様式　7（Form 7）

学　位　論　文　要　旨
Dissertation Abstract

学位請求論文題名 Dissertation Title

ビン型データ構造を持つノンパラメトリック確率密度関数の推定に関する研究

（和訳または英訳）Japanese or English Translation

Study on Estimation of Nonparametric Probability Density Functions with Binned Data Structures

人間社会環境学　専　攻（Division）

氏　　名（Name）　齊藤　実祥

主任指導教員氏名（Primary Supervisor）　寒河江　雅彦

（注）学位論文要旨の表紙
Note: This is the cover page of the dissertation abstract.

Abstract

I study that nonparametric probability density functions to analyze the structure of data. I especially focus on Histograms, which are binned density estimators and can shrink the amount of data. Through improving Histograms, I aim to avoid the loss of information in the data. I propose the smoothed polynomial density estimation that satisfies both of continuity conditions and local moment conditions. I show its large sample properties and investigate its characteristics when the model analyzes finite samples. The proposed model performs better an accuracy of estimate than some major bin type estimation methods. I also propose two improvements of Histograms to estimate the finite data. One is correcting Histograms' bin widths. The other is using a distribution-free bin width determination method independent of the population distribution. I revealed that both of these can improve the accuracy of estimation and perform better than Histograms in the finite samples.

要旨

データ構造を統計モデルを用いて推定する場合、大規模データと小規模データでそれぞれ異なる課題がある。大規模データ分析では、ハードウェアとソフトウェアの性能上の制約、計算アルゴリズムにおける計算量の増加で、従来の統計モデルのままでは対応が難しくなっている。それに対して、小規模データ分析においては、データ量の少なさに起因する情報不足や推定結果の不安定さから、データの構造を正確に捉えることが難しい。そこで、データの構造解析に対する統計理論の構築のため、大規模データ及び小規模データ解析の問題点に対処するノンパラメトリック統計手法について研究する。

ノンパラメトリック統計手法には2つのメリットが挙げられる。1つ目は、データに柔軟に対応が可能でモデルに柔軟性がある。複雑な構造を取り扱いやすく、例えば、同一のモデルで単峰から二山以上の多峰の分布まで分析が可能である。2つ目は、あらかじめ特定の母集団分布を仮定せずに推定が可能であり、母集団構造が特定できない状況において相性が良い方法と考えられる。本稿では特にデータ構造を規定する確率密度推定に焦点を当て、ビン型の確率密度関数であるHistogramについて議論する。

Histogramは、データを重複しない区間に分け、各区間に入るデータ数に比例した柱状グラフを構成して密度推定を行う方法である。この重複しない区間は「ビン」と呼ばれ、ビン幅及び推定区間がHistogramを同定するためのパラメータとみなすことができる。Histogramではビンごとに独立に推定することでデータを縮約できるが、元データを復元することは不可能であるため、重要な情報まで破棄している場合がある。そこで、ビン化によってデータを縮約化しつつ縮約に伴う情報損失を回避することを考える。

大規模データに対するHistogramの改良には2つの方向が考えられ、一つは、各ビン内の補助的な情報である局所モーメント情報を追加して用いるPolynomial Histogram型で、もう一つはHistogramを平滑化するFrequency Polygon型及びHistospline型である。先行研究から、いずれの方向でもHistogramより推定精度が改良されることが示されている。しかしながら、局所モーメント情報の利用とHistogramの平滑化を同時に行う推定量のモデルについては未整備の状態である。したがって、1次までの局所モーメント情報の保持と各ビンの境界での2次連続性の条件を同時に満たす多項式型の平滑化Polynomial Histogram推定法(S-PH$_{(2,1)}$)を提案し、その漸近的性質と有限標本における特性を明らかにする。MISE基準に基づく漸近的性質の導出から、その推定精度が従来型の主なビン型推定法であるHistogram、1次までの局所モーメント条件のみを満たす1次PH、2次連続性の条件のみを満たすHistospline、全データを用いた正規カーネルにおけるカーネル推定よりも推定精度が良いことを示した。また、S-PH$_{(2,1)}$は漸近正規性が成り立つことを示した。有限標本特性を確かめるため、二山の分布を例に用いて数値実験を行った。データ数が10^3以上の場合で、S-PH$_{(2,1)}$がHistogram、1次PH、Histosplineより優れた推定精度であることが明らかになった。各推定量の最適ビン数

に注目すると、データ数が10^5の時、今回の実験におけるS-PH$_{(2,1)}$のビン数はHistogramの約1/45である。

2次までの局所モーメント情報の保持と各ビンの境界での2次連続性の条件を同時に満たす平滑化Polynomial Histogram推定法(S-PH$_{(2,2)}$)を提案し、そのモデルがS-PH$_{(2,1)}$より更に漸近的に推定精度を改良することを導出し、漸近正規性を持つことを示した。数値実験から、大標本において従来型の主なビン型推定量と比較して推定精度が改良できることを示した。また、実データへの適用から、提案モデルではデータの標本平均の情報等を損失することなく推定できることを示した。

局所モーメント情報を利用し、各ビンの推定から構成されるS-PHは、分析時の計算負荷を緩和する方法であるデータスカッシングと似た発想であり、ビン化によってデータを縮約して計算量を軽減しつつ、元データが持つ情報を保持できる方法だと考えられる。一般的にデータ量が数十テラバイトから数ペタバイトの範囲に及ぶと言われているビッグデータの解析においては、S-PHの少ないビン数で優れた推定精度を達成できる性質が分析に有効に働くと思われる。

小規模データに対するHistogramの改良について、本稿では2つの方法を議論する。一つは、Histogramのビン幅を補正により改良する方法で、もう一つは、母集団の分布によらないdistribution-freeなビン幅決定法を用いる方法である。

前者は、Histogramが小規模データにおいて、ビン幅が広く推定されやすいことから生じる推定区間の問題に対する改良法である。閉区間の定義域が与えられているHistogramの始点を定義域の最小値にし、ビン幅を推定して構築したHistogramの推定区間は一般に定義域と一致しない。理由としては、ビン幅は未知の母集団分布とデータ数によって決められる。したがって、ビン幅を決めた後に構成される推定区間は始点となる左端点の情報のみしか用いられず、結果として定義域の右端点と推定区間の右端点が異なる。Histogramのビン幅が広い場合、この推定区間と定義域とのずれが大きく、ずれの範囲内のデータは推定に使用されないため、無視されるデータが多くなる。そこで、推定区間と定義域のずれを解消するようにビン幅を補正する方法を提案する。この補正法の漸近的性質の導出と数値実験から、補正の有用性を理論面から明らかにした。漸近的性質について、補正後ビン幅で推定したHistogramがMISE(平均積分二乗誤差)の意味で漸近一致性と漸近正規性が成り立つことを示した。数値実験の結果から、右片側tail部分の確率が大きい場合、ビン残差部分にデータが多く存在するため、ビン幅補正による効果が特に大きく、補正後ビン幅を用いたHistogramのMISEの方が、通常のビン幅を用いたHistogramよりMISEと分散が共に小さくなり、推定量は改良されることが示された。したがって、単純な補正法であるが、様々なビン幅推定法で適用可能な手法と言える。

後者は、Histogramの最適ビン幅推定において、未知の母集団分布fをその推定量$\hat{f}$で置き換える必要があるが、データを繰り返し利用することによって$\hat{f}$の推定を回避する改良法である。具体的には、Histogramの分割点を一様乱数により決定して不等間隔のヒストグラムを繰り返し推定し、この平均を推定量とするRandom Partitioned

Histogram(RPH)を提案する。データ数、分割数、繰り返し回数に関する様々なパターンのシミュレーションから、RPHの有限標本における性質と有効性を明らかにする。RPHはビン数と繰り返し回数の適切な選択をする限りHistogramより推定精度が改良でき、分散も安定化することが明らかになった。Histogramでは最適ビン幅よりも狭いビン幅では平滑化不足となり、反対に、最適なビン幅より広いビン幅では平滑化過多で、どちらの場合も推定精度が悪くなる。一方で、RPHでは選択するビン数について広い許容範囲でHistogramより優れた推定精度を持つ。また、コンピュータプログラム上で計算が容易な推定法であり、ビン幅など統計の専門的知識に乏しい場合も含めて実用的に利用できる方法だと考えられる。Histogramを改良することで、小規模データにおける情報不足による推定結果の不安定化を克服できることが本研究で示された。データ構造を見える化する単純で利用しやすいHistogramがモデルの根幹となるため、ビン幅補正法及びRPHの計算・統計処理プログラムへの実装は様々な分野に適用が期待できる。

学位論文審査報告書

２０２２年　２月　３日

１　論文提出者

金沢大学大学院人間社会環境研究科

専　攻　　人間社会環境学専攻

氏　名　　齊藤　実祥

２　学位論文題目（外国語の場合は，和訳を付記すること。）

３　審査結果

判　　定（いずれかに〇印）〇合 格　・　不合格

授与学位（いずれかに〇印）　博士（ 社会環境学・文学・法学・(経済学)・学術 ）

４　学位論文審査委員

委員長　　寒河江　雅彦　㊞

委　員　　佐藤　清和

委　員　　柳　在圭

委　員　　星野　伸明

委　員　　藤生　慎

委　員

（学位論文審査委員全員の審査により判定した。）

５　論文審査の結果の要旨

本稿では、データの構造解析に対する統計理論の構築のため、ノンパラメトリックな確率密度関数の推定に関して研究する。ビン型の確率密度関数である Histogram に焦点を当て、Histogram の改良を通して、データを縮約化しつつ縮約に伴う情報損失を回避することを考える。

大規模データに対する Histogram の改良には 2 つの方向が考えられ、一つは、各ビン内の補助的な情報である局所モーメント情報を追加して用いる Polynomial Histogram 型で、もう一つは Histogram を平滑化する Frequency Polygon 型及び Histospline 型である。局所モーメント情報の利用と Histogram の平滑化を同時に行う推定量のモデルについては未整備の状態である。したがって、1 次までの局所モーメント情報の保持と各ビンの境界での 2 次連続性の条件を同時に満たす多項式型の平滑化 Polynomial Histogram 推定法(S-PH(2, 1))を提案し、その漸近的性質と有限標本における特性を明らかにする。大標本において、S-PH(2, 1)は従来の主なビン型推定量よりも推定精度を改良することを示した。

2 次までの局所モーメント情報の保持と各ビンの境界での 2 次連続性の条件を同時に満たす平滑化 Polynomial Histogram 推定法(S-PH(2, 2))を提案し、そのモデルが漸近的に推定精度を改良し、漸近正規性を持つことを示した。数値実験から、大標本において従来型の主なビン型推定量と比較して推定精度が改良できることを示した。また、実データへの適用から、提案モデルではデータの標本平均の情報等を損失することなく推定できることを示した。

小規模データに対する Histogram の改良について、本稿では 2 つの方法を議論する。一つは、Histogram の推定区間に関して改良する方法で、もう一つは、ビン幅について母集団の分布によらない distribution-free なビン幅決定法を用いる方法である。

閉区間の定義域が与えられている Histogram の始点を定義域の最小値にし、ビン幅を推定して構築した Histogram の推定区間は一般に定義域と一致しない。この推定区間と定義域のずれを解消するようにビン幅を補正する方法を提案する。この補正法の漸近的性質の導出と数値実験から、補正の有用性を理論面から明らかにした。単純な補正法であるが、様々なビン幅推定法で適用可能な手法と言える。

Histogram の分割点を一様乱数により決定して不等間隔のヒストグラムを繰り返し推定し、この平均を推定量とする Random Partitioned Histogram(RPH)を提案する。データ数、分割数、繰り返し回数に関する様々なパターンのシミュレーションから、RPH の有限標本における性質

と有効性を明らかにする。RPH はビン数と繰り返し回数の適切な選択をする限り Histogram より推定精度が改良でき、分散も安定化することが明らかになった。RPH は分布の形状に関わらず、適用範囲が広く有効な推定法であることが示された。

- 学術論文(査読付)：6 編
- 学会報告：9 件
- 受賞歴：

① 平成 26 年 12 月：学業成績優秀者として金沢大学学生特別支援制度学業奨励支援給付

② 平成 31 年 3 月：平成 30 年度 優秀論文(修士)「秀」判定「ヒストグラムの有限区間におけるビン幅補正に関する研究」

③ 令和 3 年 10 月：科学技術振興機構 次世代研究者挑戦的研究プログラム 金沢大学次世代精鋭人材創発プロジェクト[令和 3 年度]採択