Integer-valued Autoregressive Models for Counts Showing Underdispersion

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Christian H. Weiß

Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

Background

This talk is based on the article

Weiß, C.H. (2013):

Integer-valued autoregressive models for counts showing underdispersion.

J. Appl. Statist. **40**(9), 1931–1948.

Further details and references are provided by this article.

Helmut Schmidt Universität
Universität der Bundeswehr Hamburg

MATH
STAT

Count Data Models with Underdispersion

Approaches & Properties

Let X be count data random variable with range $\mathbb{N}_0$.

Denote its mean by μ_X and its variance by σ_X^2.

The "normal distribution" for count data random variables:

Poisson distribution, Po(λ) with parameter $\lambda > 0$,

characterized by the **equi-cumulant property**:

$$\kappa_{X,1} = \kappa_{X,2} = \ldots = \lambda.$$

In particular, Poisson random variables exhibit **equidispersion**:

$\mu_X = \sigma_X^2 \; (= \lambda)$, or dispersion ratio $\dfrac{\sigma_X^2}{\mu_X} = 1$, respectively.

Count Data Models with Underdispersion

Real count data rarely exhibit perfect equidispersion.

Most common deviation from Poisson model:

overdispersion, i. e.,

$$\text{dispersion ratio} \qquad \frac{\sigma_X^2}{\mu_X} > 1.$$

Enumerable models like negative binomial, etc.

Nearly completely neglected in the literature:

opposite phenomenon, **underdispersion**, i. e.,

$$\text{dispersion ratio} \qquad \frac{\sigma_X^2}{\mu_X} < 1.$$

Data example 1: strike counts.

Number of outbreaks of strikes (in 4-week periods) in U.K. coalmining industry (1948–1959), see Kendall (1961).

Empirical mean and variance as 0.994 and 0.742, respectively, i. e., empirical variance-mean ratio ≈ 0.75: underdispersion.

More details later.

Data example 2: emergency counts.

Number of patients between
call for examination and first treatment
in emergency department of children's hospital
(July 16, 2009, 10-min intervals, 08:00:00–23:59:59),
see Weiß (2013).

Empirical mean and variance as 2.56 and 1.87, respectively,
i. e., empirical variance-mean ratio ≈ 0.73: underdispersion.

More details later.

How to model underdispersion?

- **Generalized Poisson distribution** (GP)
  could also be used for underdispersion, but
  - necessary to truncate range
    (truncation depends on actual parameters);
  - only approximate formulae for pmf, mean, variance, etc.
- **Double Poisson distribution** (DP):
  again, essential properties known only approximately.
- **Conway-Maxwell Poisson distribution** (COM):
  again, essential properties known only approximately.

1st Aim: Find models for underdispersion with

- exact formulae for properties like pmf, mean, variance, etc.;
- few model parameters, preferably two model parameters (one for mean, other for dispersion).

In the sequel, detailed description of

- Good distribution;
- power-law weighted Poisson distribution (PL distribution).

Further models discussed in Weiß (2013).

Good distribution: (also polylogarithmic distribution)

Denote **polylogarithm** (Jonquière's function) as

$$\mathrm{Li}_\nu(z) \;=\; \sum_{x=1}^{\infty} z^x \cdot x^{-\nu} \qquad \text{for } |z| < 1.$$

Two-parameter **Good distribution** has pmf

$$P(X=x) \;=\; \frac{q^{x+1} \cdot (x+1)^{-\nu}}{\mathrm{Li}_\nu(q)} \qquad \text{with } 0 < q < 1,\ \nu \in \mathbb{R};$$

probability generating function

$$\mathrm{pgf}(z) \;=\; \frac{1}{z}\,\frac{\mathrm{Li}_\nu(qz)}{\mathrm{Li}_\nu(q)};$$

(. . .)

Good distribution: (...)

moments

$$E[(X+1)^k] \;=\; \frac{\mathrm{Li}_{\nu-k}(q)}{\mathrm{Li}_{\nu}(q)};$$

in particular, mean and variance as

$$E[X] \;=\; \frac{\mathrm{Li}_{\nu-1}(q)}{\mathrm{Li}_{\nu}(q)} - 1, \qquad V[X] \;=\; \frac{\mathrm{Li}_{\nu-2}(q)}{\mathrm{Li}_{\nu}(q)} - \frac{\mathrm{Li}^2_{\nu-1}(q)}{\mathrm{Li}^2_{\nu}(q)}.$$

(Kulasekera & Tonkyn, 1992)

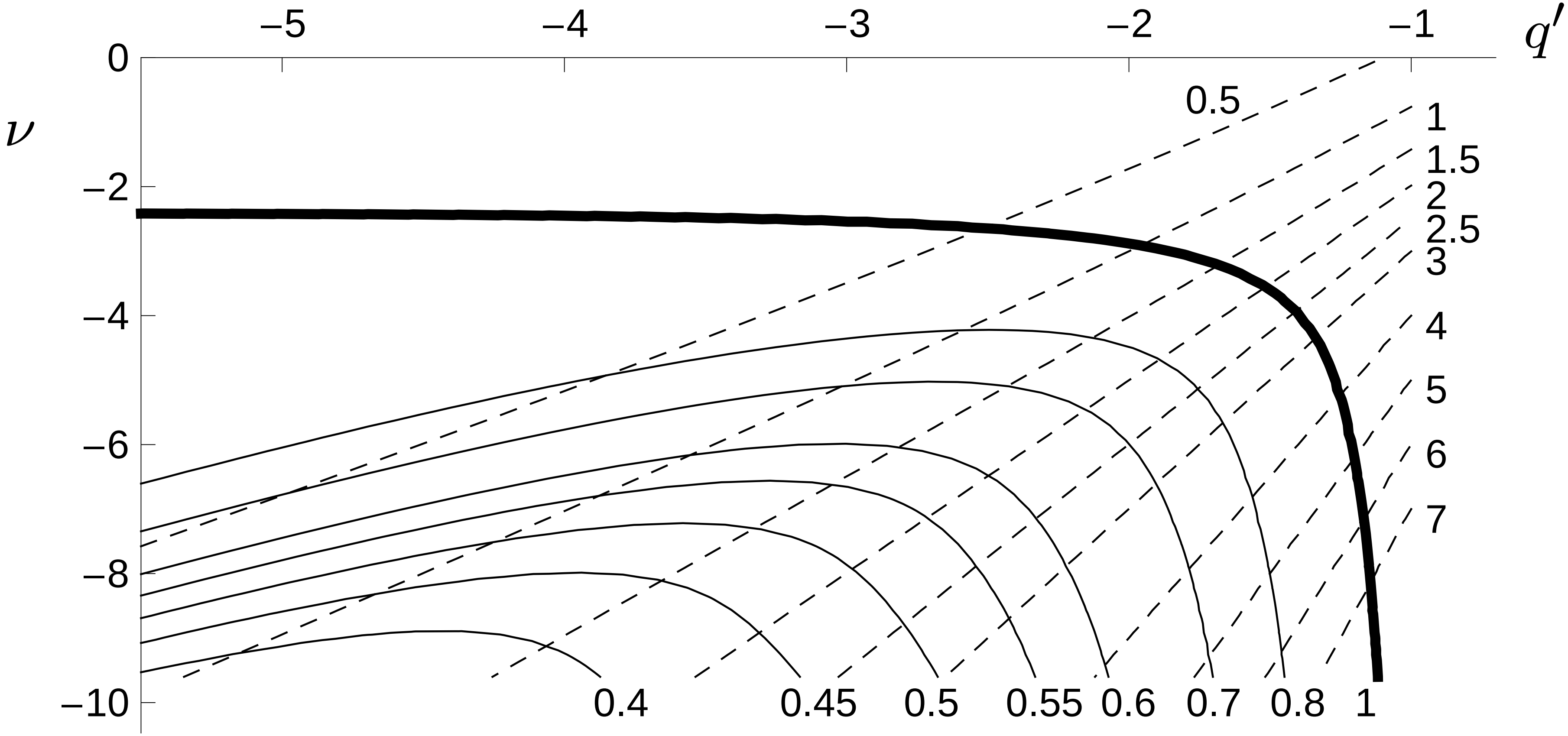
Doray & Luong (1997): **alternative parametrization**,

q replaced by $q' := \ln q \in (-\infty; 0)$,

advantageous in view of parameter estimation.

Good(q', ν) **distribution:**

Mean (dashed) and variance-mean ratio (solid):



PL distribution by del Castillo & Pérez-Casany (1998):

Starting point:

Let $Y \sim \text{Po}(\lambda)$, let $w : \mathbb{N}_0 \to [0; \infty)$ be **weight function** such that $0 < E_\lambda[w(Y)] < \infty$

(w may depend λ, and on additional parameter $\boldsymbol{\theta}$).

Define X (**weighted version** of Y) via pmf

$$P(X = x) \;=\; \frac{w(x)}{E_\lambda[w(Y)]} \cdot e^{-\lambda}\,\frac{\lambda^x}{x!} \qquad \text{for } x = 0, 1, \ldots$$

PL distribution: (...)

Particular weight function $w(x) = (x+a)^\nu$ with $a > 0$ and $\nu \in \mathbb{R}$.

Defining

$$C(\lambda, \nu, a) \;=\; e^\lambda \cdot E_\lambda[w(Y)] \;=\; \sum_{y=0}^{\infty} \frac{\lambda^y (y+a)^\nu}{y!},$$

resulting pmf as

$$P(X = x) \;=\; \frac{\lambda^x \cdot (x+a)^\nu}{C(\lambda, \nu, a) \cdot x!};$$

$$\text{pgf}(z) \;=\; \frac{C(\lambda z, \nu, a)}{C(\lambda, \nu, a)};$$

(...)

PL distribution: (...)

factorial moments

$$E\big[X\cdots(X-k+1)\big] \;=\; \lambda^k\cdot\frac{C(\lambda,\nu,a+k)}{C(\lambda,\nu,a)}.$$

del Castillo & Pérez-Casany (1998):

- equidispersion iff $\nu=0$ ($\equiv$ Po(λ)-distribution),
- overdispersion iff $\nu<0$,
- underdispersion iff $\nu>0$.

In addition: if $\nu\in\mathbb{N}$, then $C(\lambda,\nu,a)$ is e^λ times polynomial:

$$C(\lambda,\nu,a) \;=\; e^\lambda\cdot\sum_{k=0}^{\nu}\binom{\nu}{k}\,a^{\nu-k}\,E_\lambda[Y^k].$$

PL distribution for underdispersion: (...)

$\mathrm{PL}_\nu(\lambda, a)$ with given $\nu \in \mathbb{N}$.

Closed-form expressions for the pmf, pgf, mean and variance, and always underdispersion, see Weiß (2013) for examples.

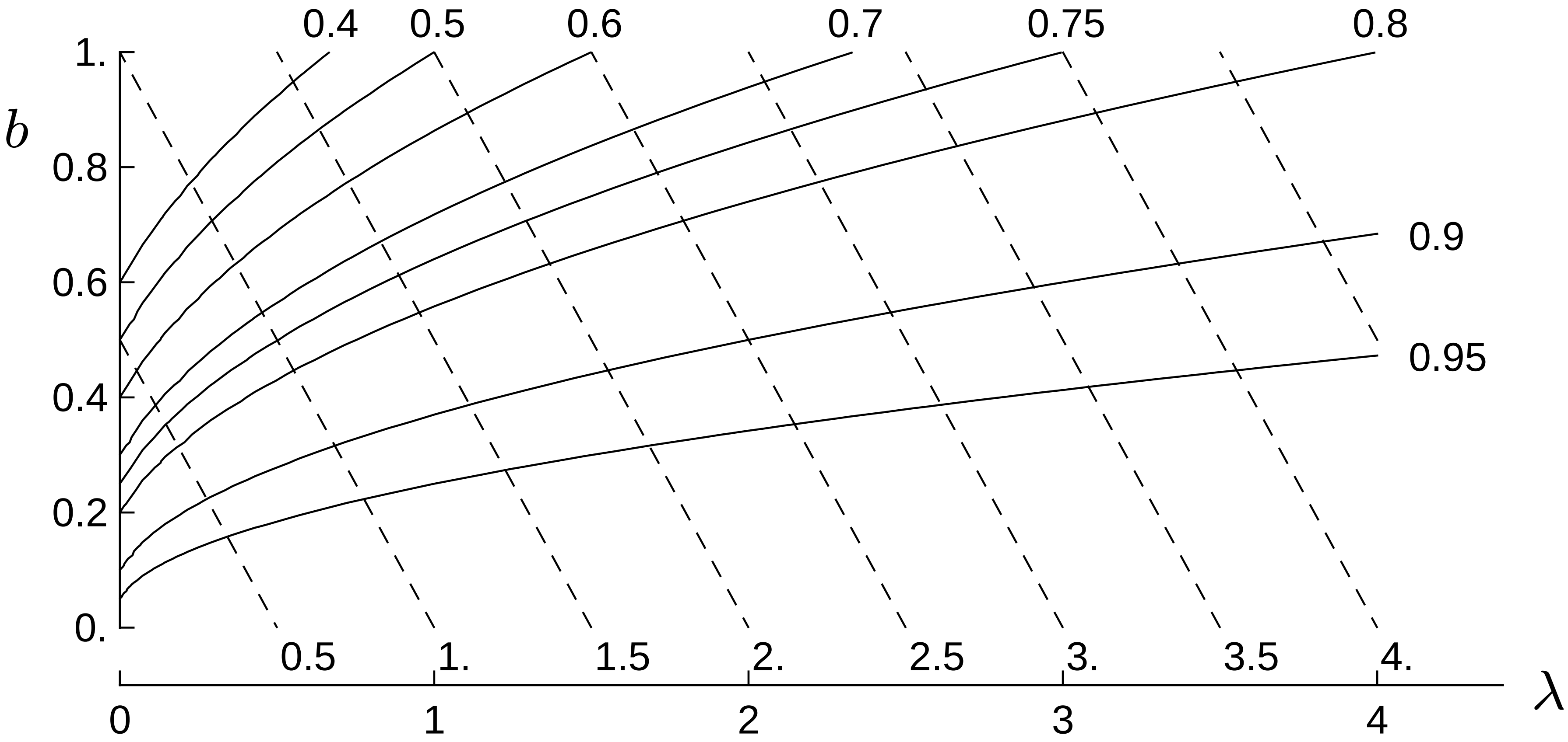
Alternative parametrization by defining

$$b := \frac{\lambda}{\lambda + a} \in (0;1), \qquad \text{equivalent to } a = \lambda \cdot \frac{1-b}{b}.$$

b more easy to interpret (Poisson corresponds to $b \to 0$), advantageous in view of parameter estimation.

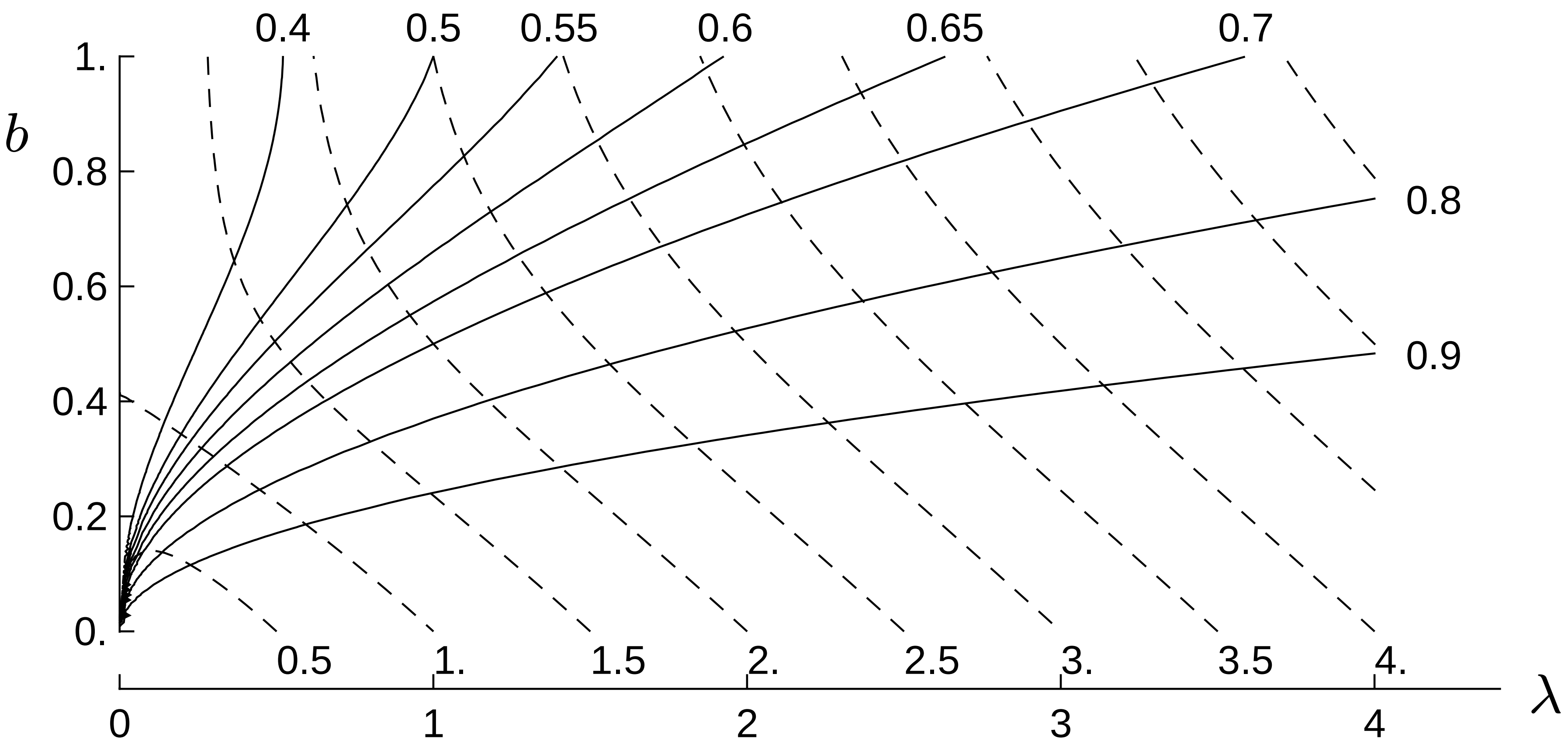
$\mathbf{PL}_1(\lambda, b)$ distribution:

Mean (dashed) and variance-mean ratio (solid):



$\mathbf{PL}_2(\lambda, b)$ distribution:

Mean (dashed) and variance-mean ratio (solid):



Data example 1: strike counts; mean 0.994, vm ratio 0.75.

ML estimates plus standard errors;

range of GP distribution truncated to ≤ 7.

		Par. 1	Par. 2	AIC	BIC
Poisson	(λ)	0.994		385.87	388.92
		(0.080)			
GP	(α_0, ϕ)	0.994	0.873	381.65	387.75
		(0.070)	(0.042)		
Good	(q', ν)	–2.866	–4.779	379.15	385.25
		(0.385)	(0.751)		
PL_1	(λ, b)	0.467	0.527	379.20	385.30
		(0.085)	(0.076)		
PL_2	(λ, b)	0.356	0.279	379.52	385.62
		(0.081)	(0.034)		

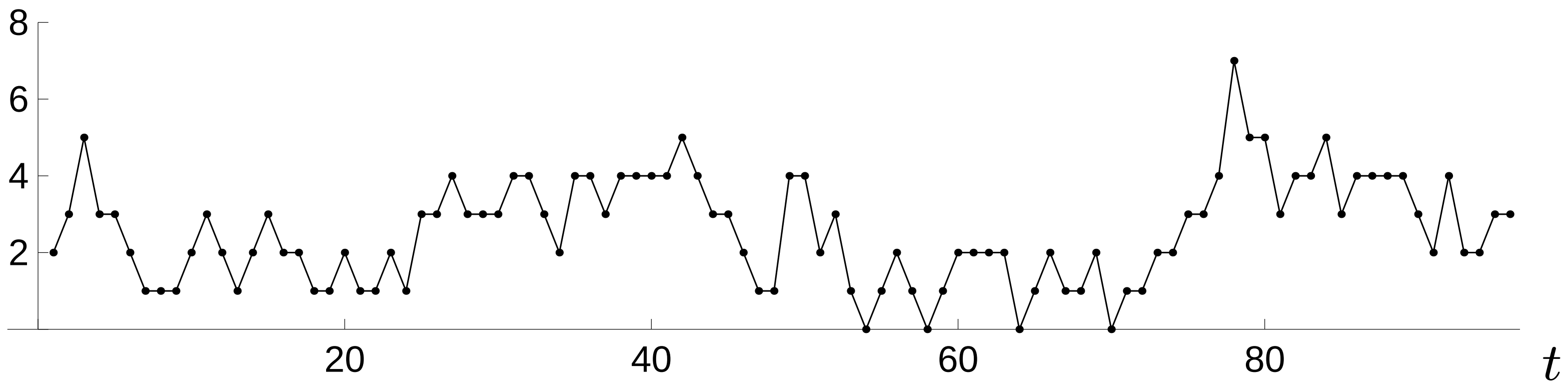
HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

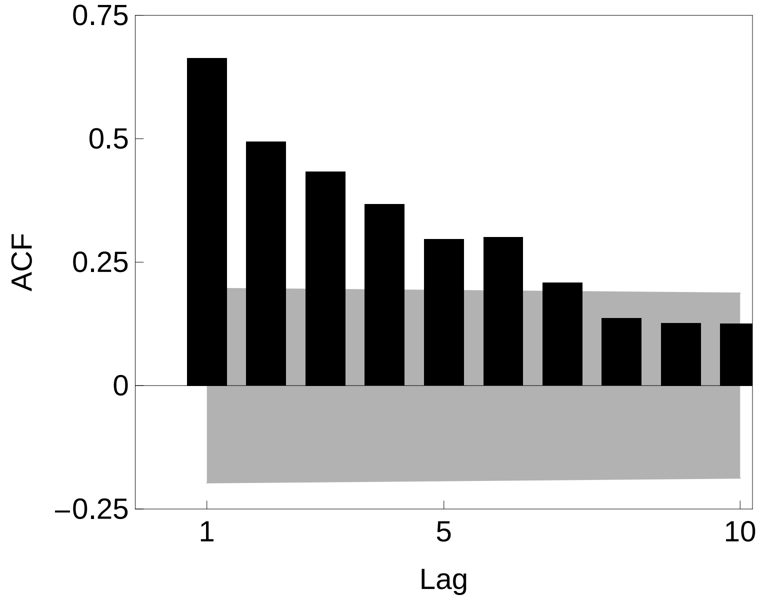
MATH STAT

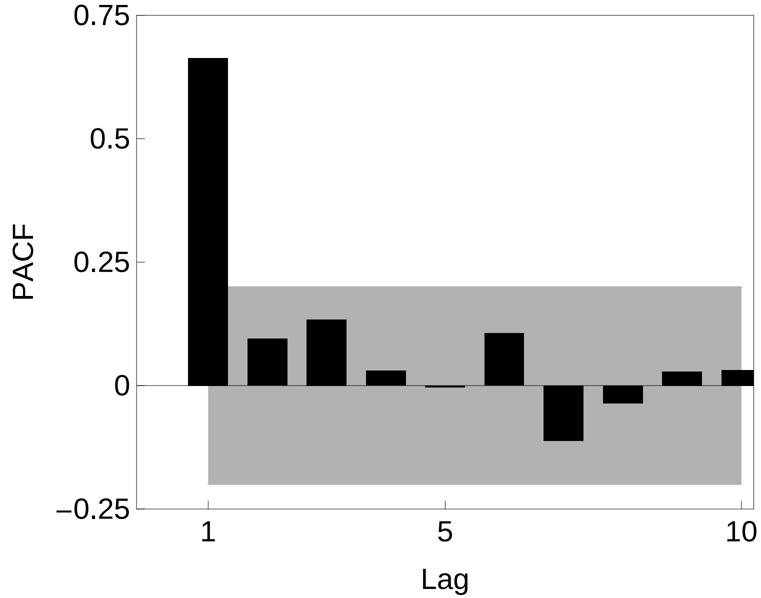
INAR(1) Processes with Underdispersion

Definition & Properties

Data example 2: emergency counts; mean 2.56, vmr 0.73.







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):

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

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

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

INAR(1) Processes with Underdispersion

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

Homogeneous Markov chain with

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

INAR(1) Processes with Underdispersion

Weiß (2013): $\mu_\epsilon < \infty$ guarantees stationary solution.

If INAR(1) process stationary, then

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

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

Note: X_t underdispersed iff ϵ_t underdispersed.

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

For further properties and references, see Weiß (2008,2013).

Data example 2: emergency counts;

mean 2.56, vmr 0.73, $\hat{\rho}(1) \approx 0.664$. We consider

- **Poisson INAR(1) model**: $\epsilon_t \sim \text{Po}(\lambda)$;
- **Good INAR(1) model**: $\epsilon_t \sim \text{Good}(q', \nu)$;
- **PL INAR(1) model**: $\epsilon_t \sim \text{PL}_\nu(\lambda, b)$;
- **GP INAR(1) model**: $\epsilon_t \sim \text{GP}_\nu(\alpha_0, \phi)$;
- **GP INARCH(1) model** by Zhu (2012).

Note: GP INAR(1) with innovations' range truncated ≤ 5.

Data example 2: emergency counts; mean 2.56, vmr 0.73.

ML estimates plus standard errors.

		Par. 1	Par. 2	Par. 3	AIC	BIC
Poisson INAR(1)	(λ, α)	0.737		0.716	279.84	284.97
		(0.132)		(0.048)		
Good INAR(1)	(q', ν, α)	-4.069	-6.887	0.643	277.48	285.18
		(1.100)	(2.360)	(0.073)		
PL_1 INAR(1)	(λ, b, α)	0.265	0.687	0.632	277.04	284.73
		(0.100)	(0.172)	(0.070)		
PL_2 INAR(1)	(λ, b, α)	0.179	0.318	0.634	277.35	285.04
		(0.080)	(0.099)	(0.072)		
GP INAR(1)	(α_0, α, ϕ)	0.837	0.851	0.677	279.59	287.28
		(0.170)	(0.091)	(0.066)		
GP INARCH(1)	$(\alpha_0, \alpha_1, \phi)$	0.918	0.691	0.645	280.00	287.69
		(0.190)	(0.040)	(0.078)		

Conclusions

- Although widely neglected, underdispersion relevant phenomenon for count data random variables.
- Two-parameter Good distribution and PL distribution attractive for modeling underdispersed counts.
- INAR(1) model constitutes simple way to generate underdispersed counts with AR(1)-like dependence.
- Marginal properties derived via factorial cumulants.
- INAR(p) model by Du & Li (1992) more problematic, since underdispersed innovations do not guarantee underdispersed observations.

Thank You for Your Interest!

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

weissc@hsu-hh.de

Weiß (2013): *Integer-valued autoregressive models for counts showing underdispersion*. J. Appl. Statist. 40(9), 1931–1948.

del Castillo & Pérez-Casany (1998): *Weighted Poisson distributions for overdispersion and underdispersion situations*, Ann. Inst. Statist. Math. 50, 567–585.

Doray & Luong (1997): *Efficient estimators for the Good family*, Comm. Statist. Simul. Comput. 26, 1075–1088.

Du & Li (1991): *The integer-valued autoregressive (INAR(p)) model*, J. Time Series Anal. 12, 129–142.

Kendall (1961): *Presidential address: natural law in the social science*, J. Royal Statist. Soc. A 124, 1–16.

Kulasekera & Tonkyn (1992): *A new distribution with applications to survival dispersal and dispersion*, Commun. Statist. Simul. Comput. 21, 499–518.

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

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.

Zhu (2012): *Modeling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models*, J. Math. Anal. Appl. 389, 58–71.