The Effect of Data Aggregation on Statistical Inference

Thomas Crow

Supervisor: Ed Cohen

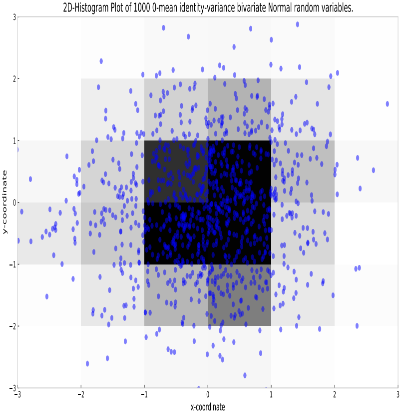
15th September 2021

Imperial College London

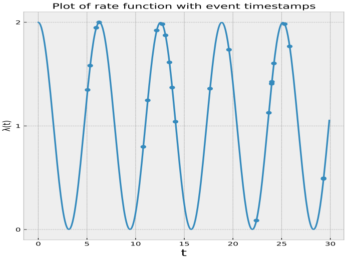
Table of Contents

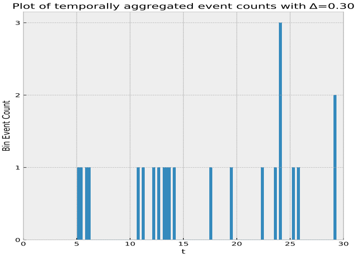
Motivation - Microscopy

- Estimating position of fluorescing molecule with microscopy.
- Images are captured using counts within pixels, not with infinite precision.
- Non-explicit likelihood functions derived in Ober, R.J. et al. (2004).



- Estimating parameters for rate function in Poisson process. E.g. estimating the number of incoming calls over the day to a telephone service.
- Can't store timestamps to infinite precision, may just have number of calls per minute, or hour.





General Approach

- Random variable X following some parametric distribution with density function $f(x;\theta)$ and distribution function $F(x;\theta)$.
- To calculate probability of an observation falling in Δ-sized bin B_k (with edges $(k\Delta, (k+1)\Delta]$):

$$\mathbb{P}(x \in B_k;\theta) = \int_{x=k\Delta}^{(k+1)\Delta} f(x)\mathrm{d}x = F((k+1)\Delta) - F(k\Delta).$$

- Use this probability with observed counts n_k in each bin to calculate binned log-likelihood:

$$\ell(\theta; n_{-\infty}, \ldots, n_{\infty}) = \sum_{k=-\infty}^{\infty} n_k \log\left(\mathbb{P}(x \in B_k;\theta)\right).$$

Likelihood Theory

- Estimating parameter θ given i.i.d data $x_1, \dots, x_n$. Maximise log-likelihood $\ell(\theta; x_1, \dots, x_n)$.
- MLE $\hat{\theta}$ at $\nabla_\theta \ell(\theta) = 0$.
- Hessian $\mathbf{H} = \nabla_\theta \nabla_\theta^T \ell(\theta)$.
- Fisher Information $\mathcal{I}(\theta) = -\mathbb{E}\left[\mathbf{H}\right]$.
- Cramér-Rao Lower Bound [Rao, C.R. (1965)]:

$$\text{Var}(\hat{\theta}) \geq \mathcal{I}^{-1}(\theta),$$

gives lower limit for asymptotic variance of MLE.

Identifiability

- Global identifiability is when there exists a unique mapping from parameter space to model space.
- Unidentifiability often occurs due to different parameters achieving the maximum likelihood.
- Fisher Information matrix is non-singular if and only if θ is locally identifiable, that is:

$$\ell(\theta; x_1, \dots, x_n) = \ell(\phi; x_1, \dots, x_n) \implies \theta = \phi,$$

in a neighbourhood around θ [Rothenberg, T.J. (1971)].

Exponential Distribution

- Random variable X follows Exponential(λ) distribution if:

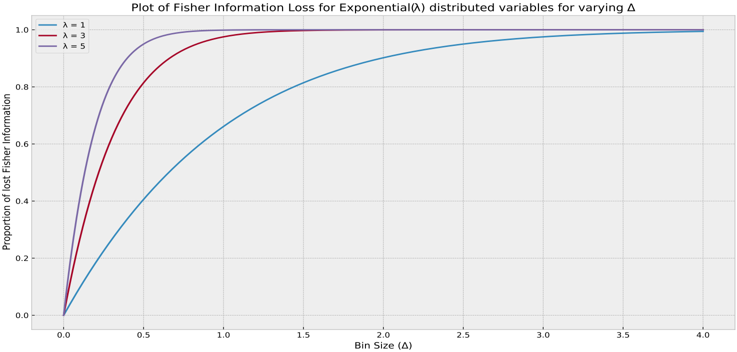
$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0, \lambda > 0.$$

- Fisher Information for λ is given in both cases as:

$$\mathcal{I}_{cts}(\lambda) = \frac{n}{\lambda^2}, \quad \mathcal{I}_{bin}(\lambda) = \frac{n\Delta^2}{(e^{\Delta\lambda} - 1)^2}.$$

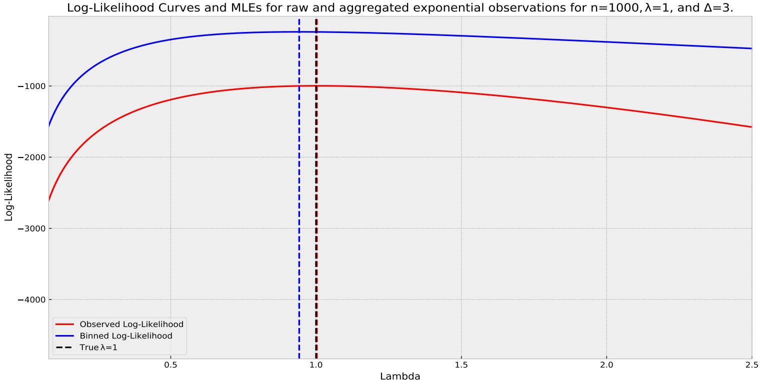
Exponential Fisher Information

- As Δ increases, proportion of lost Fisher Information quickly approaches 1.
- Smaller values of λ have greater variance, so aggregation has less relative effect.



Exponential Log-Likelihood Example

- Loss of Fisher Information means flattening of log-likelihood curve around the MLE.
- Results in a higher variance MLE due to the CRLB.



Normal Fisher Information - Aggregated Values

- Random variable X follows a $\mathcal{N}(\mu, \sigma)$ distribution if:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x, \mu \in \mathbb{R}, \sigma > 0.$$

- Fisher Information for μ given by:

$$\mathcal{I}_{cts}(\mu) = \frac{n}{\sigma^2},$$

and

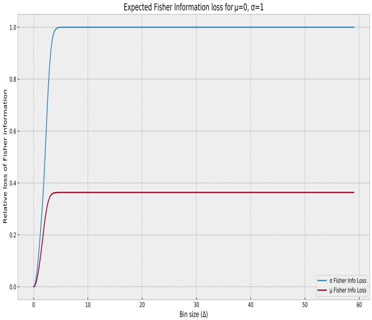
$$\mathcal{I}_{bin}(\mu) = \frac{n}{\sigma^2} \sum_{k=-\infty}^{\infty} \frac{\left(\phi\left(\frac{(k+1)\Delta-\mu}{\sigma}\right) - \phi\left(\frac{k\Delta-\mu}{\sigma}\right)\right)^2}{\Phi\left(\frac{(k+1)\Delta-\mu}{\sigma}\right) - \Phi\left(\frac{k\Delta-\mu}{\sigma}\right)}$$
$$-\frac{n}{\sigma^3} \sum_{k=-\infty}^{\infty} \left(((k+1)\Delta - \mu)\phi\left(\frac{(k+1)\Delta-\mu}{\sigma}\right) - (k\Delta-\mu)\phi\left(\frac{k\Delta-\mu}{\sigma}\right)\right).$$

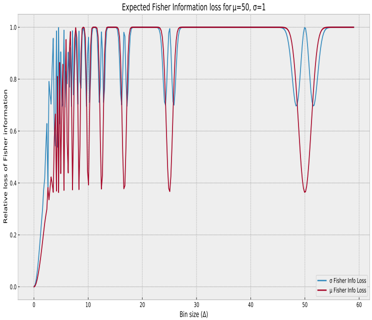
.

- Derivation also calculated for σ Fisher Information. Off-diagonals are 0.

Normal Fisher Information - Loss

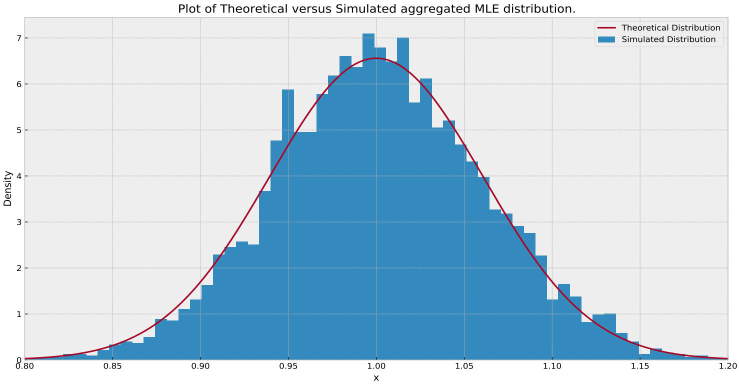
- Again, information loss increases as aggregation increases.
- Inference performance depends on both bin size (Δ) and alignment of the bin edges with μ.





Normal Aggregated MLE example

- Run 5000 simulations of toy inference problem with $X \sim \mathcal{N}(\mu = 1, \sigma = 1)$ with σ known, $\Delta = 3$, and $n = 500$.
- Variance of simulated results agree with inverse of derived Fisher Information for μ. Aggregated MLE is asymptotically efficient.



Poisson Process

- Poisson processes are a simple type of counting process defined by their rate function: $\lambda(t) \geq 0$.
- They are homogeneous if $\lambda(t) = \lambda$ is constant; and inhomogeneous if $\lambda(t)$ is deterministic and varies through time.
- They have an associated intensity, or mean, function defined over any subset A of the real line [Daley, D. J. and Vere-Jones, D. (2002)]:

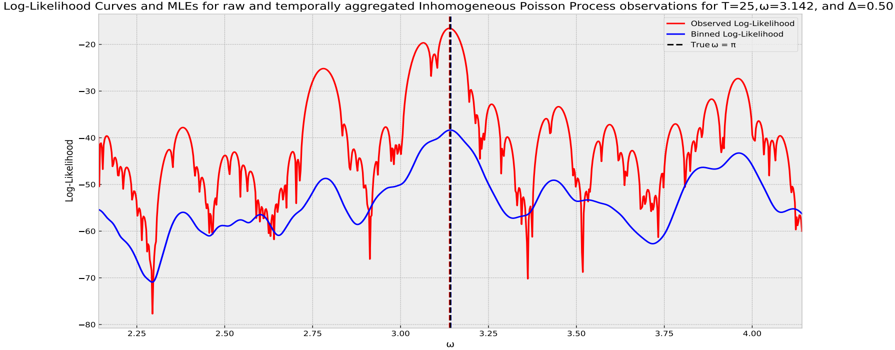
$$\Lambda(A) = \int_A \lambda(t) \mathrm{d}t.$$

Periodic Rate Poisson Process

- We investigate an inhomogeneous, periodic Poisson process with rate function:
$$\lambda(t;\omega) = 1 + \cos(\omega t).$$
- Nyquist rate is minimum sampling rate to avoid distortion in signal processing [Marks, R.J. (1991)]. Provides starting point for limit of resolution for periodic inference problems.
- For function with highest frequency B, Nyquist rate is $2B$. Frequency in this setting is $B = \omega/(2\pi)$; giving Nyquist rate of $2B = \omega/\pi$.
- We find this provides a safe upper-limit to the level of aggregation, setting $\Delta > 2B$ can result in model unidentifiability.

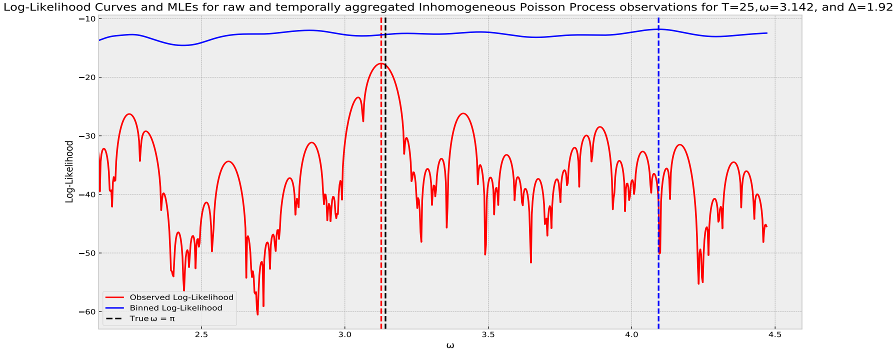
Periodic Rate Poisson Process - Identifiable

- Performing inference on $\omega = \pi$, with $T = 25, K = 50$ and so $\Delta = 0.5$.
- $\Delta = 0.5 < 2B = 1$, so we expect to be able to identify our model from the observations.
- Aggregated log-likelihood function has a clear sole maximum near the true value $\omega = \pi$, no issues with identifiability.



Periodic Rate Poisson Process - Unidentifiable

- Again we have $\omega = \pi$, with $T = 25$, now $K = 13$ and so $\Delta = 1.92$.
- $\Delta = 1.92 > 2B = 1$, expect issues with identifiability.
- Aggregated log-likelihood function has no clear sole maximum near the true value $\omega = \pi$. Curve is almost flat meaning we have an unidentifiable model. Inference can't be performed in this setting.



Conclusion

- General approach for calculating loss of Fisher Information in aggregated cases has been presented.
- Specific results derived for Poisson, Exponential, and Normal distributions, and for Poisson Processes.
- Aggregated MLE is asymptotically efficient, achieving CRLB.
- Issues exist with model identifiability due to over-aggregation with periodic functions.

Bibliography

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes*. Springer, New York, 2002.

Marks, R.J. *Introduction to Shannon Sampling and Interpolation Theory*. Springer-Verlag, Berlin, 1991.

Ober, R.J., Ram, S., and Ward, E.S. Localization accuracy in single-molecule microscopy. *Biophysical Journal*, pages 503–542, 2004.

Rao, C.R. *Linear Statistical Inference and its Applications*. John Wiley and Sons, New York, 1965.

Rothenberg, T.J. Identification in parametric models. *Econometrica*, 1971.