＜ 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

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

目次

図目次

表目次

1　はじめに

1.1　背景

実臨床で発生・蓄積されるリアルワールドデータ，すなわち医療ビッグデータの利活用が本格化している．レセプトデータ，DPC データ，電子カルテデータ等に関しては既に大規模なデータベースが整備されており，各種疾患レジストリも有用なデータ源として注目されている．また，IoT が浸透した現代では，スマートフォンほかウェアラブルデバイスや在宅機器を通じて患者から直接収集される医療データ等にも関心が向けられている．これらの医療ビッグデータは，時間の経過とともに含まれる患者の数が増えていくだけでなく，患者当たりの観測時点の数も増えていくという特徴を持つ．つまり，医療ビッグデータ時代においては，大規模かつ長期追跡のデータが一般的になり，したがって，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 (生存時間解析) の重要性が増していくだろうと見込まれる．

そのような分析手法の代表かつ基礎となるものは，やはり依然として Cox 回帰モデルであると考える．しかし，時間共変量を含む Cox 回帰モデルは，実データへの適用場面において，ほとんど常に (パラメータ推定に必要なデータが欠測しているという意味で) 不完全データの分析を強いられる，という困難を有している．この困難は，今後いかに医療ビッグデータが量的に増大し，質的に向上しようとも，変わらず付きまとい続ける本質的な問題であると考えられる．本研究はこの困難に着目するものである．

1.2　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

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

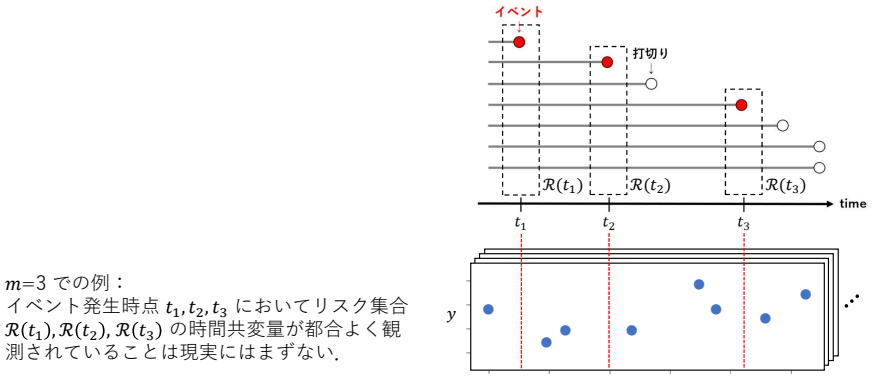


図 1　時間共変量の欠測

の最大化により行われるが [1], 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 (図 1)，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．なお，提案手法においては，なるだけ簡便であることと，納得感が高いものであることが意識されている．

1.3　関連研究

時間共変量を考慮した事象時間分析として，ジョイントモデルによるアプローチが研究されて来た [2]．これは，経時観測データと事象時間データの観測に対してそれぞれモデルをおき，それらの同時尤度に基いてパラメータ推定を行おうとするものである．典型的には，経時観測データには線形/非線形混合効果モデル，事象時間データには (経時観測の対象を時間共変量として含む)Cox 回帰モデルがおかれ，ランダム効果で条件付けたときに各観測が独立になる，という仮定において同時尤度が構成される．ジョイントモデルには様々な拡張が研究されているが，事象時間データに関しては基本的に全尤度を構成することになるため，Cox 回帰モデルをおいた場合にはベースラインハザード関数についても明示的にパラメトライズする必要がある．この点は，ベースラインハザード関数を特定しなくても部分尤度に基いて主要なパラメータを推定できるという Cox 回帰モデルの特性が，損なわれてしまっているとも受け取られる．また，ジョイントモデルでは，分析結果の妥当性についてはモデルが正特定されているかに依存する所が大きく，そして何より，同時尤度がハザード関数の積分を含むなど複雑な形になるため，パラメータ推定に計算上の工夫を要し，従って実装の容易さという点においてやや難がある．

同時尤度ではなく，段階的にそれぞれの尤度に基いてパラメータ推定を行うアプローチの方が，より見通しが良く実装も容易である．これはすなわち，まず，経時観測データのモデルのパラメー

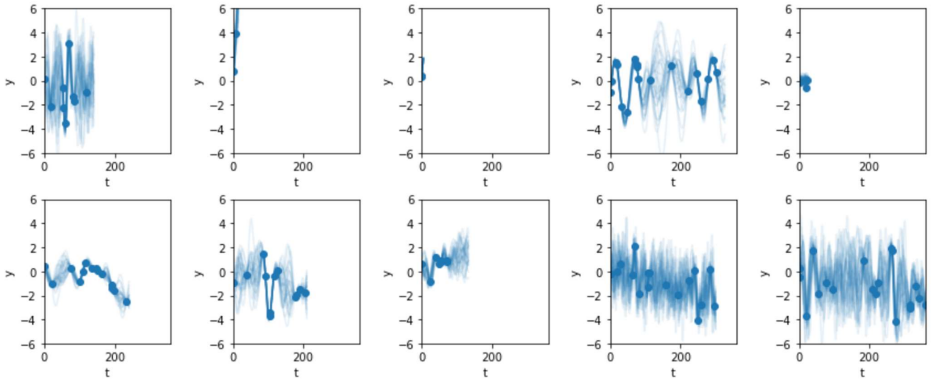


図 2　経時観測データとガウス過程回帰の事後分布から生成したサンプルパスの例

タを推定し，次に，そこから得られる予測値を利用して事象時間データの尤度 (Cox 回帰モデルであれば部分尤度) を構成することで，後者のモデルにおける主要なパラメータを推定しようとするものである．ただし，このアプローチの単純な適用は，経時観測データのモデルから得られる予測値の不確実性が，事象時間データのモデルのパラメータ推定に反映されないという問題がある．そこで先行研究 [3] は，経時観測データのモデルのパラメータ推定と予測値の利用の手続きに MICE (Multiple Imputation by Chained Equations) [4] を組み合せることを提案し，シミュレーションスタディにおいて良好な結果を得ている．また，本先行研究では，1 種類の経時観測データだけでなく，相関を持った複数の経時観測データまで考慮されている．ただし，おそらく MICE との組み合せやすさから，経時観測データのモデルはあくまで線形混合効果モデル，また，その観測時点は計画的なフォローアップを前提に予め定まっている (すなわち規則観測，ただし欠測あり)，という設定にとどまっている．

1.4　本論文の構成

次の 2 節で，提案手法に関してその考え方を中心に説明する．提案手法はガウス過程回帰と多重代入法を組み合わせて利用するものであるが，これらそれぞれの数理的詳細については，補遺において別にまとめる．続く 3 節で，提案手法の性能をシミュレーションスタディにより検証する．すなわち，パラメータの真値が既知であるモデルから生成されたデータに対して，提案手法が正しくその真値を推定できるかを調べる．そして 4 節において，提案手法の実データへの適用例を示す．最後に 5 節で，本研究の総括と今後の展望について述べる．

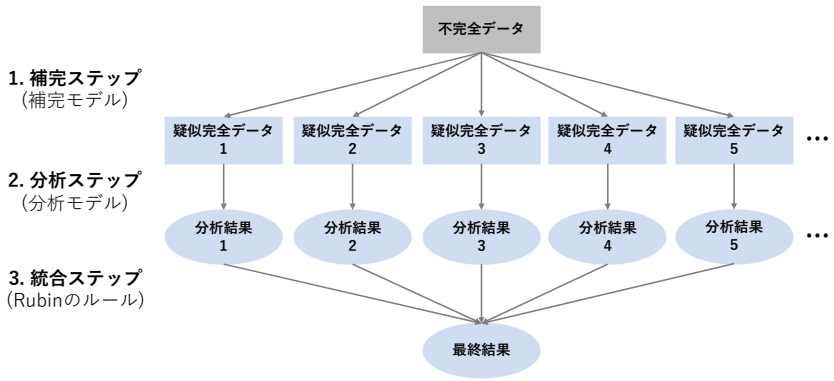


図 3　多重代入法の枠組み

2　提案手法

2.1　ガウス過程回帰の利用

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) 回帰により柔軟な当てはめを行うことが可能である [5]．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，より良い $\boldsymbol{\beta}$ の推定値が得られるだろうと期待される．また，ガウス過程回帰ではパラメトリックな非線形式等を特定する必要がなく，したがって非常に簡便でもある．

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 (図 2)．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．そこでは補完における不確実性が分かりやすく考慮されており，納得感も高いものとなっている．

2.2　多重代入法の利用

多重代入法 (Multiple imputation) は補完・分析・統合の 3 ステップからなり，各ステップにおいて補完モデル，分析モデル，Rubin のルールが用いられる (図 3)[6]．本研究では，Cox 回帰モデルが分析モデルに相当し，イベント発生時点のリスク集合における時間共変量 $y_l(t_j)$ $(l \in \mathcal{R}(t_j);$ $j = 1, \ldots, m)$ が欠測していると見なす，という意味での不完全データを考える．そしてここで，ガウス過程回帰を補完モデルとして用い，その事後分布から生成したサンプルパスを利用して補完を行う．そうして得られた疑似完全データそれぞれについて，部分尤度を構成して $\boldsymbol{\beta}$ の推定値を求め，最終的にそれらを Rubin のルールで 1 つにまとめまる，という流れになる．

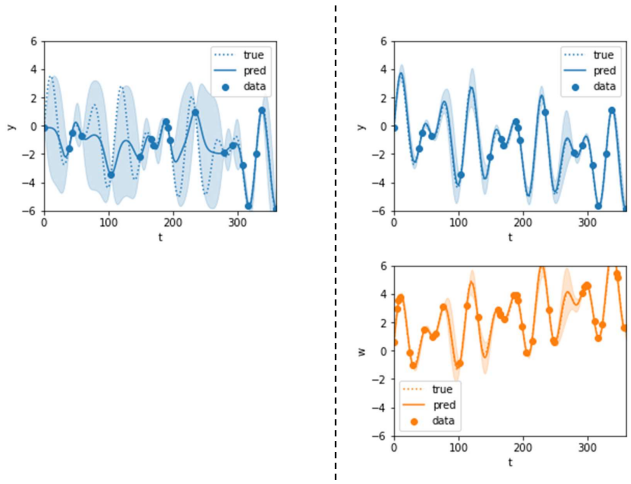


図 4　単独のガウス過程回帰 (左) と Multi-output ガウス過程回帰 (右)

2.3　補助変数の活用：Multi-output ガウス過程回帰

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 ガウス過程回帰と呼ぶ [7]．Multi-output ガウス過程回帰では，複数の系列に対し時間方向の関連性 (共分散関数) を共有しながら同時に回帰を行う．これによって，単独の系列に対し回帰を行った場合よりも回帰曲線がより真の値に近づき，また，サンプルパスのばらつきも減少する (図 4)．したがって，注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．

3　シミュレーションスタディ

3.1　データ生成モデル

ハザード関数を $h_i(t) = h_0(t)\exp\left(\beta_x x_i + \beta_z z_i + \beta_y y_i(t)\right)$, $(t \geq 0;\ i = 1, \ldots, n)$ とする．ここで x_i は連続値の固定共変量，z_i は 2 値の固定共変量であり，$y_i(t)$ が時間共変量である．また，時間共変量と相関のある補助変数として $w_i(t)$ も用意する．

ベースラインハザード関数は生存時間分布を指数分布とした下で $h_0(t) = \lambda = 0.006$ (定数) とし，推定対象である $\beta_x, \beta_z, \beta_y$ は全て 0.25 とした．固定共変量についてはそれぞれ $x_i \sim N(0, 1^2)$，$z_i \sim Bernoulli(0.5)$ とした．時間共変量と補助変数についてはそれぞれ $y_i(t) = a_{0i} + a_{1i}t + A_i\left\{\sin\left(\frac{2\pi t}{T_{1i}}\right) + \sin\left(\frac{2\pi t}{T_{2i}}\right)\right\}$，$w_i(t) = b_{0i} + b_{1i}t + B_i\left\{\sin\left(\frac{2\pi t}{T_{1i}}\right) + \sin\left(\frac{2\pi t}{T_{2i}}\right)\right\}$ とし，直線トレンドに正弦波を重ねることで複雑多様な経時変化となるようにした．ただし，周期 T_{1i}, T_{2i} は $y_i(t)$ と $w_i(t)$ で共通とすることにより，経時変化のパターンを両者で共有するようにした．直線トレンドの切片と傾き，正弦波の振幅と周期に相当するパラメータはそれぞれ

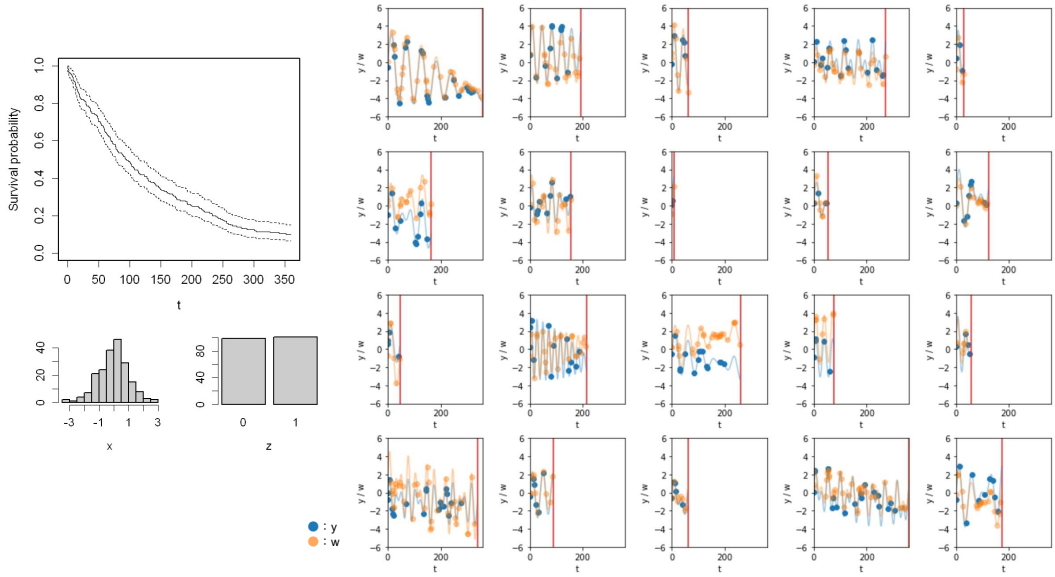


図 5　生成されたデータの見本

$a_{0i}, b_{0i} \sim U(-1, 1)$，$a_{1i}, b_{1i} \sim U(-0.01, 0.01)$，$A_i, B_i \sim U(1, 2)$，$T_{1i}, T_{2i} \sim U(20, 60)$ で与え，観測誤差は $\epsilon_{yi}(t), \epsilon_{wi}(t) \sim N(0, 0.05^2)$ とした．サンプルサイズは $n = 200$ とし，$t = 360$ で観察終了 (打切り) とした．また，$w_i(t)$ の観測頻度は $y_i(t)$ の 2 倍程度とし，観測時点はランダムに与えた．

上記の設定で生成されたデータの見本を図 5 に示す．左上が生存曲線 $S(t)$，その下が固定共変量 x_i と z_i の分布である．右に並んでいるのが一部の個体における時間共変量 $y_i(t)$ と補助変数 $w_i(t)$ のプロットであり，図中の縦のラインはイベント発生時点 (または観察終了時点) を示している．

3.2　シミュレーション

生成されたデータに対し，以下の 6 通りの方法で $\boldsymbol{\beta}$ を推定した．なお，試行回数は 1000 回とした．

LOCF:

時間共変量を LOCF により補完し，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

GP:

(1) 時間共変量のデータに対しガウス過程回帰を実行する．(2) その回帰曲線を用いて時間共変量の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

GP×MI:

(1) 時間共変量のデータに対しガウス過程回帰を実行する．(2) 得られた事後分布からサンプルパスを生成して時間共変量の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．(3)(2) を

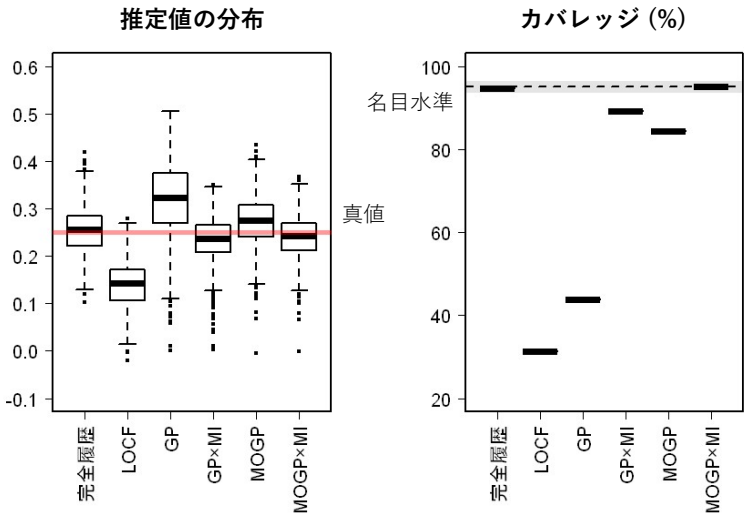


図 6　$\hat{\beta}_y$ の分布とカバレッジ (シミュレーションスタディ)

30 通り行い，Rubin のルールにより 1 つの $\hat{\boldsymbol{\beta}}$ に統合する．

MOGP:

(1) 時間共変量と補助変数のデータの組に対し Multi-output ガウス過程回帰を実行する．

(2) その回帰曲線を用いて時間共変量の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

MOGP×MI:

(1) 時間共変量と補助変数のデータの組に対し Multi-output ガウス過程回帰を実行する．

(2) 得られた事後分布からサンプルパスを生成して時間共変量の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．(3)(2) を 30 通り行い，Rubin のルールにより 1 つの $\hat{\boldsymbol{\beta}}$ に統合する．

完全履歴:

時間共変量の完全な履歴 (真値) を用いて補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

なお，ここでは最後の完全履歴における結果が正対照となる．

3.3　結果と考察

結果の詳細は補遺にまとめた．図 6 に各方法において得られた $\hat{\beta}_y$ の分布とカバレッジ (95% 信頼区間が真値を含んだ割合) を示す．図中の横のラインはそれぞれ β_y の真値と名目水準である．

本シミュレーションスタディの設定においては，LOCF は過小推定傾向となり，カバレッジも著しく不良であった．一方，GP では逆に過大推定傾向となり，カバレッジも依然として名目水準から大きく乖離していた．これに対し，提案手法である GP×MI は，推定値，カバレッジともに LOCF や GP での結果を大きく改善した．MOGP ではやや過大推定傾向となったものの，GP での結果よりは良好であり，カバレッジも遥かに名目水準に近いレベルであった．今回検討した 6 通りの方法の中では，もう 1 つの提案手法である MOGP×MI が，推定値，カバレッジともに最

も良好な結果を示していた．単一代入法的な GP および MOGP は，多重代入法を組み合わせて GP×MI および MOGP×MI となることで，推定結果がマイルドになり，カバレッジも改善する傾向が見られた．

以上の結果から，本研究が想定する状況において，ガウス過程回帰と多重代入法を組み合わせた提案手法により，従来法と比較してより妥当な推定を行える可能性が示唆された．

4 実データへの適用例

4.1 Turbofan engine degradation simulation data set

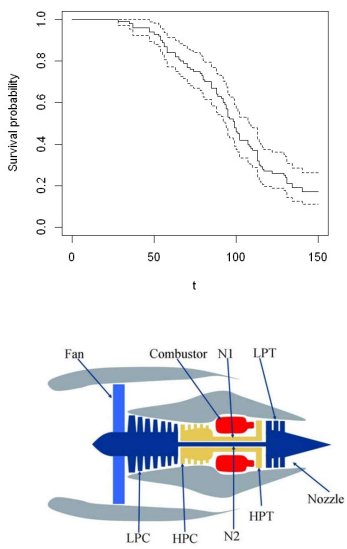
ここでは予知保全領域のオープンデータセットを用いた．本データセットは第 1 回 Prognostics and health management (PHM) コンペティションにおいても使用されたものであり，物理シミュレータから得られた 100 機分のターボファンエンジンの故障時間，および 24 種類のセンサーデータ (経時観測) からなる．本データセットは NASA のデータリポジトリより入手可能である [8]．図 7 にエンジンの生存曲線 $S(t)$ とセンサーの種類を示す．ターボファンエンジンの内部構造の模式図は図 7 左下の通りであり，各箇所に設置されたセンサーにより温度，圧力，速度，比指標等が外部条件とともにモニタリングされている，という設定である．

医療系の実データへの適用例を示すことがより望ましかったが，入手の困難さ等もあり，本経過報告書では医療系とは程遠い領域のデータへの適用例となっている．しかし，予知保全領域において考えられている問題は，本研究が想定しているものと構造的にはほぼ同様であり，従って，十分示唆に富んだ適用例になっていると考える．

4.2 分析シナリオ

分析モデルは Cox 回帰モデルである．ここでは sensor15 のバイパス比に注目し，これを時間共変量としてエンジン故障への影響 (対数ハザード比 β_y) を調べることにした．バイパス比とは，コアエンジン部を通る空気と通らない (バイパスされる) 空気の流量比であり，ターボファンエンジンに特有の指標である．一般にバイパス比が高いほど良いエンジンであるとされているが，通常時レベルからの逸脱はエンジン故障と関連していると予想される．

また，他のセンサーデータを活用して (すなわち，Multi-output ガウス過程回帰を利用して) β_y を推定することも行った．しかし，どのセンサーが補助変数として適当であるかは領域知識の不足のため不明であり，ここでは試行的に sensor2, sensor3, sensro4 (いずれも温度センサー), sensor7, sensor11 (いずれも圧力センサー)，sensor13, sensor14 (いずれも速度センサー) の 7 つを検討することとした．



変数名	記号	内容
condition1	-	altitude
condition2	-	Mach number
condition3	-	Throttle resolver angle
sensor1	T2	Total temperature at fan inlet
sensor2	T24	Total temperature at LPC outlet
sensor3	T30	Total temperature at HPC outlet
sensor4	T50	Total temperature at LPT outlet
sensor5	P2	Pressure at fan inlet
sensor6	P15	Total pressure in bypass-duct
sensor7	P30	Total pressure at HPC outlet
sensor8	Nf	Physical fan speed
sensor9	Nc	Physical core speed
sensor10	epr	Engine pressure ratio (P50/P2)
sensor11	Ps30	Static pressure at HPC outlet
sensor12	phi	Ratio of fuel flow to Ps30
sensor13	NRf	Corrected fan speed
sensor14	NRc	Corrected core speed
sensor15	BPR	Bypass ratio
sensor16	farB	Burner fuel-air ratio
sensor17	htBleed	Bleed enthalpy
sensor18	Nf_dmd	Demanded fan speed
sensor19	PCNfR_dmd	Demanded corrected fan speed
sensor20	W31	HPT coolant bleed
sensor21	W32	LPT coolant bleed

図 7　Turbofan engine degradation simulation data set

4.3　分析モデルの詳細とデータ加工

ハザード関数を $h_i(t) = h_0(t)\exp\left(\beta_{x1}x_{1i} + \beta_{x2}x_{2i} + \beta_y y_i(t)\right)$, $(t \geq 0;\ i = 1, \ldots, 100)$ とする．ここで x_{1i} と x_{2i} は連続値の固定共変量であり，プレ観察期間 $(-100 \leq t < 0)$ のセンサーデータを用いた主成分分析の結果に基いて用意した．$y_i(t)$ が時間共変量であり，バイパス比 (sensor15) が相当する．補助変数 $w_i(t)$ については，sensor2, sensor3, sensro4 (いずれも温度センサー), sensor7, sensor11 (いずれも圧力センサー)，sensor13, sensor14 (いずれも速度センサー) から 1 つを選ぶ．従って，Multi-output ガウス過程回帰は都合 7 パターン行うことになる．

センサーデータは全て標準化して用いた他に，次のデータ加工を施した上で分析を行った．まず，$t = 150$ で観察終了とした．これにより 17 機のエンジンは打切り扱いとなった．また，時間共変量 $y_i(t)$ と補助変数 $w_i(t)$ に相当するデータを，それぞれ 6 割と 4 割間引きした．元データが高頻度かつ規則観測データであったため，本研究で想定している不規則観測データへ加工する目的でこの処理を行った．この間引きの仕方は一意には決まらないため，複数通りのランダムな間引きパターンで検討することとした．そのため，本適用例は準シミュレーションのような形となっている．

4.4　準シミュレーション

間引きデータに対し，以下の 6 通りの方法で $\boldsymbol{\beta}$ を推定した．なお，試行回数は 100 回とした (すなわち，100 通りの間引きパターンで検討した)．

LOCF:

バイパス比 (sensor15) を LOCF により補完し，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

GP:

(1) バイパス比 (sensor15) のセンサーデータに対しガウス過程回帰を実行する．(2) その回帰曲線を用いてバイパス比 (sensor15) の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

GP×MI:

(1) バイパス比 (sensor15) のセンサーデータに対しガウス過程回帰を実行する．(2) 得られた事後分布からサンプルパスを生成してバイパス比 (sensor15) の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．(3)(2) を 30 通り行い，Rubin のルールにより 1 つの $\hat{\boldsymbol{\beta}}$ に統合する．

MOGP:

(1) バイパス比 (sensor15) ともう 1 種類 (温度：sensor2, sensor3, sensro4，圧力：sensor7, sensor11，速度：sensor13, sensor14 から選択) のセンサーデータの組に対し Multi-output ガウス過程回帰を実行する．(2) その回帰曲線を用いてバイパス比 (sensor15) の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

MOGP×MI:

(1) バイパス比 (sensor15) ともう 1 種類 (温度：sensor2, sensor3, sensro4，圧力：sensor7, sensor11，速度：sensor13, sensor14 から選択) のセンサーデータの組に対し Multi-output ガウス過程回帰を実行する．(2) 得られた事後分布からサンプルパスを生成してバイパス比 (sensor15) の補完を行い，部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．(3)(2) を 30 通り行い，Rubin のルールにより 1 つの $\hat{\boldsymbol{\beta}}$ に統合する．

全データ利用:

間引きは行わず元データを全て利用し，通常の方法 (つまり LOCF) により部分尤度を構成して $\hat{\boldsymbol{\beta}}$ を求める．

元データは高頻度かつ規則観測データであるため，最後の全データ利用は 3 節のシミュレーションスタディにおける完全履歴での推定とほぼ同様と考えられる．そこで，ここでは全データ利用における結果を仮の正対照と見なすことにした．なお，全データ利用に関しては試行は 1 回のみである．

4.5 結果と考察

図 8 に各方法において得られた $\hat{\beta}_y$(バイパス比の対数ハザード比) の分布を示す．左はバイパス比 (sensor15) のセンサーデータのみを用いた全データ利用，LOCF，GP，GP×MI での結果，右 3 つはその他のセンサーデータ (温度：sensor2, sensor3, sensro4，圧力：sensor7, sensor11，速度：sensor13, sensor14) も併せて用いた MOGP および MOGP×MI での結果である．

全データ利用における推定値が真値である保証はないため，確かなことは言えないが，バイパス比 (sensor15) のセンサーデータのみを用いた場合に関しては，3. シミュレーションスタディで得

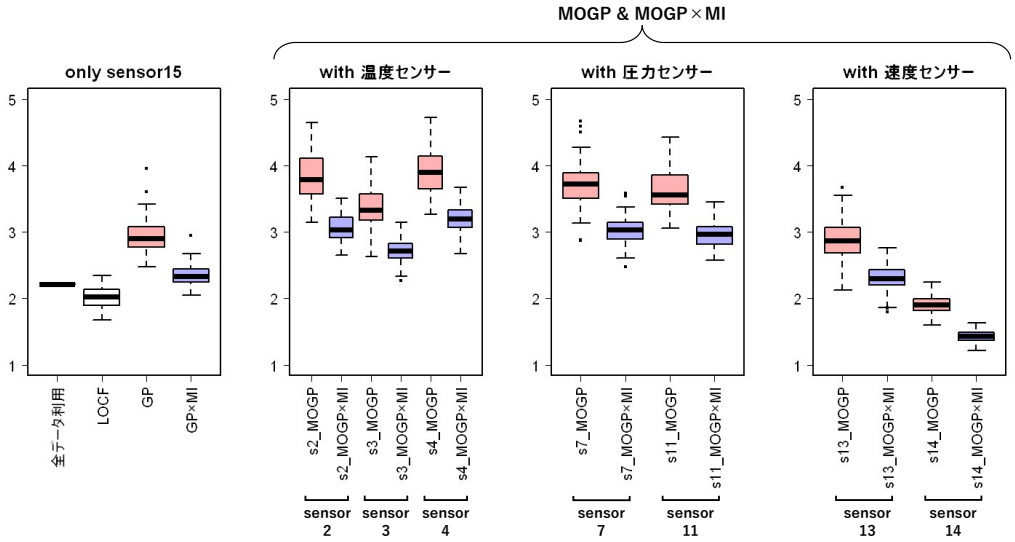


図 8　$\hat{\beta}_y$ の分布 (適用例)

られた結果とほぼ同様の傾向が認められた．すなわち，LOCF では過小推定傾向，GP では過大推定傾向が見られ，多重代入法を組み合わせた GP×MI では GP よりもマイルドな推定結果となっていた．

一方，その他のセンサーデータも併せて用いた場合に関しては，どのセンサーを補助変数として選んだかによって，推定結果が大きく異なっていた．この結果は，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推定値を得てしまう恐れもあることを示唆しており，分析データに関する領域知識が十分でない状況における課題が浮かび上がった．

5　おわりに

5.1　総括

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．

5.2 今後の展望

本研究では，ガウス過程回帰を補完モデルとして用い，多重代入法の枠組みでパラメータ推定を行う，というアプローチが取られている．ここで扱われたのはガウス過程回帰と相性の良い連続値の時間共変量のみであり，カテゴリカルな時間共変量については検討されていない．しかし，患者から得られる時間共変量には，検査値等の連続的な値をとる変数の他にも，症状や曝露の有無，重症度，併用薬の種類といったカテゴリカルな変数が数多く存在する．したがってその適切な扱い，すなわち補完法に関しても，何らかの解決策が与えられるべきである．

また，Multi-output ガウス過程回帰の利用により，複数の変数間での情報の共有の方向性は示されたが，複数の個体間での情報の共有という方向性も考えられる．個々の個体に注目したとき，ガウス過程回帰を適用して補完モデルを得るにはデータが少な過ぎる，という状況は十分にあり得る．そのような場合でも，全個体で見れば平均的な傾向を掴める可能性があり，その平均的な傾向を共有しながら個々の個体の補完モデルを得る，ということができるかもしれない．

また一般に，補完モデルには分析モデルのアウトカムに由来する変数を含めるべきとされているが，本研究ではこの観点からの検討はなされていない．分析モデルを Cox 回帰モデルとした場合においては，打切り指標と Nelson-Aalen 推定量 (累積ハザード関数の推定量) を含めることを推奨する先行研究がある [9][3]．分析モデルを時間共変量を含む Cox 回帰モデルとし，補完モデルにガウス過程回帰を用いた場合における，先行研究の拡張の方法は自明でないが，これらを上手く取り込むことでさらに推定が改善する可能性はあり，検討の価値があると考えられる．

以上を踏まえたとき，本研究のアプローチを継承した場合の次の展開としては，以下のようにまとめられる．

- 本アプローチにおいてカテゴリカルな時間共変量も扱えるよう対応し，さらに，連続変数とカテゴリカル変数の両方を含む，複数の変数間での情報の共有の方向性を示す．併せて，複数の個体間での情報の共有の方向性についても検討する．
- 本アプローチにおけるアウトカム由来変数の取り込み方について検討し，それらのパラメータ推定における有用性を検証する．

カテゴリカルな時間共変量ついては，ガウス過程回帰の自然な拡張により対応できると考えられる (例えば，カテゴリカルな観測値の背後に潜在的な連続値の時間共変量を想定する等)．変数型を限定しない複数の変数間での情報の共有，および複数の個体間での情報の共有については，Multi-output ガウス過程回帰の拡張を考えることになると思われる．これに関しては [10] などの先行研究が参考になるかもしれない．アウトカム由来変数の取り込みについては，補完モデルにガウス過程回帰を用いているがゆえの難しさもあるだろうと予想される (例えば，累積ハザード関数であれば単調増加でなくてはならないが，こういったガウス過程回帰と相性のあまり良くない制約も考慮する必要がある等)．

補遺 A　ガウス過程回帰

本節では一般論として議論を行うため，他節とは独立した表記法を用いる．なお，本節をまとめるに当たっては主に [11] を参考にした．

A.1　Single-output ガウス過程回帰

平均関数を $m(x)$，共分散関数 (カーネル関数) を $k(x, x'; \boldsymbol{\theta})$ として，関数 $f(x)$ を生成するガウス過程：

$$f(x) \sim GP\left(m(x), k(x, x'; \boldsymbol{\theta})\right) \tag{3}$$

は，任意の仮想的な入力の列 $x_{(1)}, x_{(2)}, \ldots$ とそれに対応する出力の列 $f(x_{(1)}), f(x_{(2)}), \ldots$ を考えて，以下のように多変量正規分布 (ガウス分布) として読みかえられる．

$$\begin{bmatrix} f(x_{(1)}) \\ f(x_{(2)}) \\ \vdots \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} k(x_{(1)}, x_{(1)}; \boldsymbol{\theta}) & k(x_{(1)}, x_{(2)}; \boldsymbol{\theta}) & \cdots \\ k(x_{(2)}, x_{(1)}; \boldsymbol{\theta}) & k(x_{(2)}, x_{(2)}; \boldsymbol{\theta}) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}\right) \tag{4}$$

ただし，便宜的に平均ベクトルは $\mathbf{0}$ とした．なお，ここで $\boldsymbol{\theta}$ は共分散関数のパラメータ (ハイパーパラメータ) を表す．上式は，

$$\boldsymbol{f} \sim N(\mathbf{0}, \boldsymbol{K}) \tag{5}$$

のように略記され，$\boldsymbol{K}$ はグラム行列と呼ばれる．このように任意次元のベクトル $\boldsymbol{f}$ を生成する多変量正規分布を想定した下で，ガウス過程回帰では，多変量正規分布が持つ便利な性質を利用して，実際のデータから回帰曲線を求める．

今，N 個の入力 $x_1, x_2, \ldots, x_N$ に対応して出力 $f(x_1), f(x_2), \ldots, f(x_N)$ があり，誤差 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ ($\varepsilon_n \sim N(0, \sigma^2)$) が乗った $y(x_1), y(x_2), \ldots, y(x_N)$ が観測されるとする．このとき，次のように書くことができる．

$$\boldsymbol{f}_N \sim N(\mathbf{0}, \boldsymbol{K}_N) \tag{6}$$

$$\boldsymbol{y}_N \sim N(\mathbf{0}, \boldsymbol{K}_N + \sigma^2 \boldsymbol{I}) \tag{7}$$

ここで，

$$\boldsymbol{f}_N = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix}, \quad \boldsymbol{y}_N = \begin{bmatrix} y(x_1) \\ \vdots \\ y(x_N) \end{bmatrix}, \quad \boldsymbol{K}_N = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix} \tag{8}$$

である (以降，共分散関数は $\boldsymbol{\theta}$ を省略して $k(x, x')$ と表記する)．未観測の M 個の入力 $x_{(1)}, x_{(2)}, \ldots, x_{(M)}$ とそれに対応する出力 $f(x_{(1)}), f(x_{(2)}), \ldots, f(x_{(M)})$ を想定して，観測データ $\boldsymbol{y}_N$ の下での $\boldsymbol{f}_M = \begin{bmatrix} f(x_{(1)}) & \cdots & f(x_{(M)}) \end{bmatrix}^T$ の事後分布 $p(\boldsymbol{f}_M | \boldsymbol{y}_N)$ を求めることを考える．事前

分布 $p(\boldsymbol{f}_M)$ についてもやはり $\boldsymbol{f}_M \sim N(\boldsymbol{0}, \boldsymbol{K}_M)$ と書けることから，まず，同時分布 $p(\boldsymbol{f}_M, \boldsymbol{y}_N)$ は次のようになる．

$$\begin{bmatrix} \boldsymbol{y}_N \\ \boldsymbol{f}_M \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_N + \sigma^2 \boldsymbol{I} & \boldsymbol{K}_{NM} \\ \boldsymbol{K}_{MN} & \boldsymbol{K}_M \end{bmatrix} \right) \tag{9}$$

ここで，

$$\boldsymbol{K}_M = \begin{bmatrix} k(x_{(1)}, x_{(1)}) & \cdots & k(x_{(1)}, x_{(M)}) \\ \vdots & \ddots & \vdots \\ k(x_{(M)}, x_{(1)}) & \cdots & k(x_{(M)}, x_{(M)}) \end{bmatrix} \tag{10}$$

$$\boldsymbol{K}_{NM} = \begin{bmatrix} k(x_1, x_{(1)}) & \cdots & k(x_1, x_{(M)}) \\ \vdots & \ddots & \vdots \\ k(x_N, x_{(1)}) & \cdots & k(x_N, x_{(M)}) \end{bmatrix} \tag{11}$$

$$\boldsymbol{K}_{MN} = \begin{bmatrix} k(x_{(1)}, x_1) & \cdots & k(x_{(1)}, x_N) \\ \vdots & \ddots & \vdots \\ k(x_{(M)}, x_1) & \cdots & k(x_{(M)}, x_N) \end{bmatrix} \tag{12}$$

である．多変量正規分布における条件付分布は，同時分布から公式的に直ちに導くことができ，事後分布 $p(\boldsymbol{f}_M|\boldsymbol{y}_N)$ は以下のようになる．

$$\boldsymbol{f}_M|\boldsymbol{y}_N \sim N(\boldsymbol{m}', \boldsymbol{K}'_M) \tag{13}$$

ただし，ここで，

$$\boldsymbol{m}' = \boldsymbol{K}_{MN}(\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_N \tag{14}$$

$$\boldsymbol{K}'_M = \boldsymbol{K}_M - \boldsymbol{K}_{MN}(\boldsymbol{K}_N + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{K}_{NM} \tag{15}$$

である．M 個の入力 $x_{(1)}, x_{(2)}, \ldots, x_{(M)}$ は任意に選べるため，これはつまり x の定義域全てにわたって $f(x)$ の事後分布が得られたことと同じである．この事後分布の平均が，ガウス過程回帰における回帰曲線に相当する．

ハイパーパラメータ $\boldsymbol{\theta}$ の推定については，対数周辺尤度 $\log p(\boldsymbol{y}_N|\boldsymbol{\theta})$ の最大化を考える (Type II 最尤推定)．(7) 式より，

$$\log p(\boldsymbol{y}_N|\boldsymbol{\theta}) = -\frac{1}{2} \log \left| \boldsymbol{K}_N + \sigma^2 \boldsymbol{I} \right| - \frac{1}{2} \boldsymbol{y}_N^T \left(\boldsymbol{K}_N + \sigma^2 \boldsymbol{I} \right)^{-1} \boldsymbol{y}_N + \text{const.} \tag{16}$$

であり，右辺を最大にする $\boldsymbol{\theta}$ を勾配法などにより求める．周辺尤度 (エビデンス) を基準としてハイパーパラメータを決定するこの枠組みは，ベイズモデル選択と見ることができる．

なお，以上では入力空間を 1 次元とした場合について論じたが，多次元とした場合も同様の議論が成り立つ．

A.2 Multi-output ガウス過程回帰

共分散関数 $k(x, x')$ を共有した，例えば 2 つの関数 $f_1(x)$, $f_2(x)$ を同時に扱いたい場合は，新たにハイパーパラメータ b_{ij} $(i, j = 1, 2)$ を導入して，

$$\begin{bmatrix} f_1(x_{(1)}) \\ f_1(x_{(2)}) \\ \vdots \\ f_2(x_{(1)}) \\ f_2(x_{(2)}) \\ \vdots \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \otimes \boldsymbol{K}\right) \tag{17}$$

を考えればよい．ただし，ここで $\otimes$ はクロネッカー積を表し，

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \otimes \boldsymbol{K} = \begin{bmatrix} b_{11}\boldsymbol{K} & b_{12}\boldsymbol{K} \\ b_{21}\boldsymbol{K} & b_{22}\boldsymbol{K} \end{bmatrix} \tag{18}$$

である．$f_1(x)$, $f_2(x)$ の出力を縦につないだ左辺のベクトルを $\boldsymbol{f}^{\mathrm{MO}}$ として，上式は以下のように略記される．

$$\boldsymbol{f}^{\mathrm{MO}} \sim N(\mathbf{0}, \boldsymbol{B} \otimes \boldsymbol{K}) \tag{19}$$

これは, $\boldsymbol{K}$ が $\boldsymbol{B} \otimes \boldsymbol{K}$ に置き換わっただけで, 本質的には (5) 式と同じであり, Multi-output ガウス過程回帰と言っても，実際には Single-output ガウス過程回帰をしている (従って，Single-output ガウス過程回帰と同様の手順で回帰曲線を求めることができる)．なお，さらに複数の関数を同時に扱いたければ，$\boldsymbol{f}^{\mathrm{MO}}$ をさらに縦に伸ばし，関数の数に合わせて $\boldsymbol{B}$ の次元を増やせばよい．

補遺 B　多重代入法

本節でも一般論として議論を行うため，他節とは独立した表記法を用いる．なお，本節をまとめるに当たっては主に [12] を参考にした．

B.1　補完値の生成

D_{obs} と D_{mis} で観測データと欠測データを表し，$\boldsymbol{r}$ を欠測の有無を表す変数 (欠測指標) とする．また，パラメータを $\boldsymbol{\theta}$ とする．多重代入法は基本的にベイズの枠組みであり，予測分布 $p(D_{mis}|D_{obs})$ から発生させた乱数を用いて欠測を補完すればよい，という発想である．なお，多重代入法では Missing at random (MAR)，すなわち $p(D_{mis}|D_{obs}, \boldsymbol{r}) = p(D_{mis}|D_{obs})$ が仮定されている．予測分布は，

$$p(D_{mis}|D_{obs}) = \int p(D_{mis}|D_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|D_{obs}) d\boldsymbol{\theta} \tag{20}$$

であることから，

$$\boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}|D_{obs}) \tag{21}$$

$$D_{mis}^{(m)} \sim p(D_{mis}|D_{obs}, \boldsymbol{\theta}^{(m)}) \quad (m = 1, \ldots, M) \tag{22}$$

などとして生成した $D_{mis}^{(m)}$ を用いて欠測を補完する．M 回の繰り返しの結果，M セットの疑似完全データ $(D_{mis}^{(m)}, D_{obs})$ が得られることになる．この繰り返し回数 M については，データに占める不完全データの割合のパーセンテージ以上という目安があるが，モンテカルロ誤差の制御の観点からは，可能な限り多く (100 回 $\sim$) 設定することがより望ましいとされている．

B.2　Rubin のルール

多重代入法における統計的推測は $p(\boldsymbol{\theta}|D_{obs})$ に基いて行われる．ここで，

$$p(\boldsymbol{\theta}|D_{obs}) = \int p(\boldsymbol{\theta}|D_{mis}, D_{obs}) p(D_{mis}|D_{obs}) dD_{mis} \tag{23}$$

であり，事後平均 $E[\boldsymbol{\theta}|D_{obs}]$ と事後分散 $V[\boldsymbol{\theta}|D_{obs}]$ を推定したいとする．Rubin のルールでは，M セットの疑似完全データそれぞれから得られた $\hat{\boldsymbol{\theta}}_1$，$\hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_M$ と，その分散の推定値 $\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_M$ を，

$$\hat{\boldsymbol{\theta}}_{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_m \tag{24}$$

$$\hat{V}(\hat{\boldsymbol{\theta}}_{\mathrm{MI}}) = W_{\mathrm{MI}} + \left(1 + \frac{1}{M}\right) B_{\mathrm{MI}} \tag{25}$$

のように統合し，最終的な推定値とする．ここで，W_{MI} と B_{MI} はそれぞれ補完内と補完間での分散を表したものであり，

$$W_{\mathrm{MI}} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m \tag{26}$$

$$B_{\mathrm{MI}} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{\mathrm{MI}})(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{\mathrm{MI}})^T \tag{27}$$

である．Rubin のルールは基本的に周辺平均と周辺分散の公式から導かれるものであり，

$$E[\boldsymbol{\theta}|D_{obs}] = E_{D_{mis}|D_{obs}}\left[E(\boldsymbol{\theta}|D_{mis}, D_{obs})|D_{obs}\right] \tag{28}$$

$$\begin{aligned} V[\boldsymbol{\theta}|D_{obs}] = {} & E_{D_{mis}|D_{obs}}\left[V(\boldsymbol{\theta}|D_{mis}, D_{obs})|D_{obs}\right] \\ & + V_{D_{mis}|D_{obs}}\left[E(\boldsymbol{\theta}|D_{mis}, D_{obs})|D_{obs}\right] \end{aligned} \tag{29}$$

を考えている．すなわち，(28) 式の右辺が $\frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\theta}}_m$ に，(29) 式の右辺第 1 項と第 2 項がそれぞれ W_{MI} と B_{MI} に対応している．ただし，$p(D_{mis}|D_{obs})$ より生成した $D_{mis}^{(m)}$ で補完した疑似完全データから得られる推定値 $\hat{\boldsymbol{\theta}}_m$, $\hat{V}_m$ が，それぞれ $E(\boldsymbol{\theta}|D_{mis}^{(m)}, D_{obs})$, $V(\boldsymbol{\theta}|D_{mis}^{(m)}, D_{obs})$ を近似するものと見なしている．また，(25) 式における $\left(1+\frac{1}{M}\right)$ は，補完の回数が有限回であることに対する補正である．

検定や信頼区間の構成については，$\hat{\boldsymbol{\theta}}_{\mathrm{MI}}$ と $\hat{V}(\hat{\boldsymbol{\theta}}_{\mathrm{MI}})$ の関連する要素を抜き出し，t 分布に基いて行われる．その際の自由度 υ は以下で与えられる．

$$\upsilon = (M-1)\frac{1}{\lambda^2}, \quad \lambda = \frac{\left(1+\frac{1}{M}\right)B_{\mathrm{MI}}}{\hat{V}(\hat{\boldsymbol{\theta}}_{\mathrm{MI}})} \tag{30}$$

ただし，この υ は完全データが得られた場合の自由度を超える可能性があり，次の修正自由度 υ_{adj} も提案されている．

$$\upsilon_{adj} = (M-1)\frac{\upsilon \times \upsilon_{obs}}{\upsilon + \upsilon_{obs}}, \quad \upsilon_{obs} = \frac{(n-p+1)(n-p)}{n-p+3}(1-\lambda) \tag{31}$$

ここで，p はパラメータの数，n は完全データのサンプルサイズである．

補遺 C　シミュレーションスタディ結果詳細

表 1 に 3 節で行ったシミュレーションスタディの結果の詳細を示す．時間共変量の係数の推定値 $\hat{\beta}_y$ のバイアスとカバレッジについては，本論で言及した通りである．固定共変量 (連続および 2 値) の係数の推定値 $\hat{\beta}_x$，$\hat{\beta}_z$ については，いずれの推定方法においてもバイアスはほぼなく，また，カバレッジも良好であった．なお，モンテカルロ誤差に関しては最大でも 0.005 を超えないと見積もられた．

表 1　シミュレーションスタディ結果詳細

		推定結果 (1000回試行) の要約		
推定対象	推定方法	平均	標準偏差	カバレッジ
β_x	完全履歴	0.255	0.079	94.6
(真値：0.25)	LOCF	0.250	0.078	95.3
	GP	0.249	0.082	94.6
	GP×MI	0.249	0.079	95.2
	MOGP	0.254	0.079	94.8
	MOGP×MI	0.252	0.078	95.7
β_z	完全履歴	0.256	0.147	95.1
(真値：0.25)	LOCF	0.250	0.147	95.2
	GP	0.252	0.155	94.7
	GP×MI	0.250	0.148	95.6
	MOGP	0.255	0.150	94.9
	MOGP×MI	0.253	0.148	95.0
β_y	完全履歴	0.254	0.047	94.7
(真値：0.25)	LOCF	0.140	0.047	31.4
	GP	0.318	0.076	43.7
	GP×MI	0.234	0.046	89.2
	MOGP	0.272	0.053	84.3
	MOGP×MI	0.240	0.046	95.0

謝辞

まず最初に，指導教員である杉本知之教授に感謝申し上げます．文字通り私をここまで教え導いて下さっただけでなく，小心な私の背中を押して，博士後期課程進学という大冒険へ踏み切らせて下さいました．

また，副指導教員を引受けて下さった岩山幸治准教授に感謝申し上げます．多くの学生の研究指導を抱え，ご多忙の身であるにも関わらず，魯鈍な私の愚昧な質問にも親切丁寧にお答え下さいました．

最後に，ゼミを共にした森田息吹さんと川崎大輔さん，また，大学院同期の皆さんに感謝申し上げます．皆さんは私の無能力を隅々まで映し出し，そして徹底して見つめ直すための実によく磨かれた鏡でした．皆さんの経歴と有能ぶり，皆さんの若さと未来に，羨望と嫉妬が絶える時はなかったことを告白します．

この 2 年間，私は私の小心と魯鈍と無能力とを，いずれも克服することはできませんでした．それでもこうして何とか修士論文を形にすることができたのは，ここに挙げた方々が支えとなってくれたお陰です．最後に重ねて感謝申し上げます．

参考文献

[1] 杉本知之. 生存時間解析 (統計解析スタンダード). 朝倉書店, 2021.

[2] Lang Wu, Wei Liu, Grace Y. Yi, and Yangxin Huang. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, 2012:1–17, 2012.

[3] Margarita Moreno-Betancur, John B. Carlin, Samuel L. Brilleman, Stephanie K. Tanamas, Anna Peeters, and Rory Wolfe. Survival analysis with time-dependent covariates subject to missing data or measurement error: Multiple Imputation for Joint Modeling (MIJM). *Biostatistics*, 19(4):479–496, 2018.

[4] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45:1–67, 2011.

[5] Christopher K. Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[6] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

[7] Robert Dürichen, Marco A.F. Pimentel, Lei Clifton, Achim Schweikard, and David A. Clifton. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

[8] A. Saxena and K. Goebel. Turbofan engine degradation simulation data set. *NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository)*, 2008.

[9] Ian R. White and Patrick Royston. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998, 2009.

[10] Pablo Moreno-Muñoz, Antonio Artés-Rodríguez, and Mauricio A. Álvarez. Heterogeneous multi-output Gaussian process prediction. *arXiv preprint arXiv:1805.07633*, 2018.

[11] 持橋大地・大羽成征. ガウス過程と機械学習 (機械学習プロフェッショナルシリーズ). 講談社, 2019.

[12] 高井啓二・星野崇宏・野間久史. 欠測データの統計科学——医学と社会科学への応用 (調査観察データ解析の実際 1). 岩波書店, 2016.