Phase-I Analysis of Time-Dependent Counts with Missing Observations

Christian H. Weiß

Department of Mathematics & Statistics,

Helmut Schmidt University, Hamburg

Murat C. Testik

Industrial Engineering Department,

Hacettepe University, Beytepe-Ankara

HELMUT SCHMIDT
UNIVERSITÄT
Universität der Bundeswehr Hamburg

Phase I Analysis for Process Monitoring

Introduction

Phase I analysis: retrospective analysis of in-control data.

Objective: determine reliable estimates of process parameters, establish control limits for future process monitoring.

Steps of a Phase I analysis (Montgomery, 2009):

1. Select appropriate model for set of data,
   and control statistic for process monitoring.

2. Estimate model parameters from data
   that are believed to represent in-control process.

3. Use estimated parameters for determining control limits.

Steps of a Phase I analysis (cont.):

4. Calculate control statistics by using set of data, plot these on designed control chart.
   (a) Points plot within control limits $\Rightarrow$ set of data declared to represent in-control process, go to Step 5.
   (b) Points plot beyond control limits but assignable cause of variation $\Rightarrow$ corrective actions to improve process.
   (c) Points plot beyond control limits but no assignable cause $\Rightarrow$ discard observations ("outliers"); return to Step 2.
5. Parameter estimates for Phase II application.

Phase I analysis for attributes data:

- binomial counts $X_1, \ldots, X_T$ according to $\text{Bin}(n, \pi)$ with fixed sample size n known, or
- Poisson counts $X_1, \ldots, X_T$ according to $\text{Pois}(\mu)$.

Recommendation (Montgomery, 2009):

Shewhart-type control charts, i. e., np charts and c charts, respectively, for binomial and Poisson counts.

Initial assumption ("textbook case"): i.i.d. data.

Then, **parameter estimation** during Phase I analysis via mean $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$ of set of data:

$$\hat{\pi} = \frac{1}{n}\bar{X} \qquad \text{vs.} \qquad \hat{\mu} = \bar{X}.$$

Note: ML estimator = LS estimator = MM estimator.

If some observations discarded during Step 4:

Compute mean of remaining observations, say $X_{t_1}, \ldots, X_{t_K}$ with $1 \leq t_1 < \ldots < t_K \leq T$.

Now allow for **serially dependent counts**:

- binomial counts $X_1, \ldots, X_T$ according to binomial AR(1) model (McKenzie, 1985), or
- Poisson counts $X_1, \ldots, X_T$ according to Poisson INAR(1) model (McKenzie, 1985).

If data complete: ML, CLS and MM estimator known from literature, e. g., Weiß & Kim (2013) or Weiß (2011), resp.

But: ML estimator $\neq$ LS estimator $\neq$ MM estimator.

Now allow for **serially dependent counts**:

- binomial counts $X_1, \ldots, X_T$ according to binomial AR(1) model (McKenzie, 1985), or
- Poisson counts $X_1, \ldots, X_T$ according to Poisson INAR(1) model (McKenzie, 1985).

If data incomplete (e. g., during Step 4 of Phase I analysis):

ML, CLS and MM estimator?

Reliability of such estimators?

Or simply ignore outliers?

HELMUT SCHMIDT
UNIVERSITÄT
Universität der Bundeswehr Hamburg

Parameter Estimation with Incomplete Data from Poisson INAR(1) Process

Approaches

Poisson INAR(1) model of McKenzie (1985),
using binomial thinning operator of Steutel & van Harn (1979):

Let $\mu > 0$ and $\alpha \in (0;1)$. Let $X_0 \sim \text{Pois}(\mu)$,
let $(\epsilon_t)_{\mathbb{N}}$ be i.i.d. according to $\text{Pois}\big(\mu(1-\alpha)\big)$.

Then $(X_t)_{\mathbb{N}_0}$ defined via

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

plus appropriate independence assumptions.

$\Rightarrow$ Stationary Markov chain with $\text{Pois}(\mu)$-marginals,
autocorrelation function equals $\rho(k) = \alpha^k$.

h-step regression properties (Freeland & McCabe, 2004):

For time lag $h \geq 1$,

$$\begin{aligned} p^{(h)}(k|l) &:= \mathbb{P}(X_{t+h} = k \mid X_t = l) \\ &= \sum_{m=0}^{\min\{k,l\}} \binom{l}{m} \left(\alpha^h\right)^m \left(1-\alpha^h\right)^{l-m} \\ &\quad \cdot \frac{\left(\mu\,(1-\alpha^h)\right)^{k-m}}{(k-m)!} \exp\left(-\mu\,(1-\alpha^h)\right), \end{aligned}$$

$$m^{(h)}(x) := \mathbb{E}[X_{t+h} \mid X_t = x] = \alpha^h x + \mu\,(1-\alpha^h).$$

Yajima & Nishino (1999): $o(t) = 0$ if missing observation.

$$\tilde{x} := \left(\sum_{t=1}^{T} o(t)\, x_t\right) \Big/ \left(\sum_{t=1}^{T} o(t)\right),$$

$$\tilde{\gamma}(k) := \left(\sum_{t=1}^{T-k} o(t)o(t+k)\,(x_t-\tilde{x})(x_{t+k}-\tilde{x})\right) \Big/ \left(\sum_{t=1}^{T} o(t)o(t+k)\right).$$

Available observations: $x_{t_1}, \ldots, x_{t_K}$, $1 \leq t_1 < \ldots < t_K \leq T$.

Missing observations between x_{t_k} and $x_{t_{k-1}}$ when $t_k - t_{k-1} > 1$.

Indicator $o(t) = 1$ iff $t \in \{t_1, \ldots, t_K\}$ and 0 otherwise.

- **Modified ML approach:** $\tilde{\mu}_{\mathrm{ML}}, \tilde{\alpha}_{\mathrm{ML}}$ by maximizing

$$\tilde{L}(\mu, \alpha) \;=\; p(x_{t_1}) \, \Pi_{k=2}^{K} \, p^{(t_k - t_{k-1})}(x_{t_k} | x_{t_{k-1}}).$$

- **Modified CLS approach:** $\tilde{\mu}_{\mathrm{CLS}}, \tilde{\alpha}_{\mathrm{CLS}}$ by minimizing

$$\tilde{S}(\mu, \alpha) \;=\; \Sigma_{k=2}^{K} \, \left(x_{t_k} - m^{(t_k - t_{k-1})}(x_{t_{k-1}}) \right)^2.$$

- **Modified MM approach:**

$$\tilde{\mu}_{\mathrm{MM}} \;=\; \tilde{x}, \qquad \tilde{\alpha}_{\mathrm{MM}} \;=\; \tilde{\gamma}(1) \,/\, \tilde{\gamma}(0).$$

HELMUT SCHMIDT
UNIVERSITÄT
Universität der Bundeswehr Hamburg

Parameter Estimation and Chart Design with Incomplete Data or with Outliers

Simulation Study

Simulation study:

Poisson INAR(1) time series of lengths $T = 50,\ 200$;

means $\mu = 1.44,\ 4$; autocorrelations $\alpha = 0.3,\ 0.5,\ 0.7$.

For rates r=0.02, 0.05, 0.10, 0.25, $\lfloor r \cdot T \rfloor$ observations discarded or contaminated as positive additive outliers (AO$^+$):

$O_t = X_t + \kappa_t$, where $\kappa_t \sim \text{Pois}(4\sqrt{\mu})$ (Barczy et al., 2012).

10,000 replications for each scenario,

ML and CLS estimates numerically via R's `nlminb`,

ARL computations with Markov chain approach in Matlab.

Selected results from simulation study:

α-estimate invalid if outside $(0; 1)$: If missings . . .

			$\mu = 1.44$			$\mu = 4$		
α	T	Rate	Inv_{ML}	Inv_{CLS}	Inv_{MM}	Inv_{ML}	Inv_{CLS}	Inv_{MM}
0.5	50	0	15	18	16	13	15	15
		0.02	10	10	10	12	14	14
		0.05	23	25	25	17	18	18
		0.1	37	40	39	36	40	40
		0.25	106	120	120	109	119	122
0.5	200	0	0	0	0	0	0	0
		0.02	0	0	0	0	0	0
		0.05	0	0	0	0	0	0
		0.1	0	0	0	0	0	0
		0.25	0	0	0	0	0	0

Selected results from simulation study:

α-estimate invalid if outside $(0; 1)$: If outliers . . .

			$\mu = 1.44$			$\mu = 4$		
α	T	Rate	Inv_{ML}	Inv_{CLS}	Inv_{MM}	Inv_{ML}	Inv_{CLS}	Inv_{MM}
0.5	50	0	15	18	16	13	15	15
		0.02	188	189	183	157	157	156
		0.05	518	521	502	435	436	426
		0.1	1588	1588	1564	1489	1488	1472
		0.25	3177	3166	3145	3019	3003	2981
0.5	200	0	0	0	0	0	0	0
		0.02	0	1	0	0	0	0
		0.05	4	4	4	4	4	4
		0.1	117	116	113	61	60	58
		0.25	1004	994	979	849	838	810

Selected results from simulation study:

Certainly, μ-estimate increased iff AO^+ contamination.

Means of α-estimates (for $\mu = 1.44$):

			Missing			Outlier		
α	T	Rate	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{CLS}}$	$\hat{\alpha}_{\mathrm{MM}}$	$\hat{\alpha}_{\mathrm{ML}}$	$\hat{\alpha}_{\mathrm{CLS}}$	$\hat{\alpha}_{\mathrm{MM}}$
0.5	200	0	0.494	0.485	0.483	0.494	0.485	0.483
		0.02	0.494	0.484	0.484	0.379	0.353	0.351
		0.05	0.493	0.484	0.483	0.252	0.251	0.250
		0.1	0.493	0.484	0.484	0.151	0.175	0.174
		0.25	0.492	0.482	0.481	0.077	0.110	0.109

Outliers mimic much less serial dependence,

ML estimator more sensitive to outliers,

while less biased (and less s.d.) otherwise.

Selected results from simulation study:

Design of c-chart with estimates,

choose smallest UCL such that $\mathrm{ARL}_0 \geq 200$:

Theoretically:

μ	α	UCL	ARL_0
1.44	0.5	6	323.3

Practice:

			$\mu = 1.44$			
α	T	Rate	$\widehat{\mathrm{UCL}}_{\mathrm{miss}}$	$\widehat{\mathrm{UCL}}_{\mathrm{out}}$	$\widehat{\mathrm{ARL}}_{0;\mathrm{miss}}$	$\widehat{\mathrm{ARL}}_{0;\mathrm{out}}$
0.5	200	0	6.112		467.7	
		0.02	6.118	6.387	476.2	787.7
		0.05	6.118	6.810	477.0	1324.0
		0.1	6.112	7.194	469.9	2862.0
		0.25	6.117	8.653	476.9	37140.0

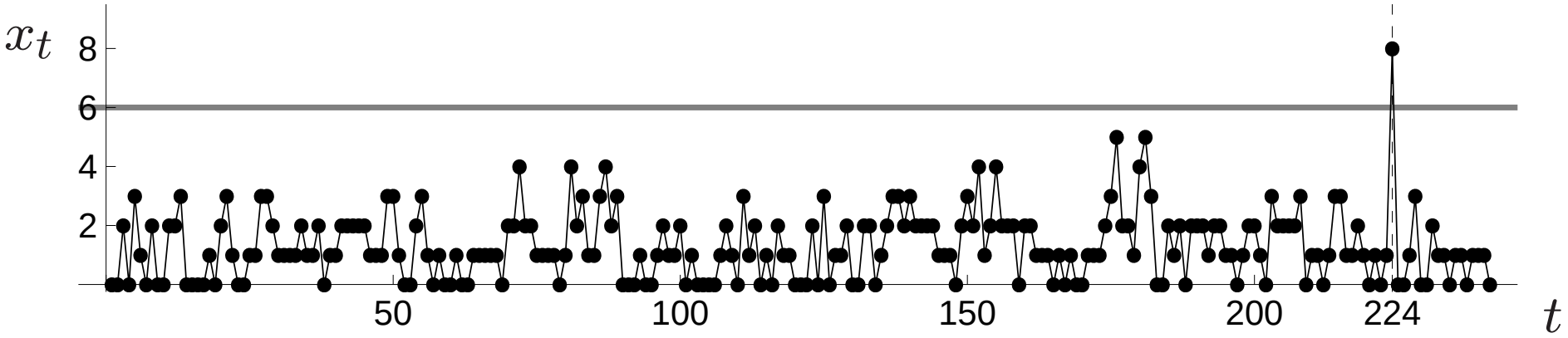
A Real-Data Example

IP Counts

Data set from Weiß (2007): [ENBIS-6, Wroclaw, 2006]

Number of different IP addresses accessesing web server per 2-min on November 29, 2005 (10 a.m. – 6 p.m.)

$\Rightarrow$ time series $x_1, \ldots, x_{241}$ of length $T = 241$.



Mean close to variance (1.315 vs. 1.392) $\Rightarrow$ Poisson model.

SACF indicates AR(1) structure (with $\widehat{\rho}(1) \approx 0.219$).

Phase I analysis:

1. Model: Poisson INAR(1); use c chart (plot x_t).

2. Estimates $\widehat{\mu}_{\mathrm{ML}} = 1.312$, $\widehat{\alpha}_{\mathrm{ML}} = 0.235$.

3. Control limit UCL $= 6$, "believed ARL_0": 441.5.

4. Plot data on c chart (see before);

   alarm at time period 224;

   assume that no assignable cause (Step 4.c)

   $\Rightarrow x_{224}$ as AO^+, discard from data, return to Step 2.

Phase I analysis:

2. Revised estimates $\hat{\mu}_{\text{ML}} = 1.281$, $\hat{\alpha}_{\text{ML}} = 0.290$.

3. Control limit UCL $= 6$, "believed ARL_0": 503.0.

4. Plot data on c chart (see before);

   no further alarm.

5. In-control model found, start with Phase II monitoring.

However, for particular data example,

we can associate **assignable cause** with time period 224:

replace $x_{224} = 8$ by corrected value $x_{224} := 1$ (Weiß, 2007).

Estimates	Complete data	without x_{224}	Corrected $x_{224} := 1$
$\widehat{\mu}_{\mathrm{ML}}$	1.312	1.281	1.282
$\widehat{\alpha}_{\mathrm{ML}}$	0.235	0.290	0.291
ARL of c chart with UCL = 6 and ML estimates	441.5	503.0	501.3

Estimates for corrected data $\approx$ estimates for incomplete data.

Outlier: μ-estimate increased, α-estimate decreased.

Conclusions

. . . and Next Steps

Conclusions

- Even few outliers may have severe effect
  on estimates and chart design
  $\Rightarrow$ diligent removal of outliers from Phase I data.
- Above estimation approaches for time-dependent counts
  for incomplete data (without outliers) during Phase I.
- Even 25 % missing observations (due to outliers)
  no serious problem in chart design in Phase I:
  only minor effect on ARL_0 performance.
- Approach successful for real application.

- Parameter estimation from incomplete data
  for binomial AR(1) processes:
  Approaches by Yajima & Nishino (1999)
  and Weiß & Pollett (2012).
- Simulations for incomplete binomial data,
  performance of estimators.
- Additive outliers for binomial data with their finite range.
- Effect of missing observations and outliers on np chart.

Thank You

for Your Interest!

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

weissc@hsu-hh.de

Literature

Barczy et al. (2012): *Additive outliers in INAR(1) models*. Statist. Papers 53, 935-949.

Freeland & McCabe (2004): *Forecasting discrete valued low count time series*. Int. Journal of Forecasting 20, 427-434.

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

Montgomery (2009): *Introduction to statistical quality control*. 6th ed., Wiley, New York.

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

Weiß (2007): *Controlling correlated processes of Poisson counts*. Qual. Reliab. Engng. Int. 23, 741-754.

Weiß (2011): *Simultaneous confidence regions for the parameters of a Poisson INAR(1) model*. Statist. Methodology 8, 517-527.

Weiß & Kim (2013): *Parameter estimation for binomial AR(1) models with applications in finance and industry*. Statist. Papers 54, 563-590.

Weiß & Pollett (2012): *Chain binomial models and binomial autoregressive processes*. Biometrics 68, 815-824.

Yajima & Nishino (1999): *Estimation of the autocorrelation function of a stationary time series with missing observations*. Sankhyā A 61, 189-207.