

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年8月27日(27.08.2020)



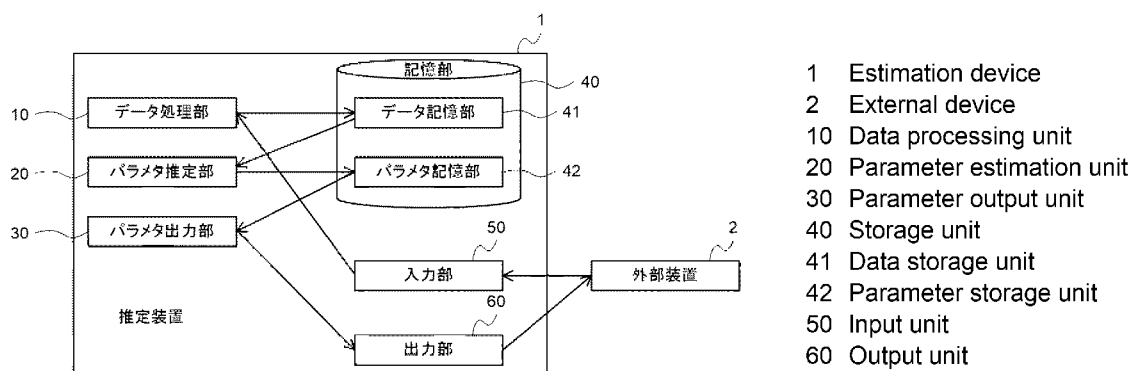
(10) 国際公開番号

WO 2020/170867 A1

- (51) 国際特許分類:
G06N 20/00 (2019.01)
- (21) 国際出願番号: PCT/JP2020/004909
- (22) 国際出願日: 2020年2月7日(07.02.2020)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2019-029769 2019年2月21日(21.02.2019) JP
- (71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: 幸島 匡宏 (KOJIMA, Masahiro); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP). 松林 達史 (MATSUBAYASHI, Tatsushi); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP). 戸田 浩之 (TODA, Hiroyuki); 〒1808585 東京都武蔵野市緑町三丁目9番11号 NTT 知的財産センタ内 Tokyo (JP).
- (74) 代理人: 特許業務法人太陽国際特許事務所 (TAIYO, NAKAJIMA & KATO); 〒1600022 東京都新宿区新宿4丁目3番17号 Tokyo (JP).
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH,

(54) Title: ESTIMATION DEVICE, ESTIMATION METHOD, AND PROGRAM

(54) 発明の名称: 推定装置、推定方法、及びプログラム



(57) Abstract: According to the present invention, a calculation time is suppressed, and parameters of a model, which represents a probability distribution of truncated data, are enabled to be accurately estimated. According to the present invention, a parameter estimation unit 20 estimates the parameters of the model by optimizing an objective function which is divergence between the model and a probability distribution of the truncated data, which is obtained from the truncated data, wherein the model expresses the probability distribution of the truncated data that includes observation data of samples, observation values of which are observed, observation data of samples, observation values of which are not observed, and variables that express whether the observation values for the respective samples are observed, the probability distribution being expressed by using a probability density function of the observation data, which is expressed as a mixed model of each component, which expresses a distribution of the observation values that correspond to the respective samples of the truncated data.

CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,
HN, HR, HU, ID, IL, IN, IR, IS, JO, KE, KG, KH,
KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY,
MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,
NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT,
QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 一 国際調査報告 (条約第21条(3))

(57) 要約: 計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができるようにする。パラメタ推定部20が、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、観測データの確率密度関数を用いて表される、当該打ち切りデータの確率分布を表すモデルと、当該打ち切りデータから得られる、当該打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、モデルのパラメタを推定する。

明 細 書

発明の名称：推定装置、推定方法、及びプログラム

技術分野

[0001] 本発明は、推定装置、推定方法、及びプログラムに係り、特に、打ち切りデータから混合モデルのパラメタを推定するための推定装置、推定方法、及びプログラムに関する。

背景技術

[0002] 打ち切りデータとは、観測値がある閾値以上（またはある閾値以下）であるサンプルについては、値が観測されず、閾値以上である、という情報しか得られないデータのことを指す。病気の発症や人の死亡などを記述する臨床データや、インターネット回線利用者の契約履歴データ、Eコマースサイトのサービス利用履歴データなど多くのデータが打ち切りデータとして表現される。打ち切りデータを用いる問題の代表例は、例えばある機器が故障するまでに要する時間の分布を推定する、生存時間分析である。生存時間分布は、初期故障や劣化故障などの存在によりしばしば多峰性をもつため、混合モデルを用いた分布推定が広く用いられる。

[0003] 打ち切りデータから混合モデルのパラメタを推定するには、文献（非特許文献1）で提案されているEMCM（Expectation Maximization for Censored Mixture models）アルゴリズムが適用できる。

先行技術文献

非特許文献

[0004] 非特許文献1：Didier Chauveau. A stochastic em algorithm for mixtures with censored data. Journal of statistical planning and inference, Vol.46, No.1, 1995, pp.1-25.

発明の概要

発明が解決しようとする課題

- [0005] しかし、このアプローチには以下の2つの問題があった。
- [0006] 1つ目は、局所最適解の存在である。すなわち、EMCMは目的関数の単調減少は保証されるものの収束先は初期値に依存して変わるため、異なる初期値からの繰り返し実行が必要となってしまう、という問題があった。
- [0007] 2つ目は、モデルで利用する確率分布の切断分布（ある一定の範囲でしか値をとらないように変更した分布）の統計量の計算の必要性である。すなわち、1次元正規分布の切断分布である1次元切断正規分布などの例外的なものを除き、この統計量は解析的に計算することができず、モンテカルロ法などの数値計算を利用することが必要となる、という問題があった。EMCMは、パラメタの更新と統計量の計算を何度も繰り返すアルゴリズムであるから、各反復で数値計算を繰り返すことは避けることが望ましいと考えられる。
- [0008] 本発明は上記の点に鑑みてなされたものであり、計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができる推定装置、推定方法、及びプログラムを提供することを目的とする。

課題を解決するための手段

- [0009] 本発明に係る推定装置は、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する推定装置であって、前記打ち切りデータの入力を受け付ける入力部と、前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定するパラメタ推定部と、を備えて構成される。
- [0010] また、本発明に係る推定方法は、観測値が観測されたサンプルの観測デー

タと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する推定方法であって、入力部が、前記打ち切りデータの入力を受け付け、パラメタ推定部が、前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定する。

[0011] また、本発明に係るプログラムは、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する処理をコンピュータに実行させるプログラムであって、入力部が、前記打ち切りデータの入力を受け付け、パラメタ推定部が、前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定することを含む処理をコンピュータに実行させるプログラムである。

[0012] 本発明に係る推定装置、推定方法及びプログラムによれば、入力部が、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの入力を受け付ける。

[0013] そして、パラメタ推定部が、入力部が受け付けた打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表さ

れる、観測データの確率密度関数を用いて表される、打ち切りデータの確率分布を表すモデルと、入力部が受け付けた打ち切りデータから得られる、打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、モデルのパラメタを推定する。

[0014] このように、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、観測データの確率密度関数を用いて表される、当該打ち切りデータの確率分布を表すモデルと、当該打ち切りデータから得られる、当該打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、モデルのパラメタを推定することにより、計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができる。

[0015] また、本発明に係る推定装置の前記打ち切りデータの確率分布を表すモデルは、前記観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さを用いて表される、前記変数の確率分布と、前記観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さを用いて表される、前記変数が与えられた下での、前記観測データの確率分布と、を用いて表されることができる。

[0016] また、本発明に係る推定装置の前記目的関数は、前記打ち切りデータの確率分布を表すモデルと、前記打ち切りデータの確率分布とのカルバックライブラーダイバージェンス、又は L_2 ダイバージェンスであることができる。

発明の効果

[0017] 本発明の推定装置、推定方法、及びプログラムによれば、計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができる。

図面の簡単な説明

[0018] [図1] 1次元の打ち切りデータの例を表すイメージ図である。

[図2] 2次元の打ち切りデータの例を表すイメージ図である。

[図3] 1次元の打ち切りデータの生存時間表現の例を表すイメージ図である。

[図4] 2次元の打ち切りデータの生存時間表現の例を表すイメージ図である。

[図5] 本発明の実施の形態に係る推定装置として機能するコンピュータの概略構成を示すブロック図である。

[図6] 本発明の実施の形態に係る推定装置の構成を示すブロック図である。

[図7] 本発明の実施の形態に係る推定装置の推定処理ルーチンを示すフローチャートである。

発明を実施するための形態

[0019] 以下、本発明の実施の形態について図面を用いて説明する。

[0020] <本発明の実施の形態に係る推定装置の原理>

まず、本発明の実施形態の原理について説明する。

[0021] 本発明の実施の形態では、次に示す目的関数の異なる2つの手法を構築した。1つ目は確率分布間のKullback-Leibler (KL:カルバック-ライブラー) ダイバージェンス最小化に基づく推定手法、2つ目は確率分布間の L_2 ダイバージェンス最小化に基づく推定手法である。1つ目の推定手法では非常にシンプルな繰り返し計算でパラメタの推定が可能であり、2つ目の推定手法では、繰り返しの必要すらく解析的にパラメタの推定が可能である。

[0022] 本発明の実施の形態に係る手法を構築する上でポイントとなるのは、`exemplar based model` (以下、`emb`) と呼ばれるアプローチ (参考文献1) の利用である。

[参考文献1] Danial Lashkari and Polina Golland, "Convex clustering with exemplar-based models", In Advances in neural information processing systems, 2008, pp.825-832.

[0023] `emb` アプローチでは、混合モデルの各コンポーネントのパラメタを陽にパラメタとしては扱わず、データ点の存在する点に各コンポーネントを配置する。これによって、混合比 (各コンポーネントの重みパラメタ) のみの推

定を行う定式化をすることで大域的最適解へ収束するアルゴリズムが構築できる。上記の手法は、このアプローチに対して打ち切りデータを入力として利用できるよう発展させたものとみなすことができる。

[0024] 更に、上述したEMCMで必要となる切断分布の統計量はコンポーネントのパラメタの推定に必要なものであったため、このアプローチの採用により数値計算の繰り返し実行の必要のない手法を構築することができる。

[0025] なお、文献（参考文献1）の方法では、通常のデータを用いてKLダイバージェンス最小化によるembを行う手法を提案しているが、打ち切りデータを扱うことができない。

[0026] 確率分布の誤差の関数に L_2 ダイバージェンスを用いるアプローチは文献（参考文献2）でも考えられているが、打ち切りデータではなく通常のデータを入力する状況を考えている。

[参考文献2] DavidW Scott, "Parametric statistical modeling by minimum integrated square error", Technometrics, Vol.43, No.3, 2001, pp.274-285.

[0027] また、打ち切りデータを利用し、かつ L_2 ダイバージェンスを用いる研究には文献（参考文献3）が存在するが、指数分布、ワイブル分布のように単峰の単純な分布をモデルとして用いる場合のみが考えられている。

[参考文献3] Srabashi Basu, Ayanendranath Basu, and MCJones, "Robust and efficient parametric estimation for censored survival data", Annals of the Institute of Statistical Mathematics, Vol.58, No.2, 2006, pp. 341-355.

[0028] <<準備>>

<<<打ち切りデータ>>>

まず、打ち切りデータについて説明する。図1に、1次元の打ち切りデータの例を示す。図1の例に示すように、打ち切りデータの代表例である機器故障データを用いて説明する。

[0029] 機器故障データでは、各機器の設置された時刻と故障が起こった際の故障

の時刻が記録されている。機器 1 及び 2 に関しては観察期間中に故障が起きているため、その時刻が記録されている。一方、機器 3 及び 4 に関しては観察期間中には故障せずに、観察が打ち切られたため、故障の時刻が記録されていない。

[0030] しかしながら、機器 3 及び機器 4 の何れも、いつかは必ず故障するとして、故障の時刻は観察終了時刻以降にあるということは読み取れる。このように機器 1 と 2 のように観測値（故障の時刻）が分かるデータと機器 3 と 4 のように観測値（故障の時刻）がある値以上であることが分かるデータの組からなるデータを、打ち切りデータと呼ぶ。

[0031] 図 1 の機器故障データは、1 次元の打ち切りデータであるが、本手法では 2 次元以上の打ち切りデータも扱うことができるためここで説明する。

[0032] 図 2 は、ある 2 つのサービスの利用者の利用期間を表す 2 次元の打ち切りデータである。図 2 のケースでは、利用者の（少なくとも 1 つの）サービスを利用開始した時刻と各サービスを追加契約または解約した時刻が記録されている。

[0033] 利用者 1 は両サービスを同時に利用開始し、観察期間中に同時に解約しているため両方のサービスの解約時刻が記録されている。利用者 2 は両サービスを同時に利用開始し、観察期間中にサービス 2 のみを解約している。利用者 3 は両サービスを同時に利用開始し、観察期間中にサービス 1 のみを解約している。利用者 4 はサービス 2 をまず利用開始し、観察期間中にサービス 1 を追加契約している。よって、観測の打ち切りによって、利用者 2 のサービス 1 の解約時刻、利用者 3 のサービス 2 の解約時刻、利用者 4 のサービス 1 と 2 の解約時刻は記録されていない。

[0034] このように 2 次元の打ち切りデータにおいては、打ち切られている値の次元が異なる、3 通りの打ち切りの種類が存在する。一般に n 次元の打ち切りデータでは $2^n - 1$ 種類の打ち切りが存在する。なお、以下、各次元に観測が打ち切られるかが決まる例を用いて説明するが、どれか一つの要素が打ち切られたときに全ての要素が観測されなくなるという状況でも同様のアプロー

チで対応することができる。

[0035] ここで、打ち切りデータの定義を与える。扱いを簡単にするため、観測データは図1のようなカレンダー時刻ではなく、生存時間（機器の設置から故障までの時間、サービスの契約から解約までの時間）を用いて表現する。図1の観測データを生存時間で表現したものを図3、図2の観測データを生存時間で表現したものを図4に示す。多次元の打ち切りデータとして定義するため、以下では図4を説明の例に用いる。

[0036] 打ち切りデータを

$$\mathcal{D} = \{x_i, w_i\}_{i=1}^n$$

と書く。ここで、

$$x_i$$

と、

$$w_i$$

とは共に d_x 次元のベクトルであり、

$$x_i \in \mathbb{R}^{d_x}$$

、

$$w_i \in \{0, 1\}^{d_x}$$

である。 x_{ij} が利用者 i のサービス j の利用時間、 w_{ij} が利用者 i のサービス j の解約時刻が記録されたか ($w_{ij} = 1$)、打ち切りにより記録されなかったか ($w_{ij} = 0$) を表す。同様に、 i 番目の利用者の、観測終了時刻までの時間の長さを

$$v_i \in \mathbb{R}^{d_x}$$

と書く。 v_{ij} が利用者 i のサービス j 利用開始時刻から観測終了時刻までの長さを表す。打ち切りにより観測値が観測されなかったとき ($w_{ij} = 0$) には、 $x_{ij} = v_{ij}$ と設定されているとする。

[0037] <<<混合モデル>>>

次に、本発明の実施の形態で用いるモデルについて説明する。混合モデルで表される、観測値の確率密度関数は一般に下記式（１）で定義される。

[0038] [数1]

$$f(x|\theta) = \sum_{k=1}^K \theta_k \psi_k(x) = \theta^T \psi(x), \quad \dots (1)$$

[0039] ここで、Kは混合数、

$$\psi_k(x)$$

がk番目コンポーネントの確率分布を表す。コンポーネントの確率分布

$$\psi_k(x)$$

には、例えば下記式で表されるガウス分布が利用できる。

[0040] [数2]

$$\psi_k(x) = \mathcal{N}(x|\mu_k, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma^2)^{d_x}} \exp\left(-\frac{\|x - \mu_k\|^2}{2\sigma^2}\right)$$

[0041] ここで、

$$\mu_k$$

と σ とは、ガウス分布の平均と標準偏差を表す。ただし、e b m（参考文献1）のアプローチに従い、コンポーネントの確率分布のパラメタ

(ガウス分布の場合, $\mu_1, \dots, \mu_K$)

は $K=n$ として、それぞれが観測データ点に対応するよう

$$(\mu_1, \dots, \mu_n) = (x_1, \dots, x_n)$$

と設定されたもの等であるとする。この場合には、観測値の確率密度関数は、観測データの各々に対する、各コンポーネントの混合モデルで表され、各コンポーネントに含まれる、ガウス分布の平均を、対応する観測値とする。本手法は打ち切りデータを扱うため、 $K=n$ とし、値が観測されていれば（ $w_{ij}=1$ ）、 $\mu_{ij}=x_{ij}$ 、そうでなければ（ $w_{ij}=0$ ）、 $\mu_{ij}=x_{ij}+\varepsilon$

と設定してもよい。 ε は0以上の値を取る確率分布（例えば指数分布）からランダムに生成した値を表す。また、データ数が多い場合には、例えばランダムに選んだ100個のデータのみを用いてもよいし、事前知識に基づいて設定したコンポーネントを用いてもよい。標準偏差 σ は交差検証法等により決定することが可能である。

[0042] 上記モデルを用いたときの打ち切りデータ

$$\mathcal{D} = \{x_i, w_i\}$$

の生成過程は次のように記述できる。まず、各サンプルについて観測終了までの時間の長さ

$$v_i$$

が既知のもと、打ち切りが起こるか否かを表す変数

$$w_i$$

が、下記式（2）の確率分布に従い生成される。

[0043] [数3]

$$\begin{aligned} P(w_i|\theta) &= F(w_i|\theta; v_i) = \int_{v_i^c}^{\infty} \left\{ \int_{-\infty}^{v_i^o} f(x_i|\theta) dx_i^o \right\} dx_i^c \\ &= \sum_{k=1}^K \theta_k \Psi_k^w(w_i; v_i) = \theta^T \Psi_i^w, \end{aligned} \quad \dots (2)$$

$$\Psi_k^w(w_i; v_i) = \int_{v_i^c}^{\infty} \left\{ \int_{-\infty}^{v_i^o} \psi_k(x) dx^o \right\} dx^c. \quad \dots (3)$$

[0044] ただし、

$$x_i$$

と、

$$v_i$$

とのうち、 $w_{ij} = 1$ である、観測値が観測された要素の集合を

$$x_i^o = \{x_{ij}|w_{ij} = 1\}, v_i^o = \{v_{ij}|w_{ij} = 1\}$$

とし、 $w_{ij} = 0$ である、観測値が観測されなかった集合を

$$x_i^c = \{x_{ij}|w_{ij} = 0\}, v_i^c = \{v_{ij}|w_{ij} = 0\}$$

とする。また、観測値が観測された要素と、観測値が観測されなかった要素とを区別しない場合に、観測データと呼ぶこととする。コンポーネントの確率分布

$$\psi_k(x)$$

に、上記式（２）のガウス分布のように標準的な確率分布を利用すると、

$$\Psi_k^w(w_i; v_i)$$

を累積密度関数を用いて解析的に計算できる。

[0045] $w_i \neq 0$

であり、少なくとも一つ観測された要素がある場合、

$$x_i$$

は下記式（４）で表される分布にしたがって生成される。

[0046] [数4]

$$P(x_i|w_i, \theta) = \delta(x_i^c - v_i^c) f_{tr}(x_i^o|w_i, \theta) \quad \dots (4)$$

[0047] ここで、

$$\delta(\cdot)$$

はデルタ関数であり、 f_{tr} は f を観測されなかった要素に関して周辺化した分布の切断分布であり、下記式で表される。

[0048] [数5]

$$f_{tr}(x_i^o|w_i, \theta) = \begin{cases} \frac{f^o(x_i^o|\theta)}{F(w_i|\theta; v_i)} & (\text{if } x_i^o \leq v_i \text{ and } x_i^c \geq v_i) \\ 0 & (\text{otherwise}) \end{cases}$$

[0049] ただし、 f_{tr} は下記式（５）～（６）に従う。

[0050]

[数6]

$$\begin{aligned}
 f^{x^o}(x_i^o|\theta) &= \int_{v^o}^{\infty} f(x_i|\theta) dx_i^c \\
 &= \sum_{k=1}^K \theta_k \Psi_{ik}^{x^o} = \theta^T \Psi_i^{x^o} \quad \dots (5)
 \end{aligned}$$

$$\Psi_{ik}^{x^o} = \Psi_k^{x^o}(x_i^o; v_i) = \int_{v^o}^{\infty} \psi_k(x_i) dx_i^c \quad \dots (6)$$

[0051] コンポーネントが上記式(2)のガウス分布である場合、

 $\Psi_{ik}^{x^o}$

は

$$\Psi_{ik}^{x^o} = \mathcal{N}(x_i^o|\mu_k^{x_i^o}, \sigma^2) \int_{v_i^o}^{\infty} \mathcal{N}(x_i^c|\mu_k^{x_i^o}, \sigma^2) dx_i^c$$

となる。ただし、

 $\mu_k^{x_i^o}, \mu_k^{x_i^c}$

はそれぞれ

 μ_k

から

 x_i^o

と

 x_i^c

に対応する次元の要素を抜き出したベクトルである。

[0052] $w_i = 0$

であり、一つも観測された要素がない場合、下記式(7)で示すように、デルタ関数のみの表現となる。

[0053] [数7]

$$P(x_i|w_i = 0, \theta) = \delta(x_i^c - v_i) \quad \dots (7)$$

[0054] よって、上記をまとめると、各打ち切りデータの生成確率は、下記式（８）で与えられる。

[0055] [数8]

$$P(x_i, w_i | \theta) = P(x_i | w_i, \theta) P(w_i | \theta) \quad \dots (8)$$

[0056] <<KL ダイバージェンスと L_2 ダイバージェンス>>

次に、提案手法の目的関数を定義する際に利用するダイバージェンスについて記す。良く知られるように、確率分布 $p(x)$ と $q(x)$ とに対するカルバックライブラー（KL）ダイバージェンスは下記式（９）で定義される。

[0057] [数9]

$$\begin{aligned} KL(p, q) &= \mathbb{E}_q \left[\log \frac{q(x)}{p(x)} \right] \quad \dots (9) \\ &= \int q(x) \log q(x) dx - \int q(x) \log p(x) dx \end{aligned}$$

[0058] これに加え、本発明の実施の形態では、下記で定義される L_2 ダイバージェンス（参考文献２）も利用する場合についても説明する（下記式（１０））。

[0059] [数10]

$$\begin{aligned} L_2(p, q) &= \int \{p(x) - q(x)\}^2 dx \quad \dots (10) \\ &= \int \{p(x)\}^2 dx + \int \{q(x)\}^2 dx - 2 \int p(x) q(x) dx \end{aligned}$$

[0060] L_2 ダイバージェンスは、２つの確率密度関数の２乗誤差として定義されている。何れのダイバージェンスを用いるべきかについては、問題に応じて異なる。よって、本発明の実施の形態では、何れのダイバージェンスを利用する場合にも適用できるように、２種類の手法を構築することとした。

[0061] <<KL ダイバージェンスの最適化による推定>>

まず、目的関数としてKLダイバージェンスを利用する場合の提案手法を示す。打ち切りデータの確率分布を表すモデル

$$P(x, w|\theta)$$

と、打ち切りデータから得られる真の確率分布

$$P^*(x, w)$$

とのKLダイバージェンスは、上記式(9)の定義に従い、下記式(11)で与えられる。

[0062] [数11]

$$\mathcal{D}_{KL}(\theta) = \mathbb{E}_{P^*(x_i, w_i)} \left[\log \frac{P^*(x_i, w_i)}{P(x_i, w_i|\theta)} \right] \quad \dots (11)$$

[0063] これは、下記式(12)のように式変形することができる。

[0064] [数12]

$$\begin{aligned} \mathcal{D}_{KL}(\theta) &= \sum_{w_i \neq 0} P^*(w_i) \mathbb{E}_{P^*(x_i|w_i)} \left[\log \frac{P^*(x_i|w_i)P^*(w_i)}{P(x_i|w_i, \theta)P(w_i|\theta)} \right] \\ &\quad + P^*(w_i = 0) \log \frac{P^*(w_i = 0)}{P(w_i = 0|\theta)} \\ &= \sum_{w_i \neq 0} P^*(w_i) \mathbb{E}_{P^*(x_i^o|w_i)} \left[\log \frac{P^*(x_i^o|w_i)}{f^{x^o}(x^o|\theta)} \right] \\ &\quad + P^*(w_i = 0) \log \frac{P^*(w_i = 0)}{P(w_i = 0|\theta)} \\ &= \sum_{w_i \neq 0} P^*(w_i) \mathbb{E}_{P^*(x^o|w_i)} \left[-\log f^{x^o}(x^o|\theta) \right] \\ &\quad - P^*(w_i = 0) \log F(w_i = 0|\theta; v_i) + Const. \end{aligned} \quad \dots (12)$$

[0065] 定数項を除去し、未知である真の分布

$$P^*$$

に関する期待値を標本平均で置き換えれば、下記の式(13)が導かれる。

[0066]

[数13]

$$\begin{aligned}
\hat{D}_{KL}(\theta) &= -\frac{1}{n} \left\{ \sum_{\{i|w_i \neq 0\}} \log f^{x^o}(x_i^o | \theta) \right. \\
&\quad \left. + \sum_{\{i|w_i = 0\}} \log F(0 | \theta; v_i) \right\} \\
&= -\frac{1}{n} \left\{ \sum_{\{i|w_i \neq 0\}} \log(\theta^T \Psi_i^{x^o}) \right. \\
&\quad \left. + \sum_{\{i|w_i = 0\}} \log(\theta^T \Psi_i^w) \right\}.
\end{aligned} \quad \dots (13)$$

[0067] ここで、

$$\hat{D}_{KL}(\theta)$$

は、データから計算できる量であり、これを目的関数とすることでアルゴリズムを導く。具体的には、下記式（14）の最適化問題を解けばよい。

[0068] [数14]

$$\hat{\theta} = \arg \min_{\theta} \hat{D}_{KL}(\theta) \text{ s.t. } \theta_k \geq 0, |\theta| = 1.0. \quad \dots (14)$$

[0069] ここで、制約条件（パラメタの要素が0以上、かつ、和が1）は、混合モデル f が確率分布となるためのものである。ラグランジュの未定乗数法を用いると、上記の最適化問題の解は、下記式（15）を満たすことが分かる。

[0070] [数15]

$$\hat{\theta}_k = \frac{1}{n} \left\{ \sum_{\{i|w_i \neq 0\}} \frac{\theta_k \Psi_k^{x^o}(x_i^o; v_i)}{\theta^T \Psi_i^{x^o}} + \sum_{\{i|w_i = 0\}} \frac{\theta_k \Psi_k^w(w_i; v_i)}{\theta^T \Psi_i^w} \right\} \quad \dots (15)$$

[0071] よって、上記式（15）に基づいて

$$\hat{\theta}$$

の更新を繰り返すことにより、最適化が可能となる。なお、参考文献1の手法と同様に、計算量の削減と収束を早めるため、パラメタ更新の際に θ_k がある閾値（例えば $10^{-3}/n$ ）より小さい場合には、 $\theta_k = 0$ と設定した後に、全体を和が1になるように調整する、再正規化操作を行ってもよい。

[0072] <<L₂ダイバージェンスの最適化による推定>>

次に、 L_2 ダイバージェンスを利用する場合の提案手法を示す。KLダイバージェンスとは異なり、 L_2 ダイバージェンスの定義から直接目的関数を定義することはせず、KLダイバージェンスを用いた際の目的関数に注目することにより新たな目的関数を定義する。

[0073] KLダイバージェンスを用いた際の目的関数（式（12））に注目すると、混合モデル f の周辺分布

$$f^{x^o}(x_i^o|\theta)$$

と、変数が与えられた下での観測値の真の分布

$$P^*(x|w_i)$$

のKLダイバージェンスに対応する項と、観測値が観測されなかったことを表す変数のモデルの分布

$$P(w_i = 0|\theta)$$

と、真の分布

$$P^*(w_i = 0)$$

を用いた対数尤度比に対応する項との2つの項をそれぞれ

$$P(w_i)$$

で重み付き和を取ることで構成されていることが分かる。

[0074] この洞察に基づき、これら2つの各項でKLダイバージェンス／対数尤度比を用いている箇所を L_2 ダイバージェンスに置き換えることで、目的関数を下記式のように設計することができる。

[0075] [数16]

$$\begin{aligned} \mathcal{D}_{L2}(\theta) = & \sum_{w_i \neq 0} P^*(w_i) \int \left(f^{x^o}(x_i^o|\theta) - P^*(x_i^o|w_i) \right)^2 dx_i^o \\ & + P^*(w_i = 0) \left(F(w_i = 0|\theta; v_i) - P^*(w_i = 0) \right)^2 \end{aligned}$$

[0076] これを式変形すれば、

$$\begin{aligned}\mathcal{D}_{L2}(\theta) = & \sum_{w_i \neq 0} P^*(w_i) \int \{f^{x^o}(x_i^o|\theta)\}^2 dx_i^o \\ & - 2 \sum_{w_i \neq 0} \int P^*(x_i^o, w_i) f^{x^o}(x_i^o|\theta) dx_i^o \\ & + P^*(w_i = 0) \{F(w_i = 0|\theta; v_i)\}^2 \\ & - 2 \{P^*(w_i = 0)\}^2 F(w_i = 0|\theta; v_i) + Const\end{aligned}$$

となる。更に、定数項（Const）を除去し、真の分布に関する平均を標本平均で置き換えた下記式の最適化を考える。

[0077] [数16]

$$\begin{aligned}\hat{\mathcal{D}}_{L2}(\theta) = & \frac{1}{n} \left\{ \sum_{\{i|w_i \neq 0\}} \int \{f^{x^o}(x_i^o|\theta)\}^2 dx_i^o \right. \\ & \left. - 2 \sum_{\{i|w_i \neq 0\}} f^{x^o}(x_i^o|\theta) + \sum_{\{i|w_i = 0\}} F(w_i|\theta; v_i)^2 \right\} \\ & - 2 \frac{n_0}{n^2} \sum_{\{i|w_i = 0\}} F(w_i|\theta; v_i).\end{aligned}$$

[0078] ただし、

$$n_w = \sum_{i=1}^n I(w_i = w)$$

とおいた。これを目的関数とすることで、アルゴリズムを導出する。

$$\hat{\mathcal{D}}_{L2}(\theta)$$

は、行列・ベクトル形式で、下記式（16）、（17）のように表すことができる。

[0079] [数17]

$$\hat{\mathcal{D}}_{L2}(\theta) = \theta^T \hat{G} \theta - 2 \hat{g}^T \theta + \theta^T \hat{H} \theta - 2 \hat{h}^T \theta \quad \dots (16)$$

$$= \theta^T (\hat{G} + \hat{H}) \theta - 2(\hat{g}^T + \hat{h}^T) \theta, \quad \dots (17)$$

[0080] ただし、下記式（１８）～（２０）に従う。

[0081] [数18]

$$\hat{G} = \frac{1}{n} \sum_{\{i|w_i \neq 0\}} \int \Psi_i^{x^o} (\Psi_i^{x^o})^T dx_i^o \quad \dots (18)$$

$$\hat{H} = \frac{1}{n} \sum_{\{i|w_i=0\}} \Psi^w(w_i=0; v_i) \Psi^w(w_i=0; v_i)^T, \quad \dots (19)$$

$$\hat{g} = \frac{1}{n} \sum_{\{i|w_i \neq 0\}} \Psi_i^{x^o}, \quad \hat{h} = \frac{n_0}{n^2} \sum_{\{i|w_i=0\}} \Psi_i^w. \quad \dots (20)$$

[0082] また、

$$\hat{G}$$

は、コンポーネントの分布にガウス分布を用いる場合、解析的に計算できる値であり、下記式（２１）のように表すことができる。

[0083] [数19]

$$\begin{aligned} \hat{G}_{kk'} &= \frac{1}{n} \sum_{\{i|w_i \neq 0\}} \left(\int \Psi_i^{x^o} (\Psi_i^{x^o})^T dx_i^o \right)_{kk'} \\ &= \frac{1}{n} \sum_{\{i|w_i \neq 0\}} \mathcal{N}(\mu_k^{x^o} | \mu_k^{x^o}, 2\sigma^2) \int_{x_k^o}^{\infty} \mathcal{N}(x_k^o | \mu_k^{x^o}, \sigma^2) dx_k^o \int_{x_{k'}^o}^{\infty} \mathcal{N}(x_{k'}^o | \mu_{k'}^{x^o}, \sigma^2) dx_{k'}^o \quad \dots (21) \end{aligned}$$

[0084] 上記式（２１）中で、

$$w$$

に関して総和をとる際、

$$d_{x^o}$$

や

$$\mu_k^{x^o}$$

は

$$w$$

の値によって異なる値であることを明記しておく。これにより、目的関数は

θ

に関する2次の形式で表現されることが分かる。よって、パラメタの推定値は下記で示す制約付きの2次最適化問題を解くことで得ることができる。

[0085] [数20]

$$\hat{\theta} = \arg \min_{\theta} \hat{\mathcal{D}}_{L2}(\theta) \text{ s.t. } \theta_k \geq 0, |\theta| = 1.0. \quad \dots (22)$$

[0086] 数値ソルバーを用いて直接上記式(22)の最適化問題を解くことができる。その際、正則化項を加えた下記の最適化問題を解くことにしてもよい。

[0087] [数21]

$$\hat{\theta} = \arg \min_{\theta} [\hat{\mathcal{D}}_{L2}(\theta) + \beta \theta^T \theta] \text{ s.t. } \theta_k \geq 0, |\theta| = 1.0$$

[0088] ただし、 β がハイパーパラメタを表す。また、例えば次に示す近似的な方法を用いてもよい。上記式(22)の最適化問題から制約条件を外した問題の最適解は下記式(23)、(24)のように求められる。

[0089] [数22]

$$\tilde{\theta} = \arg \min_{\theta} [\tilde{\mathcal{D}}_{L2}(\theta) + \beta \theta^T \theta] \quad \dots (23)$$

$$= (G + \hat{H} + \beta I_K)^{-1} (\hat{g} + \hat{h}) \quad \dots (24)$$

[0090] ただし、

 $\theta^T \theta$

は正則化項、 β はハイパーパラメタであり、パラメタの発散を防ぐ効果がある。

 θ

は、値が0以上かつ和が1であるから、下記式(25)の処理で条件を満たす

 $\hat{\theta}$

が得られる。

[0091] [数23]

$$\hat{\theta} = \frac{\max(\tilde{\theta}, 0)}{|\max(\tilde{\theta}, 0)|} \quad \dots (25)$$

[0092] ただし、

$$|\cdot|$$

は、

$$\ell_1$$

ノルムを表す

[0093] 本発明の実施の形態では、以上説明した2つの手法の何れかにより、モデルのパラメタを推定することにより、計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができる。

[0094] <本発明の実施の形態に係る推定装置の構成>

次に、図5及び図6を参照して、本発明の実施の形態に係る推定装置1の構成について説明する。図5は、本発明の実施の形態に係る推定装置1として機能するコンピュータの概略構成を示すブロック図である。図6は、本発明の実施の形態に係る推定装置1の構成を示すブロック図である。

[0095] 図5に示すように、推定装置1は、CPU110と、RAM等のメモリ120と、通信インターフェース(IF)部130と、キーボード等の入力部140と、ディスプレイ等の表示部150と、後述する推定処理ルーチンを実行するためのプログラム170を記憶したROM等の記憶部160とを備えたコンピュータで構成されている。また、CPU110、メモリ120、通信IF部130、入力部140、表示部150、及び記憶部160は、バス100を介して接続されている。また、通信IF部130は、LANケーブル等の通信回線により外部装置2と接続されている。なお、ネットワーク(図示しない)を介して外部装置2と接続される構成としてもよい。

[0096] 図6に示すように、本実施形態に係る推定装置1は、データ処理部10と、パラメタ推定部20と、パラメタ出力部30と、記憶部40と、入力部50と、出力部60とを備えて構成される。

[0097] データ処理部10は、入力部50が受け付けた観測値が観測されたサンプルの観測データ

$$x_i^o$$

と、観測値が観測されなかったサンプルの観測データ

$$x_i^c$$

と、各サンプルについて観測値が観測されたか否かを表す変数

$$w_i$$

とを含む打ち切りデータ

$$\mathcal{D} = \{x_i, w_i\}$$

をデータ記憶部41に格納する。サンプルは、例えば、上述の図1、3の例では各機器のこと、図2、4の例では各利用者のことである。

[0098] パラメタ推定部20は、入力部50が受け付けた打ち切りデータ

$$\mathcal{D}$$

の各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、観測データの確率密度関数を用いて表される、打ち切りデータ

$$\mathcal{D}$$

の確率分布を表すモデル

$$P(x, w|\theta)$$

と、入力部50が受け付けた打ち切りデータ

$$\mathcal{D}$$

から得られる、当該打ち切りデータ

$\mathcal{D}$

の真の確率分布

$P^*(x, w)$

とのダイバージェンスである目的関数を最適化することにより、モデルのパラメタ

θ

を推定する。

[0099] 具体的には、パラメタ推定部20は、まず、当該打ち切りデータ

$\mathcal{D}$

の真の確率分布

$P^*(x, w)$

を求める。

[0100] 次に、パラメタ推定部20は、目的関数を、当該打ち切りデータ

$\mathcal{D}$

の確率分布を表すモデル

$P(x, w|\theta)$

と、当該打ち切りデータ

$\mathcal{D}$

の真の確率分布

$P^*(x, w)$

とのKLダイバージェンス、又は L_2 ダイバージェンスとして、パラメタを推定する。

[0101] ここで、打ち切りデータの確率分布を表すモデル

$$P(x, w|\theta)$$

は、上記式（８）で示すように、観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さとを用いて表される、観測データの確率分布

$$P(x_i|w_i, \theta)$$

と、観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さとを用いて表される、変数が与えられた下での、変数の確率分布

$$P(w_i|\theta)$$

と、を用いて表される。

[0102] パラメタ推定部２０は、KLダイバージェンスを用いる場合、上記式（１５）のパラメタ更新を繰り返すことにより、パラメタを推定する。また、パラメタ推定部２０は、 L_2 ダイバージェンスを用いる場合、上記式（２４）、（２５）を用いてパラメタを推定する。

[0103] そして、パラメタ推定部２０は、推定したパラメタ

$$\theta$$

をパラメタ記憶部４２に格納する。

[0104] パラメタ出力部３０は、パラメタ記憶部４２のパラメタ

$$\theta$$

を取得し、パラメタ出力部３０に当該パラメタ

$$\theta$$

を渡す。

[0105] 記憶部４０は、データ記憶部４１と、パラメタ記憶部４２とを含む。データ記憶部４１には、打ち切りデータ

$$D$$

が格納される。また、パラメタ記憶部 42 には、モデルのパラメタ

$$\theta$$

が格納される。

[0106] 入力部 50 は、外部装置 2 から、打ち切りデータ

$$\mathcal{D}$$

の入力を受け付ける。そして、入力部 50 は、受け付けた打ち切りデータ

$$\mathcal{D}$$

を、データ処理部 10 に渡す。

[0107] 出力部 60 は、パラメタ出力部 30 から受け取ったモデルのパラメタ

$$\theta$$

を、外部装置 2 へ出力する。

[0108] <本発明の実施の形態に係る推定装置の作用>

図 7 は、本発明の実施の形態に係る推定処理ルーチンを示すフローチャートである。

[0109] 入力部 50 に打ち切りデータ

$$\mathcal{D} = \{x_i, w_i\}$$

が入力されると、推定装置 1 において、図 7 に示す推定処理ルーチンが実行される。

[0110] まず、ステップ S100 において、入力部 50 は、観測値が観測されたサンプルの観測データ

$$x_i^o$$

と、観測値が観測されなかったサンプルの観測データ

$$x_i^o$$

と、各サンプルについて観測値が観測されたか否かを表す変数

$$w_i$$

とを含む打ち切りデータ

$$\mathcal{D} = \{x_i, w_i\}$$

の入力を受け付ける。また、入力部 50 は、各サンプル i についての観測終了時刻までの時間の長さの入力を受け付ける。

[0111] ステップ S 110 において、パラメタ推定部 20 は、目的関数を、当該打ち切りデータ

$$\mathcal{D}$$

の確率分布を表すモデル

$$P(x, w|\theta)$$

と、当該打ち切りデータ

$$\mathcal{D}$$

の確率分布

$$P^*(x, w)$$

との KL ダイバージェンス、又は L_2 ダイバージェンスとして、上記式 (15) のパラメタ更新を繰り返すことにより、パラメタを推定するか、あるいは、上記式 (24)、(25) を用いてパラメタを推定する。

[0112] ステップ S 120 において、出力部 60 は、上記ステップ S 110 により推定されたパラメタ

$$\theta$$

を出力する。

[0113] 以上説明したように、本発明の実施形態に係る推定装置によれば、観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、観測データの確率密度関数を用いて

表される、当該打ち切りデータの確率分布を表すモデルと、当該打ち切りデータから得られる、当該打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、モデルのパラメタを推定するため、計算時間を抑えて、打ち切りデータの確率分布を表すモデルのパラメタを精度良く推定することができる。

[0114] なお、本発明は、上述した実施の形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。

[0115] 例えば、上述の実施の形態では、ダイバージェンスとしてKLダイバージェンス又は L_2 ダイバージェンスを用いる場合について説明したが、これに限定されるものではなく、他のダイバージェンスを用いることも可能である。

[0116] また、上述の実施の形態では、時系列データである打ち切りデータを前提に記載しているが、これに限定されるものではなく、時系列データではない任意の打ち切りデータに対しても本発明の適用が可能である。

[0117] また、上述の実施の形態に係る推定装置1は、各部の処理をプログラムとして構築し、推定装置として利用されるコンピュータにインストールして実行させる構成として説明したが、ネットワークを介して流通させる構成としてもよい。

[0118] また、本願明細書中において、プログラムが予めインストールされている実施形態として説明したが、当該プログラムを、コンピュータ読み取り可能な記録媒体に格納して提供することも可能である。

符号の説明

- [0119] 1 推定装置
2 外部装置
10 データ処理部
20 パラメタ推定部
30 パラメタ出力部
40 記憶部
41 データ記憶部

4 2 パラメタ記憶部

5 0 入力部

6 0 出力部

1 0 0 バス

1 1 0 C P U

1 2 0 メモリ

1 3 0 通信 I F 部

1 4 0 入力部

1 5 0 表示部

1 6 0 記憶部

1 7 0 プログラム

請求の範囲

- [請求項1] 観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する推定装置であって、
- 前記打ち切りデータの入力を受け付ける入力部と、
- 前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定するパラメタ推定部と、
- を含む推定装置。
- [請求項2] 前記打ち切りデータの確率分布を表すモデルは、
- 前記観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さを用いて表される、前記変数の確率分布と、
- 前記観測データの確率密度関数と、各サンプルについて予め与えられた、観測終了までの時間の長さを用いて表される、前記変数が与えられた下での、前記観測データの確率分布と、
- を用いて表される請求項1記載の推定装置。
- [請求項3] 前記目的関数は、前記打ち切りデータの確率分布を表すモデルと、前記打ち切りデータの確率分布とのカルバックライブラーダイバージェンス、又は L_2 ダイバージェンスである
- 請求項1又は2記載の推定装置。
- [請求項4] 観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測さ

れたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する推定方法であって、

入力部が、前記打ち切りデータの入力を受け付け、

パラメタ推定部が、前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定する

推定方法。

[請求項5]

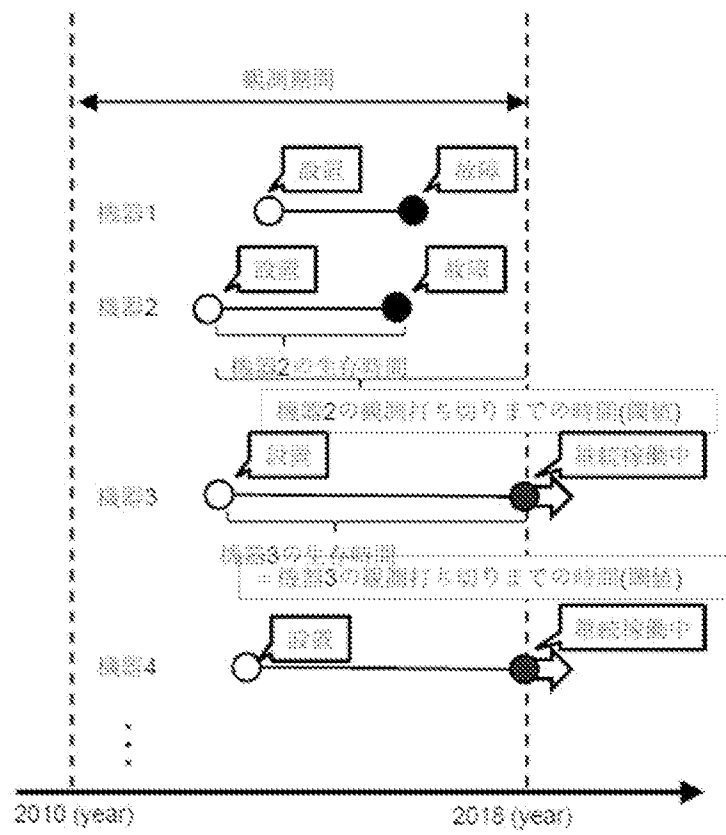
観測値が観測されたサンプルの観測データと、観測値が観測されなかったサンプルの観測データと、各サンプルについて観測値が観測されたか否かを表す変数とを含む打ち切りデータの確率分布を表すモデルのパラメタを推定する処理をコンピュータに実行させるプログラムであって、

入力部が、前記打ち切りデータの入力を受け付け、

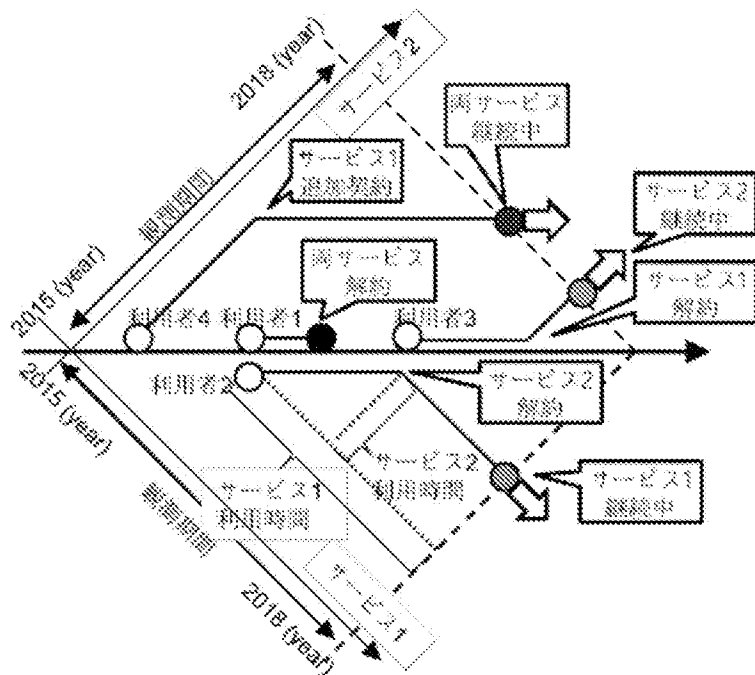
パラメタ推定部が、前記入力部が受け付けた前記打ち切りデータの各サンプルに対応する、観測値の分布を表す各コンポーネントの混合モデルで表される、前記観測データの確率密度関数を用いて表される、前記打ち切りデータの確率分布を表すモデルと、前記入力部が受け付けた前記打ち切りデータから得られる、前記打ち切りデータの確率分布とのダイバージェンスである目的関数を最適化することにより、前記モデルのパラメタを推定する

ことを含む処理をコンピュータに実行させるプログラム。

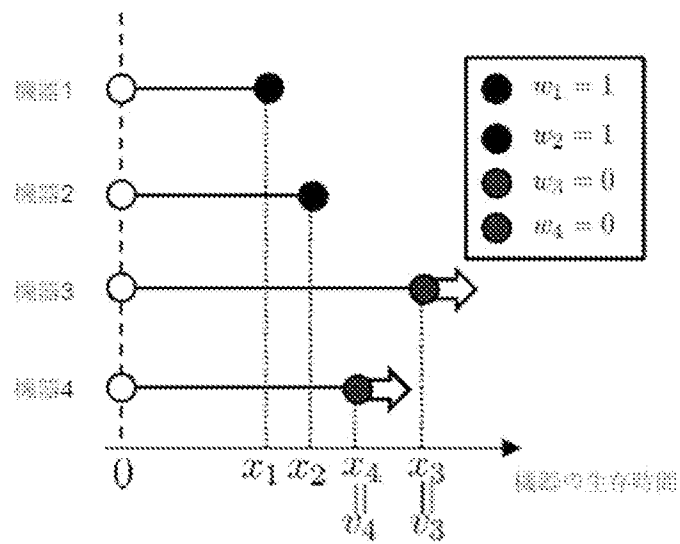
[図1]



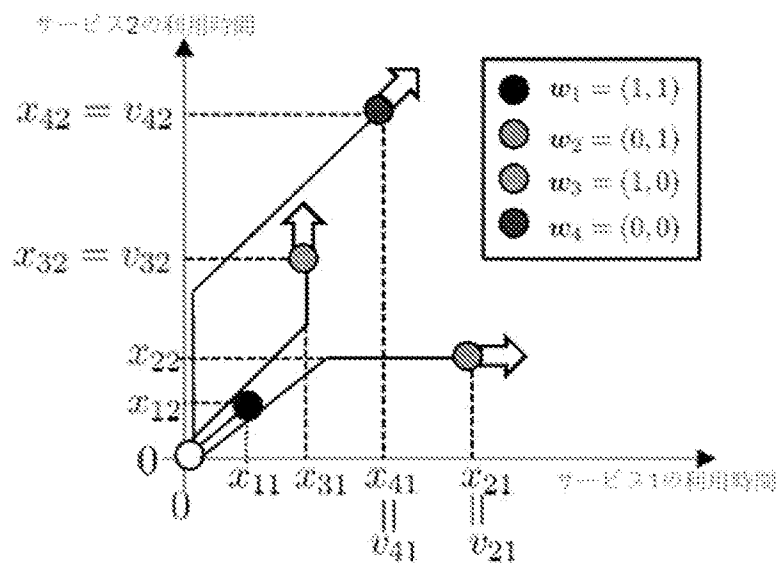
[図2]



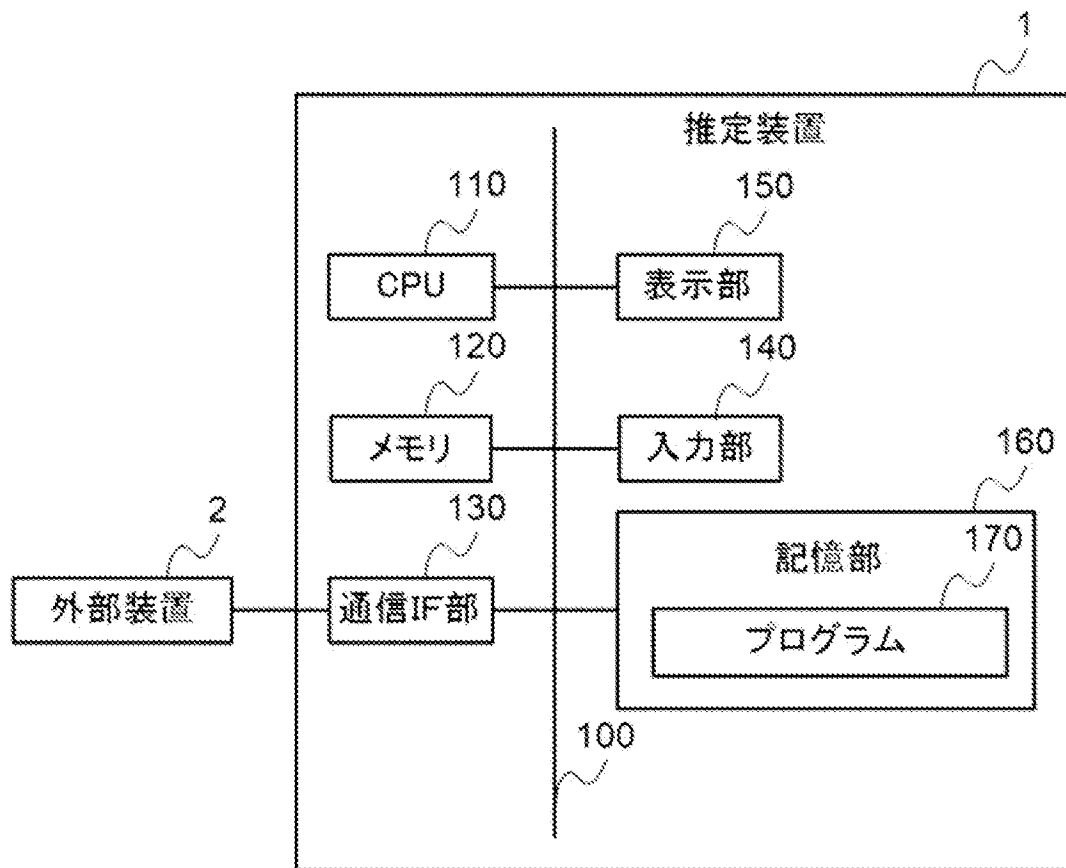
[図3]



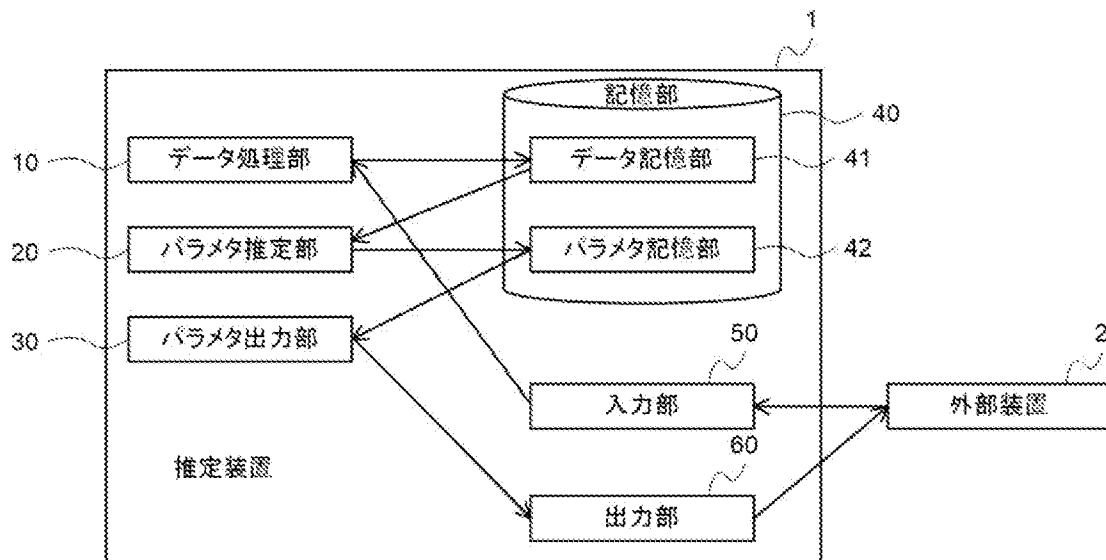
[図4]



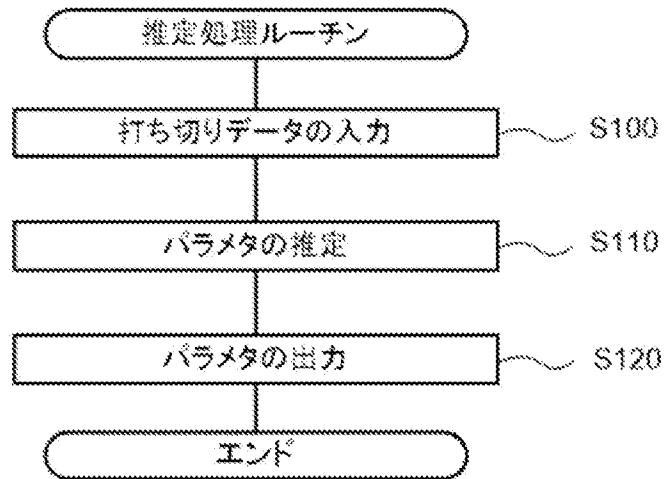
[図5]



[図6]



[図7]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/004909

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G06N20/00 (2019.01) i

FI: G06N20/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int. Cl. G06N20/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BASU, Srabashi et al., Robust and efficient parametric estimation for censored survival data, Annals of the Institute of Statistical Mathematics [online], 2006 [retrieved on 18 March 2020], vol. 58, no. 2, pp. 341-355, Retrieved from the Internet: <URL: https://www.ism.ac.jp/editsec/aism/pdf/058_2_0341.pdf >, ISSN 0020-3157, in particular, pp. 343-347	1-5



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

17.03.2020

Date of mailing of the international search report

31.03.2020

Name and mailing address of the ISA/

Japan Patent Office

3-4-3, Kasumigaseki, Chiyoda-ku,

Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/004909

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	幸島匡宏ほか，打ち切りデータに対する混合モデルのオンライン EM 法の導出と大規模集客イベントにおける到着時間分布推定，第 10 回データ工学と情報マネジメントに関するフォーラム (第 16 回日本データベース学会年次大会)，March 2018，in particular，"Mixed Models for Censored Data"，non-official translation (KOHJIMA, Masahiro et al. Introduction of Online EM Method for Mixed Models for Censored Data, Arrival Time Estimation for Large-Scale Gathering Events. DEIM Forum 2018 (16th Annual Meeting of the Database Society of Japan).)	1-5

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 20/00(2019.01)i FI: G06N20/00														
B. 調査を行った分野														
調査を行った最小限資料（国際特許分類（IPC）） G06N20/00														
最小限資料以外の資料で調査を行った分野に含まれるもの <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922 - 1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971 - 2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996 - 2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994 - 2020年</td> </tr> </table>			日本国実用新案公報	1922 - 1996年	日本国公開実用新案公報	1971 - 2020年	日本国実用新案登録公報	1996 - 2020年	日本国登録実用新案公報	1994 - 2020年				
日本国実用新案公報	1922 - 1996年													
日本国公開実用新案公報	1971 - 2020年													
日本国実用新案登録公報	1996 - 2020年													
日本国登録実用新案公報	1994 - 2020年													
国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）														
C. 関連すると認められる文献														
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号												
Y	BASU, Srabashi ほか, Robust and efficient parametric estimation for censored survival data, Annals of the Institute of Statistical Mathematics[online], 2006 [retrieved on 2020.03.18], Vol. 58, No. 2, pp.341-355, Retrieved from the Internet: <URL: https://www.ism.ac.jp/editsec/aism/pdf/058_2_0341.pdf>, ISSN 0020-3157 特にpp.343-347	1-5												
Y	幸島 匡宏 ほか, 打ち切りデータに対する混合モデルのオンラインEM法の導出と大規模集客イベントにおける到着時間分布推定, 第10回データ工学と情報マネジメントに関するフォーラム（第16回日本データベース学会年次大会）, 2018.03 特に「打ち切りデータに対する混合モデル」の欄	1-5												
☐ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの</td> </tr> <tr> <td>“A” 特に関連のある文献ではなく、一般的技術水準を示すもの</td> <td>“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの</td> </tr> <tr> <td>“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの</td> <td>“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの</td> </tr> <tr> <td>“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）</td> <td>“&” 同一パテントファミリー文献</td> </tr> <tr> <td>“O” 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの	“A” 特に関連のある文献ではなく、一般的技術水準を示すもの	“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの	“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの	“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	“&” 同一パテントファミリー文献	“O” 口頭による開示、使用、展示等に言及する文献		“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献	
* 引用文献のカテゴリー	“T” 国際出願日又は優先日後に公表された文献であって出願と抵触するものではなく、発明の原理又は理論の理解のために引用するもの													
“A” 特に関連のある文献ではなく、一般的技術水準を示すもの	“X” 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの													
“E” 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	“Y” 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの													
“L” 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）	“&” 同一パテントファミリー文献													
“O” 口頭による開示、使用、展示等に言及する文献														
“P” 国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献														
国際調査を完了した日 17.03.2020	国際調査報告の発送日 31.03.2020													
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 桜井 茂行 5B 1157 電話番号 03-3581-1101 内線 3545													