＜ 修 士 論 文 ＞

時間共変量を含む Cox 回帰モデルにおけるガウス過程回帰と多重代入法の利用について
（要　旨）

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：2021 年度
学籍番号：6020111
氏　　名：田中　健太
指導教員：杉本　知之
提出年月日：2022 年 1 月 12 日

背景と問題意識

ゼロ時点における年齢，性別等の固定共変量の他に，経時的に観測される検査値等の時間共変量を考慮した事象時間分析を Cox 回帰モデルで行いたい．ただし，時間共変量の観測時点については特に定まっていない (不規則観測) という状況を考える．つまり，個体間で観測時点が揃っている必然性はなく，また，時間共変量が複数ある場合には変数間で揃っている必然性もないという，より一般的な状況を想定する．これはリアルワールドデータの分析においてしばしば遭遇し得る状況であり，また，近年の mHealth (モバイルヘルス) の潮流等も念頭に置いている．

時間共変量を含む Cox 回帰モデルの個体 i におけるハザード関数は

$$h(t|\boldsymbol{x}_i, \boldsymbol{y}_i(t)) = h_0(t) \exp\left(\boldsymbol{\beta}_x' \boldsymbol{x}_i + \boldsymbol{\beta}_y' \boldsymbol{y}_i(t)\right) \tag{1}$$

である ($t \geq 0$; $i = 1, \ldots, n$)．ここで，$h_0(t)$ はベースラインハザード関数，$\boldsymbol{x}_i$ は固定共変量，$\boldsymbol{y}_i(t)$ は時間共変量である．なお，時点 t における時間共変量の値は同時点のハザードにのみ影響するとしている．パラメータ $\boldsymbol{\beta}$ の推定は，t_j ($t_1 < t_2 < \cdots < t_m$) を m 個体において観測されたイベント発生時点，$\mathcal{R}(t_j)$ を t_j におけるリスク集合として，部分尤度

$$L_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\exp\left(\boldsymbol{\beta}_x' \boldsymbol{x}_j + \boldsymbol{\beta}_y' \boldsymbol{y}_j(t_j)\right)}{\sum_{l \in \mathcal{R}(t_j)} \exp\left(\boldsymbol{\beta}_x' \boldsymbol{x}_l + \boldsymbol{\beta}_y' \boldsymbol{y}_l(t_j)\right)} \tag{2}$$

の最大化により行われるが，これを完全に構成するには全ての $\mathcal{R}(t_j)$ において $\boldsymbol{y}_l(t_j)$ ($l \in \mathcal{R}(t_j)$) が得られている必要がある．しかし，そのように都合よく (しかも測定誤差もなく) 時間共変量が観測されていることは現実にはほとんどあり得ず，何かしらの方法で補完する必要がある．通常の統計ソフトでは LOCF (Last Observation Carried Forward) により補完がなされるが，これは明らかに $\boldsymbol{\beta}$ の推定にバイアスをもたらすと考えられる．本研究では，これに代わるより妥当な補完と推定の方法を提案する．

提案手法の概要

経時的に観測された時間共変量 $y_i(t)$ に対しては，ガウス過程 (Gaussian process: GP) 回帰により柔軟な当てはめを行うことが可能である．そして素朴には，その回帰曲線を利用して補完することにより，LOCF よりは良い $\boldsymbol{\beta}$ の推定値が得られるだろうと期待される．本研究ではさらに，ガウス過程回帰の結果として，時間共変量 $y_i(t)$ の経時変化曲線そのものの事後分布が得られる点に注目する．これはつまり，経時変化曲線のサンプルパスを望む数だけ生成できるということである．このサンプルパスを利用して補完した場合，その補完値は，各時点における当てはめの不確実性を反映して，サンプルパス毎に異なった値となる．不確実性を反映した複数の補完値が得られるならば，多重代入法 (Multiple imputation: MI) の枠組みで分析することにより，より妥当な $\boldsymbol{\beta}$ の推定値が得られるだろうと期待される．また，ガウス過程回帰は観測時点が不揃いな複数の系列を同時に扱うこともでき，これを Multi-output ガウス過程 (MOGP) 回帰と呼ぶ．したがって，注

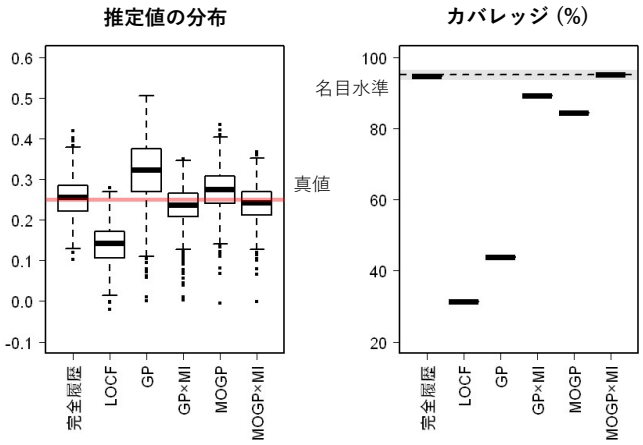

図 1　数値実験の結果

目している時間共変量 $y_i(t)$ 他に，それと相関のある変数 $w_i(t)$ を補助変数として活用することにより，さらに良い $\boldsymbol{\beta}$ の推定値が得られるだろうと期待される．

数値実験と適用例

LOCF による従来の方法では $\hat{\beta}_y$ が過小推定傾向となり，カバレッジ (95% 信頼区間が真値を含んだ割合) も著しく不良となるような設定において，ガウス過程回帰に多重代入法を組み合わせた方法 (GP×MI) と，そこにさらに補助変数の情報も活用した方法 (MOGP×MI) は，$\hat{\beta}_y$ のバイアスとカバレッジをともに大きく改善し得ることが，数値実験で示された (図 1)．これらの改善の度合いは，多重代入法を組み合わせない単一代入法的な方法 (GP および MOGP) よりも優れたものであった．

実データ (Turbofan engine degradation simulation data set) への適用例においても，数値実験と同様の傾向が見られた．一方，分析データに関する領域知識が十分でない状況においては，もし適切でない補助変数を選択してその情報を用いた場合，むしろバイアスのある $\hat{\beta}_y$ を得てしまう恐れもあることが示唆された．

考察

分析モデルを時間共変量を含む Cox 回帰モデルとするとき，ガウス過程回帰と多重代入法を組み合わせることで，より妥当な (バイアスが小さく，カバレッジが良好な) 推定を行える可能性が示唆された．提案手法のメリットとしては，補完モデルにおいてパラメトリックな非線形式等を特定する必要がなく簡便であること，また，補完における不確実性が分かりやすく考慮されており納得感が高いこと等が挙げられる．また，提案手法は不規則観測データへの対応も容易であり，補助変数の情報を活用することで更に推定が改善する可能性も示唆された．一方で，領域知識が十分でない状況においては，適切な補助変数の選択という課題があることも明らかとなった．