The INARCH(1) Model for Overdispersed Time Series of Counts

TECHNISCHE UNIVERSITÄT DARMSTADT

Christian H. Weiß

Department of Mathematics

Darmstadt University of Technology

Some introductory words . . .

This talk is based on the articles

Weiß, C.H. (2009a). *Modelling time series of counts with overdispersion*. Appears in Statistical Methods and Applications.

Weiß, C.H. (2009b). *The INARCH(1) Model for Overdispersed Time Series of Counts*. Submitted.

All references mentioned in this talk correspond to the references in these articles.

Processes of Counts with Overdispersion

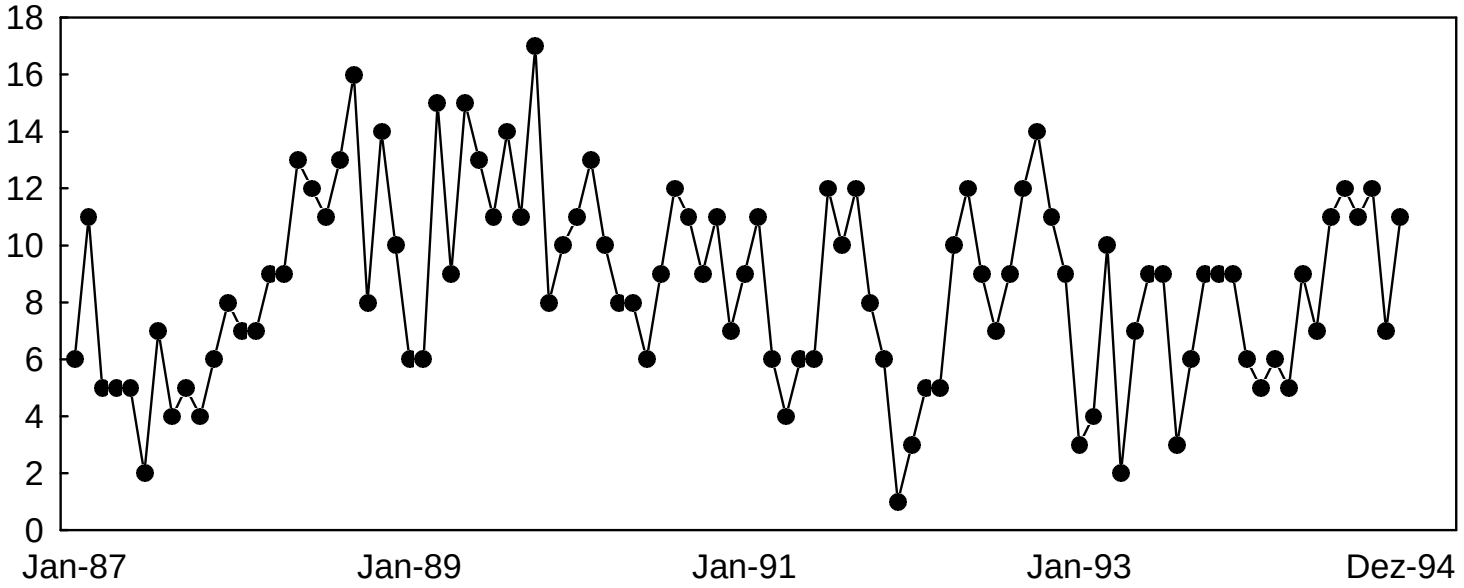
Motivation

Counts with Overdispersion

Processes of counts commonly observed in real-world applications. Examples from diverse fields in practice:

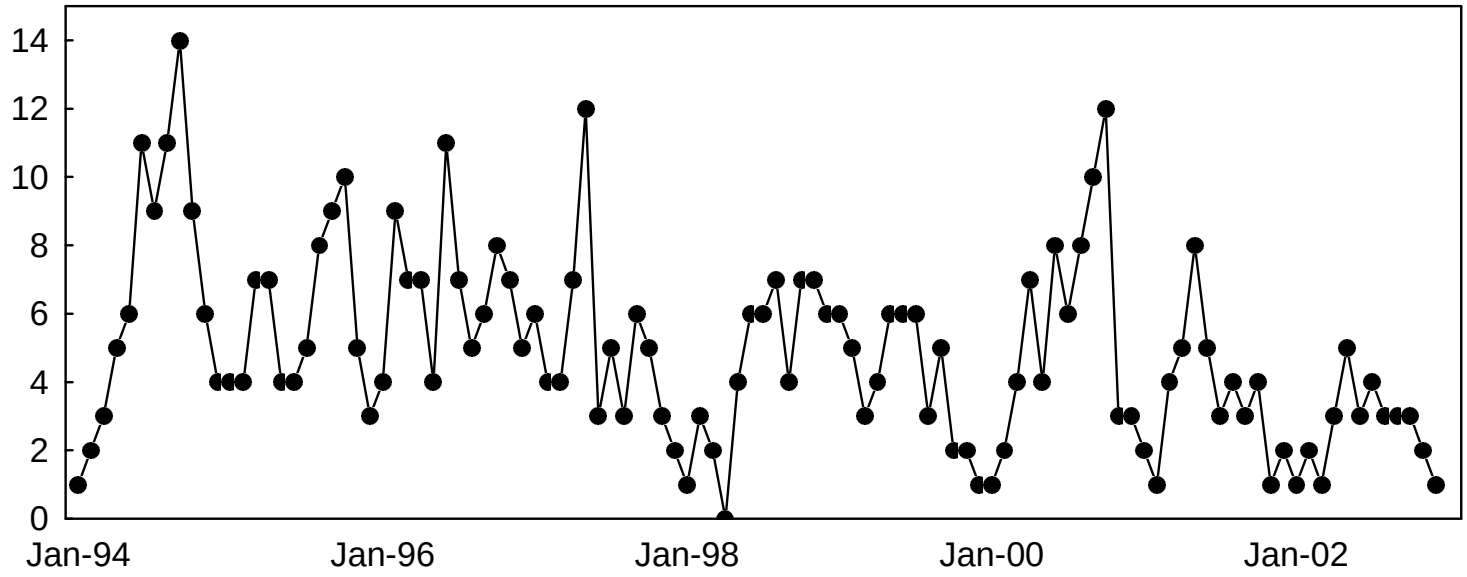
- insurance (e. g., time series of claim counts),
- economics (e. g., counts of price changes),
- statistical process control (e. g., counts of defects),
- traffic (e. g., counts of accidents),
- network monitoring (e. g., intrusion detection system),
- epidemiology (e. g., counts of diseases), and others.

Example 1: Monthly claims counts (1987 to 1994): burn related injuries in heavy manufacturing industry. Source: Freeland (1998).



Example 2: Monthly strike data (1994 to 2002): number of work stoppages leading to 1000 workers or more being idle in effect in the period.

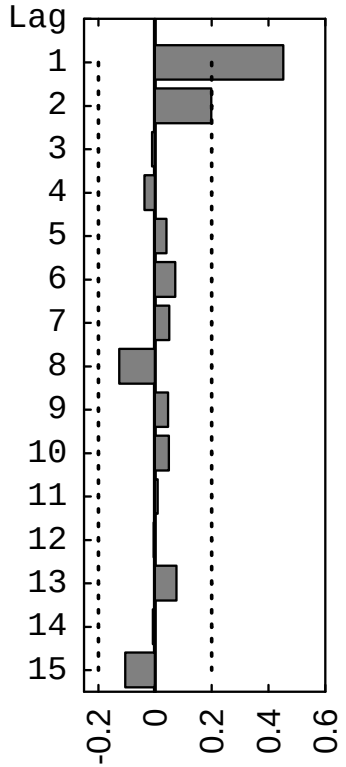
Source: U.S. Bureau of Labor Statistics.



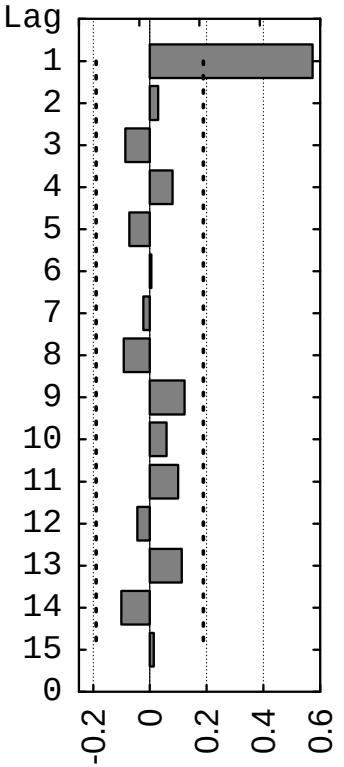
Analysis of both time series:

Partial autocorrelation function of

Example 1:



Example 2:



Analysis of both time series: (continued)

For both examples, AR(1)-like dependence structure

$\Rightarrow$ Popular Poisson INAR(1) model appropriate?

Analysis of both time series: (continued)

Marginal properties:

- Example 1: mean 8.60 (SE$\approx$ 0.49) and variance 11.36;
- Example 2: mean 4.94 (SE$\approx$ 0.41) and variance 7.92.

$\Rightarrow$ **Overdispersion** for both examples!

$\Rightarrow$ The popular Poisson INAR(1) model cannot be used!

Counts with Overdispersion

Overdispersion commonly observed in practice.

Typical reasons:

- presence of positive correlation between monitored events (Friedman, 1993; Poortema, 1999; Paroli et al., 2000);
- variation in probability of monitored events (Heimann, 1996; Poortema, 1999; Christensen et al., 2003);
- further potential causes of overdispersion discussed by Jackson (1972).

In a nutshell:

Modeling of real-world time series of counts

usually requires to consider both

serial dependence and overdispersion!

Recent approach:

INGARCH models, the *in*teger-valued *g*eneralized *a*uto-*r*egressive *c*onditional *h*eteroskedasticity models.

The INGARCH Models

Definition & Properties

INGARCH(p, q) model introduced by Heinen (2003):

$(X_t)_{\mathbb{Z}}$ with range $\mathbb{N}_0$ follows **INGARCH(p, q) model** with $p \geq 1$ and $q \geq 0$ if

(i) X_t, conditioned on $X_{t-1}, \ldots$, is Poisson-distributed according to $Po(M_t)$, where

(ii) conditional mean $M_t := E[X_t \mid X_{t-1}, \ldots]$ fulfills

$$M_t = \beta_0 + \sum_{i=1}^{p} \alpha_i \cdot X_{t-i} + \sum_{j=1}^{q} \beta_j \cdot M_{t-j},$$

where $\beta_0 > 0$ and $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q \geq 0$.

Condition (i) $\Rightarrow$

conditional distribution of $X_t\big|_{X_{t-1},\ldots}$ **equi**dispersed, i. e.,

$E[X_t \mid X_{t-1},\ldots] = M_t = V[X_t \mid X_{t-1},\ldots]$.

But **unconditional** distribution shows **over**dispersion:

$$\mu_t := E[X_t] = E[M_t],$$

$$V[X_t] = E[M_t] + V[M_t] = \mu_t + V[M_t].$$

$\Rightarrow$ INGARCH(p, q) model useful for

modeling time series of overdispersed counts!

Ferland et al. (2006):

For $\alpha_\bullet + \beta_\bullet := \sum_{i=1}^{\mathrm{p}} \alpha_i + \sum_{j=1}^{\mathrm{q}} \beta_j < 1$,

the INGARCH(p, q) model exists and is strictly stationary,

with finite first and second order moments.

Mean $\mu := E[X_t] = \beta_0/(1 - \alpha_\bullet - \beta_\bullet)$.

Weiß (2009a):

Yule-Walker-type equations for $\gamma_X(k) := Cov[X_t, X_{t-k}]$:

$$V[X_t] = \mu + \sum_{i=1}^{\mathsf{p}} \alpha_i \cdot \gamma_X(i) + \sum_{j=1}^{\mathsf{q}} \beta_j \cdot \gamma_M(j),$$

$$\gamma_X(k) = \sum_{i=1}^{\mathsf{p}} \alpha_i \cdot \gamma_X\big(|k-i|\big) + \sum_{j=1}^{\min\{k-1,\mathsf{q}\}} \beta_j \cdot \gamma_X(k-j) + \sum_{j=k}^{\mathsf{q}} \beta_j \cdot \gamma_M(j-k) \quad \text{for } k \geq 1,$$

where $\gamma_M(k) := Cov[M_t, M_{t-k}]$ with

$$\gamma_M(k) = \sum_{i=1}^{\min\{k,\mathsf{p}\}} \alpha_i \cdot \gamma_M(k-i) + \sum_{i=k+1}^{\mathsf{p}} \alpha_i \cdot \gamma_X(i-k) + \sum_{j=1}^{\mathsf{q}} \beta_j \cdot \gamma_M\big(|k-j|\big) \quad \text{for } k \geq 0.$$

Further simplification for q = 0, i. e.,

for INARCH(p) family with $\alpha_\bullet < 1$:

$$\gamma_X(k) \;=\; \sum_{i=1}^{\mathsf{p}} \alpha_i \cdot \gamma_X(|k-i|) \;+\; \delta_{k0} \cdot \mu \qquad \text{for } k \geq 0.$$

Nearly identical to YW equations of standard AR(p) models.

Consequence:

Model order p can be identified

via usual partial autocorrelation function $\rho_{\mathsf{part}}(k)$:

$\rho_{\mathsf{part}}(k) = 0$ for $k > \mathsf{p}$.

In the following,
we concentrate on most simple INARCH model,
the **INARCH(1) model**.

As seen above,
INARCH(1) model has AR(1)-like dependence structure
(like in above examples), i. e.,
INARCH(1) model is alternative to
popular Poisson INAR(1) model,
but is able to model overdispersion.

The INARCH(1) Model

Definition & Properties

Definition:

Let $(X_t)_{\mathbb{Z}}$ be a process with range $\mathbb{N}_0 = \{0, 1, \ldots\}$,

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

$(X_t)_{\mathbb{Z}}$ is said to follow an **INARCH(1) model**

if X_t, conditioned on $X_{t-1}, X_{t-2}, \ldots$,

is Poisson distributed according to $Po(\beta + \alpha \cdot X_{t-1})$.

Basic properties:

- Stationary Markov chain with transition probabilities

$$\begin{aligned} p_{i|j} &:= P(X_t = i \mid X_{t-1} = j) \\ &= \exp(-\beta - \alpha \cdot j) \cdot \frac{(\beta + \alpha \cdot j)^i}{i!} > 0; \end{aligned}$$

- autocorrelation function $\rho_X(n) := Corr[X_t, X_{t-n}] = \alpha^n$.

Weiß (2009a): (Marginal Cumulants)

The cumulants follow recursively from

$$\kappa_1 = \frac{\beta}{1-\alpha}, \qquad \kappa_n = -(1-\alpha^n)^{-1} \cdot \sum_{j=1}^{n-1} s_{n,j} \cdot \kappa_j \quad \text{for } n \geq 2,$$

where $s_{n,j}$ are Stirling numbers of first kind:

$$s_{n,0} = 0 \quad \text{and} \qquad s_{n,n} = 1 \quad \text{for } n \geq 1,$$
$$s_{n+1,j} = s_{n,j-1} - n \cdot s_{n,j} \qquad \text{for } j = 1, \ldots, n \text{ and } n \geq 1.$$

Weiß (2009a): (Marginal Cumulants) (continued)

In particular,

$$\kappa_1 = \frac{\beta}{1-\alpha} = E[X_t], \quad \kappa_2 = \frac{\beta}{(1-\alpha)(1-\alpha^2)} = V[X_t],$$

i. e., **overdispersion**,

$$\kappa_3 = \frac{1+2\alpha^2}{1-\alpha^3} \cdot \kappa_2, \quad \kappa_4 = \frac{1+6\alpha^2+5\alpha^3+6\alpha^5}{(1-\alpha^3)(1-\alpha^4)} \cdot \kappa_2,$$

i. e., skewness and excess of X_t are given by $\frac{1+2\alpha^2}{1+\alpha+\alpha^2} \cdot \sqrt{\frac{1+\alpha}{\beta}}$ and $\frac{1+6\alpha^2+5\alpha^3+6\alpha^5}{\beta(1+\alpha+\alpha^2)(1+\alpha^2)}$, respectively.

The INARCH(1) Model

Parameter Estimation

Conditional maximum likelihood approach:

$$L(\beta,\alpha) := P(X_T = x_T,\ldots,X_2 = x_2 \mid X_1 = x_1)$$

$$= \frac{e^{-(T-1)\beta}\cdot e^{-\alpha\sum_{t=2}^{T} x_{t-1}}\cdot\prod_{t=2}^{T}(\beta+\alpha x_{t-1})^{x_t}}{\left(\prod_{t=2}^{T} x_t!\right)}.$$

ML estimates via numerical maximization.

Observed Fisher information given by

$$\mathbf{J}(\beta,\alpha) = \begin{pmatrix} \sum_{t=2}^{T}\frac{x_t}{(\beta+\alpha x_{t-1})^2} & \sum_{t=2}^{T}\frac{x_t x_{t-1}}{(\beta+\alpha x_{t-1})^2} \\ \sum_{t=2}^{T}\frac{x_t x_{t-1}}{(\beta+\alpha x_{t-1})^2} & \sum_{t=2}^{T}\frac{x_t x_{t-1}^2}{(\beta+\alpha x_{t-1})^2} \end{pmatrix}.$$

$\rightarrow$ Approximate asymptotic SE of ML estimators.

INARCH(1) model performs very well for initial examples:

- **Example 1:**

  ML-estimates $\widehat{\beta} = 4.3796$ and $\widehat{\alpha} = 0.4911$,

  model mean 8.61 and variance 11.34,

  empirical mean 8.60 and variance 11.36.

- **Example 2:**

  ML-estimates $\widehat{\beta} = 1.8114$ and $\widehat{\alpha} = 0.6364$,

  model mean 4.98 and variance 8.37,

  empirical mean 4.94 and variance 7.92.

Conditional least squares approach:

$$\begin{aligned} CSS(\beta,\alpha) &:= \Sigma_{t=2}^{T} \left(x_t - E[X_t \mid X_{t-1} = x_{t-1}]\right)^2 \\ &= \Sigma_{t=2}^{T} \left(x_t - \beta - \alpha x_{t-1}\right)^2. \end{aligned}$$

Minimization leads to

$$\widehat{\alpha}_{\mathsf{CLS}} = \frac{\Sigma_{t=2}^{T} X_t X_{t-1} - \frac{1}{T-1} \cdot \Sigma_{t=2}^{T} X_t \cdot \Sigma_{s=2}^{T} X_{s-1}}{\Sigma_{t=2}^{T} X_{t-1}^2 - \frac{1}{T-1} \cdot \left(\Sigma_{t=2}^{T} X_{t-1}\right)^2},$$

$$\widehat{\beta}_{\mathsf{CLS}} = \frac{1}{T-1}\left(\Sigma_{t=2}^{T} X_t - \widehat{\alpha}_{\mathsf{CLS}} \cdot \Sigma_{t=2}^{T} X_{t-1}\right).$$

Weiß (2009b): (Asymptotic Distribution CLS)

Using a result of Klimko & Nelson (1978), it follows that

$$\sqrt{T-1}\,(\hat{\beta}_{\text{CLS}}-\beta,\hat{\alpha}_{\text{CLS}}-\alpha)^{\top} \xrightarrow{D} N(\mathbf{0},\boldsymbol{\Sigma}_{\beta,\alpha}),$$

where $\boldsymbol{\Sigma}_{\beta,\alpha} =$

$$\begin{pmatrix} \frac{\beta}{1-\alpha}\big(\beta(1+\alpha)+\frac{1+2\alpha^4}{1+\alpha+\alpha^2}\big) & -\beta(1+\alpha)-\frac{(1+2\alpha)\alpha^3}{1+\alpha+\alpha^2} \\ -\beta(1+\alpha)-\frac{(1+2\alpha)\alpha^3}{1+\alpha+\alpha^2} & (1-\alpha^2)\big(1+\frac{\alpha(1+2\alpha^2)}{\beta(1+\alpha+\alpha^2)}\big) \end{pmatrix}.$$

$\rightarrow$ Approximate asymptotic SE of CLS estimators, derivation of simultaneous confidence regions.

Weiß (2009b): (Simultaneous Confidence Regions)

Two types of confidence region based on CLS estimators:

$$\Big\{(\beta,\alpha) \mid (\hat{\beta}_{\text{CLS}}-\beta, \hat{\alpha}_{\text{CLS}}-\alpha)\boldsymbol{\Sigma}^{-1}_{\beta,\alpha}(\hat{\beta}_{\text{CLS}}-\beta, \hat{\alpha}_{\text{CLS}}-\alpha)^{\top} < \tfrac{z}{T-1}\Big\},$$

where z denotes γ-quantile of χ^2_2-distribution, and

$$\Big\{(\beta,\alpha) \mid$$

$$(\hat{\beta}_{\text{CLS}}-\beta)^2 < \tfrac{z^2}{T-1}\cdot\frac{\hat{\beta}_{\text{CLS}}\big(1+2\hat{\alpha}^4_{\text{CLS}}+\hat{\beta}_{\text{CLS}}(1+\hat{\alpha}_{\text{CLS}})(1+\hat{\alpha}_{\text{CLS}}+\hat{\alpha}^2_{\text{CLS}})\big)}{1-\hat{\alpha}^3_{\text{CLS}}},$$

$$(\hat{\alpha}_{\text{CLS}}-\alpha)^2 < \tfrac{z^2}{T-1}\cdot\big(1-\hat{\alpha}^2_{\text{CLS}}+\tfrac{\hat{\alpha}_{\text{CLS}}+\hat{\alpha}^3_{\text{CLS}}-2\hat{\alpha}^5_{\text{CLS}}}{\hat{\beta}_{\text{CLS}}(1+\hat{\alpha}_{\text{CLS}}+\hat{\alpha}^2_{\text{CLS}})}\big)\Big\},$$

where z denotes $(3+\gamma)/4$-quantile of $N(0,1)$-distribution.

Performance for finite values of T via simulation study using Mathematica 5, with 50,000 replications per (β, α), γ, T.

Main results:

For $\alpha \leq 0.2$, CLS_1 region reliable already if $T \geq 100$.

For $0.4 \leq \alpha \leq 0.6$, CLS_1 region requires $T \geq 500$,

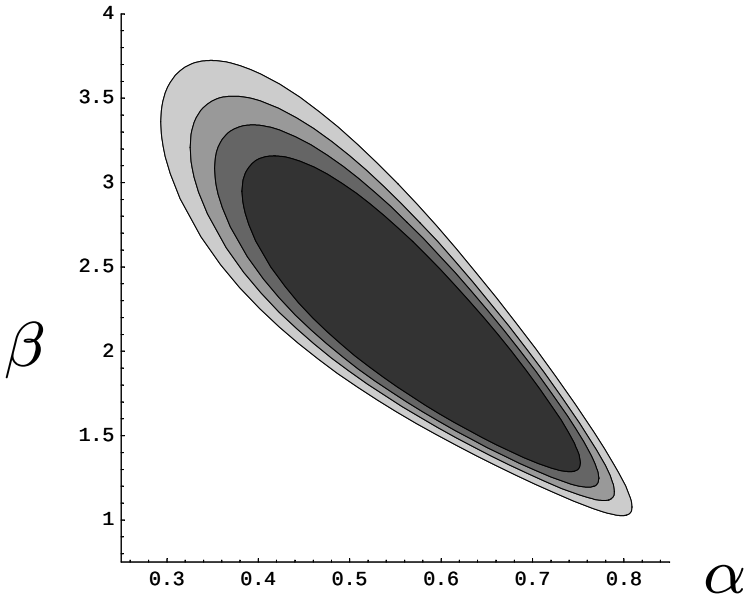
while $\alpha \approx 0.8$ even requires $T \geq 1000$.

CLS_2 region conservative, but can be used if $\gamma \geq 0.95$ independent of (β, α) and T.

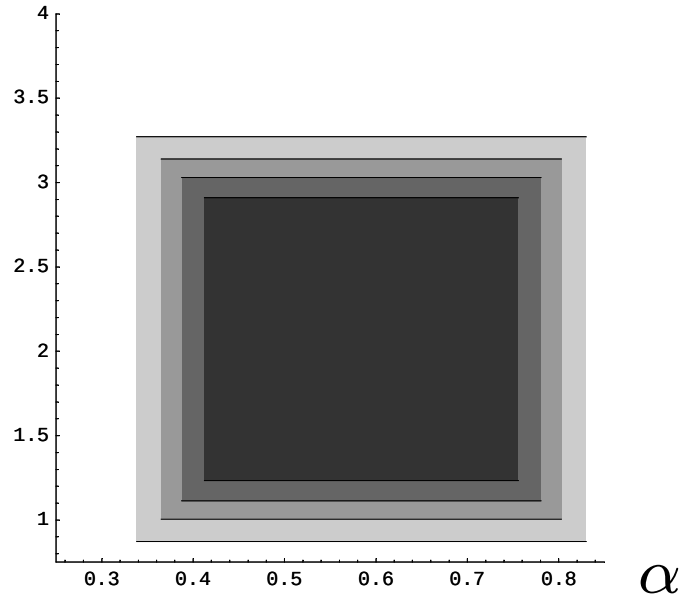
Confidence regions applied to Example 2:

($T = 108$, levels 90 %, 95 %, 97.5 %, 99 %)

(a)



(b)



Method of moments:

$$\widehat{\alpha}_{\mathrm{MM}} = \frac{\sum_{t=2}^{T}(X_t - \bar{X}_T)(X_{t-1} - \bar{X}_T)}{\sum_{t=1}^{T}(X_t - \bar{X}_T)^2},$$

$$\widehat{\beta}_{\mathrm{MM}} = \bar{X}_T \cdot (1 - \widehat{\alpha}_{\mathrm{MM}}).$$

Applying arguments of Theorem 3 in Freeland & McCabe (2005), it follows that MM estimator has same asymptotic properties as CLS estimator.

The INARCH(1) Model

Marginal Distribution

Problem:

As shown above, **marginal cumulants** can be computed recursively up to any order.

But explicit expression for **marginal probabilities** $p_i := P(X_t = i)$ of INARCH(1) process not known!

Feasible approximation of p_i possible?

First approach: Markov chain approximation

$(X_t)_{\mathbb{Z}}$ is ergodic Markov chain, it follows that

$$p_i \;=\; \lim_{n\to\infty} \; p_{i|j}(n) \qquad \text{for all } i,j \in \mathbb{N}_0,$$

where n-step transition probabilities

$$p_{i|j}(n) := P(X_t = i \mid X_{t-n} = j)$$

follow recursively via

$$p_{i|j}(n) \;=\; \sum_{r=0}^{\infty} \; p_{i|r} \cdot p_{r|j}(n-1).$$

INARCH(1) Model: Marginal Distribution

These relations allow to determine marginal probabilities numerically:

Choosing $M, N \in \mathbb{N}$ sufficiently large, we approximate

$$p_i \approx p_{i|j}(N), \qquad \text{where } p_{i|j}(n) \approx \sum_{r=0}^{M} p_{i|r} \cdot p_{r|j}(n-1)$$

for arbitrary $i, j \in \mathbb{N}_0$, e. g., choose $j := \lceil \mu_X \rceil$.

Summary: Precise but computationally expensive, requires appropriate choice of M, N.

Simpler approximation possible?

Second approach:

Poisson-Charlier expansion of Barbour (1987).

Probability generating function (pgf) of X: $p_X(z) := E[z^X]$.

Factorial cumulant generating function (fcgf):

$$k_X(z) \;:=\; \ln\big(p_X(1+z)\big) \;=\; \ln E[(1+z)^X] \;=:\; \sum_{r=1}^{\infty} \frac{\kappa_{(r)}}{r!} \cdot z^r,$$

with factorial cumulants $\kappa_{(r)}$. Relation between pgf and fcgf:

$$p_X(z) \;=\; \exp\big(k_X(z-1)\big) \;=\; \exp\big(\textstyle\sum_{r=1}^{\infty} \frac{\kappa_{(r)}}{r!} \cdot (z-1)^r\big).$$

Basic idea: $p_X(z) \;\approx\; \exp\big(\sum_{r=1}^{m} \frac{\kappa_{(r)}}{r!} \cdot (z-1)^r\big)$.

Poisson-Charlier expansion of Barbour (1987) further refinement of this approach.

Let $\pi_i := e^{-\kappa_1} \cdot \kappa_1^i / i!$ denote Poisson probabilities.
Let ∇ denote difference operator: $\nabla \pi_i = \pi_i - \pi_{i-1}$.

m^{th} order **PC approximation**: $p_i \approx f_m(\nabla) \cdot \pi_i$,
where f_m is $(m-1)^{\text{th}}$ order Taylor polynomial
around $z = 0$ and evaluated in $z = 1$ of

$$f(z, \nabla) \;:=\; \exp\left(\frac{1}{z} \cdot \sum_{r=2}^{\infty} \frac{\kappa_{(r)}}{r!} \cdot (-z\nabla)^r \right).$$

INARCH(1) Model: Marginal Distribution

The first four Poisson-Charlier approximations:

$$
\begin{aligned}
f_1(\nabla) &= 1, \qquad (\rightarrow \text{Poisson approximation}) \\
f_2(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2, \\
f_3(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2 - \tfrac{1}{6}\kappa_{(3)}\nabla^3 + \tfrac{1}{8}\kappa_{(2)}^2\nabla^4, \\
f_4(\nabla) &= 1 + \tfrac{1}{2}\kappa_{(2)}\nabla^2 - \tfrac{1}{6}\kappa_{(3)}\nabla^3 + \left(\tfrac{\kappa_{(2)}^2}{8} + \tfrac{\kappa_{(4)}}{24}\right)\nabla^4 \\
&\quad - \tfrac{1}{12}\kappa_{(2)}\kappa_{(3)}\nabla^5 + \tfrac{1}{48}\kappa_{(2)}^3\nabla^6.
\end{aligned}
$$

So only knowledge about first few factorial cumulants of X required!

Weiß (2009b): (Marginal Factorial Cumulants)

Factorial cumulants of INARCH(1) process determined from usual cumulants via

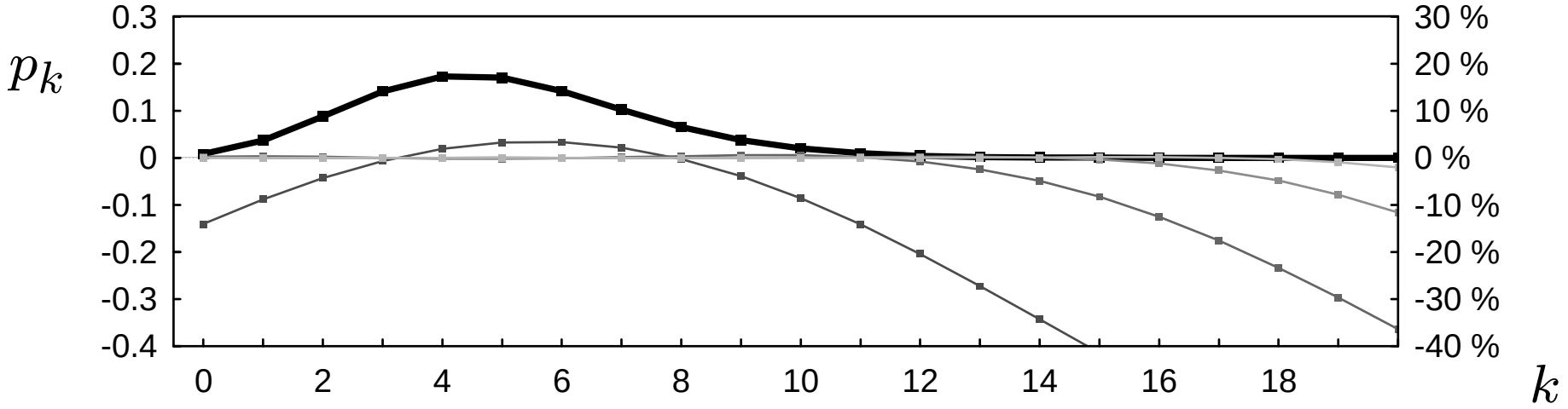
$$\kappa_{(1)} = \kappa_1, \qquad \kappa_{(n)} = \alpha^n \cdot \kappa_n \quad \text{for } n \geq 2.$$

$\Rightarrow$ Poisson-Charlier approximation of any order m easily implemented for INARCH(1) model!

Performance?

INARCH(1) Model: Marginal Distribution

True marginal distribution and relative errors of Poisson-Charlier approximations for $(\beta, \alpha) = (3.75, 0.25)$:



$\Rightarrow$ Approximations only good in central region:

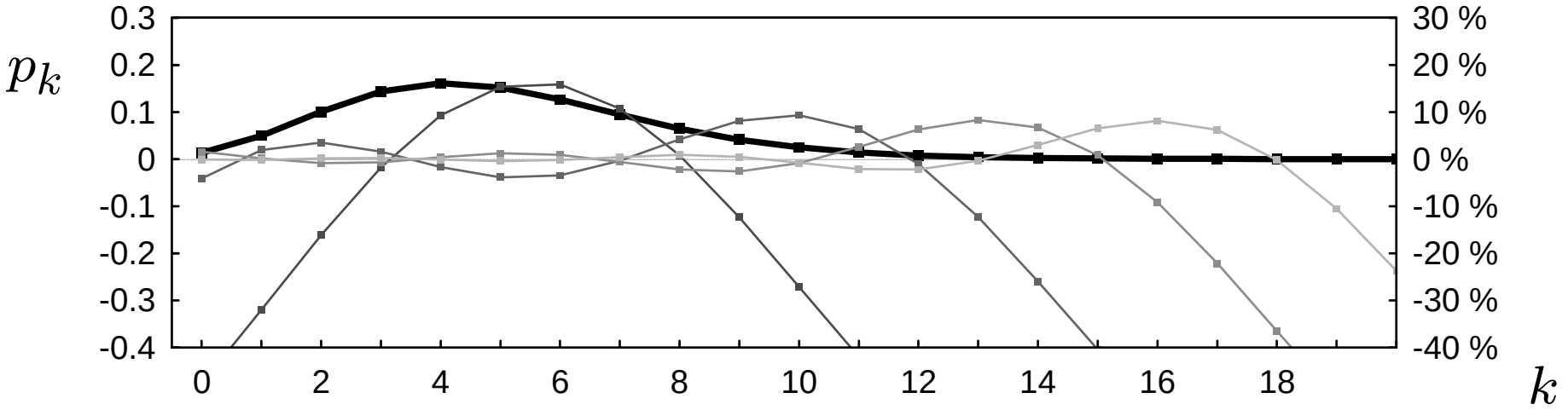
For $k \to \infty$, relative error tends to $-100\,\%$.

Usual Poisson approximation always bad.

INARCH(1) Model: Marginal Distribution

Performance becomes worse for increasing α.

The fourth order approximation still works well for $\alpha = 0.5$ (graph for $\beta = 2.5$):



For $\alpha \geq 0.6$, no satisfactory approximation for the marginal probabilities.

Conclusions and Future Research

- INGARCH models: simple ARMA-like models for time series of overdispersed counts,
  Yule-Walker equations for autocorrelation function.
- INARCH(1) model: Explicit expressions for marginal (factorial) cumulants, autocorrelation function, transition probabilities. Asymptotic distribution for CLS estimators, simultaneous confidence regions.
- Marginal probabilities not explicitly known.
  MC approximation possible but expensive.
  PC approximation better than Poisson approximation, really satisfactorily only for moderate autocorrelation.

Thank You

for Your Interest!

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Christian H. Weiß

Department of Mathematics

Darmstadt University of Technology