氏　　名	野津 昭文	
学位(専攻分野)	博士(統計科学)	
学 位 記 番 号	総研大甲第 1677 号	
学位授与の日付	平成26年3月20日	
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当	
学 位 論 文 題 目	Statistical Analysis via Local Learning with Gamma-Divergence	
論 文 審 査 委 員	主　査	教　授　藤澤 洋徳
		教　授　江口 真透
		准教授　川崎 能典
		教　授　西井 龍映　九州大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

A divergence measure describes discrepancy between two probability distributions. We present a local learning approach with a specific form of the measures called the gamma-divergence. Learning algorithms are divided into two types, global learning algorithms and local learning algorithms. Global learning algorithms employ all the data simultaneously to estimate the whole data structure, while local learning algorithms employ a part of the data to capture the local structure. Estimation with the gamma-divergence has the local learning capability, which is the main topic of this thesis. The gamma-divergence is a generalization of the Kullback-Leibler divergence with the power index gamma. It employs the power transformation of density functions, instead of the logarithmic transformation employed by the Kullback-Leibler divergence. We consider the gamma-divergence between the underlying distribution and a distribution in a parametric family, where the underlying distribution means the one which data follow. When the gamma-divergence is used for standard parameter estimation problems, the global minimum point of the gamma-divergence with respect to the parameter is employed as an estimator. We, however, focus on another aspect of the gamma-divergence. The gamma-divergence has an interesting property, i.e. it has some local minimum points corresponding to the local structure in the data set. If the underlying distribution is represented by a mixture of some distributions, there exist some local minimum points of the gamma-divergence corresponding to the mixture components. Therefore, we can capture the local structure by the local minimum points, and this means the gamma-divergence has the local learning capability. We show that the existence of the local minimum points theoretically in some simple settings. The local learning capability of estimation with the gamma-divergence is applied with respect to cluster analysis and detection of heterogeneous correlation structure. Cluster analysis is aimed to divide data into some groups called clusters. Finding clusters can be regarded as investigation of the local structure of the data set, so we can apply the local learning capability to cluster analysis. We propose a new method for clustering with local minimization of the gamma-divergence based on the normal distribution, which we call “spontaneous clustering”. The greatest advantage of the spontaneous clustering is that it automatically detects the number of clusters that adequately reflect the data structure. In contrast, existing methods, such as K-means, fuzzy c-means, or model-based clustering need to prescribe the number of clusters. Instead of the number of clusters, the value of gamma should be determined for the spontaneous clustering. We propose two methods for this purpose. One is a heuristic choice similar to the bandwidth selection in kernel density estimation. The other is based on Akaike Information Criterion (AIC). We detect all the local minimum points of the gamma-divergence, by which we define the cluster centers. As for the second

application we discuss a parameter estimation problem for a Gaussian copula model. A copula is a multivariate distribution function with uniformly distributed marginals on the unit interval and it determines the correlation structure of a multivariate distribution. We consider the heterogeneous correlation structure, that is, the copula of the underlying distribution might be a mixture of some Gaussian copulas. This situation can occur, for example, when we consider the relation between the movement of stock prices and interest rates in finance. This heterogeneity can be captured by finding the local minimum points of the gamma-divergence based on the Gaussian copula model. We propose a fixed point algorithm to obtain the local minimum points of the gamma-divergence. It is also shown that the gamma-estimation is robust against outliers in terms of the influence function. A feasible form of the gamma-divergence is given that suites the Gaussian copula model. In both applications, we consider the situation where the underlying distribution might deviate from the statistical model we fit. The statistical model is a single parametric model, while the underlying distribution is represented by a mixture of some distributions in the model. This is not the standard situation where the statistical model includes the underlying distribution. In this thesis, we show that even in such a situation the estimation is possible by using the gamma-divergence. One of the advantages of this method is that it works well for mixtures of any number of distributions if they are “distinct” enough.

(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

博士論文原稿ではガンマ･ダイバージェンスの援用による局所学習について考察されている. 論文は全 5 章 100 ページで構成されている. 第 1 章ではダイバージェンスの紹介と論文の全体の構成について説明されている. 第 2 章では最小ダイバージェンス法について総説を行っている. 特にガンマ･ダイバージェンスの最小化推定について焦点を当て, 以下の展開のために推定方程式, 推定値アルゴリズムなどの準備をしている. 第 3 章と第 4 章で主要結果を述べている. 第 5 章でまとめとディスカッションを行っている.

第 3 章では主要結果の一つであるガンマ･クラスタリングを提案している. これは正規モデルの下でのガンマ推定に基づく方法である. データが正規モデルから大きく乖離している状況におけるガンマ推定の振る舞いを考察している. この考察の中で最尤推定量と際立って異なる振る舞いが指摘されている. データが K 個のクラスターを成し, クラスターが十分に互いに離れているとき, 正規分布の平均に対するガンマ･ロス関数が K 個の極小解を持つことに着目し, クラスターの中心としてこの極小解を提案している. これは既存のガンマ推定の外れ値に対するロバスト性に関する研究とは異なる新しい性質の指摘である. そして極小解を具体的に得るためのアルゴリズムを提案している. 同時にそれぞれのクラスターの分散を得る方法も提案している. その結果としてクラスタリングが可能となる. 特別な正規混合分布の場合には極小解の存在と一致性について理論的保証も与えている. 次に K 平均法などの既存のクラスタリング手法の文献のレビューの下に提案手法と従来法の比較をしている. 特に提案手法はチューニングパラメータのガンマが適切に選択されていれば自発的にクラスター数を求める良好な性能を持つことが示された. データに基づきガンマを選択する方法として, 簡便なデータレンジに基づく方法と正規混合モデルの仮定からの AIC 規準による方法が提案されている. 最後にシミュレーションと実データ解析を行い, 提案手法の妥当性を結論している.

第 4 章では, もう一つの主要結果として, 異なる相関構造の探索法を提案している. コピュラモデルの枠組みでデータに異なる相関構造が混合されている場合を考察している. 第 3 章と同様な観点から, ガウシアン･コピュラの下でのガンマ･ロス関数について着目して, その複数の局所最小解が存在する条件について詳しく議論している. 特に, 2 つの相関行列が十分に離れているならば, それぞれのコピュラモデルの混合分布からデータが得られたと考えたとき, ガンマ推定は適切に 2 つの相関行列を極小解として推定することが明らかにされている. その探索性能の限界についても考察されている. そして, データから極小解を具体的に得るためのアルゴリズムを提案している. 最後にシミュレーションによって提案手法の妥当性を確認している. 特に, 混合数が分かっていた場合の最尤法に比較しても, 二乗誤差が小さくなっている. これは, 提案手法が, シミュレーション中にたまに現れる外れ値にもうまく対処しているためと考えられ, その有用性を提示している.

ロバスト推定とクラスタリングは関係ないように見えるが, あるグループから他のグループを見ると外れ値であると考えることもできる. そこでロバストに関連したダイバージェンスを使ってクラスタリングを行うことを考えている. 特に, そこで生じる極小解について考察を行い, データから極小解を具体的に得るためのアルゴリズムを提案し,

(Separate Form 3)

シミュレーションや実データ解析できちんとパフォーマンスを示しているところが,オリジナルな点であり評価できる.なお,第 3 章に対応する論文“Spontaneous clustering via minimum gamma-divergence”は Neural Computation (2014) Vol.26, No.2, 421-448, に掲載されており,第 4 章に対応する論文“Detection of heterogeneous structures on the Gaussian copula model using projective power entropy”は ISRN Probability and Statistics (2013), Article ID 787141, 10 pages, に掲載されている.

総合研究大学院大学複合科学研究科における課程博士及び修士の学位の学位授与に係る論文審査等の手続き等に関する規程第１０条に基づいて,口述による試験を実施した.口述による試験を実施した結果,出願者はその博士論文を中心としてそれに関連がある専門分野及びその基礎となる分野について博士(統計科学)の学位の授与に十分な学識を有するものと判断し,合格と判定した.