} = \sum_{k=1}^{\infty} \frac{(1-\pi)^k}{-k \ln \pi}\, z^k.$$

Important properties: (also see lecture)

- $\kappa_{\epsilon,r} \;=\; \lambda \cdot \mu_{Y,r}$ (Aki, 1985)

  $\Rightarrow$

  ϵ equidispersed iff $\nu = 1$,

  ϵ overdispersed iff $\nu > 1$.

- If X_1, X_2 independent with $X_i \sim \mathrm{CP}_\nu(\lambda_i, H_i)$ (common ν), then $X_1 + X_2$ $\mathrm{CP}_\nu(\lambda, H)$-distributed.

- If $X \sim \mathrm{CP}_\nu(\lambda, H)$, then $\alpha \circ X \sim \mathrm{CP}_\nu(\mu, G)$.

INAR(1) process $(X_t)_{t\in\mathbb{Z}}$ referred to as

CPINAR(1) process

if innovations $(\epsilon_t)_{\mathbb{Z}}$ i.i.d. $\text{CP}_\nu(\lambda, H)$ (possibly $\nu = \infty$).

$\nu = 1$: Poisson INAR(1) model.

NB-innovations: Jung et al. (2005), Pedeli & Karlis (2011).

We showed (also see lecture):

If $H'(1) < \infty$, then unique **stationary** marginal distribution of $(X_t)_{t\in\mathbb{Z}}$ is $\text{CP}_\nu(\mu, G)$, where

$$\mu\big(G(z) - 1\big) \;=\; \lambda \textstyle\sum_{i=0}^{\infty} \big(H(1-\alpha^i+\alpha^i z) - 1\big).$$

Preliminary summary:

Stationary CPINAR(1) process with $\text{CP}_\nu(\lambda, H)$-innovations

has $\text{CP}_\nu(\mu, G)$-observations (same ν), i. e.,

$(X_t)_{t\in\mathbb{Z}}$ is **equi**dispersed **iff** $\nu = 1$,

$(X_t)_{t\in\mathbb{Z}}$ is **over**dispersed **iff** $\nu > 1$.

Aim: Develop test to distinguish between

- null hypothesis of equidispersion ($\nu = 1$), and
- alternative hypothesis of overdispersion ($\nu > 1$).

Index of dispersion: $I_X := \dfrac{\sigma_X^2}{\mu_X}, \qquad \hat{I}_X := \dfrac{S_X^2}{\bar{X}}.$

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Compound Poisson INAR(1) Processes

Testing for Overdispersion

Testing for Overdispersion

Index of dispersion: $I_X := \frac{\sigma_X^2}{\mu_X}, \qquad \hat{I}_X := \frac{S_X^2}{\bar{X}}.$

Theorem:

For CPINAR(1) process $(X_t)_{\mathbb{Z}}$ with $H'(1) < \infty$,

$$\sqrt{T} \cdot (\hat{I}_X - I_X) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0,\ \sigma^2\right) \quad \text{as} \quad T \longrightarrow \infty,$$

$$\sigma^2 = \frac{1+\alpha}{1-\alpha}\,(\mu_X - \sigma_X^2)\left(\frac{\bar{\mu}_{X,3}}{\mu_X^3} - \frac{\sigma_X^4}{\mu_X^4}\right)$$

$$+ \frac{1+\alpha^2}{1-\alpha^2}\left(\frac{\bar{\mu}_{X,4}}{\mu_X^2} - \frac{\bar{\mu}_{X,3}}{\mu_X^3}\,(\mu_X + \sigma_X^2) + \frac{\sigma_X^4}{\mu_X^3}\,(1 - \mu_X)\right).$$

Design of **test for overdispersion**:

H_0: $(X_t)_{t\in\mathbb{Z}}$ follows *Poisson* INAR(1) model with parameters (λ, α).

In this case, we have

$$\sqrt{T}\cdot(\hat{I}_X - 1) \xrightarrow{\mathcal{D}} \mathrm{N}\left(0,\ 2\,\frac{1+\alpha^2}{1-\alpha^2}\right) \quad \text{as} \quad T \longrightarrow \infty.$$

$\Rightarrow$ critical value

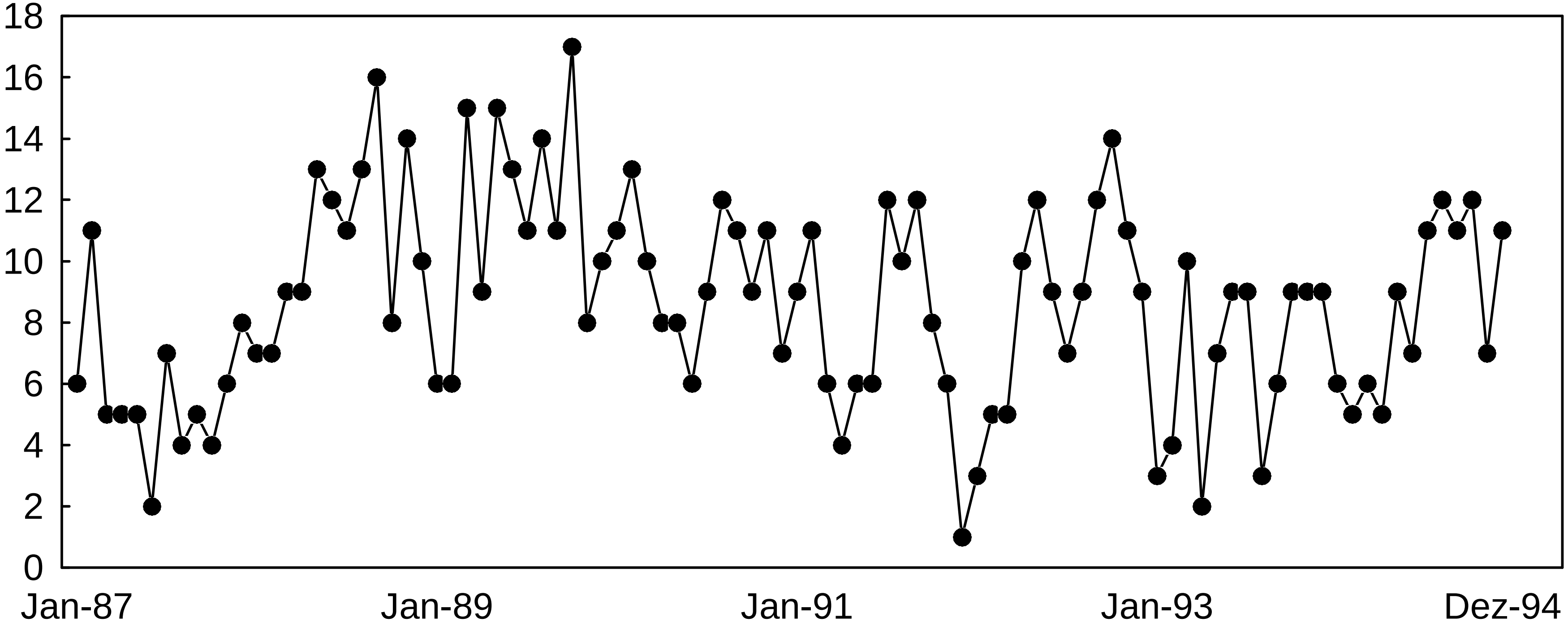
$$1 + z_{1-\psi}\cdot\sqrt{\frac{2}{T}\,\frac{1+\alpha^2}{1-\alpha^2}},$$

where $z_{1-\psi}$: $(1-\psi)$-quantile of N$(0,1)$-distribution.

Example: Monthly claims counts (1987 to 1994):

burn related injuries in heavy manufacturing industry.

Source: Freeland (1998).

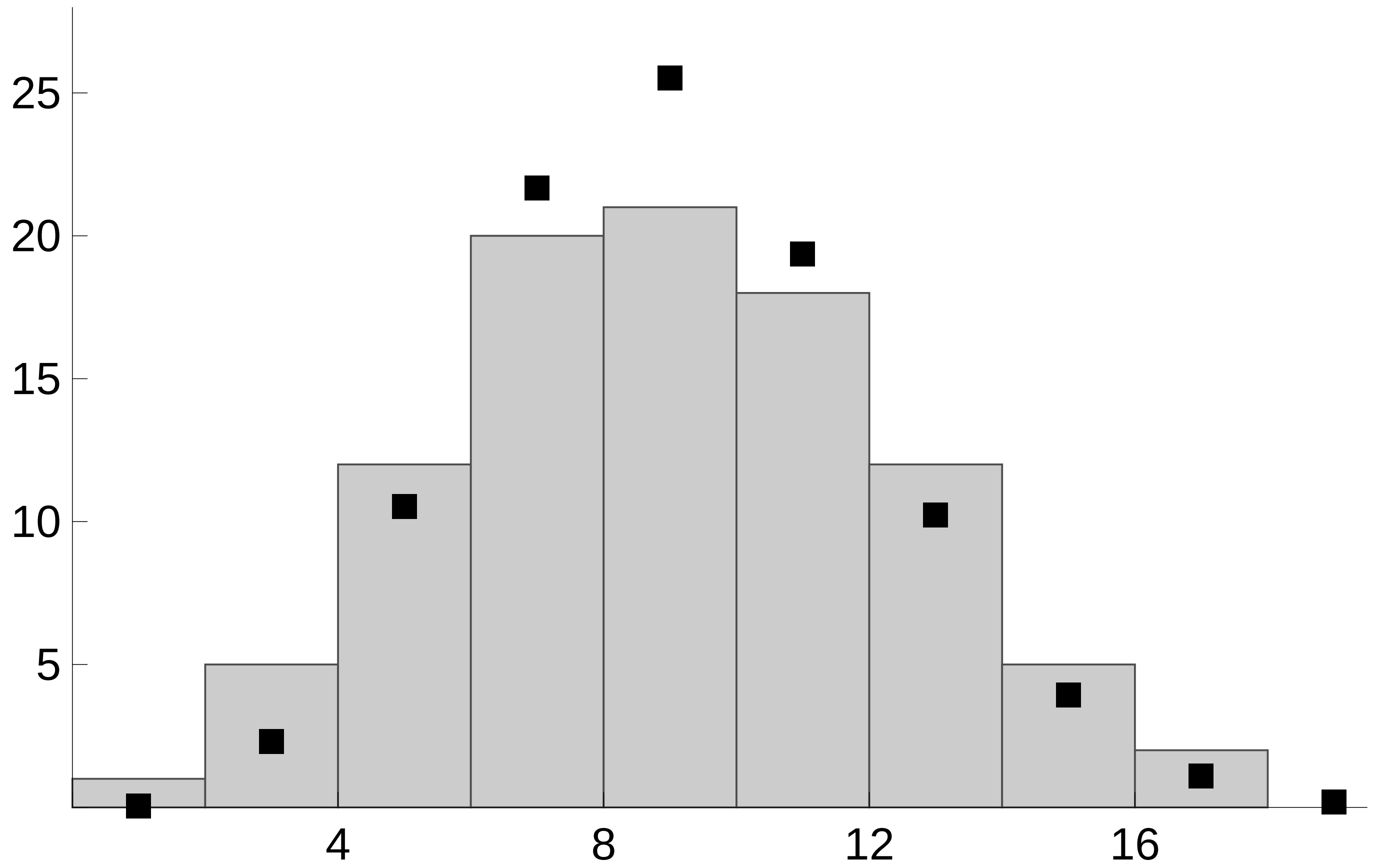


Example (continued): Freeland's claims counts data.

Length $T = 96$, $\bar{x} \approx 8.604$, $s_x^2 \approx 11.24$
$\Rightarrow \hat{I}_y \approx 1.306$, i. e., about 31 % of empirical overdispersion.

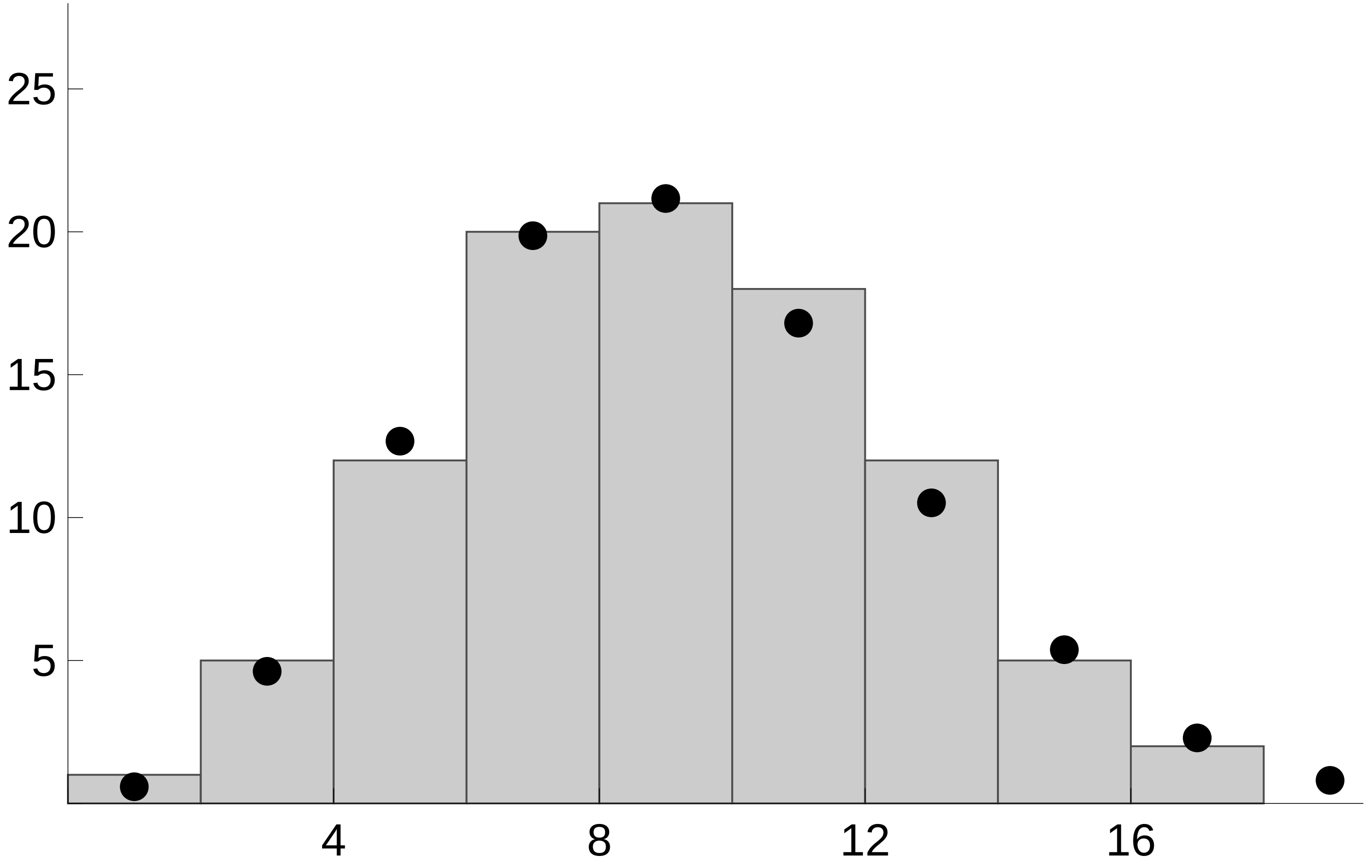
However, plugging-in $\hat{\rho}(1) \approx 0.452$ instead of α,
critical value 1.292 (significance level $\psi = 0.05$),
i. e., quite narrow decision.

Example (continued): Freeland's claims counts data.



Histogram with fitted Poi-marginal distribution.

Example (continued): Freeland's claims counts data.



Histogram with fitted Poi_2-marginal distribution.

Our above result

$$\sqrt{T}\cdot(\hat{I}_X - I_X) \xrightarrow{\mathcal{D}} \mathsf{N}\left(0,\ \sigma^2\right) \quad \text{as} \quad T \longrightarrow \infty,$$

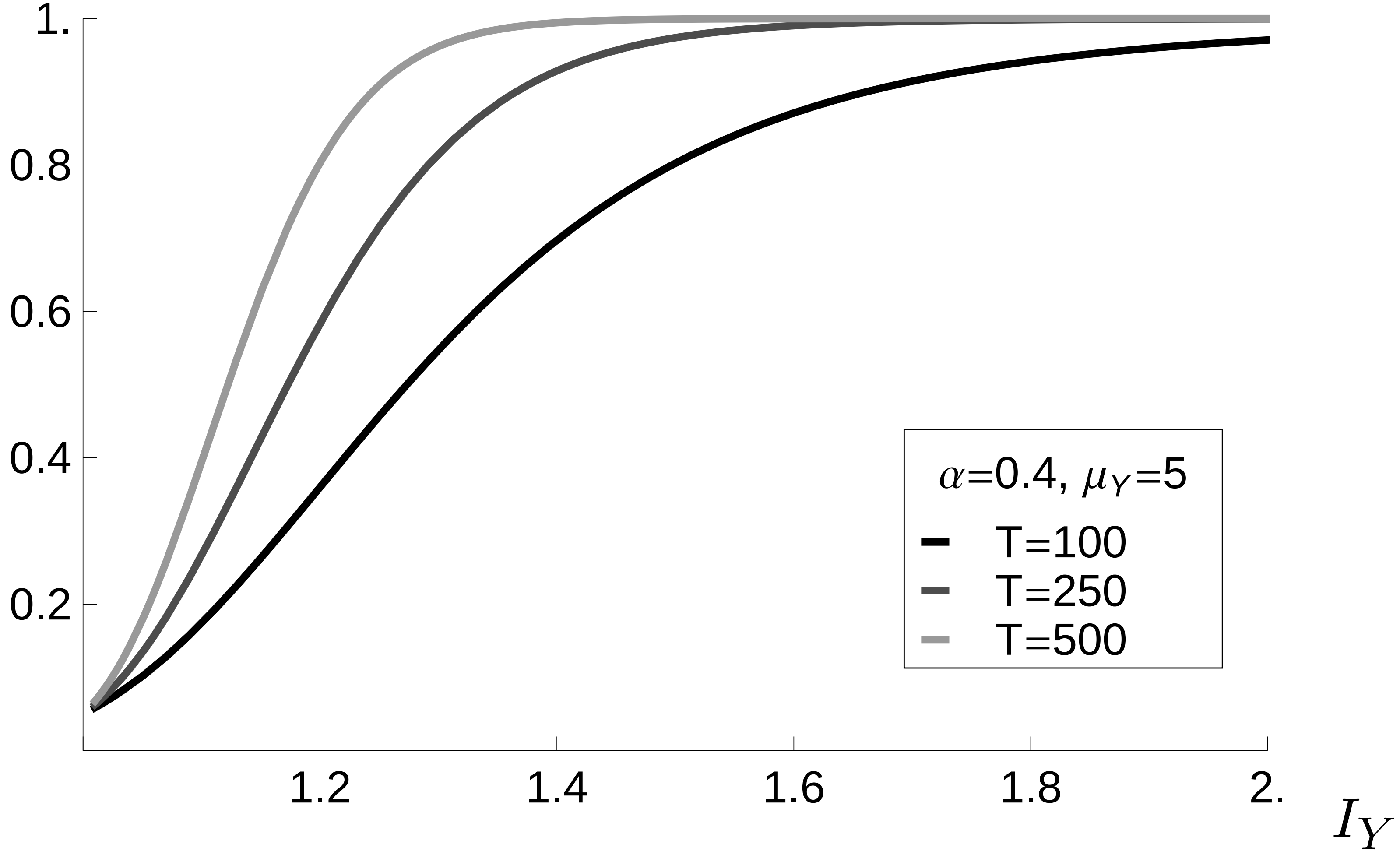
$$\sigma^2 = \frac{1+\alpha}{1-\alpha}\,(\mu_X - \sigma_X^2)\left(\frac{\bar{\mu}_{X,3}}{\mu_X^3} - \frac{\sigma_X^4}{\mu_X^4}\right) + \frac{1+\alpha^2}{1-\alpha^2}\left(\frac{\bar{\mu}_{X,4}}{\mu_X^2} - \frac{\bar{\mu}_{X,3}}{\mu_X^3}\,(\mu_X + \sigma_X^2) + \frac{\sigma_X^4}{\mu_X^3}\,(1-\mu_X)\right),$$

also applicable to overdispersed alternative!

$\Rightarrow$ **Power analysis**, e. g., w.r.t. NB(n,π)-innovations:

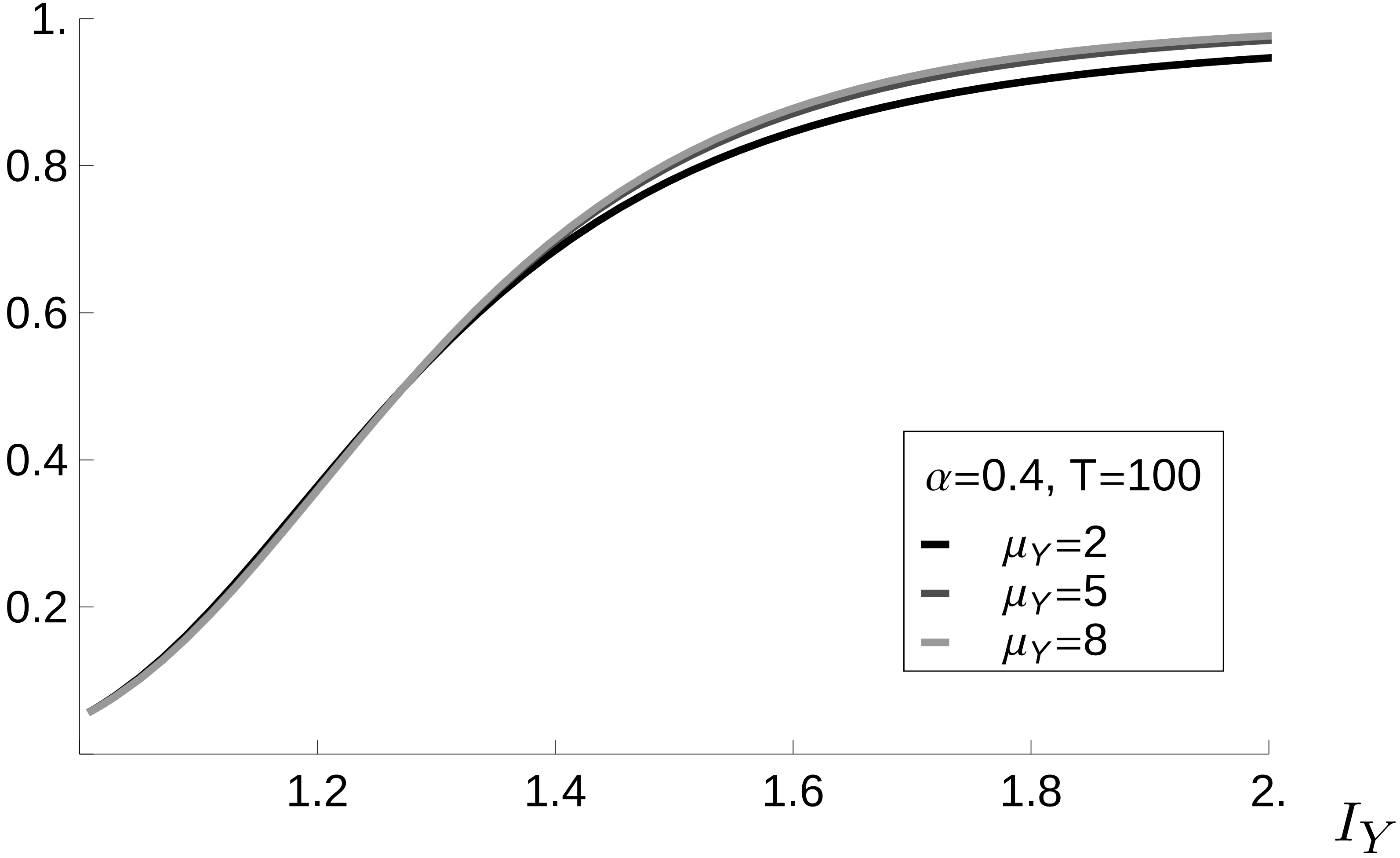
n controls mean, π overdispersion, α autocorrelation.

Power analysis w.r.t. NB(n, π)-innovations:

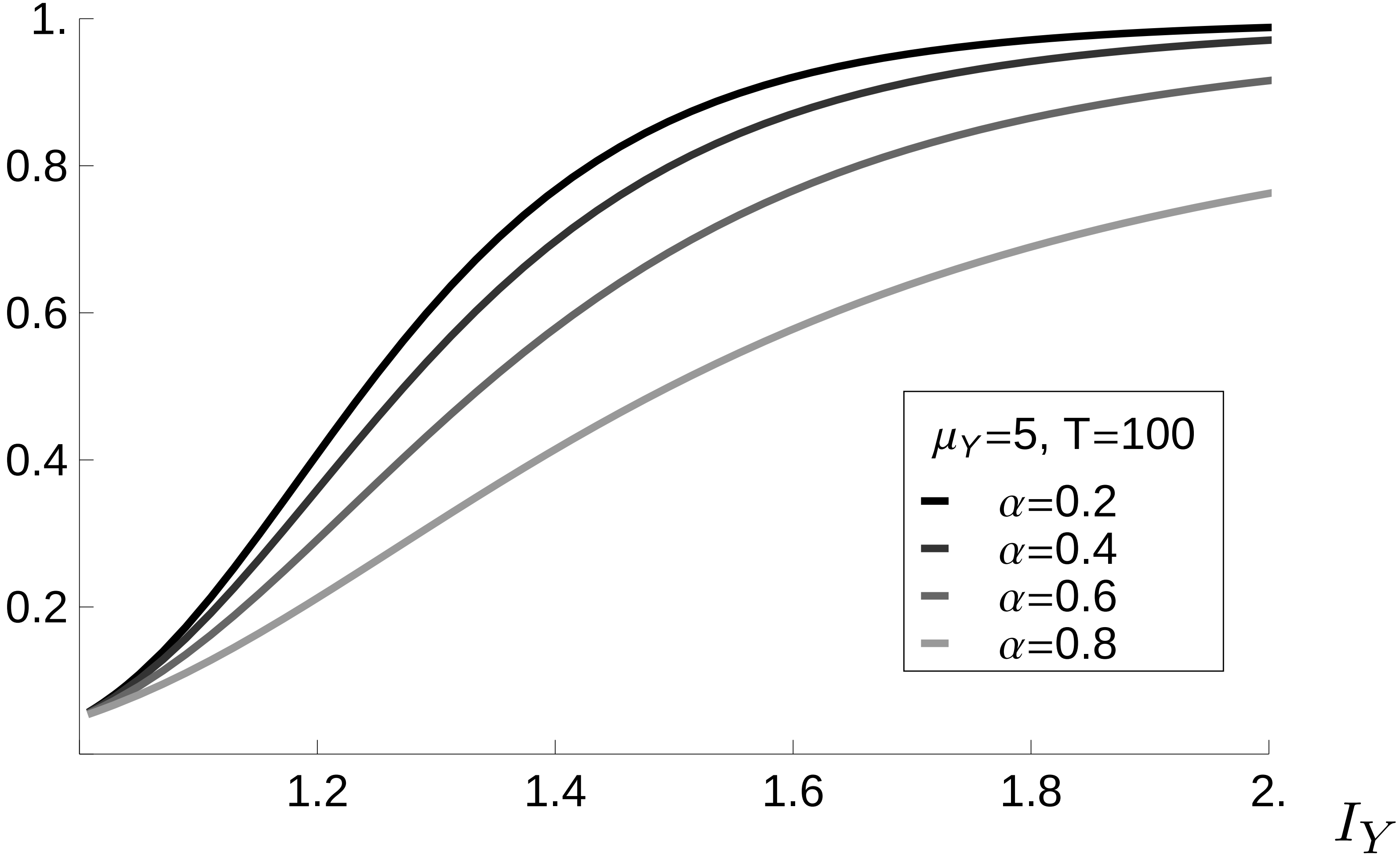


MATH STAT

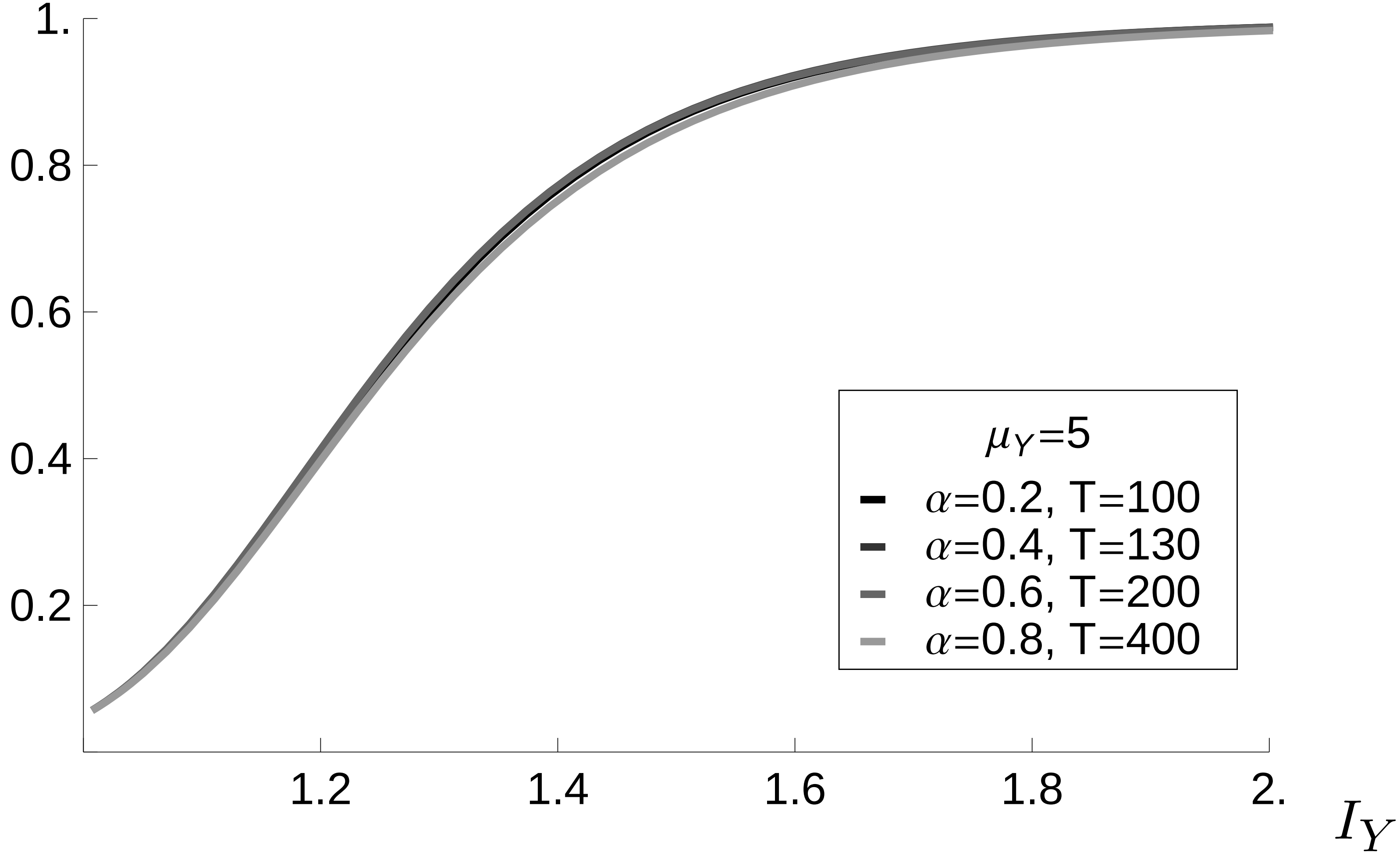
Power analysis w.r.t. NB(n, π)-innovations:



Power analysis w.r.t. NB(n, π)-innovations:



Power analysis w.r.t. NB(n, π)-innovations:



Finite-sample properties:

Simulated Poi-INAR(1) with $(\alpha, \lambda) = (0.40, 5.1)$,

i. e., $\mu_Y = 8.5$ and $I_Y = 1$,

and significance level $\psi = 0.05$ (100,000 repl.):

empir. and asympt. properties of $\hat{I}_Y$; empir. rejection rates.

T	mean	s.d.	s.d.$_{\text{a}}$	skew.	$\hat{q}_{0.05}$	$q_{0.05,\text{a}}$	$\hat{q}_{0.95}$	$q_{0.95,\text{a}}$	r.r.$_{\alpha}$	r.r.$_{\text{a}}$
100	0.977	0.162	0.166	0.421	0.731	0.727	1.262	1.273	0.044	0.050
250	0.991	0.104	0.105	0.265	0.828	0.827	1.169	1.173	0.047	0.050
500	0.995	0.074	0.074	0.191	0.878	0.878	1.120	1.122	0.047	0.050
1000	0.998	0.052	0.053	0.143	0.914	0.914	1.086	1.086	0.049	0.050

Generally rather good approximation,

but visible negative bias for small T.

"Work in Progress"

Based on second-order Taylor expansion,

following **bias-corrected** approximation for mean of $\hat{I}$

in case of Poisson INAR(1) model:

$$E[\hat{I}] \approx 1 - \frac{1}{T}\frac{1+\alpha}{1-\alpha}.$$

Successfully corrects the negative bias for large α and small T.

$\Rightarrow$ **Improved critical value**

$$1-\frac{1}{T}\frac{1+\alpha}{1-\alpha} + z_{1-\psi}\cdot\sqrt{\frac{2}{T}\frac{1+\alpha^2}{1-\alpha^2}},$$

where $z_{1-\psi}$: $(1-\psi)$-quantile of N$(0,1)$-distribution.

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

MATH
STAT

INARCH(1) Model for Time Series of Counts

Motivation & Properties

INARCH(1) Processes

Another popular model for
AR(1)-like counts with overdispersion:

two-parametric (Poisson) **INARCH(1) model**,

part of INGARCH family (Heinen, 2003; Ferland et al., 2006),
where "multiplication problem" solved by
applying ARMA recursion to conditional means.

INARCH(1) model:

$$X_t \;\Big|\; X_{t-1}, X_{t-2}, \ldots \quad \sim \ \text{Poi}(\beta + \alpha \cdot X_{t-1}),$$

where $\beta > 0$ and $0 < \alpha < 1$.

Properties: (also see lecture)

Homogeneous Markov chain with

$$\mathbb{P}(X_t = k \mid X_{t-1} = l) \;=\; \exp\left(-\beta - \alpha \cdot l\right) \cdot \frac{(\beta + \alpha \cdot l)^k}{k!}.$$

If INARCH(1) process stationary, then

$$\mu \;=\; \frac{\beta}{1-\alpha} \qquad \text{and} \qquad \sigma^2 \;=\; \frac{\beta}{(1-\alpha)(1-\alpha^2)} \;>\mu.$$

Autocorrelation function: $\rho(k) = \alpha^k$, i. e., AR(1)-type.

(Degree of overdispersion depends on autocorrelation level.)

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Index of Dispersion for INARCH(1) Processes

"Work in Progress"

Index of Dispersion for INARCH(1) Processes

Theorem:

For stationary INARCH(1) process $(X_t)_{\mathbb{Z}}$,

empirical index of dispersion satisfies

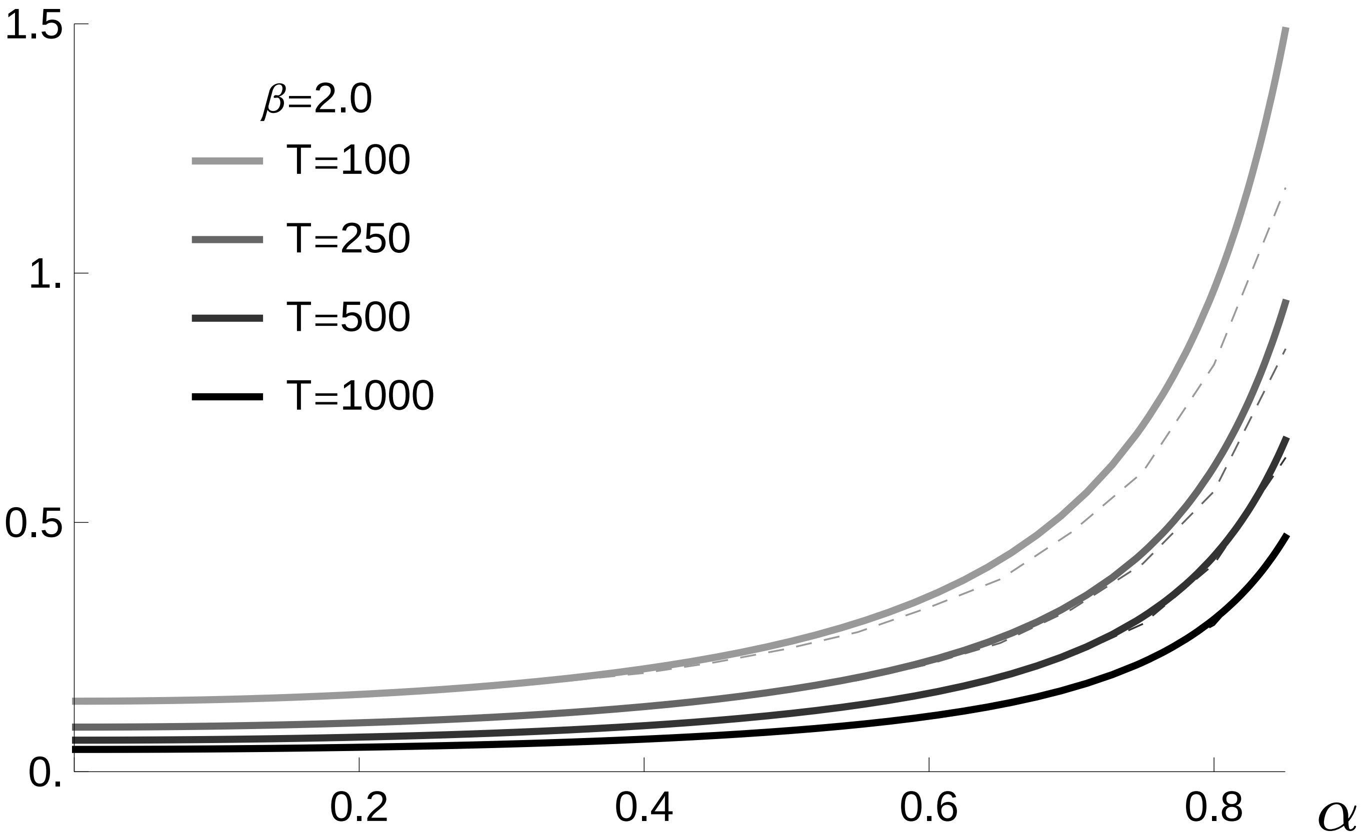
$$\sqrt{T}\left(\hat{I} - \frac{1}{1-\alpha^2}\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0,\ \sigma_{\mathrm{I}}^2),$$

where

$$\sigma_{\mathrm{I}}^2 = \frac{1}{\mu} \cdot \frac{2\alpha^2(1+2\alpha+3\alpha^3)}{(1-\alpha^2)^3(1-\alpha^3)} + \frac{2(1+\alpha^2)}{(1-\alpha^2)^3}.$$

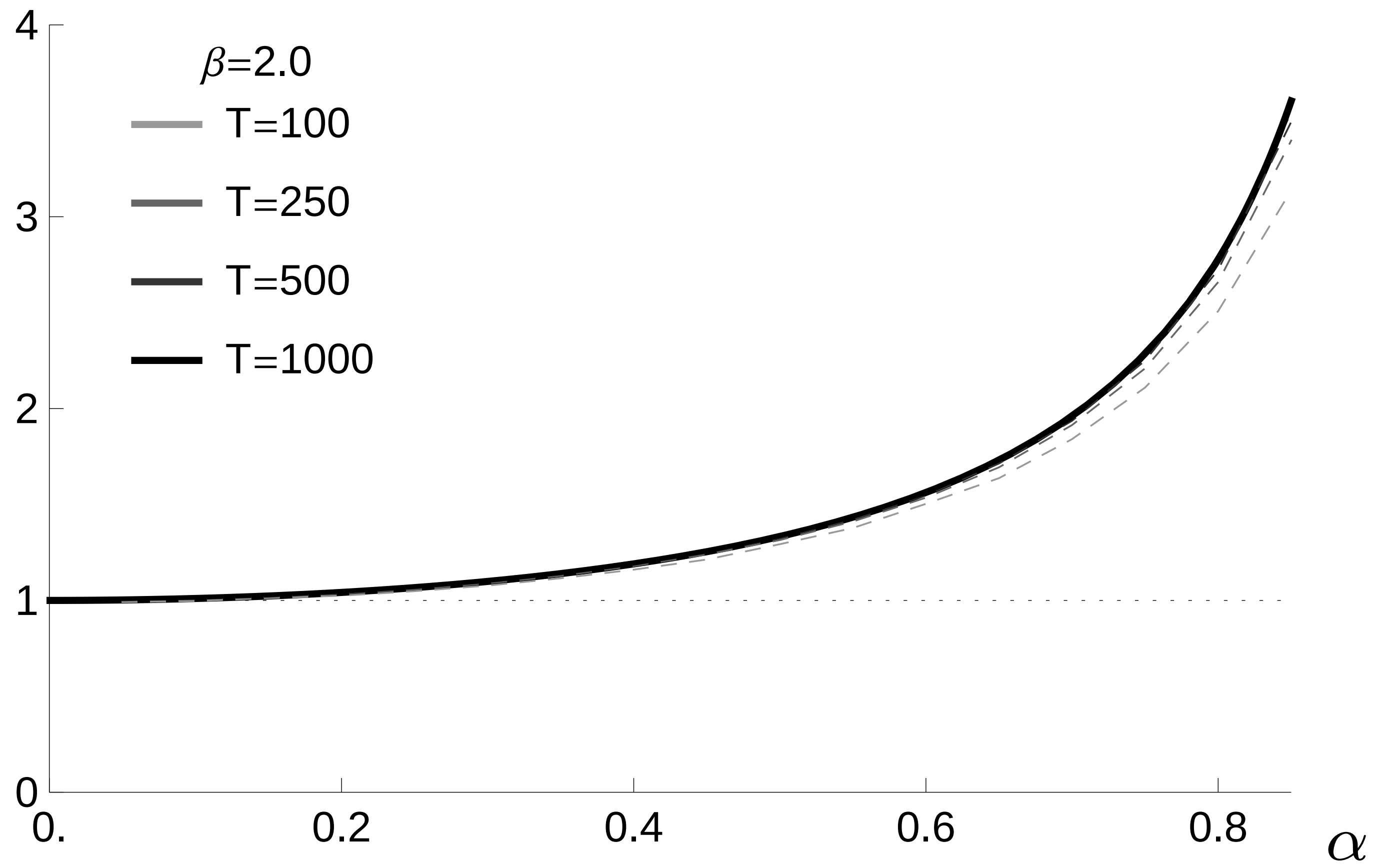
Finite-sample performance: (10,000 replications)

standard deviations of $\hat{I}$ against α,



Finite-sample performance: (10,000 replications)

means of $\widehat{I}$ against α,



Index of Dispersion for INARCH(1) Processes

Based on second-order Taylor expansion,

following **bias-corrected** approximation for mean of $\hat{I}$

in case of INARCH(1) model:

$$E\big[\hat{I}\big] \;\approx\; \frac{1}{1-\alpha^2} \;-\; \frac{1}{T}\cdot\left(\frac{1}{(1-\alpha)^2} \;+\; \frac{2}{\beta}\cdot\frac{\alpha^2}{(1-\alpha^2)^2}\right).$$

Successfully corrects the negative bias for large α and small T:

	T	100		250		500		1000	
α	$\frac{1}{1-\alpha^2}$	b.c.	sim.	b.c.	sim.	b.c.	sim.	b.c.	sim.
0.2	1.042	1.026	1.025	1.035	1.036	1.038	1.038	1.040	1.040
0.4	1.190	1.160	1.160	1.178	1.180	1.184	1.185	1.187	1.187
0.6	1.563	1.491	1.504	1.534	1.536	1.548	1.549	1.555	1.554
0.8	2.778	2.478	2.508	2.658	2.659	2.718	2.723	2.748	2.747

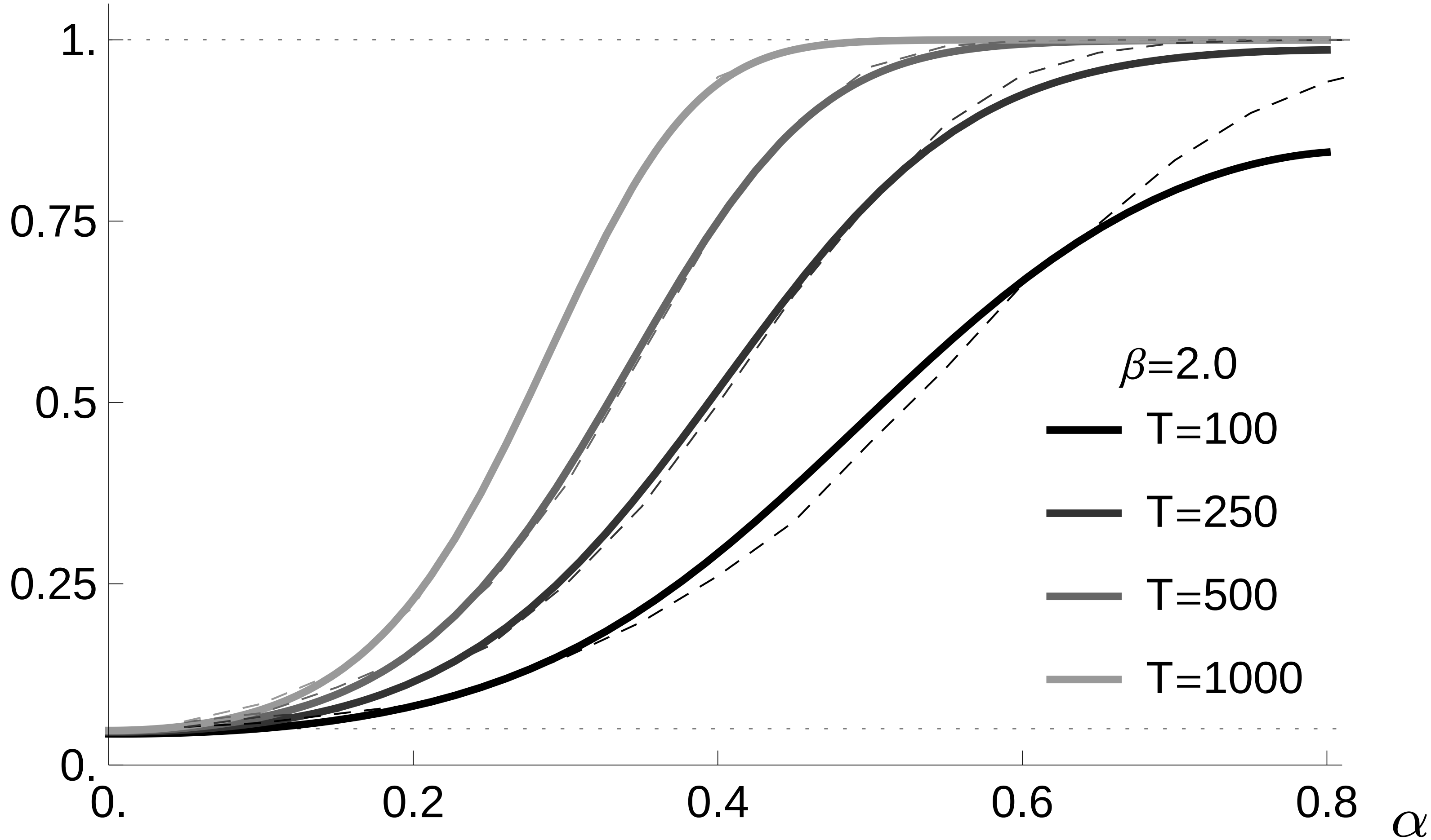
1^{st} application:

H_0: Poisson INAR(1) model with α and $\mu_\epsilon = \beta$,

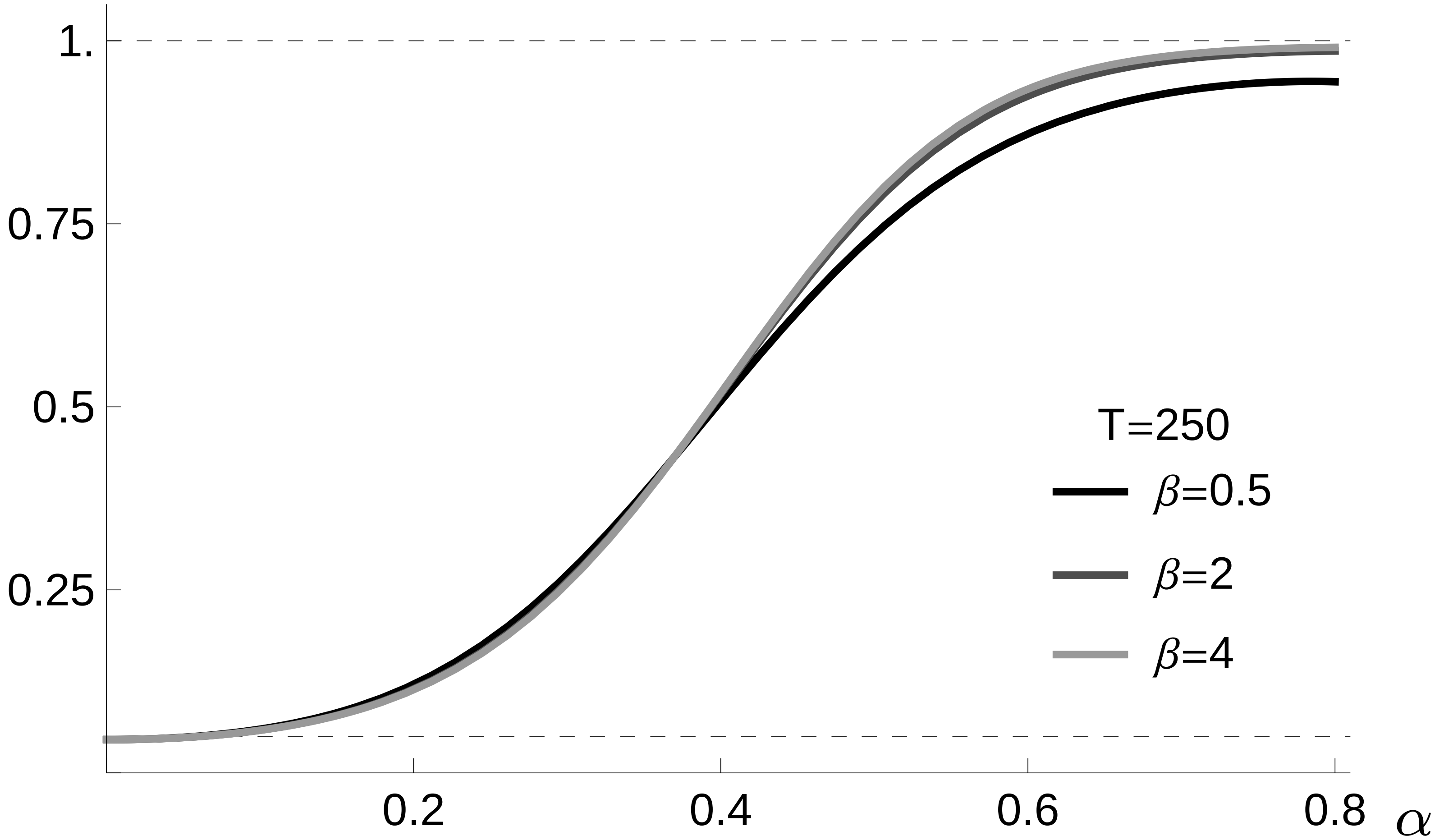
H_1: INARCH(1) model with α and β,

power of $\widehat{I}$ test in detecting INARCH(1) model?

Power analysis w.r.t. INARCH(1) model:



Power analysis w.r.t. INARCH(1) model:



Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

MATH
STAT

INARCH(1) Processes with Conditional Overdispersion

"Work in Progress"

Dispersed INARCH models (DINARCH) by Xu et al. (2012):

$$\mathbb{E}[X_t \mid X_{t-1}, X_{t-2}, \ldots] \;=\; \beta + \alpha \cdot X_{t-1},$$

$$\mathrm{Var}[X_t \mid X_{t-1}, X_{t-2}, \ldots] \;=\; \theta(\beta + \alpha \cdot X_{t-1}).$$

Unconditonal mean and variance:

$$\mu \;=\; \frac{\beta}{1-\alpha} \qquad \text{and} \qquad \sigma^2 \;=\; \frac{\theta}{1-\alpha^2} \cdot \frac{\beta}{1-\alpha},$$

i. e., θ controls degree of dispersion independently of α.

NB-DINARCH(1) model:

$$X_t \mid X_{t-1}, X_{t-2}, \ldots \quad \sim \; \mathrm{NB}\left(\frac{\beta + \alpha \cdot X_{t-1}}{\theta - 1},\; \frac{1}{\theta}\right) \quad \text{with } \theta > 1.$$

2nd application:

H_0: INARCH(1) model with α and β,

H_1: NB-DINARCH(1) model with $\theta > 1$.

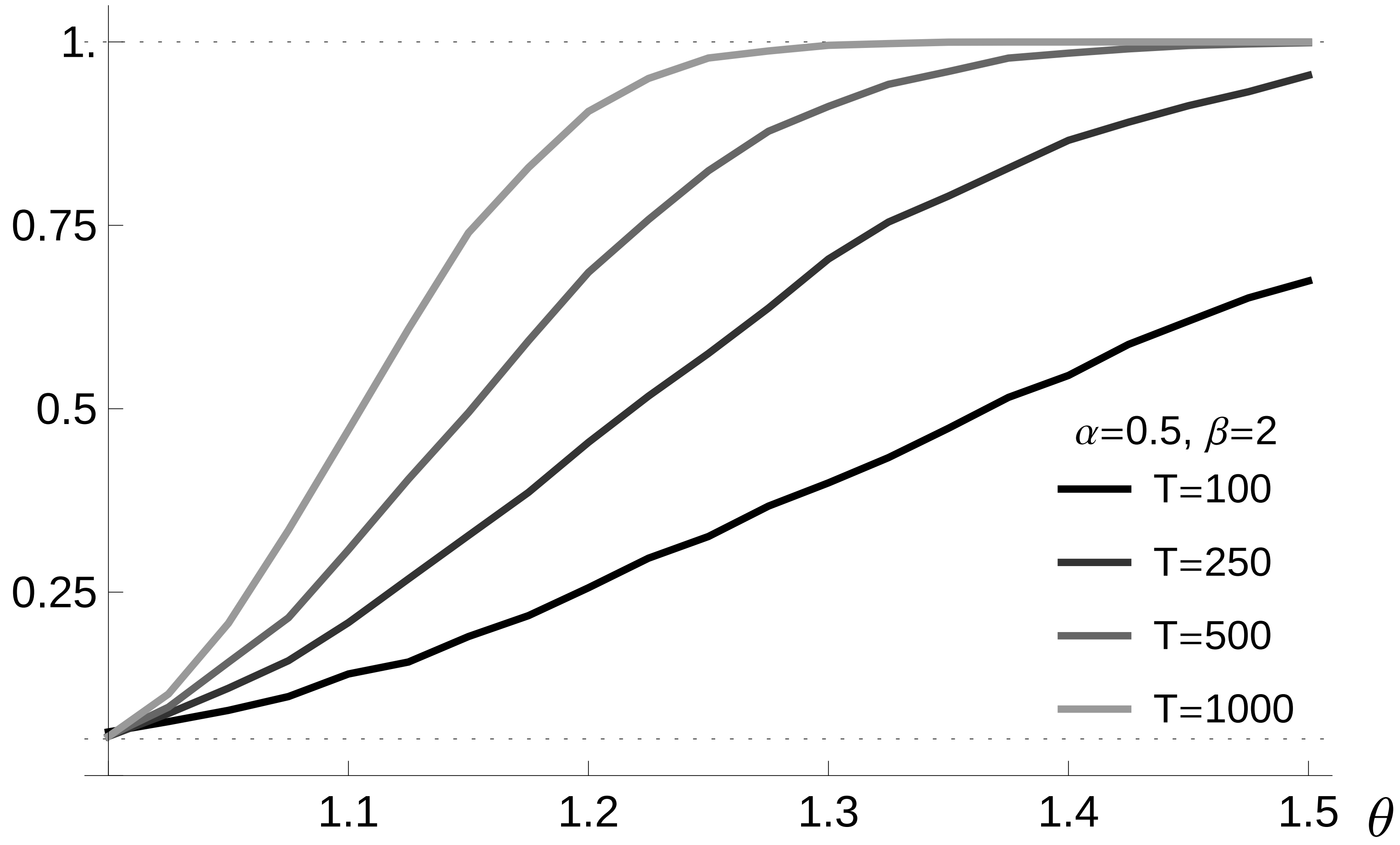
$\hat{I}$-test adapted for INARCH(1) leads to **critical value**

$$\frac{1}{1-\alpha^2} - \frac{1}{T}\cdot\left(\frac{1}{(1-\alpha)^2} + \frac{2}{\beta}\cdot\frac{\alpha^2}{(1-\alpha^2)^2}\right)$$

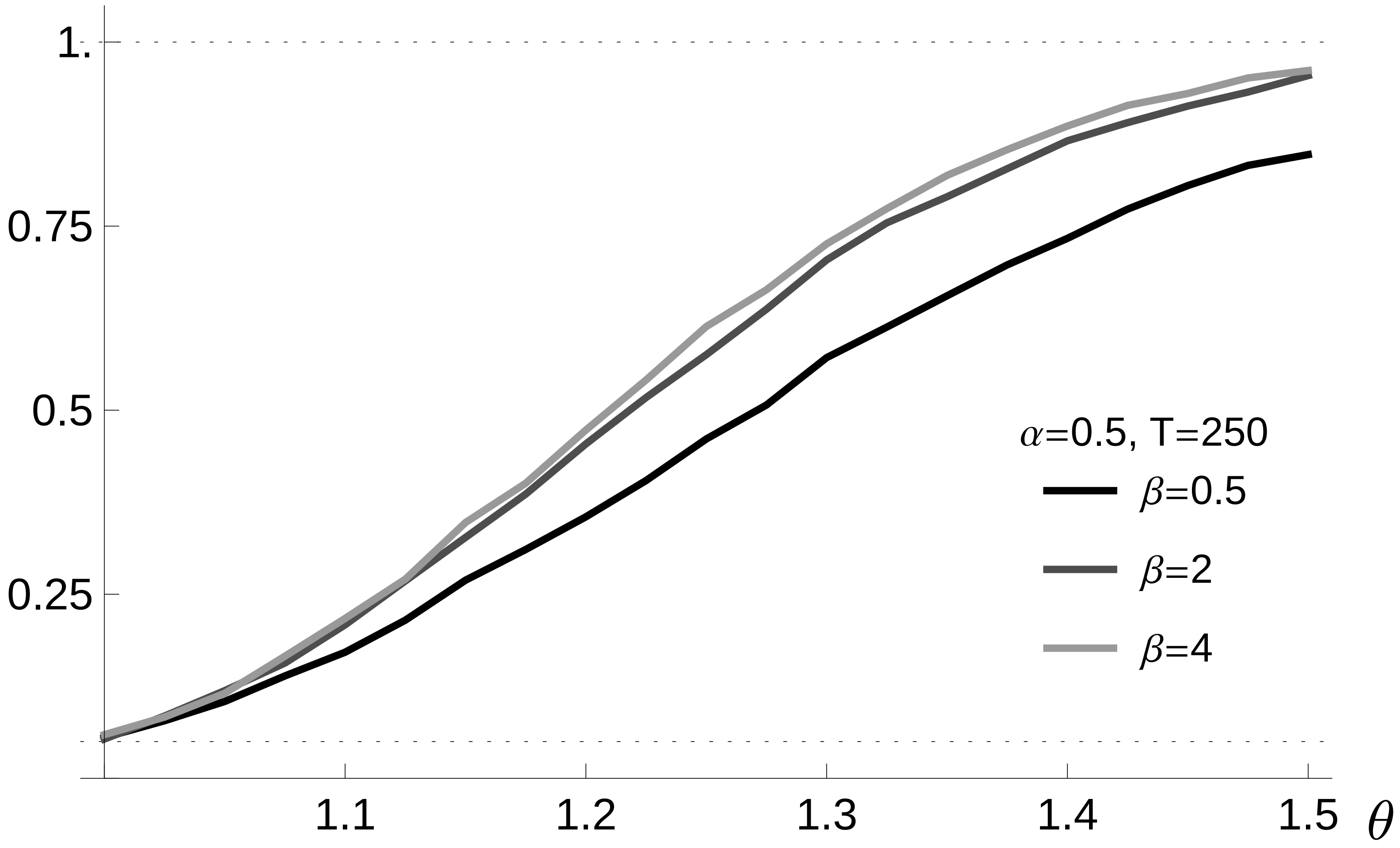
$$+ z_{1-\psi}\cdot\sqrt{\frac{1}{T}\left(\frac{1}{\mu}\cdot\frac{2\alpha^2(1+2\alpha+3\alpha^3)}{(1-\alpha^2)^3(1-\alpha^3)} + \frac{2(1+\alpha^2)}{(1-\alpha^2)^3}\right)}.$$

Power of $\hat{I}$ test in detecting NB-INARCH(1) model?

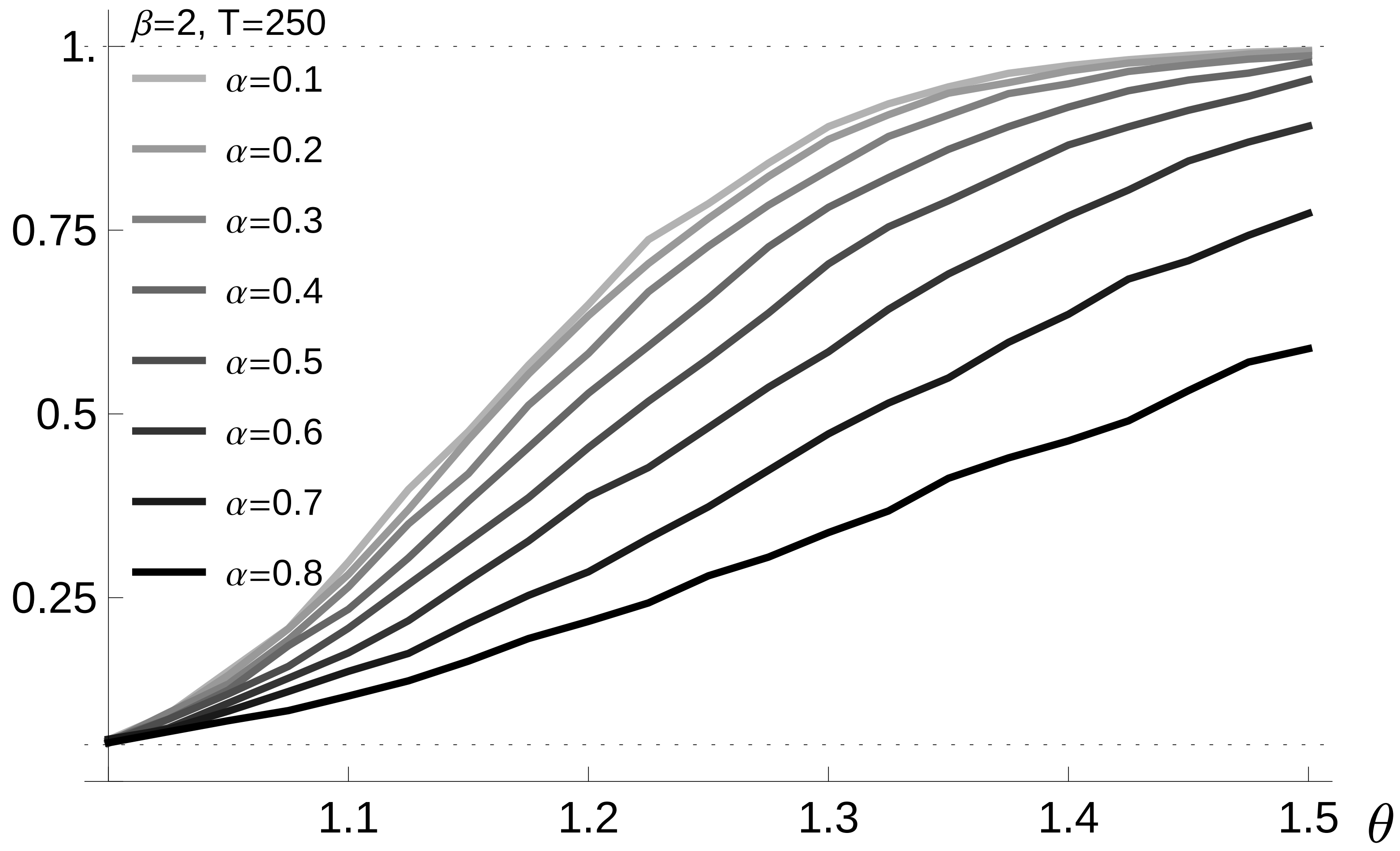
Power analysis w.r.t. NB-DINARCH(1) model:



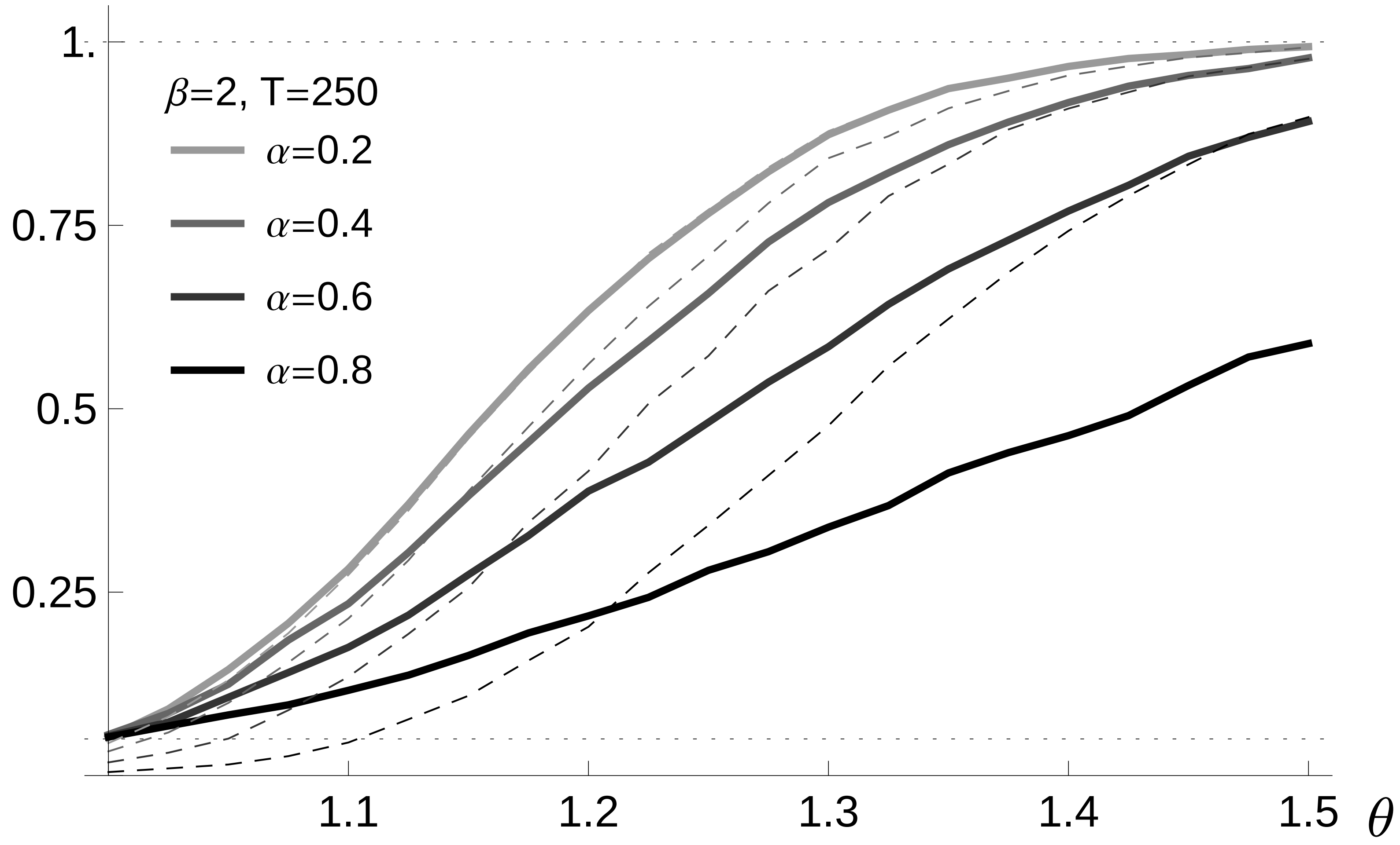
Power analysis w.r.t. NB-DINARCH(1) model:



Power analysis w.r.t. NB-DINARCH(1) model:



Power analysis w.r.t. NB-DINARCH(1) model:



Conclusions

... and Future Research

Conclusions

- Index of dispersion $\hat{I}$ is simple and well-interpretable statistic for uncovering overdispersion.
- Distribution of $\hat{I}$ severely influenced by autocorrelation
  $\Rightarrow$ choose critical values carefully according to null model,
  power analysis has to consider degree of autocorrelation.
- Poisson INAR(1) vs. CPINAR(1) ✓,
  Poisson INAR(1) vs. INARCH(1) ✓,
  INARCH(1) vs. DINARCH(1):
  rather strong degradation of power for large α.

Future research: adjustment for $\hat{I}$-test, or novel test.

Thank You

for Your Interest!

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

`weissc@hsu-hh.de`

MATH
STAT

Literature

Schweer & Weiß (2014): *Compound Poisson INAR(1) processes: stochastic properties and testing for overdispersion*. Comp. Stat. Data Anal. 77, 267–284.

Weiß & Schweer (2014): *Detecting overdispersion in INARCH(1) processes*.

Aki (1985): *Discrete distributions of order k . . .* Ann. Inst. Stat. Math. 37, 205–224.

Feller (1968): *Introduction to probability theory and its applications, Vol. I*. 3rd ed., Wiley.

Ferland et al. (2006): *Integer-valued GARCH processes*. J. Time Ser. Anal. 27(6), 923–942.

Freeland (1998): *Statistical analysis of discrete time series . . .* PhD, Univ. Brit. Columbia.

Heinen (2003): *Modelling time series count data: . . .* CORE 2003-63.

Jung et al. (2005): *Estimation in conditional first order autoregression with discrete support*. Stat. Papers 46, 195–224.

McKenzie (1985): *Some simple models for discrete variate time series*. Water Resources Bull. 21, 645–650.

Pedeli & Karlis (2011): *Bivariate INAR(1) process with application*. Stat. Mod. 11, 325–349.

Steutel & van Harn (1979): *Discrete analogues of self-decomposability and stability*. Ann. Probab. 7, 893–899.

Weiß (2008): *Thinning operations for modelling time series of counts — a survey*. Adv. Stat. Anal. 92, 319–341.

Xu et al. (2012): *A model for integer-valued time series with conditional overdispersion*. Comp. Stat. Data Anal. 56, 4229–4242.