Statistical inference for the $M/G/\infty$ queue

Alexander Goldenshluger

Department of Statistics

University of Haifa

Israel

Workshop on
"Data Driven Operations Management"

Eurandom, October 24–26, 2016

Outline

I. The $M/G/\infty$ estimation problem
   - problem formulation, examples
   - existing literature

II. Estimation of G from the arrival–departure data
   - bivariate arrival–departure and superposed point processes
   - estimators and their accuracy

III. Estimation of G from the queue–length data
   - the queue–length random process
   - estimator and its properties

IV. Comparison of estimators of G

V. Estimation of service time expectation and arrival rate

VI. Concluding remarks

I. The $M/G/\infty$ estimation problem: formulation and background

The $M/G/\infty$ estimation problem

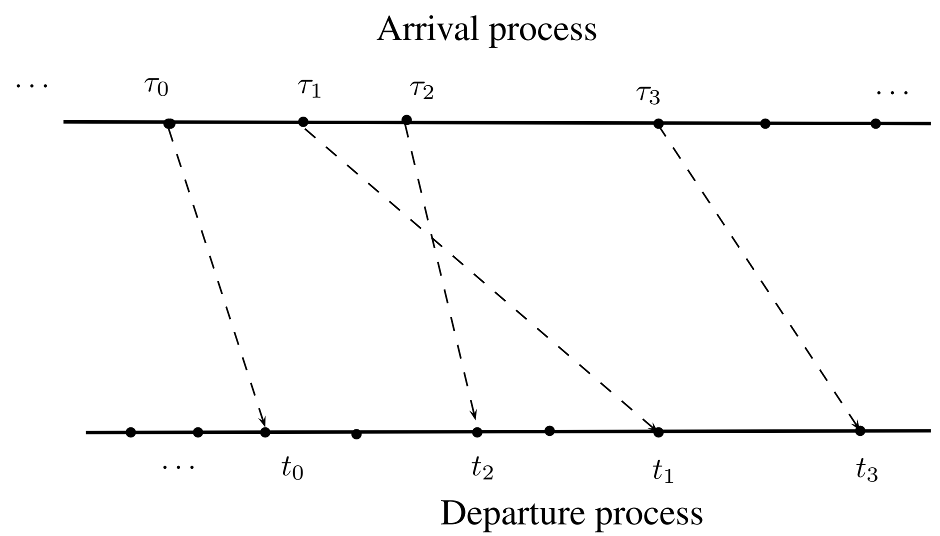
- Arrival process: customers come to a system according to homogeneous Poisson process of intensity λ.
- Service times: upon arrival, every customer obtains service and leaves the system after service completion. The service times are i.i.d. random variables, independent of the arrival process, with common distribution G.
- Observations: during some time period incomplete "arrival–departure" data or "number–of–busy–servers" recordings are given .
- Goal: estimate (make inference on) the service time distribution G and/or functionals thereof.

Observation schemes

- $(\tau_j)_{j\in\mathbb{Z}}$ are arrival epochs: homogeneous Poisson process of intensity λ on $\mathbb{R}$;
- $(\sigma_j)_{j\in\mathbb{Z}}$ are serivice times: i.i.d. random variables, independent of $(\tau_j)_{j\in\mathbb{Z}}$, with common distribution G.
- $(t_j)_{j\in\mathbb{Z}}$ are departure epochs: $t_j = \tau_j + \sigma_j$, $j \in \mathbb{Z}$.
- Arrival–departure data: for a given time interval we observe
  - (A): *arrival and departure epochs without matchings;*
  - (B): *superposed arrival–departure epochs without identification of the epoch type;*
- Queue–length data: for a given time interval we observe
  - (C): *queue–length (number–of–busy servers) process.*

Arrival–departure data

- The departure point process is obtained by translating the input points by i.i.d. random variables with distribution G.
  It is also Poisson process of intensity λ.

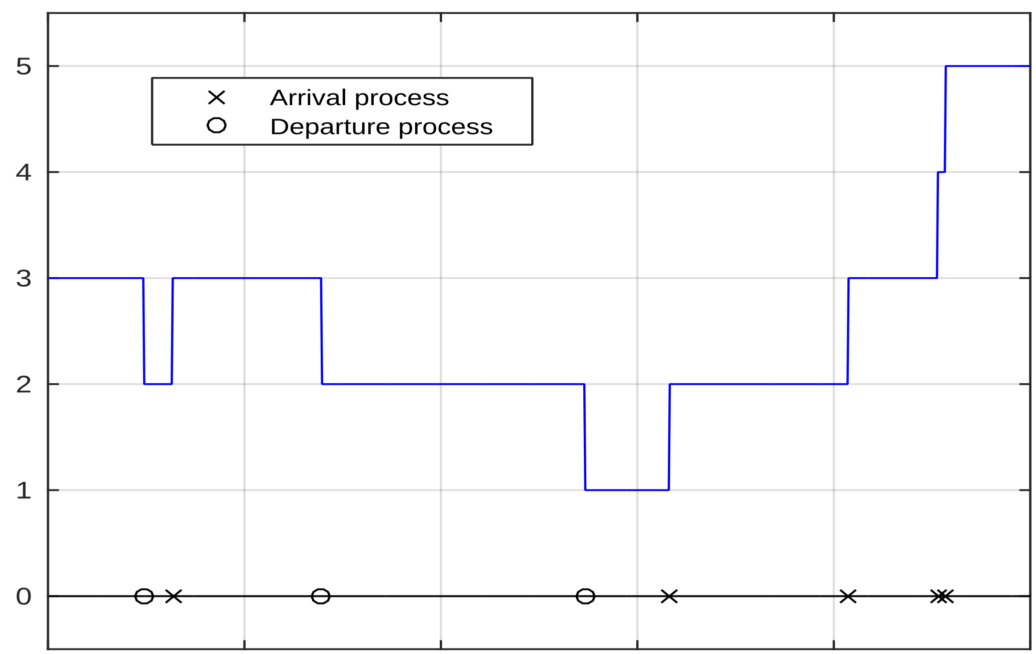


- (A): (τ_j), (t_j) are observed without correspondences (arrows);
- (B): epochs (s_j) of the superposed process are recorded without the epoch type.

Queue–length data

- (C): queue–length (number of busy servers) process $X(t)$,

$$X(t) = \sum_{j\in\mathbb{Z}} \mathbf{1}\{\tau_j \le t,\ \sigma_j > t - \tau_j\}, \quad t \in \mathbb{R}.$$



- Assume that $\{X(k\delta),\, k = 1, \ldots, n,\, T = n\delta\}$ is observed...

Applications

- The $M/G/\infty$ model is used in many applications:
  - Communication systems

    Beneš (1957), Mandjes & Zuraniewski (2011),...
  - Mobility of particles

    dates back to *Smoluchowski (1906);*

    Rothschild (1953), Lindley (1956), Bingham & Dunham (1997),...
  - Modelling a low density traffic

    Renyi (1964), Brown (1970), Petty et al. (1998),...

1. Existing literature: arrival–departure data

- "Sequence of differences" estimator of *Brown (1970)*
  - Associate each output point t_j in $[t_0, t_n]$ with the nearest input point τ_k to the left of t_j. Call the corresponding distances z_j, $j = 1, \ldots, n$.
  - The sequence $\{z_j\}$ is stationary and ergodic, z_j has distribution D:

$$D(x) = 1 - (1 - G(x))e^{-\lambda x} \quad \Leftrightarrow \quad G(x) = 1 - (1 - D(x))e^{\lambda x}.$$

  - Estimate D empirically using $z_1, \ldots, z_n$, and invert for G.
  - Consistency of the estimator is proved.

2. Existing literature: arrival–departure data

- Recent variations on Brown's idea
  - *Blanghaps, Nov & Weiss (2013):* an estimator can be based on distances to the rth nearest input point; consistency of the estimator is shown...
  - *Schweer & Wichelhaus (2015):* a Brown–type estimator is considered for a discrete queue model, and a functional central limit theorem is proved...

Existing literature: queue–length data

- Methods based on the relationship between correlation function of $\{X(t)\}$ and G: correlation function of $\{X(t)\}$ equals to the normalized integrated tail of G.
  - *Pickands & Stine (1997):* discrete model, standard time series methods for estimating correlations;
  - *Bingham & Pitts (1999):* standard time series methods for estimating the integrated normalized tail of G.
- Other observation schemes:
  - *Hall & Park (2004):* observations of durations of the busy periods.

Research questions

- Only consistency results in the setting (A) are available.
- Research questions partially answered in this talk:
  - ∗ how to construct estimators of G and/or functionals thereof under different observation schemes?
  - ∗ what is the achievable estimation accuracy in the original $M/G/\infty$ problem?

II. Estimation from the arrival–departure data

A model of random translations

- Input process: M is homogeneous Poisson of intensity λ

$$M := \sum_{j \in \mathbb{Z}} \epsilon_{\tau_j}, \quad \epsilon_x(A) := \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases} \quad x \in \mathbb{R},\ A \in \mathscr{B}.$$

- Output process: for (σ_j) independent of M,

$$N := \sum_{j \in \mathbb{Z}} \epsilon_{t_j}, \quad t_j = \tau_j + \sigma_j, \quad \sigma_j \stackrel{iid}{\sim} G,$$

$(\sigma_j)_{j \in \mathbb{Z}}$ are not necessarily non–negative random variables.

- Superposed process:

$$S = \sum_{j \in \mathbb{Z}} \epsilon_{s_j} := M + N.$$

Estimation problem

- Problem: estimate G on the basis of

  (A): a realization of the bivariate point process $(M,N)|_{\mathcal{T}}$, restricted to a time "window" $\mathcal{T} = \mathcal{T}_M \times \mathcal{T}_N$.

  (B): a realization of the superposed process $S|_{\mathcal{T}}$, restricted to a time window $\mathcal{T}$.

- In terms of the marked point process $\{s_j, \varkappa_j\}_{j\in\mathbb{Z}}$ where

$$\varkappa_j = \begin{cases} 1, & s_j \text{ is an input point,} \\ 2, & s_j \text{ is an output point} \end{cases}$$

  (A): corresponds to observation of $\{s_j, \varkappa_j\}$;

  (B): $\{s_j\}$ are observed, but the marks $\{\varkappa_j\}$ are not.

Some properties of (M, N)

- **Proposition 1:** *Let* $\{A_i\}_{i=1,\ldots,m}$ *and* $\{B_l\}_{l=1,\ldots,n}$ *be two families of disjoint intervals of the real line; then*

$$\log \mathrm{E}_G \exp\left\{\sum_{i=1}^m \eta_i M(A_i) + \sum_{l=1}^n \xi_l N(B_l)\right\} = \lambda \sum_{i=1}^m (e^{\eta_i} - 1)|A_i|$$

$$+ \lambda \sum_{l=1}^n (e^{\xi_l} - 1)|B_l| + \lambda \sum_{i=1}^m \sum_{l=1}^n (e^{\eta_i} - 1)(e^{\xi_l} - 1) Q(A_i, B_l),$$

where $|\cdot|$ *is the Lebesgue measure, and*

$$Q(A, B) := \int_A G(B - x)\mathrm{d}x. \tag{1}$$

- **Notation:** $G(I) := G(b) - G(a)$, for $I = (a, b]$, $a < b$.

Some properties of S

- Proposition 2: *Let* $\{A_i\}_{i=1,\ldots,m}$ *be disjoint intervals of* $\mathbb{R}$*; then for any* $\eta = (\eta_1, \ldots, \eta_m) \in \mathbb{R}^m$

$$\log \mathrm{E}_G \exp\Big\{\sum_{i=1}^m \eta_i S(A_i)\Big\} = 2\lambda \sum_{i=1}^m (e^{\eta_i} - 1)|A_i| + \lambda \sum_{i=1}^m \sum_{l=1}^m (e^{\eta_i} - 1)(e^{\eta_l} - 1) Q(A_i, A_l),$$

where $Q(\cdot,\cdot)$ *is defined in (1).*

- Remark: S is the Gauss–Poisson process, its p.g.f. is

$$\mathscr{G}_S(\eta) := \mathrm{E}_G\Big\{\prod_{j\in\mathbb{Z}} \eta(s_j)\Big\} = \mathrm{E}_G \exp\Big\{\int \log \eta(t) \mathrm{d}S(t)\Big\}$$

$$= \exp\Big\{2\lambda \int [\eta(t) - 1]\mathrm{d}t + \lambda \iint [\eta(t) - 1][\eta(\tau) - 1] Q(\mathrm{d}\tau, \mathrm{d}t)\Big\} .$$

Proof outline

- Step 1: conditioning on (τ_j):

$$\mathrm{E}_G\left[e^{\sum_{i=1}^n \eta_i M(A_i)+\sum_{l=1}^m \xi_l N(B_l)}\middle|(\tau_j)\right] = \exp\left\{\sum_{j\in\mathbb{Z}} f(\tau_j)\right\},$$

where

$$f(x) := \sum_{i=1}^n \eta_i \mathbf{1}_{A_i}(x) + \log\left[\sum_{l=1}^m (e^{\xi_l}-1)G(B_l-x)+1\right].$$

- Step 2: the use of Campbell's formula

$$\mathrm{E}_G \exp\left\{\sum_j f(\tau_j)\right\} = \exp\left\{\lambda\int_0^\infty [e^{f(x)}-1]\mathrm{d}x\right\}.$$

Covariance measures

- Corollary 1: *For any two intervals A and B one has*

$$\mathrm{E}_G\big[M(A)N(B)\big] = \lambda^2|A|\cdot|B| + \lambda Q(A,B),$$

and for $\mathrm{d}M(\tau) = M((\tau,\tau+\mathrm{d}\tau])$ *and* $\mathrm{d}N(t) = N((t,t+\mathrm{d}t])$

$$\mathrm{E}_G[\mathrm{d}M(\tau)\mathrm{d}N(t)] = \lambda^2\mathrm{d}\tau\mathrm{d}t + \lambda\mathrm{d}G(t-\tau)\mathrm{d}\tau.$$

- Corollary 2: *Let A_1 and A_2 be disjoint intervals; then*

$$\mathrm{E}_G[S(A_1)S(A_2)] = 4\lambda^2|A_1||A_2| + \lambda\big[Q(A_1,A_2) + Q(A_2,A_1)\big].$$

For $A_1 = (\tau,\tau+\mathrm{d}\tau]$ *and* $A_2 = (t,t+\mathrm{d}t]$ *with* $\tau \neq t$ *one has*

$$\mathrm{E}_G[\mathrm{d}S(\tau)\mathrm{d}S(t)] = 4\lambda^2\mathrm{d}\tau\mathrm{d}t + \lambda\big[\mathrm{d}G(t-\tau)\mathrm{d}\tau + \mathrm{d}G(\tau-t)\mathrm{d}t\big].$$

- The proof is by differentiation of formulas in Propositions 1, 2.

Important corollary

- Corollary 3: *For any function φ for which the integrals on the RHS are defined*

$$\mathrm{E}_G\Big[\iint \varphi(\tau,t)\mathrm{d}M(\tau)\mathrm{d}N(t)\Big] = \mathrm{E}_G\Big[\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\varphi(\tau_j,t_k)\Big]$$

$$= \lambda^2 \iint \varphi(\tau,t)\mathrm{d}\tau\mathrm{d}t + \lambda\iint \varphi(\tau,t)\mathrm{d}G(t-\tau)\mathrm{d}\tau, \qquad (2)$$

and for $\Omega := \{(\tau,t) : \tau \neq t\}$

$$\mathrm{E}_G\Big[\iint_\Omega \varphi(\tau,t)\mathrm{d}S(\tau)\mathrm{d}S(t)\Big] = 4\lambda^2\iint_\Omega \varphi(\tau,t)\mathrm{d}\tau\mathrm{d}t$$

$$+ \lambda\iint_\Omega \varphi(\tau,t)[\mathrm{d}G(t-\tau)\mathrm{d}\tau + \mathrm{d}G(\tau-t)\mathrm{d}t].$$

- Expression (2) appeared in *Mori (1975)* who attributes it to *Cox & Lewis (1972)*. We will call the resulting estimator [based on (2)] *the Cox–Lewis estimator*.

Estimator based on $(M, N)|_{\mathcal{T}}$

- Goal: estimate $G(I) = G(b) - G(a)$, $I := (a, b]$.
- Data: realization of $(M, N)|_{\mathcal{T}}$ restricted to

$$\mathcal{T} = [\tau_{\min}, \tau_{\max}] \times [\tau_{\min} + a, \tau_{\max} + b], \quad T := \tau_{\max} - \tau_{\min},$$

$$\mathscr{D}_{\mathcal{T}} = \Big\{(\tau_j : \tau_{\min} \leq \tau_j \leq \tau_{\max}),\ (t_k : \tau_{\min} + a \leq t_k \leq \tau_{\max} + b)\Big\}.$$

- Estimator: Let $\varphi_0(\tau, t) := \mathbf{1}_{[\tau_{\min}, \tau_{\max}]}(\tau)\mathbf{1}_I(t - \tau)$; and

$$\begin{aligned}\widehat{G(I)} &:= \frac{1}{\lambda T} \iint \varphi_0(\tau, t)\mathrm{d}M(\tau)\mathrm{d}N(t) - \lambda|I| \\ &= \frac{1}{\lambda T} \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \mathbf{1}_{[\tau_{\min}, \tau_{\max}]}(\tau_j)\mathbf{1}_I(t_k - \tau_j) - \lambda|I|.\end{aligned}$$

Accuracy of $\widehat{G(I)}$

- Theorem 1: *For any G the estimator $\widehat{G(I)}$ of $G(I)$ is unbiased, and*

$$\mathrm{var}_G\{\widehat{G(I)}\} = \tfrac{2\lambda|I|}{T}\Big\{|I| + \int_{-T}^{T} G(I+u)\big(1 - \tfrac{|u|}{T}\big)\mathrm{d}u - \tfrac{|I|^2}{6T}\Big\}$$

$$+ \tfrac{|I|}{T} + \tfrac{2}{T}|I|G(I) + \tfrac{1}{T}\int_{-T}^{T} G(I+u)G(I-u)\big(1 - \tfrac{|u|}{T}\big)\mathrm{d}u$$

$$+ \tfrac{2}{T}\int_{0}^{|I|} [G(I) + G(b-u) - G(a+u)]\big(1 - \tfrac{u}{T}\big)\mathrm{d}u \; + \; \tfrac{G(I)}{\lambda T}.$$

In the $M/G/\infty$ setting...

- In the $M/G/\infty$ setting $G(0)=0$, $[\tau_{\min},\tau_{\max}]=[0,T]$, $I=(0,x_0]$, so that the estimator is given by

$$\hat{G}(x_0) \;=\; \frac{1}{\lambda T}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\mathbf{1}_{[0,T]}(\tau_j)\mathbf{1}_{[0,x_0]}(t_k-\tau_j)-\lambda x_0. \qquad (3)$$

- Theorem 2: *The estimator* $\hat{G}(x_0)$ *is unbiased and*

$$\begin{aligned}\mathrm{var}_G\{\hat{G}(x_0)\} \;=\; & \frac{2\lambda x_0}{T}\Big\{x_0+\int_{-T}^{T}[G(x_0+u)-G(u)]\big(1-\tfrac{|u|}{T}\big)\mathrm{d}u-\frac{x_0^2}{6T}\Big\} \\ & +\tfrac{2}{T}x_0G(x_0)+\tfrac{1}{T}\int_{-T}^{T}[G(x_0+u)-G(u)][G(x_0-u)-G(-u)]\big(1-\tfrac{|u|}{T}\big)\mathrm{d}u \\ & +\frac{x_0}{T}+\frac{2}{T}\int_0^{x_0}[G(x_0)+G(x_0-u)-G(u)]\big(1-\tfrac{u}{T}\big)\mathrm{d}u+\frac{G(x_0)}{\lambda T}.\end{aligned}$$

Estimator based on $S|_{\mathcal{T}}$

- **Theorem 3:**

Let $G(0)=0$, $\varphi_*(\tau,t) := \mathbf{1}_{[0,T]}(\tau)\mathbf{1}_{(0,x_0]}(t-\tau)$, *and*

$$\begin{aligned}\tilde{G}(x_0) &= \frac{1}{\lambda T}\iint \varphi_*(\tau,t)\mathrm{d}S(\tau)\mathrm{d}S(t) - 4\lambda x_0 \\ &= \frac{1}{\lambda T}\sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \mathbf{1}_{[0,T]}(s_j)\mathbf{1}_{(0,x_0]}(s_k - s_j) - 4\lambda x_0.\end{aligned}$$

Then $\tilde{G}(x_0)$ *is an unbiased estimator of* $G(x_0)$, *and*

$$\mathrm{var}_G\{\tilde{G}(x_0)\} = \frac{1}{T}R_T^{(1)}(\lambda,x_0;G) + \frac{1}{T^2}R_T^{(2)}(\lambda,x_0;G),$$

where $R_T^{(1)}$ *and* $R_T^{(2)}$ *are positive functions satisfying*

$$R_T^{(1)}(\lambda,x_0;G) \le 76\lambda x_0^2 + 36x_0 + \frac{1}{\lambda}G(x_0), \quad R_T^{(2)}(\lambda,x_0;G) \le 36\lambda x_0^3.$$

Remarks

- Exact expressions for $R_T^{(1)}(\lambda, x_0; G)$ and $R_T^{(2)}(\lambda, x_0; G)$ are available...

- Qualitatively the behavior of $\tilde{G}(x_0)$ is similar to that of $\hat{G}(x_0)$:

  – the rate of convergence is parametric $O\left(\frac{1}{T}\right)$ in terms of dependence on the observation horizon T;

  – the accuracy deteriorates with growth of λ and x_0;

  – if λ and T are large, the leading term is $\sim \lambda x_0^2/T$.

- No conditions on G: e.g., it can have infinite expectation.

Can we do better?

III. Estimation from the queue–length data

Queue–length data

- The queue–length (number of busy servers) process $X(t)$:

$$X(t) = \sum_{j\in\mathbb{Z}} \mathbf{1}\{\tau_j \le t,\ \sigma_j > t-\tau_j\},\ \ t\in\mathbb{R}.$$

- The queue–length data contains more information than the arrival–departure data:
  - from observation of $\{X(t)\}$ one can reconstruct the arrival and departure epochs;
  - if the arrival and departure epochs and the initial state of the system are known then the queue–length process can be reconstructed.
- The available data: $\{X(k\delta),\, k=1,\ldots,n,\, T=n\delta\}$ for some small $\delta>0$.

Properties of the queue–length process

Define:

$$\tfrac{1}{\mu} := \int_0^\infty [1 - G(t)]\mathrm{d}t, \quad \rho := \tfrac{\lambda}{\mu}, \quad H(t) := \mu \int_t^\infty [1 - G(x)]\mathrm{d}x.$$

Proposition 3:

(a) $X(t) \sim \mathrm{Poisson}(\rho)$, $\forall t \in \mathbb{R}$, *and* $\{X(t), t \in \mathbb{R}\}$ *is stationary with*

$$\mathrm{cov}_G\{X(t), X(s)\} = \rho H(t - s), \quad \forall t, s \in \mathbb{R}.$$

(b) Let $X = (X(\delta), \ldots, X(n\delta))$*; then for any* $\theta \in \mathbb{R}^n$

$$\log \mathrm{E}_G\left[\exp\{\theta^T X\}\right] = \rho S_n(\theta),$$

$$S_n(\theta) := \sum_{k=1}^{n} (e^{\theta_k} - 1) + \sum_{k=1}^{n-1} H(k\delta) \sum_{m=k}^{n-1} \left(e^{\theta_{m-k+1}} - 1\right) e^{\sum_{i=m-k+2}^{m} \theta_i} \left(e^{\theta_{m+1}} - 1\right).$$

Remarks

- Statement (a) is well known...
- Statement (b) for $n = 2, 3$ appears in *Beneš (1957)*, *Lindley (1956)*. To the best of our knowledge, the case of general n is new.
- There is a nice structure in the formula:
  Let $1 \le i \le j \le k \le m \le n$; then

$$\frac{1}{\rho} \ln \mathrm{E}_G \big[\exp\{\theta_1 X_i + \theta_2 X_j + \theta_3 X_k + \theta_4 X_m\} \big] = \textstyle\sum_{i=1}^{4} (e^{\theta_\ell} - 1)$$

$$+ H_{j-i}(e^{\theta_1} - 1)(e^{\theta_2} - 1) \; + \; H_{k-i}(e^{\theta_1} - 1)e^{\theta_2}(e^{\theta_3} - 1)$$

$$+ H_{m-i}(e^{\theta_1} - 1)e^{\theta_2+\theta_3}(e^{\theta_4} - 1) \; + \; H_{k-j}(e^{\theta_2} - 1)(e^{\theta_3} - 1)$$

$$+ H_{m-j}(e^{\theta_2} - 1)e^{\theta_3}(e^{\theta_4} - 1) + H_{m-k}(e^{\theta_3} - 1)(e^{\theta_4} - 1),$$

  where the suffix notation is used $X_i := X(i\delta)$, $H_i := H(i\delta)$.

Proposition 3: proof outline

- Step 1: conditioning on $\{\tau_j, j \in \mathbb{Z}\}$ one can show that

$$\mathrm{E}_G\big[\exp\{\theta^T X\}\big] = \mathrm{E}_G\Big[\exp\Big\{\sum_{j\in\mathrm{Z}} f(\tau_j)\Big\}\Big],$$

$$f(x) := \ln\Big\{1 + \sum_{k=1}^{m}\Big[\exp\Big\{\sum_{i=1}^{k}\theta_i \mathbf{1}(x \le t_i)\Big\} - 1\Big]\mathrm{P}_G[\sigma_j \in I_k(x)]\Big\},$$

  where $I_0(x) = (-\infty, t_1 - x]$, $I_m(x) = (t_m - x, \infty)$, and $I_k(x) = (t_k - x, t_{k+1} - x]$ for $k = 1, \ldots, m-1$, $t_i := i\delta$.

- Step 2: Application of Campbell's formula for Poisson processes,

$$\mathrm{E}_G\Big[\exp\Big\{\sum_{j\in\mathbb{Z}} f(\tau_j)\Big\}\Big] = \exp\Big\{\lambda\int_{-\infty}^{\infty}[e^{f(x)} - 1]\mathrm{d}x\Big\}.$$

Idea of estimator construction

- Covariance function of $\{X(t)\}$:

$$\begin{aligned} R(t) &:= \operatorname{cov}_G\{X(s), X(s+t)\} = \rho H(t) \\ &= \rho \cdot \frac{\int_t^\infty [1 - G(u)]\mathrm{d}u}{\int_0^\infty [1 - G(u)]\mathrm{d}u} = \lambda \int_t^\infty [1 - G(u)]\mathrm{d}u. \end{aligned}$$

Hence,

$$1 - G(t) = -\tfrac{1}{\lambda} R'(t), \quad t \in \mathbb{R}_+. \tag{4}$$

- The idea is to estimate the first derivative of the covariance function of $X(t)$ at point x_0, and then recover $G(x_0)$ from (4).

1. Estimator construction

- Estimators of $R_k := R(k\delta)$: $\hat\rho = \frac{1}{n}\sum_{i=1}^n X_i$, $X_i := X(i\delta)$,

$$\hat R_k := \frac{1}{n-k}\sum_{t=1}^{n-k}(X_t - \hat\rho)(X_{t+k} - \hat\rho), \qquad k = 0, 1, \dots, n-1.$$

- Local window: for $h > 0$ let $D_x := [x-h, x+h]$, $\forall x \in [h, T-h]$, and $M_{D_x} = \{k : k\delta \in D_x\}$, $N_{D_x} = \#\{M_{D_x}\}$.

- Differentiating filter: fix integer $\ell > 0$, real $h \geq \frac{1}{2}(\ell+2)\delta$, and let $\{a_k(x), k \in M_{D_x}\}$ be the solution to

$$\begin{aligned}
\min \quad & \sum_{k\in M_{D_x}} a_k^2(x) \\
\text{s.t.} \quad & \sum_{k\in M_{D_x}} a_k(x) = 0, \\
& \sum_{k\in M_{D_x}} a_k(x)(k\delta)^j = jx^{j-1}, \quad j = 1, \dots, \ell.
\end{aligned} \qquad (\mathsf{Opt}(x))$$

2. Estimator construction

- Remarks
  - $h \geq \frac{1}{2}(\ell+2)\delta$ ensures at least $\ell+1$ grid points in M_{D_x}.
  - The filter reproduces the first derivative of any polynomial of degree $\leq \ell$:

$$\sum_{k \in M_{D_x}} a_k(x) p(k\delta) = p'(x), \quad \forall p : \deg(p) \leq \ell.$$

- Estimator of $G(x_0)$:

$$\tilde{G}_h(x_0) = 1 + \frac{1}{\lambda} \sum_{k \in M_{D_{x_0}}} a_k(x_0) \hat{R}_k.$$

- Two design parameters to be chosen:

  degree of the fitted polynomial ℓ and window width h.

Functional class

- Local smoothness: let $\beta > 0$, $L > 0$ and $I \subset (0, \infty)$ be a closed interval containing x_0. We say that $G \in \mathcal{H}_\beta(L, I)$ if

$$|G^{(\lfloor\beta\rfloor)}(x) - G^{(\lfloor\beta\rfloor)}(y)| \leq L|x-y|^{\beta-\lfloor\beta\rfloor}, \ \ \forall x, y \in I,$$

where $\lfloor\beta\rfloor := \max\big\{k \in \{0,1,2,\dots\big\} : k < \beta\}$.

- Tail (moment) conditions: we say that $G \in \mathcal{M}_p(K)$ with $p \geq 1$, $K > 0$ if

$$\mathrm{E}_G[\sigma^p] = \int_0^\infty p x^{p-1}[1 - G(x)]\mathrm{d}x \leq K < \infty.$$

If $G \in \mathcal{M}_2(K)$ then $\{H(i\delta)\}$ is summable $\Rightarrow$ *short–range dependence*.

- Functional class: we consider

$$\Sigma_\beta = \Sigma_\beta(L, I, K) := \mathcal{H}_\beta(L, I) \cap \mathcal{M}_2(K).$$

Upper bound

- Theorem 4:

Let $I = [x_0 - d, x_0 + d] \subset [0, (1 - \varkappa)T]$ *for some* $\varkappa \in (0, 1)$. *Let* $\tilde{G}_*(x_0)$ *be the estimator* $\tilde{G}_{h_*}(x_0)$ *associated with*

$$\ell \geq \lfloor \beta \rfloor + 1, \quad h_* = \left[\frac{K}{L^2 \varkappa T} \left(1 + \tfrac{1}{\lambda}\right) \right]^{1/(2\beta+2)}.$$

If

$$\frac{K}{L^2 \varkappa} \left(1 + \tfrac{1}{\lambda}\right) d^{-2\beta-2} \;\leq\; T \;\leq\; \frac{K}{L^2 \varkappa} \left(1 + \tfrac{1}{\lambda}\right) \left[\frac{2}{(\ell + 2)\delta} \right]^{2\beta+2}$$

then

$$\sup_{G \in \Sigma_\beta} \left[\mathrm{E}_G |\tilde{G}_*(x_0) - G(x_0)|^2 \right]^{1/2} \;\leq\; C(\ell) L^{1/(\beta+1)} \left[\frac{K}{\varkappa T} \left(1 + \tfrac{1}{\lambda}\right) \right]^{\beta/(2\beta+2)}.$$

Remarks

- Under local smoothness and second moment conditions:

$$\mathrm{Risk}_{x_0}[\tilde{G}_*;\Sigma_\beta] := \sup_{G\in\Sigma_\beta}\left[\mathrm{E}_G|\tilde{G}_*(x_0)-G(x_0)|^2\right]^{1/2}$$

$$\asymp C\left[\tfrac{1}{T-x_0}(1+\tfrac{1}{\lambda})\right]^{\beta/(2\beta+2)}, \quad T\to\infty.$$

- The rate of convergence is nonparametric, but dependence on λ and x_0 is "weak".

- What about optimality of this estimator?

Gaussian approximation in heavy traffic

- Corollary to Proposition 3:

  Let $\{M_\ell/G/\infty,\ \ell = 1, 2, \ldots\}$ *be a sequence of the* $M/G/\infty$ *systems with fixed* G *and arrival rates* $\lambda_\ell = \ell\lambda$, $\lambda > 0$. *Let* $X_\ell^n = (X_\ell(\delta), \ldots, X_\ell(n\delta))$ *be observations of the queue–length process in the* ℓ*th system; then*

$$\frac{X_\ell^n - \ell\rho e_n}{\sqrt{\ell\rho}} \xrightarrow{d} \mathcal{N}_n(0, V(H)), \ \ \ell \to \infty,$$

  where $\rho = \frac{\lambda}{\mu}$, $e_n = (1, \ldots, 1) \in \mathbb{R}^n$, $V(H) = \{H((i-j)\delta)\}_{i,j=1,\ldots,n}$.

- This result is in line with general results of *Borovkov (1967), Iglehart (1973) and Whitt (1974)* on weak convergence for queues.

A Gaussian model

- In heavy traffic $\{X(t)\}$ is close to a stationary Gaussian process. By (4), in the heavy traffic limit, $1-G$ is the negative derivative of the covariance function.
- A problem for stationary Gaussian process:

  Let $\{X(t),\ t\in\mathbb{R}\}$ be a stationary Gaussian process with zero mean and covariance function γ. We observe $X^n=(X(\delta),\dots,X(n\delta))$, $t_i=i\delta$, $i=1,\dots,n$, $n\delta=T$.
- The goal is to estimate $\theta=\gamma'(x_0)$ using observation X^n. We are mainly interested in lower bounds on the minimax risk

$$\mathrm{Risk}^*_{x_0}[\Gamma]=\inf_{\hat\theta}\sup_{\gamma\in\Gamma}\left[\mathrm{E}_\gamma\left|\hat\theta-\gamma'(x_0)\right|^2\right]^{1/2},$$

  where Γ is a suitable class of covariance functions.

Lower bound in the Gaussian problem

- **Definition:** Let $I = [x_0 - d, x_0 + d]$, $L > 0$ and $\beta > 0$. We say that $\gamma \in \Gamma_\beta := \Gamma_\beta(L, I, K)$ if

  (i) $\int_{-\infty}^{\infty} |\gamma(t)|\mathrm{d}t \le K < \infty$;

  (ii) γ is $\ell := \max\{k \in \mathbb{N} : k < \beta + 1\}$ times continuously differentiable on I and

$$|\gamma^{(\ell)}(x) - \gamma^{(\ell)}(y)| \le L|x - y|^{\beta+1-\ell}, \;\; \forall x, y \in I.$$

- **Theorem 5:**

  There exist positive constants C_1, C_2 and c depending on β, x_0, d and K only such that if

$$C_1\delta^{-2} \le T, \qquad L^2T \le C_2\delta^{-2\beta-2}$$

  then

$$\liminf_{T\to\infty} \left\{ L^{-1/(\beta+1)} T^{\beta/(2\beta+2)} \, \mathrm{Risk}^*_{x_0}[\Gamma_\beta] \right\} \ge c > 0.$$

Remarks

- The lower bound in the Gaussian model strongly suggests that the "queue–length"–based estimator is rate optimal in the heavy traffic regime...
- In the original model derivation of lower bounds on the risk is difficult because of the dependence structure. The distribution of observations is not available in a usable form.

IV. Comparison of estimators of G

Three estimators of G

Monotonized and confined to $[0,1]$ versions of the estimators:

- The Cox–Lewis estimator

$$\hat{G}_{\mathrm{CL}}(x_0) = \frac{1}{\lambda T} \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \mathbf{1}_{[0,T]}(\tau_j) \mathbf{1}_{[0,x_0]}(t_k - \tau_j) - \lambda x_0.$$

- Local polynomial estimator:

$$\hat{G}_{\mathrm{LP}}(x_0) = 1 + \frac{1}{\lambda} \sum_{k \in M_{D_{x_0}}} a_k(x_0) \hat{R}_k.$$

- Brown's estimator:

$$\hat{G}_B(x_0) = 1 - e^{\lambda x_0} \frac{\sum_{k \in \mathbb{Z}} \mathbf{1}_{(x_0,\infty)}(z_k) \mathbf{1}_{[0,T]}(t_k)}{\sum_{k \in \mathbb{Z}} \mathbf{1}_{[0,T]}(t_k)} \; .$$

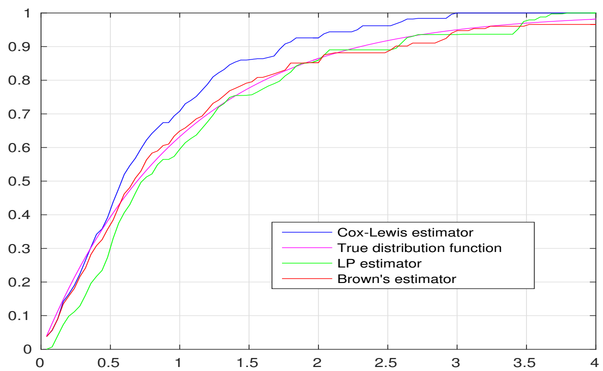
Numerical experiments

- The goal: study influence of the arrival rate λ and the tail of G on accuracy.
- Experiment 1: $T = 1000$, $G(x) = 1 - e^{-x}$, $\lambda \in \{0.5, 1, 5, 15\}$. The distribution G is estimated at equidistant points on $[0, 4]$.
- Experiment 2: $T = 1000$, $\lambda = 1$, $G(x) = 1 - e^{-\mu x}$ with $\mu \in \{\frac{1}{2}, \frac{1}{5}, \frac{1}{10}, \frac{1}{15}\}$. The distribution G is estimated at equidistant points on $[0, 10]$.
- The bandwidth of $\hat{G}_{\rm LP}(x_0)$ was chosen minimal, $h = 3\delta$.
- In both experiments we compute the maximal error

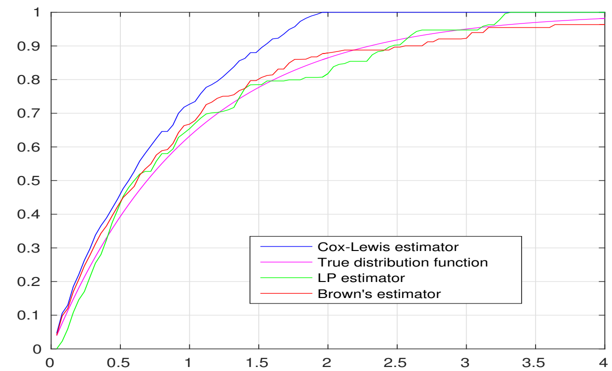
$$\mathrm{Err}\{\hat{G}\} = \max_{x \in \{x_i\}} |\hat{G}(x) - G(x)|, \quad \hat{G} \in \{\hat{G}_{\rm CL}, \hat{G}_{\rm LP}, \hat{G}_{\rm B}\}$$

over 100 simulation runs.

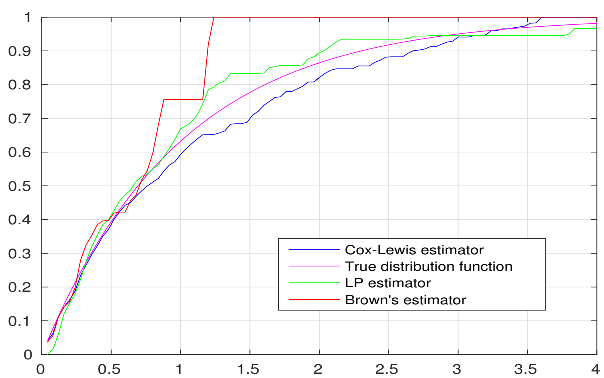
Experiment 1: typical realizations



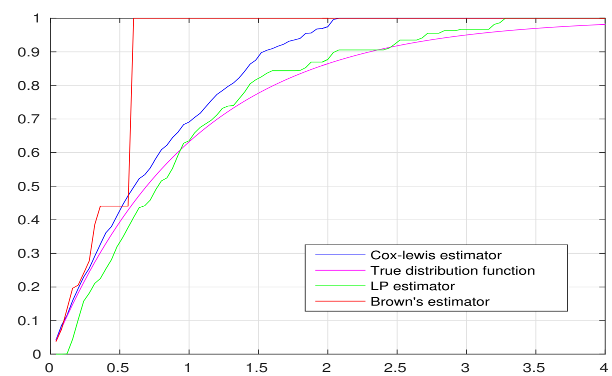
(a) $\lambda = 0.5$



(b) $\lambda = 1$

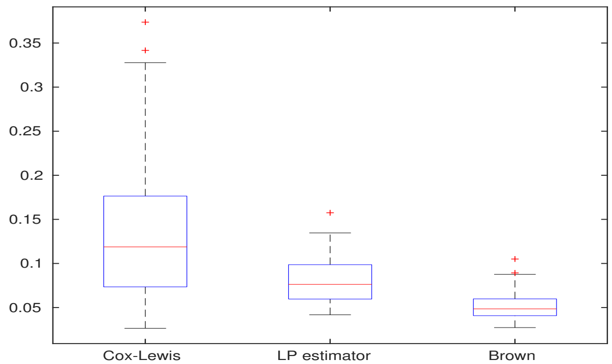


(c) $\lambda = 5$

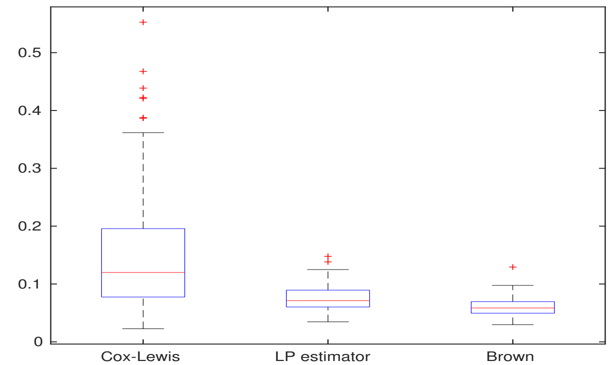


(d) $\lambda = 15$

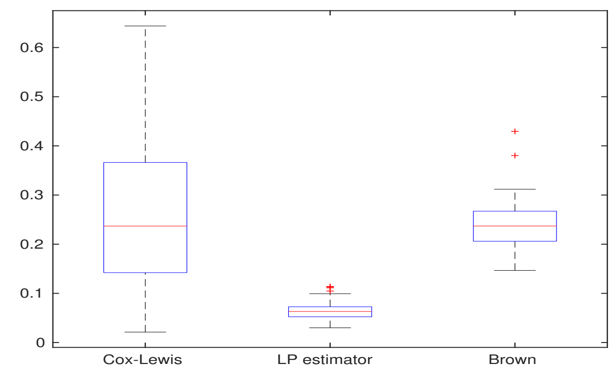
Experiment 1: accuracy boxplots



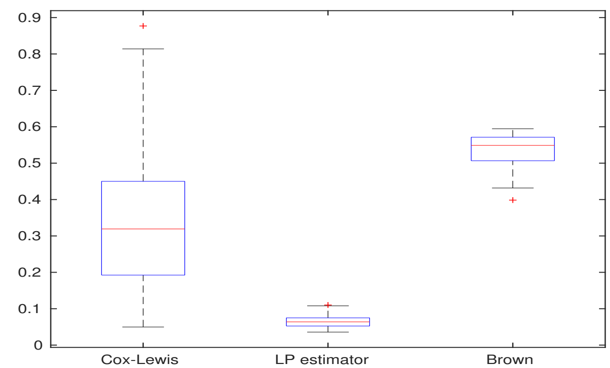
(a) $\lambda = 0.5$



(b) $\lambda = 1$

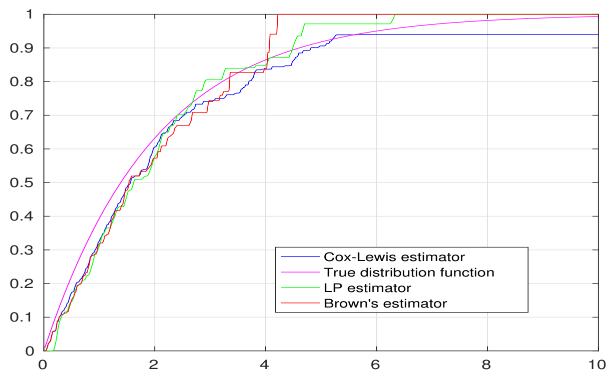


(c) $\lambda = 5$

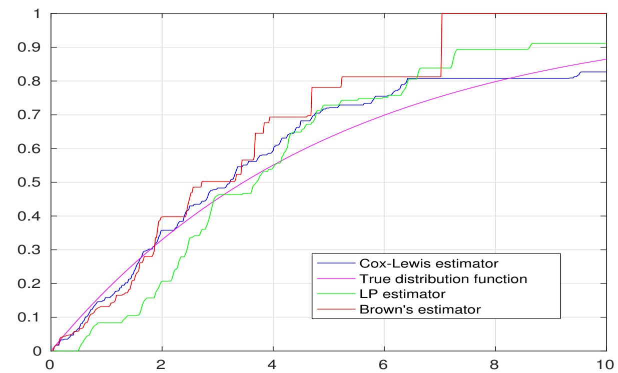


(d) $\lambda = 15$

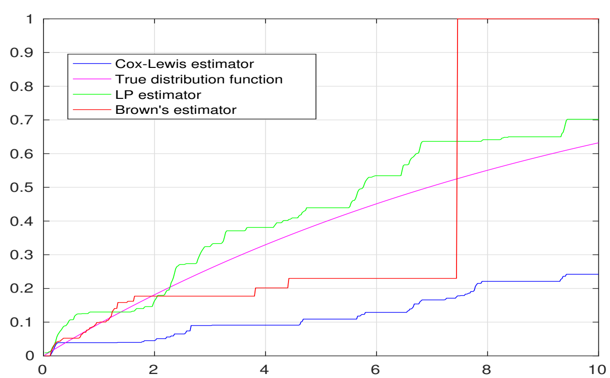
Experiment 2: typical realizations



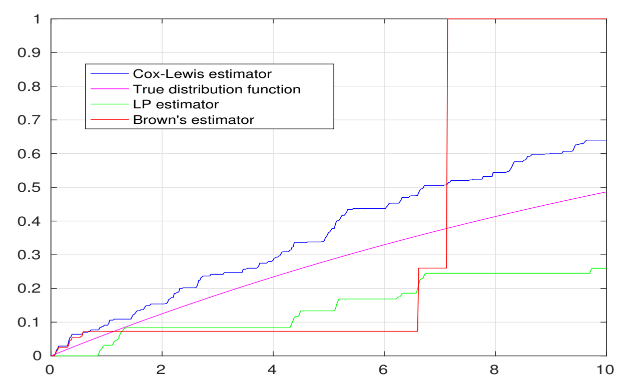
(a) $\mu = \frac{1}{2}$



(b) $\mu = \frac{1}{5}$

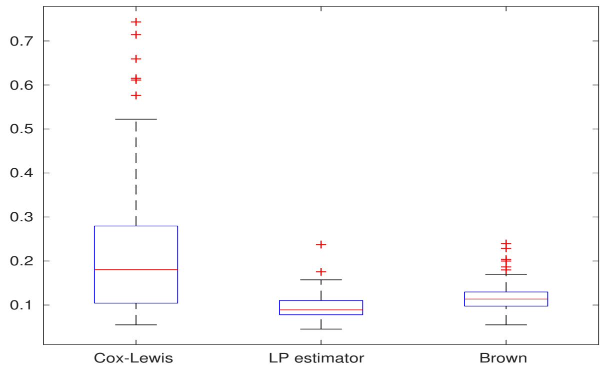


(c) $\mu = \frac{1}{10}$

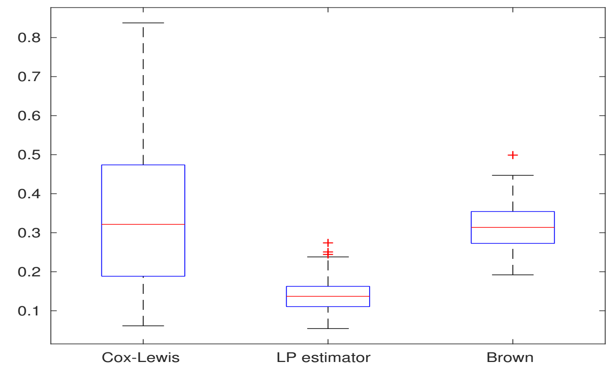


(d) $\mu = \frac{1}{15}$

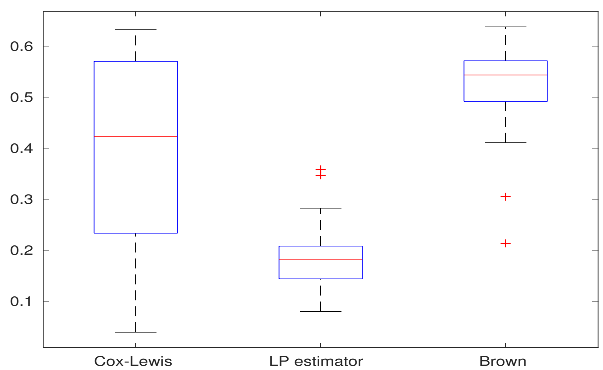
Experiment 2: accuracy boxplots



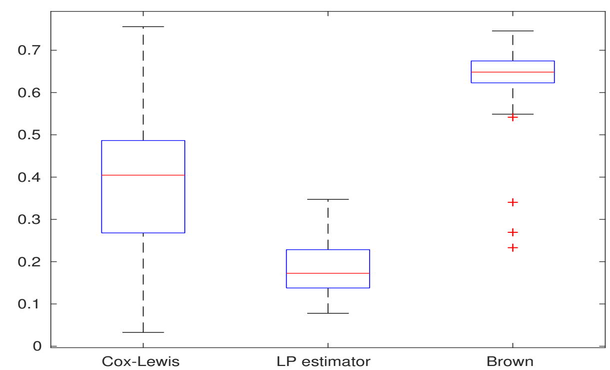
(a) $\mu = \frac{1}{2}$



(b) $\mu = \frac{1}{5}$



(c) $\mu = \frac{1}{10}$



(d) $\mu = \frac{1}{15}$

Comparison of estimators

- Experiment 1:
  - $\hat{G}_{\mathrm{CL}}$ and $\hat{G}_{\mathrm{B}}$ behave poorly for large values of λ and x_0. In particular, $\hat{G}_{\mathrm{B}}$ is completely upset for $\lambda = 15$, although it behaves well for small λ. $\hat{G}_{\mathrm{CL}}$ exhibits more variability than $\hat{G}_{\mathrm{B}}$.
  - $\hat{G}_{\mathrm{LP}}$ is stable and even improves with growth of λ.
- Experiment 2:
  - Accuracy of all estimators is badly affected by heavy tails of G. $\hat{G}_{\mathrm{B}}$ is most sensitive, and $\hat{G}_{\mathrm{LP}}$ is most stable.

V. Estimation of service time expectation and arrival rate

Estimation of service time expectation

- Estimation from the arrival–departure data
  Let $\alpha := \frac{1}{\mu} = \mathrm{E}_G[\sigma]$, and $\hat{G}$ be the Cox–Lewis estimator of G.
  For $b > 0$ let
  $$\hat{\alpha} = \int_0^b [1 - \hat{G}(t)]\mathrm{d}t.$$
  Let $\mathcal{M}_p(A)$ be the set of distributions with pth moment $\leq A$,
  $$\mathcal{M}_p(A) := \Big\{ G : p \int_0^\infty x^{p-1}[1 - G(x)]\mathrm{d}x \leq A < \infty \Big\},\ p > 1.$$

- Theorem 6: *Let $\hat{\alpha}_*$ be the estimator associated with $b = b_* := (A/p)^{1/(p+1)}(T/\lambda)^{1/(2p+2)}$. Then for all $T \geq \lambda(1 \vee \lambda^{-2})^{2p+2}(A/p)^2$ one has*
  $$\sup_{G \in \mathcal{M}_p(A)} \mathrm{E}_G|\hat{\alpha}_* - \alpha|^2 \leq C\big(\tfrac{A}{p}\big)^{4/(p+1)}\big(\tfrac{\lambda}{T}\big)^{(p-1)/(p+1)},$$
  where C is an absolute constant.

Remarks

- The rate of convergence is nonparametric, $O(T^{-\frac{p-1}{2p+2}})$.
  We were not able to construct an estimator whose risk converges at faster rate.
- Estimation of α from the queue–length data is immediate:

$$\tilde{\alpha} = \frac{\hat{\rho}}{\lambda} = \frac{1}{\lambda n}\sum_{i=1}^{n} X_i, \quad X_i := X(i\delta).$$

  It is easy to verify that

$$\sup_{G \in \mathcal{M}_p(A)} \mathrm{E}_G|\tilde{\alpha} - \alpha|^2 \le \frac{C(A,p)\alpha}{\lambda T}.$$

- Even though the difference between observations schemes (A) and (C) is only in the initial state of the queue, the results are completely different!

Estimation of arrival rate

- Arrival–departure data

  the problem is trivial: equivalent to estimating parameter of exponential distribution from i.i.d. samples.

- Continuous–time queue–length data

  the same as for arrival–departure data:

$$\begin{aligned}\hat{\lambda}^{+} &= \tfrac{1}{T}\cdot\#\big\{t\in(0,T] : X(t)-X(t-)=1\big\},\\ \hat{\lambda}^{-} &= \tfrac{1}{T}\cdot\#\big\{t\in(0,T] : X(t)-X(t-)=-1\big\}.\end{aligned}$$

  It is immediate to show that

$$\mathrm{E}_{G,\lambda}\big|\hat{\lambda}^{+}-\lambda\big|^{2}=\mathrm{E}_{G,\lambda}\big|\hat{\lambda}^{-}-\lambda\big|^{2}=\lambda T^{-1},\quad \forall\lambda,\forall G.$$

What about discrete observations of the queue–length process?

1. Estimation of arrival rate λ

- Problem: estimate λ from $X^n = \{X(k\delta), k = 1, \ldots, n\}$.
- Recall (4): if $R(x) = \mathrm{cov}_G\{X(t), X(t+x)\}$ then

$$1 - G(x) = -\tfrac{1}{\lambda} R'(x) \quad \Rightarrow \quad \lambda = -R'(0).$$

- Estimator: If $\hat{R}_k := \frac{1}{n-k} \sum_{t=1}^{n-k} (X_t - \hat{\rho})(X_{t+k} - \hat{\rho})$, $D_0 = [0, 2h]$, and $\{a_k(0), k \in M_{D_0}\}$ is the solution to $(\mathrm{Opt}(0))$ then we put

$$\hat{\lambda} := - \sum_{k \in M_{D_0}} a_k(0) \hat{R}_k.$$

Estimator depends on design parameters h and ℓ.

2. Estimation of arrival rate λ

- Theorem 7: Let $I = [0, 2d]$, and let $\hat{\lambda}_*$ be the estimator associated with $\ell \geq \lfloor\beta\rfloor + 1$, and $h = h_* := \left[\frac{K}{L^2 T}\right]^{1/(2\beta+2)}$. Let $KL^{-2}d^{-2\beta-2} \leq T \leq KL^{-2}\left[\frac{2}{(\ell+2)\delta}\right]^{2\beta+2}$; then

$$\sup_{G\in\Sigma_\beta(L,I,K)} \left[\mathrm{E}_{\lambda,G}|\hat{\lambda}-\lambda|^2\right]^{1/2} \leq C(\ell)(\lambda^2+\lambda)^{1/2}L^{1/(\beta+1)}\left[\frac{K}{T}\right]^{\beta/(2\beta+2)}.$$

- The proof coincides with that of Theorem 4.

- For discrete observations accuracy of $\hat{\lambda}^{\pm}$ is poor: one can show that

$$\mathrm{E}_{\lambda,G}|\hat{\lambda}^{\pm}-\lambda|^2 \leq C\left\{\lambda^4\delta^2 + \lambda^2\left[\frac{1}{\delta}\int_0^\delta G(x)\mathrm{d}x\right]^2 + \frac{\lambda}{T}\right\}.$$

Thus, $\hat{\lambda}$ may be preferable...

VI. Conclusion

A quote

- David R. Cox in "Some problems of statistical analysis connected with congestion", published in 1965, reviews some statistical problems for queues and writes:

  "There are a very large number of papers on particular probabilistic models for queues and, by comparison, extremely few papers on the corresponding problems of statistical analysis."

  "When a simple mathematical model is investigated primarily to get qualitative insight..., the statistical problems are not so relevant. When, however,... a practical congestion problem is tackled..., non–trivial statistical problems arise."

- *The first statement is still true...*
- *Statistical research for queueing models is important, today even much more important than in the past...*

1. Concluding remarks

- With abundance of data on service systems, statistical inference for stochastic models becomes more and more important...

- Statistical problems are challenging even for simplest queueing models:
  - observations are dependent;
  - joint distributions of observations are not available in a usable form...

- Fundamental statistical problem

  How to judge optimality of estimators in models in which joint distribution of observations is not available?

2. Concluding remarks

- There are very few results on accuracy of statistical procedures for queueing models.

- Even for simplest models many questions are unanswered.

  For instance, in the $M/G/\infty$ model

  – test for exponentiality of G ($M/M/\infty$ queue);

  – detect changes in arrival rate to the $M/G/\infty$ queue;

  – estimate nonparametrically the arrival rate to the $M_t/G/\infty$ queue;

  – ...

- Other queueing models?