A DETAILS ON LEMMA 1

Before we proceed, we state a technical result:

Lemma 1. *Let $y \sim \mathcal{N}(\mu, \sigma^2)$ and $\varphi = (\mu/\sigma)^2$. Then*

$$\mathbb{E}_y[\ln y^2] = \ln(2\sigma^2) + \sum_{j=0}^{\infty} \frac{(\varphi/2)^j \exp(-\varphi/2)}{j!} \psi(j+1/2), \tag{1}$$

where $\psi(\cdot)$ is the digamma function.

Proof. Let $\tilde{y} = y/\sigma$, then the expectation can be calculated as

$$\begin{aligned}\mathbb{E}_y[\ln y^2] &= \int_{-\infty}^{\infty} \ln y^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big) dy \\ &= \ln \sigma^2 + \int_{-\infty}^{\infty} \ln \tilde{y}^2 \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(\tilde{y}-\mu/\sigma)^2}{2}\Big) d\tilde{y}.\end{aligned} \tag{2}$$

The second part has the form of $\mathbb{E}_{\bar{y}}[\ln \bar{y}^2]$, where $\bar{y} \sim \mathcal{N}(\mu/\sigma, 1)$. Let $w = \bar{y}^2$ and w follows a standard non-central chi-squared distribution with parameter $\varphi = (\mu/\sigma)^2$ (Famoye, 1995). The distribution of w is given as follows:

$$p(w) = \frac{e^{-\frac{w+\varphi}{2}}}{\sqrt{2w}} \sum_{j=0}^{\infty} \frac{(w\varphi/4)^j}{j!\Gamma(j+1/2)}. \tag{3}$$

The expectation of $\ln w$ then is

$$\begin{aligned}\mathbb{E}_w[\ln w] &= \int_0^{\infty} \ln w \frac{e^{-\frac{w+\varphi}{2}}}{\sqrt{2w}} \sum_{j=0}^{\infty} \frac{(w\varphi/4)^j}{j!\Gamma(j+1/2)} dw \\ &= \sum_{j=0}^{\infty} \frac{(\varphi/2)^j e^{-\varphi/2}}{j!} (\ln 2 + \psi(j+1/2)).\end{aligned} \tag{4}$$

Substituting this back yields the answer. □

B DETAILS ON LEMMA 2

Let us recall that

$$g_m(x) = \sum_{j=0}^{\infty} \frac{x^j \exp(-x)}{j!} \psi(j+m). \tag{5}$$

The derivative of $g_m(x)$ with respect to x is

$$\begin{aligned}g'_m(x) &= \sum_{j=0}^{\infty} \frac{(jx^{j-1} - x^j)\exp(-x)}{j!} \psi(j+m) \\ &= \sum_{j=0}^{\infty} \frac{x^j \exp(-x)}{j!} \frac{1}{j+m}.\end{aligned} \tag{6}$$

To prove the Lemma 2 in Section 4, we first present two results:

Lemma 2. *(Moser, 2007)*

$$g'_m(x) \geq \frac{1}{x+m},\ m \in \mathbb{N}^+, x > 0.$$

Note that the inequality holds when $m \in \mathbb{N}_+$. However, following the same lines of the proof, one can generalized their results for $m \in \mathbb{R}^+$, hence the proof is elided. In our case, we are interested in a bound when $m = \frac{1}{2}$. We state the following:

Lemma 3. *The following inequality holds:*

$$g_m(x) \geq \ln(x+m) + \psi(m) - \ln(m). \tag{7}$$

Proof. Since

$$\frac{1}{x+m} \leq g'_m(x), \tag{8}$$

integrating both sides yield

$$\begin{aligned}&\ln(x+m) - \ln m = \int_0^x \frac{1}{y+m} dy \\ &\leq \int_0^x g'_m(y) dy = g_m(x) - g_m(0) = g_m(x) - \psi(m).\end{aligned}$$

□

Lemma 4. *Let $x \sim \mathcal{N}(\mu, \sigma^2)$. Then we have*

$$\mathbb{E}_x[\ln x^2] \geq \ln(\mu^2 + b\sigma^2) - C - \ln 2,\ b \in [0,1], \tag{9}$$

where C is Euler's constant and takes the value ≈ 0.5772.

Proof. Invoking Lemma 2, it is obvious that the inequality holds true for $b = 1$,

$$\begin{aligned}\mathbb{E}_x[\ln x^2] &= \ln(2\sigma^2) + g_{0.5}\Big(\frac{\mu^2}{2\sigma^2}\Big) \\ &\geq \ln(2\sigma^2) + \ln\Big(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\Big) + \psi(1/2) + \ln(2) \\ &= \ln(\mu^2 + \sigma^2) - C - \ln 2.\end{aligned} \tag{10}$$

This implies that the inequality holds true for all values of $b \in [0,1]$. □

C DETAILS ON THEOREM 2

Theorem 2 can be obtained by applying Theorem 1 on the ELBO $\mathcal{L}$. In Theorem 2, there are two expectations $\mathbb{E}_q^2 f(x)$ and $\mathrm{Var}_q f(x)$ which can be computed as follows (Lloyd et al., 2015):

$$\mathbb{E}_q^2 f(x) = \mathrm{tr}(K_{RR}^{-1} \Phi K_{RR}^{-1} (\boldsymbol{\mu}\boldsymbol{\mu}^\top)), \tag{11}$$

$$\mathrm{Var}_q f(x) = \gamma |\mathcal{X}^{(k)}| - \mathrm{tr}(K_{RR}^{-1}\Phi) + \mathrm{tr}(K_{RR}^{-1}\Phi K_{RR}^{-1}\Sigma). \tag{12}$$

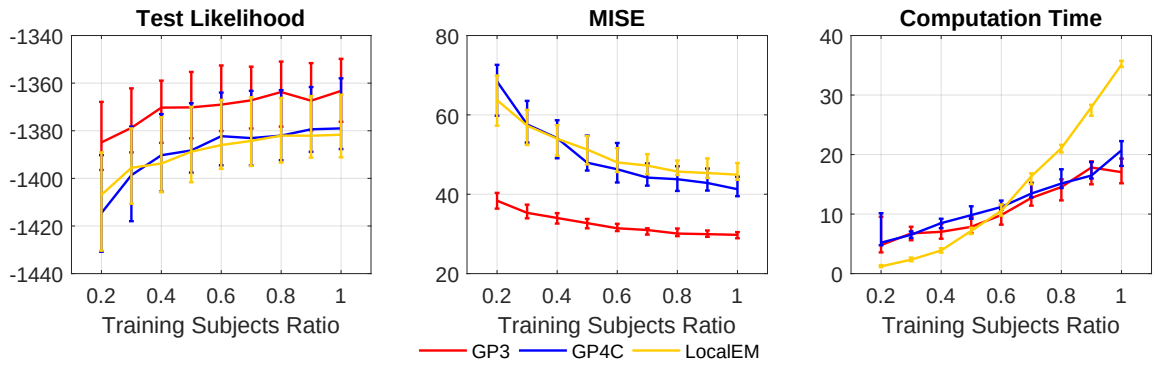

Figure 1: **Synthetic Data Set**. Comparison of performance of GP3, GP4C and LocalEM in terms of $\mathcal{L}_{\text{test}}$, MISE and T when varying the ratio of training subjects and the test set is the same. For MISE and the computation time, the 0.25 and 0.75 quantiles of the statistics in 40 experiments are shown with error bars. All methods benefit from the increase of the number of training subjects. The computation time of GP3 and GP4C grow linearly with the increase of the number of training subjects.

D TEST LIKELIHOOD OF GP4C and GP3

Recall that during the sth trial, the test likelihood is

$$\begin{aligned}
\mathcal{L}_{\text{test}}(s) &\triangleq \ln \int p(\mathcal{D}_{\text{test}}^{(s)}|f)p(f|\mathcal{D}_{\text{train}}^{(s)})df \\
&\approx \ln \frac{1}{U}\sum_{u=1}^{U} p(\mathcal{D}_{\text{test}}^{(s)}|f^{(s,u)}) \qquad (13) \\
&= \ln \sum_{u=1}^{U} \exp\Big(\ln p(\mathcal{D}_{\text{test}}^{(s)}|f^{(s,u)})\Big) - \ln U \\
&= \ln \sum_{u=1}^{U} \exp\Big(\sum_{k=1}^{K_{\text{test}}}\sum_{i=1}^{N_k}\Big(m_i^{(k)} \ln r_{ik}^{(s,u)} - \ln(m_i^{(k)}!)\Big) \\
&- \sum_{k=1}^{K_{\text{test}}} \int_{\mathcal{X}^{(k)}} \Big(f^{(s,u)}(x)\Big)^2 dx\Big) - \ln U. \qquad (14)
\end{aligned}$$

In the above derivation, we use

$$f^{(s,u)} \sim \mathcal{N}(\mu^{(s)}, \Sigma^{(s)}), \qquad (15)$$

$$r_{ik}^{(s,u)} = \int_{\mathcal{X}_i^{(k)}} \Big(f^{(s,u)}(x)\Big)^2 dx. \qquad (16)$$

We can calculate the test likelihood for each subject similarly. In Equation (13), we draw $U = 50$ samples of the function $f^{(s,u)}$ from the variational distribution $q^{(s)}(f)$ on a vector of 3001 evenly-spaced points on $\mathcal{X}$ and we approximate points at an arbitrary position on $\mathcal{X}$ with the linear interpolation. The log-exp-sum trick is used to calculate the $\mathcal{L}_{\text{test}}(s)$. We calculate all integrals in $p(\mathcal{D}_{\text{test}}^{(s)}|f)$ using Simpson's rule with 501 evenly-spaced points.

In Equation (14), the term $\sum_k \sum_i \ln(m_i^{(k)}!)$ can be extracted out and treated as a constant. [1]

E ADDITIONAL SYNTHETIC EXPERIMENTS

E.1 THE DEPENDENCE OF THE LIKELIHOOD ON THE NUMBER OF INTERVALS

The likelihood of the panel count data for the kth subject depends on the disjoint intervals $\{\mathcal{X}_i^{(k)}\}_{i=1}^{N_k}$, where $\bigcup \mathcal{X}_i^{(k)} = \mathcal{X}^{(k)}$. One phenomenon is that as the number of disjoint intervals N_k increases, the likelihood tends to decrease. This is because as we use finer disjoint intervals, we are less uncertain about the position of the time-stamps.

We conduct an experiment to show this phenomenon. First we draw a time-sequence from the intensity function $\lambda(t) = 5$ on $\mathcal{X} = [0, 60]$ and then censor the time-sequence using N disjoint intervals. We vary the number of disjoint intervals and calculate the likelihood of the generated panel count data set. The result is given in Figure 2. We see that the logarithm of the likelihood decreases with the increase of the number of intervals.

E.2 RATIO OF THE TRAINING OBJECTS

We vary the number of training subjects by adjusting the ratio relative to full training subjects. We expect all methods will benefit from the increase of the training

[1] In first versions of this paper, we omitted the constant term. However, to compute the standard deviations of the test likelihood and fairly compare the test likelihood among all algorithms, we add back the constant term in the final version.

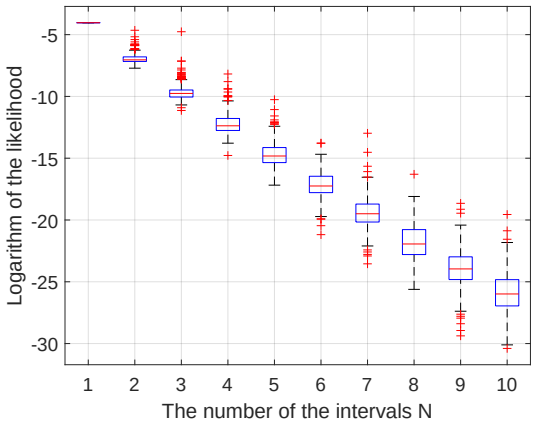

Figure 2: The logarithm of the likelihood of the same time-sequence when varying the number of disjoint intervals. As more disjoint intervals are used, the logarithm of the likelihood decreases. Even for the same number of disjoint intervals, the logarithm of the likelihood has a large variance.

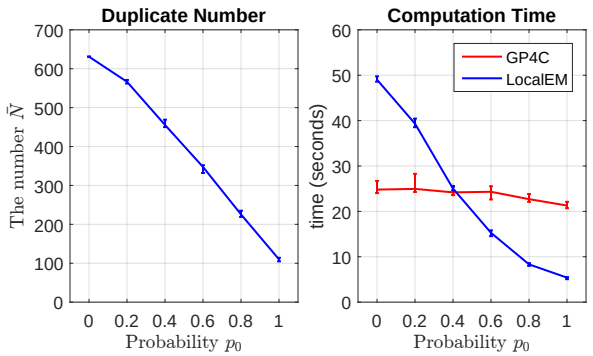

Figure 3: **Synthetic A Data Set**. Comparison of the computation time of GP4C and LocalEM algorithms. LocalEM algorithm achieves a worse computation time as the probability p_0 gets smaller.

subjects.

The result for the Synthetic A data set is given in Figure 1. We see that all three methods benefit from the increase of the number of training subjects. The computation time of GP3 and GP4C grow linearly with the increase of the number of training subjects but LocalEM grows more rapidly.

E.3 THE DEPENDENCE OF THE COMPUTATION TIME ON THE SIZE $\bar{N}$

The computational complexity of LocalEM during one iteration is $\mathcal{O}(\bar{N}^2\bar{M}^2)$ while for GP4C it is $\mathcal{O}(NM^3)$, where N and $\bar{N}$ denote the number of different intervals in the data set and the size of the merged set X. We conduct an experiment to show the influence of the size $\bar{N}$.

We generate $U = 70$ subjects from the same intensity function $\lambda(t) = h_1(t)$, which is the same as Synthetic A data set. We generate the corresponding panel count data set by censoring each subject with 10 intervals. Then we vary the number of $\bar{N}$ by rounding each end point to the next smaller integer with the probability p_0. As p_0 get larger, more end points are rounded and the value of $\bar{N}$ decreases. The experiment result is given in Figure 3. We see that the number $\bar{N}$ decreases linearly with the probability p_0 and computational time decreases much more faster. We can conclude that when the probability p_0 is small and the number of duplicates is large, LocalEM is less efficient than GP4C.

F A BRIEF DESCRIPTION OF THE REAL-WORLD DATA SET

Sun and Zhao (2016) provided three panel count data sets. A brief introduction can be found as follows.

Nausea data set. This data set contains the visiting times from 113 patients during 52 weeks. The panel count data were obtained by recording the reported count of vomits from each patient between two subsequent visits. Patients were divided into two groups, which are the treatment group (65 patients) and the placebo group (48 patients). We denote the two groups as the Nausea A (Na-A) and B (Na-B) sets.

Bladder cancer data set. This data set arises from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group. It records the counts of new tumors that occurred between subsequent visits from 85 patients during 53 weeks, who were divided into the placebo group (47 patients) and the treatment group (38 patients). We denote the two groups as the Bladder A (Bl-A) and B (Bl-B) sets.

Skin cancer data set. This data set was recorded during a skin cancer experiment conducted by the University of Wisconsin Comprehensive Cancer Center and the numbers of new skin cancers of two different types between two subsequent visits from 290 patients were recorded during five years. The visiting time was recorded in the form of days since the first visit and we divided the days by 30. Patients were divided into treatment and placebo groups. We denote the four groups from two types of cancer as the Skin A (Sk-A), Skin B (Sk-B), Skin C (Sk-C) and Skin D (Sk-D) sets.

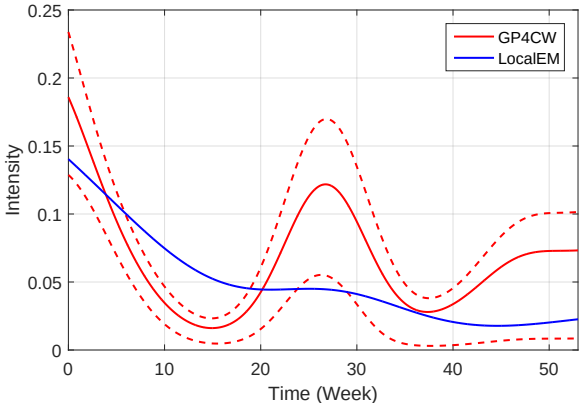

Figure 4: **Bladder Cancer Data Set**. Inferred intensity function by the LocalEM and GP4CW methods. For GP4CW, a 75% credible interval is given by dotted lines.

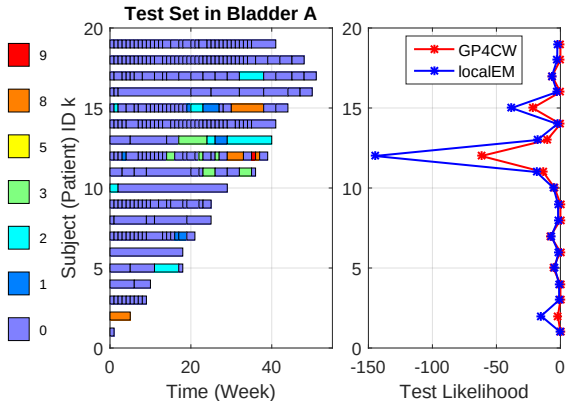

Figure 5: **Bladder A Data Set**. An illustration of the panel count data in the test set (Left) and the test likelihood from GP4CW and LocalEM of each subject (Right). GP4CW mainly outperforms LocalEM on two subjects whose numbers of newly-occurred cancers are large (No. 12 and 15).

G GP4C MODEL WITH INDIVIDUAL WEIGHT

G.1 MODEL

It is practical to assume that the k'th subject has an individual weight parameter υ_k multiplied to the basic intensity function, because in traditional panel count data sets, each subject is a patient whose personal information, such as age, is not the same and the count data from each patient may vary greatly. Such a modification is called the unobservable independent random effects in Cook and Lawless (2007). In the simplest case, we consider the following model for the underlying intensity function:

$$\lambda_k(x) = \upsilon_k f^2(x), \ f \sim \mathcal{GP}(g(x), \kappa(x, x')), \qquad (17)$$

where $\upsilon_k \in \mathbb{R}^+$ is a deterministic and positive real number. The likelihood is as follows.

$$p(\mathcal{D}, f) = \Big[\prod_{k=1}^{K} p(\boldsymbol{d}_k|\lambda(x); \upsilon_k)\Big] p(f; g, \kappa). \qquad (18)$$

We call this model **GP4C** model with individual **W**eight (GP4CW).

We can further generalize this model by assuming that the intensity function of the k'th subject is a linear combination of basis intensity functions (Lloyd et al., 2016) and the mixture weights are also deterministic.

G.2 INFERENCE

The inference of GP4CW is almost the same as GP4C. We only need to modify GP4C by adding the inference of the point estimate of υ_k in M-step of the vEM framework as follows.

$$\upsilon_k = \max\left\{\epsilon, \frac{\sum_{i=1}^{N_k} m_i^{(k)}}{\int_{\mathcal{X}^{(k)}} \mathbb{E}_q[f^2(x)]dx}\right\}, \qquad (19)$$

where $\epsilon = 10^{-6}$ is a small number to guarantee the positiveness of υ_k.

G.3 EXPERIMENT ON THE REAL WORLD DATA SET

On the three real world data sets. The test likelihood $\mathcal{L}_{\text{test}}$ and the computation time T are given in Table 1. We also plot the test likelihood of each subject and the inferred intensity function from GP4CW in Figures 5 and 4. We can notice that GP4CW provides more accurate estimation on the patient No. 12 and No. 15.

H AN EXPERIMENT TO REDUCE THE STANDARD DEVIATION ON THE REAL WORLD DATA SET

In each trial, we randomly split the whole data set into two halves $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, one for training and the other for testing. However, as is introduced in Appendix E.1, if the subjects in $\mathcal{D}_{\text{test}}$ do not share the same time window and the same set of censoring intervals, the test likelihood $\mathcal{L}_{\text{test}}^{(1)}$ will vary greatly from subject to subject. To reduce the large standard deviation caused by different random splits, we perform another round of training for each split, we train on $\mathcal{D}_{\text{test}}$ and calculate the test likelihood on the $\mathcal{D}_{\text{train}}$. The test likelihood is denoted

Table 1: Mean and standard deviations of the test likelihood ($\mathcal{L}_{\text{test}}$) and the computation time (T) on the three panel count data sets for GP4C, GP4CW and LocalEM over 40 runs. GP4CW outperforms GP4C and LocalEM.

Data Set	**METHOD**	$\mathcal{L}_{\text{test}}$	$T[s]$
Na-A	LocalEM	-492.1±306.1	1±0
	GP4C	-484.9±201.8	10±10
	GP4CW	**-179.2±81.3**	8±9
Na-B	LocalEM	-473.2±212.2	1±0
	GP4C	-411.0±184.3	10±7
	GP4CW	**-152.7±60.6**	16±13
Bl-A	LocalEM	-201.8±46.9	1±0
	GP4C	-182.2±47.3	25±9
	GP4CW	**-95.5±29.0**	29±12
Bl-B	LocalEM	-313.1±54.2	1±0
	GP4C	-310.4±54.9	26±21
	GP4CW	**-212.4±50.1**	36±23
Sk-A	LocalEM	-259.1±27.3	39±3
	GP4C	-258.7±26.7	33±6
	GP4CW	**-183.0±21.6**	35±8
Sk-B	LocalEM	-198.1±47.1	39±3
	GP4C	-191.2±42.5	24±4
	GP4CW	**-105.7±19.7**	27±5
Sk-C	LocalEM	-358.0±35.8	47±4
	GP4C	-355.7±36.0	21±12
	GP4CW	**-243.6±26.9**	19±11
Sk-D	LocalEM	-200.9±31.9	46±3
	GP4C	-198.9±30.6	27±4
	GP4CW	**-118.9±14.3**	31±4

as $\mathcal{L}^{(2)}_{\text{test}}$. This can be viewed as adding an additional reverse split. The final test likelihood is $\mathcal{L}^{(1)}_{\text{test}} + \mathcal{L}^{(2)}_{\text{test}}$. The result is given in Table 2.

We see that the variances of the test likelihood in Table 2 are reduced comparing to results in Table 1.

Table 2: Mean and standard deviations of the test likelihood ($\mathcal{L}^{(1)}_{\text{test}} + \mathcal{L}^{(2)}_{\text{test}}$) and the computation time (T) on the three panel count data sets for GP4C, GP4CW and LocalEM after performing another round of training to reduce the variance caused by random split. GP4CW outperforms GP4C and LocalEM.

Data Set	**METHOD**	$\mathcal{L}^{(1)}_{\text{test}} + \mathcal{L}^{(2)}_{\text{test}}$	$T[s]$
Na-A	LocalEM	-1272.7±288.6	2±0
	GP4C	-1205.5±157.1	20±11
	GP4CW	**-417.8±72.5**	25±16
Na-B	LocalEM	-957.2±116.1	1±0
	GP4C	-844.0±85.5	20±10
	GP4CW	**-307.1±34.2**	29±15
Bl-A	LocalEM	-421.4±44.3	2±0
	GP4C	-378.6±17.9	48±14
	GP4CW	**-206.3±5.2**	63±16
Bl-B	LocalEM	-684.7±54.4	2±0
	GP4C	-664.1±19.2	50±28
	GP4CW	**-461.4±19.3**	71±31
Sk-A	LocalEM	-519.8±6.2	77±3
	GP4C	-519.1±2.7	65±8
	GP4CW	**-366.4±1.5**	69±11
Sk-B	LocalEM	-392.5±28.5	77±3
	GP4C	-375.8±4.2	45±4
	GP4CW	**-210.9±2.8**	54±6
Sk-C	LocalEM	-733.6±11.4	91±4
	GP4C	-728.4±6.3	42±16
	GP4CW	**-498.2±2.3**	43±15
Sk-D	LocalEM	-404.2±14.8	90±4
	GP4C	-400.2±6.8	53±4
	GP4CW	**-241.5±1.9**	60±4

References

Cook, R. J. and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media.

Famoye, F. (1995). Continuous Univariate Distributions, Volume 1.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, pages 1814–1822.

Lloyd, C., Gunter, T., Osborne, M., Roberts, S., and Nickson, T. (2016). Latent point process allocation. In *Artificial Intelligence and Statistics*, pages 389–397.

Moser, S. M. (2007). Some expectations of a non-central chi-square distribution with an even number of degrees of freedom. In *TENCON 2007-2007 IEEE Region 10 Conference*, pages 1–4. IEEE.

Sun, J. and Zhao, X. (2016). *Statistical Analysis of Panel Count Data*. Springer.