COM-Poisson分布の効率的なパラメータ推定と cure rate modelの適用

An Efficient Method of Parameter Estimation for the COM-Poisson Distribution and its Application in Cure Rate Model

経営システム工学専攻　冨尾燿平

1　序論

人間の死亡数や事故発生数等，ある現象が一定時間内に起こった回数を数え上げたデータのことをカウントデータと呼ぶ．カウントデータが従う分布として，Poisson 分布は代表的な分布の一つであり, 製品の製造過程の不良品の個数の調査などで用いられている [8]. さらに, 新薬開発の臨床試験においても用いられ [10], 例として, ある一定期間内での発作の回数などに Poisson 分布が仮定される場合が多い. しかし, Poisson 分布には平均と分散が等しくなるという強い制約があり, 過小分散, 過分散を表現することが出来ない.

過小分散, 過分散を表現できる Poisson 分布として Conway-Maxwell-Poisson 分布（以下，COM-Poisson 分布と記載する）が導出された [2]．COM-Poisson 分布は, Poisson 分布を一般化したモデルあることに加えて, 二項分布, 幾何分布を含む. また，競合リスクの数を COM-Poisson 分布と仮定した cure rate model の研究が盛んであることが知られている [1][5][9]．しかし，COM-Poisson 分布は正規化定数が無限級数によって表現されており，計算が極めて困難である．そのため，ラプラス近似, 回帰, 有限和による近似等によって，正規化定数の計算を回避する手法が提案されている．しかし，ラプラス近似は特定の範囲でしか有効ではなく [10]，回帰による方法は共変量が無い場合，適用することが出来ない [1][5][9]. また, 有限和による打ち切りは, $\zeta = \lambda^{\frac{1}{\nu}}$($\lambda > 0$ は尺度パラメーター, $\nu \in \mathbb{R}$ は形状パラメーター) の値が大きいとき, 打ち切り誤差が大きくなり, 近似精度が悪くなると [4] で指摘されている.

そこで，本研究では COM-Poisson 分布の計算が極めて困難である無限級数を回避したパラメータ推定方法を 3 つ提案する. 提案手法 1 では, 条件付き最尤法を提案する [11]. また, 提案推定量の存在性と一意性についての定理を与える. さらに ν の条件付き尤度関数の, *strictly log-concavity* に関する定理も与える. また, シミュレーションにより, 提案法は, 従来の最尤法よりも bias, RMSE において推定精度が高いことも示す. 提案手法 2 では, Reversed Hazard Function を用いた最小二乗法を提案する. 提案手法 3 では, COM-Binomial 分布を競合リスクの数に当てはめた時の cure rate model の最尤法を提案する. さらに, 提案手法 1, 2, 3 に対して, 実データへの適用を行う.

2　提案手法

本研究では, COM-Poisson 分布の計算が極めて困難である無限級数を回避したパラメータ推定方法を 3 つ提案する.

2.1　提案手法 1: 条件付き最尤法

提案手法 1 では ν と, λ の十分統計量に基づく, 条件付き最尤法を提案する [11]. また, ν と λ の提案推定量の存在性と一意性についての定理を与える．さらに ν の条件付き尤度関数の, *strictly log-concavity* に関する定理も与える．これらの定理は，2 分法のような頑健法だけでなく，1 次導関数を用いた最適化手法により，唯一の最適解を求めることができることを保証する.

2.1.1　ν の条件付き尤度関数と推定量の存在性と一意性

確率変数 $X_i \overset{i.i.d.}{\sim} CMP(\lambda,\nu),\ i=1,\ldots,n$ のとき, $S_1 = s_1$ の下での条件付き尤度関数は以下である.
ただし, $S_1 = \sum_{i=1}^{n} X_i\ ,\ s_1 = \sum_{i=1}^{n} x_i$ であり, x_i は確率変数 X_i の実現値である.

$$\begin{aligned} L_{\boldsymbol{x}|s_1}(\nu) &= P\left(X_1 = x_1, \cdots, X_{n-1} = x_{n-1} | S_1 = s_1\right) \\ &= \binom{s_1}{x_1,\ldots,x_n}^{\nu} \Big/ \mathcal{C}, \end{aligned} \tag{1}$$

$$\mathcal{C} = \sum_{z_1=0}^{s_1} \cdots \sum_{z_{n-1}=0}^{s_1 - \sum_{i=1}^{n-2} z_i} \binom{s_1}{z_1,\ldots,z_n}^{\nu}.$$

定理 1 $\nu \in \mathbb{R}, x_i \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}, i = 1,\ldots,n, \boldsymbol{x} = (x_1,\ldots,x_n), n \in \mathbb{N} \setminus \{1\}$ について, $s_1 = \sum_{i=1}^{n} x_i$ が与えられた下での ν の条件付き尤度方程式の解は, 条件

1, 条件 2 場合を除いて, 一意に存在する.
条件 1 のとき，ν の条件付き尤度方程式の解は ∞ に発散する.
条件 2 のとき，ν の条件付き尤度方程式の解は $-\infty$ に発散する.

$$\text{条件 1 } :\boldsymbol{x} = \arg\max_{(z_1,\ldots,z_n)\in D(s_1)} \log\binom{s_1}{z_1,\ldots,z_n}$$

$$D(s_1) = \left\{(z_1,\ldots,z_n)\in\mathbb{N}_0^n : \textstyle\sum_{i=1}^n z_i = s_1\right\}.$$

条件 2 ：$n-1$ 個の観測値 x_i が 0 である.

定理 2 $\nu \in \mathbb{R}, x_i \in \mathbb{N}_0, i = 1,\ldots,n, \boldsymbol{x} = (x_1,\ldots,x_n), n\in\mathbb{N}\setminus\{1\}$ について , $s_1 = \sum_{i=1}^n x_i$ が与えられた下での ν の条件付き尤度関数は，*strictly log-concavity* である.

2.1.2 λ の条件付き尤度関数及び推定量の存在性と一意性

$X_i \overset{i.i.d.}{\sim} CMP(\lambda,\nu)$, $i = 1,\ldots,n$, のとき, $S_2 = s_2$ の下での条件付き尤度関数は以下である.
ただし, $S_2 = \prod_{i=1}^n X_i$, $s_2 = \prod_{i=1}^n x_i$, であり, x_i は確率変数 X_i の実現値である.

$$\begin{aligned} L_{\boldsymbol{x}|s_2}(\lambda) &= P\left(X_1 = x_1,\cdots,X_{n-1} = x_{n-1}|S_2 = s_2\right) \\ &= 1\Big/\sum_{\boldsymbol{z}\in\Omega_n(s_2)} \lambda^{\sum_{i=1}^n z_i - \sum_{i=1}^n x_i}. \end{aligned} \tag{2}$$

ただし, $\Omega_n(s_2) = \{(z_1,\ldots,z_n)\in\mathbb{N}_0^n : \prod_{i=1}^n z_i! = s_2\}$

定理 3 $\lambda > 0, \boldsymbol{x} = (x_1,\ldots,x_n), x_i \in \mathbb{N}_0, i = 1,\ldots,n, n\in\mathbb{N}\setminus\{1\}$ について，$s_2 = \prod_{i=1}^n x_i$ が与えられた下での λ の条件付き尤度方程式の解は，条件 1，条件 2 場合を除いて，一意に存在する.
条件 1 のとき，λ の尤度方程式の解は ∞ に発散する.
条件 2 のとき，λ の尤度方程式の解は 0 になる.

$$\text{条件 1 } :\boldsymbol{x} = \arg\max_{\boldsymbol{z}\in\Omega_n(s_2)} \sum_{i=1}^n z_i,$$

$$\text{条件 2 } :\boldsymbol{x} = \arg\min_{\boldsymbol{z}\in\Omega_n(s_2)} \sum_{i=1}^n z_i,$$

2.2 提案手法 2: Reversed Hazard Function を用いた最小二乗法

提案手法 2 では, Reversed Hazard Function を用いた最小二乗法を提案する.

COM-Poisson 分布の Reversed Hazard Function(discrete) は, 以下のようになる.

$$r_{COM}(t) = \frac{f_{COM}(t)}{F_{COM}(t)} = \frac{\frac{\lambda^t}{(t!)^\nu}}{\sum_{j:t_j\le t}\frac{\lambda^{t_j}}{(t_j!)^\nu}}, \quad t = 0,1,\ldots \tag{3}$$

ただし，$f_{COM}(t)$，$F_{COM}(t)$ はそれぞれ COM-Poisson 分布の確率関数と累積分布関数を表す.

また，COM-Poisson 分布の Cumulative Reversed Hazard Function は以下のように与えられる.

$$RH_{COM}(t) = \sum_{j:t_j\le t} r_{COM}(t_j), \quad t = 0,1,\ldots$$

時刻 $t_1 \le t_2 \le \ldots \le t_n$ と, 各時刻に対応する故障数 $f_1 \le f_2 \le \ldots \le f_n$(カウントデータ) が与えられたとき, Nelson-Aalen 推定値に基づく Reversed Hazard Function($\widehat{RH}_{np}(t)$) は以下のように与えられる.

$$\widehat{RH}_{np}(t) = \sum_{i:t_i\le t}\frac{f_i}{n_i}, \quad i = 1,2,\ldots$$

ただし, $n_i = \sum_{i:t_i\ge t} f_i$.

最小二乗推定量は, 以下のように与えられる.

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n \left(RH_{COM}(t_i) - \widehat{RH}_{np}(t_i)\right)^2, \tag{4}$$

ただし, $\boldsymbol{\theta} = (\lambda,\nu)$.

2.3 提案手法 3: COM-Binomial を競合リスクの数に当てはめた時の cure rate model の最尤法

提案手法 3 では, COM-Binomial 分布を競合リスクの数に当てはめた時の cure rate model の最尤法を提案する.

n 組のイベントが起こるまでの時間, 右打ち切りを示す変数 (t_1,δ_1), ..., (t_n,δ_n) が与えられた下で, 競合リスクの数に COM-Binomial 分布を当てはめた時の cure rate model の対数尤度関数は以下のようになる.

$$\begin{aligned} l(\boldsymbol{\theta}) &= -n_1\log(\gamma_1) - \sum_{i=1}^n \delta_i\log(t_i) + \frac{n_1\log(\gamma_2)}{\gamma_1} \\ &+ \frac{1}{\gamma_1}\sum_{i=1}^n \delta_i\log(t_i) + A - B \end{aligned} \tag{5}$$

ただし,

$A = \sum_{i=1}^{n} \delta_i \log \left(\sum_{m=0}^{r} m \binom{r}{m}^{\nu} \left(\exp(-(\gamma_2 t_i)^{\frac{1}{\gamma_1}}) \exp(\boldsymbol{x_i}\boldsymbol{\beta}) \right)^{m} \right)$

$B = \sum_{i=1}^{n} \delta_i \log \left(\sum_{m=0}^{r} \binom{r}{m}^{\nu} \left(\exp(-(\gamma_2 t_i)^{\frac{1}{\gamma_1}}) \exp(\boldsymbol{x_i}\boldsymbol{\beta}) \right)^{m} \right)$

ただし, r は競合リスクの数を表し, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma_1, \gamma_2, \nu)$, $n_1 = \sum_{i=1}^{n} \delta_i$ である. さらに, $\boldsymbol{x_i}$ は共変量を表し, $\boldsymbol{\beta}$ は回帰係数ベクトルを表す. また, 本研究では, 従来研究 [8] と同様に, 生存関数を Weibull 分布として右打ち切りを仮定している.

3 評価

ここでは, 提案手法 1 と, 有限和による近似を用いた最尤法について, シミュレーションによる評価を行う. なお, 本研究での有限和による近似を用いた最尤法は R のパッケージ "compoisson" [6] を用いる. サンプルサイズ $n = 5, 7, 10$, $\lambda = 0.5, 1.0, 1.5$, ν=0.5, 1.0, 1.5 の設定で bias, RMSE, 解が得られる割合を求める. また, シミュレーションの実行回数は 1000 回とした.

表 1, 2 より, bias は λ, ν 共に最尤法よりも, 条件付き最尤法の方が低いことがわかる, 表 3, 4 から, RMSE は λ は最尤法よりも, 条件付き最尤法の方が低く, ν はほぼ同等であることがわかる. また, 図 1 より, 最尤法よりも, 条件付き最尤法の方が解が得られる割合が高いことが分かる.

表 1: ν の bias

	ν=0.5, λ=0.5		ν=1.0, λ=1.0		ν=1.5, λ=1.5	
	ML	Prop.	ML	Prop.	ML	Prop.
$n = 5$	0.681	0.101	0.607	0.016	0.389	-0.193
$n = 7$	0.731	-0.016	0.738	0.141	0.623	0.134
$n = 10$	0.666	0.186	0.587	0.187	0.702	0.252

表 2: λ の bias

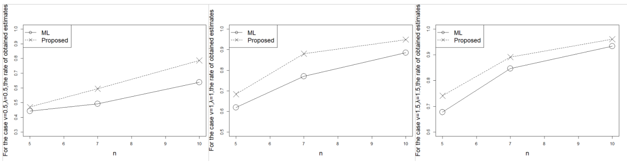
	ν=0.5, λ=0.5		ν=1.0, λ=1.0		ν=1.5, λ=1.5	
	ML	Prop.	ML	Prop.	ML	Prop.
$n = 5$	0.411	0.348	0.730	-0.285	0.713	-0.588
$n = 7$	0.302	-0.017	-0.851	-0.249	1.091	-0.576
$n = 10$	0.212	-0.132	0.540	-0.376	0.997	-0.736

表 3: ν の RMSE

	ν=0.5, λ=0.5		ν=1.0, λ=1.0		ν=1.5, λ=1.5	
	ML	Prop.	ML	Prop.	ML	Prop.
$n = 5$	1.068	0.925	1.138	1.101	1.077	1.100
$n = 7$	1.137	1.165	1.281	1.166	1.310	1.109
$n = 10$	1.091	1.449	1.175	1.033	1.961	1.122

表 4: λ の RMSE

	ν=0.5, λ=0.5		ν=1.0, λ=1.0		ν=1.5, λ=1.5	
	ML	Prop.	ML	Prop.	ML	Prop.
$n = 5$	0.906	0.811	2.202	0.849	2.267	1.096
$n = 7$	0.998	0.450	2.195	0.861	2.536	1.131
$n = 10$	0.561	0.301	1.502	0.737	2.762	1.093



(a) ν=0.5, λ=0.5 (b) ν=1.0, λ=1.0 (c) ν=1.5, λ=1.5

図 1: 解が得られる割合

4 実データへの適用

4.1 カートンデータへの適用

ここでは, 提案手法 1 と, 有限和による近似を用いた最尤法について, カートンデータ [8] の適用を行う. カートンデータは 10 個の観測値があり, それぞれの観測値は航空機が目的地に到着するまでにプラスチックの容器を輸送した回数を示す. KS 検定統計量によって, 条件付き最尤法と無限級数での有限和に基づく最尤法のデータの当てはまりの評価, 比較を行った. 表 5 より, KS 検定統計量のから, 最尤法よりも条件付き最尤法の方が当てはまりが良いことが分かる. また, 図 2, 3 より定理 1, 2, 3 が成立していることが分かる.

表 5: カートンデータの $\hat{\nu}, \hat{\lambda}$ 及び KS 検定統計量

	$\hat{\nu}$	$\hat{\lambda}$	KS
提案手法 (条件付き最尤法)	0.687	0.766	0.0637
無限級数での有限和に基づく最尤法	0.888	0.937	0.0646

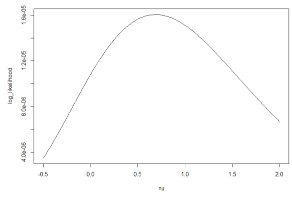


図 2: ν の条件付き対数尤度

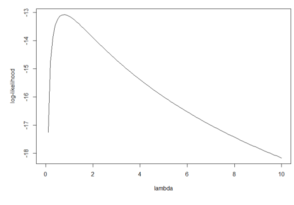


図 3: λ の条件付き対数尤度

4.2 白血病データへの適用

ここでは, 提案手法 2 に対し, 白血病患者のデータセット [3] の適用を行う. このデータは, サンプルサイズ 21 のデータセットであり, プラセボ群と投薬群の白血病の寛解状態から再発時間 or 打ち切りまでの時間（週）が観測されている. 本研究では, プラセボ群のデータのみを使用している. 図 4 から, 提案法による生存関数が Kaplan-Meier 曲線と比較しても, 遜色ないことが確認出来る.

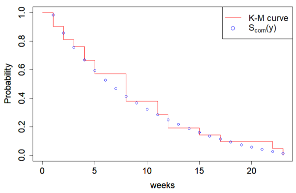


図 4: 最小二乗法に基づく生存関数及び Kaplan-Meier 曲線

4.3 ぶどう膜黒色腫データへの適用

ここでは, 提案手法 3 と, COM-Poisson 分布を競合リスクに当てはめた時の cure rate model の最尤法について, 対数尤度による比較を行う. 今回用いる実データは, サンプルサイズ 63 のぶどう膜黒色腫のデータセット [5] であり, 黒色腫が転移するまでの時間 or 打ち切りまでの時間（月）が観測されている. また, 共変量は性別 (男性=39 人, 女性=24 人) である. COM-Poisson 分布の正規化定数は有限和による近似を行う.COM-Poisson 分布による cure rate model のパラメータの同時推定は, 正規化定数の計算に非常に時間がかかるため, 従来研究 [8] と同様に, プロファイル尤度を用いたパラメータ推定を行う. COM-Binomial 分布を当てはめた時の競合リスクの数 r はそれぞれ 1 から 20 まで値を指定し, パラメータ推定を行う. 表 6 から, 競合リスクの数 $r = 3$ のときの COM-Binomial(COMB) を当てはめた時の対数尤度が最も高く, COM-Poisson(COMP) を当てはめた時の対数尤度よりも高くなっていることが分かる.

表 6: ぶどう膜黒色腫データを当てはめた時のパラメータの推定値及び対数尤度

	γ_0	γ_1	β_0	β_1	ν	対数尤度
COMB r=2	0.982	0.022	0.445	-0.389	0.640	-182.7463
COMB r=3	0.971	0.019	0.062	-0.356	0.593	-182.7204
COMB r=4	0.961	0.023	-0.797	-0.413	1.230	-182.7408
COMP	1.002	0.030	1.316	-0.422	∞	-182.8054

5 結論と今後の課題

本研究では，COM-Poisson 分布の計算が極めて困難である無限級数を回避したパラメータ推定方法を 3 つ提案した. 提案手法 1 に関して, 提案推定量の存在性と一意性に関する定理を与えた．さらに ν の条件付き尤度関数の, *strictly log-concavity* に関する定理も与えた. これらの定理は，2 分法のような頑健法だけでなく，1 次導関数を用いた最適化手法により唯一の最適解を求めることができることを保証する. また, カートンデータによる評価について, KS 検定統計量の観点から, 提案手法の方が, 有限和による近似を用いた最尤法よりもデータの当てはまりが良いことが分かった. シミュレーションの結果から, 解が得られる割合, bias, RMSE の観点において, 提案法が, 最尤法より, 推定精度が優れていることが分かった. 提案手法 2 に関しては, Reversed Hazard Function を用いた最小二乗法を提案した. 提案手法 3 に関して, COM-Binomial 分布を競合リスクに当てはめた時の cure rate model の最尤法を提案した.

今後の課題として, 提案手法 1 の, データサイズが大きいときの, 計算爆発問題の解決である. 提案手法 2 では, Reversed Hazard Function に基づく一般化最小二乗法の構築である. また, 提案手法 3 では, EM アルゴリズムを用いた競合リスクの数 r を含んだ同時推定である.

参考文献

[1] Balakrishnan, N, and Pal, S. (2018). Gamma lifetimes and associated inference for interval-censored cure rate model with COM-Poisson competing cause. Communications in Statistics–Theory and Method, Vol.47, pp.1491-1509.

[2] Conway, R. and Maxwell, W. (1962). A queuing model with state dependent service rates. Journal of Industrial Engineering, Vol.12 , pp.132-136.

[3] Gehan, E.A.(1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, Vol.52, pp.203-223.

[4] Gillispie, S. B. and Green, G. C. (2015). Approximating the Conway-Maxwell-Poisson distribution normalization constant. Statistics, Vol.49, pp.953-960.

[5] He, Z. and Emura, T. (2019). The COM-PoissonCure rate Model for Survival Data Computational Aspects. Journal of the Chinese Statistical Association, Vol.57 , pp.1–42.

[6] Jeffrey, D. (2012). compoisson: Conway-Maxwell-Poisson Distribution. R package version 0.3.URL https://CRAN.R-project.org/package=compoisson

[7] Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2003). Applied Linear Regression Models, 4th ed. McGraw-Hill, New York.

[8] Lambert, D.(1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, Technometrics, Vol.34, pp.1-14.

[9] Rodrigues, J., M. de Castro, V. G. Cancho, and N. Balakrishnan. (2009). COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, Vol.139, pp.3605–3611.

[10] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P.(2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution.Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol54, pp.127-142.

[11] Tomio, Y. and Nagatsuka, H. (2022). A Conditional Maximum Likelihood Estimation of the COM-Poisson Distribution and its Uniqueness and Existence.Total Quality Science (to appear).