氏　　名　木村 正成

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2491 号

学位授与の日付　2024 年 3 月 22 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　Identification of Importance-Weighting and Geodesics on Statistical Manifolds

論 文 審 査 委 員　主　査　福水 健次
統計科学コース 教授
矢野 恵佑
統計科学コース 准教授
田中 未来
統計科学コース 准教授
竹之内 高志
政策研究大学院大学 政策研究科 教授

# 博士論文の要旨

氏　　名：木村　正成

論文題目：Identification of Importance-Weighting and Geodesics on Statistical Manifolds

A set of probability distributions forms a Riemannian manifold, namely, the statistical manifold. By studying the geometric properties of this statistical manifold, we can expect to gain various insights. The framework for analyzing the geometric properties of statistical manifolds, known as information geometry, is notably used in various applied studies.

Within the framework of information geometry when considering statistical procedures, we can identify several statistical concepts with geometric objects, which leads to a deeper understanding of statistical procedures, improvements to existing methods, and geometrically natural generalizations of algorithms. In this study, we demonstrate the equivalence between the weighting of probability distributions and the selection of curves on a manifold. Furthermore, we show that this identification leads to the following two generalizations: (i) a generalization of algorithms designed to handle differences in the distribution of input variables between training and testing data, known as the covariate shift adaptation, and (ii) a generalization of a stable version of a divergence, known as the skew divergence, used to measure differences between probability distributions. In the following, we introduce the background for each problem setting.

In this study, we particularly focus on supervised learning, which is an approach that uses pairs of input data and output labels to learn the rules between inputs and outputs. Supervised learning aims to acquire a function that best approximates the relationship between input vectors and outputs using pairs of input vectors and corresponding outputs as labeled training data. Since he procedures of supervised learning algorithms all depend on the provided training dataset, the similarity in distribution between the training and test data seems to be a necessary assumption in supervised learning. As a stronger assumption, the well-known independent and identically distributed (i.i.d.) assumption is often imposed, requiring the independence of each instance. Supervised learning, which optimizes with respect to the empirical risk of the training data, is often formulated within the framework of empirical risk minimization (ERM). The reason why it can be concluded that ERM works well is that this framework exhibit statistically desirable properties namely consistency and unbiasedness. These properties are based on the i.i.d. assumption between the training and test data.

However, such assumptions are often violated in real-world problem settings. The situation where the distributions of the training and test datasets differ is referred to as distribution shift, also known as dataset shift. The covariate shift assumption is one of the problem settings within the context of distribution shift. Covariate shift assumes that the marginal distributions of input vectors between the training and test data are different. The most fundamental strategy for covariate shift adaptation is importance weighting. The key idea behind importance weighting is to adjust the weights of data points, assigning lower weights to the data generated from the training distribution and higher weights to the data generated from the test distribution, to perform parameter estimation effectively. The most typical weighting is based on the density ratio between the training distribution and the test distribution. Moreover, for reasons such as stability and model constraints, several variants exist. In this study, we focus on the observation that a certain group of these covariate shift adaptation methods can be identified with a collection of geometric objects on the statistical manifold formed by sets of probability distributions. We demonstrate that the selection of covariate shift adaptation methods can be identified with the choice of curves that connects the training and test distributions.

Through this identification, we introduce the pair of parameters that determine the shape of the curve and its position on the curve. This enables us to generalize multiple covariate shift adaptation methods. Furthermore, we experimentally demonstrate that the performance of the trained model changes relatively smoothly with variations in these parameters. On the basis of these observations, we show that through appropriate parameter optimization, our generalized covariate shift adaptation outperforms existing methods in terms of performance.

Moreover, we introduce that the equivalence between weighting and curves on the manifold leads to a generalization of divergence. Here, in statistics, divergence refers to a type of pseudo-distance used to evaluate the differences between probability distributions. The term pseudo-distance is used because divergence does not satisfy symmetry or triangle inequality. In particular, one of the most important divergences in statistics

and machine learning is the KL-divergence. The KL-divergence plays a central role in many contexts. For instance, cross-entropy, which commonly appears as an objective function in machine learning, is closely related to the KL-divergence. Mathematically, the KL-divergence is given as the expectation of the logarithm of the density ratio of two probability distributions with respect to one of the distributions. Owing to the inclusion of the density ratio in its definition, the KL-divergence has the potential to diverge to infinity if there is a discrepancy in the supports of the two probability distributions. Furthermore, there are cases in which the asymmetry of the KL-divergence becomes problematic. To address these limitations, there are various variants and symmetrizations of the KL-divergence.

The skew divergence is one of the variants of the KL-divergence and has numerous applications. The skew divergence addresses the issue of the KL-divergence by using a weighted average of the densities of the two distributions as the denominator, which alleviates the problem of divergence to infinity that was present in the KL-divergence. Here, we introduce how the generalization of the skew divergence can also be derived through the equivalence between averaging operations and the selection of curves on the manifold. This generalization exhibits the properties that the divergence should satisfy.

In Chapter 3, we introduce the generalization of covariate shift adaptation methods from the viewpoint of information geometry. The main contributions of this chapter are: (i) generalization of a class of covariate shift adaptation algorithms by identifying importance weighting and the selection of the curve connecting the training distribution and the test distribution, (ii) geometric interpretation of algorithms through generalization (e.g., an algorithm called Adaptive Importance Weighted ERM connects two distributions with a straight curve), and (iii) improvement of existing methods through optimization of parameters introduced by generalization. In Chapter 4, we generalize the skew-divergence family on the basis of the identified weighting and geodesics. In this chapter, we show that generalized skew divergence satisfies several properties that make it desirable as a divergence.

（別紙１ Separate Sheet 1）

**Results of the doctoral thesis defense**

# 博士論文審査結果

Name in Full
氏 名　　木村　正成

Title
論文題目　　Identification of Importance-Weighting and Geodesics on Statistical Manifolds

2024 年 1 月 29 日午後 1 時から約 2 時間にわたり木村正成氏の博士論文審査委員会を開催した．出願者による 1 時間の公開発表による概要説明と質疑応答，さらに約 1 時間の審査委員のみによる審査を行った結果，審査委員会は本論文が学位の授与に値すると判断した．

[論文の概要]

論文は付録 3 章を含む 8 章 172 ページからなり，英語で書かれている．機械学習手法の実問題への応用においては，学習時とテスト時でデータ分布が異なるという共変量シフトが生じている状況が頻出する．本論文の目的は，共変量シフトに対応した学習アルゴリズムの情報幾何学的な理解により，共変量シフト下での学習問題に新たな視座を与え，それに基づく高性能な新規アルゴリズムを開発することである．

第 1 章では統計的推論と幾何学的な概念を同一視することで幾何学的な視点から統計学の様々な問題にアプローチができることを述べ，特に本論文の主題となる共変量シフトと divergence に関して，統計学および機械学習におけるそれらの重要性と動機を述べた上で，本論文の貢献と研究全体の概要を示している．

第 2 章は本論文を通して利用する概念と記法の導入である．統計モデルの記述と，教師あり学習の代表的なアプローチである経験リスク最小化を説明し，確率分布同士を比較する際に用いる divergence を定義している．さらに，統計モデルを多様体としてみなす情報幾何の基本的な考え方を説明している．

第 3 章では既存の共変量シフト下での代表的な学習アルゴリズムを複数紹介し，それらが情報幾何学の観点から統一的に表現できることを示している．まず，重み付き経験リスク最小化による共変量シフト下での学習アルゴリズムの基本形を示し，その拡張手法である適応的重み付き経験リスク最小化及び相対的重み付き経験リスク最小化手法を紹介している．次に，情報幾何学のアイディアに基づき 2 つのパラメータで指定される曲線上の点で指定される重みをつけた経験リスク最小化手法を提案し，それが既存の手法を含むより一般的な方法になっていることを示している．さらに，提案手法が有するパラメータを選択する方法として，情報量規準に基づくものと，ベイズ最適化に基づくものを提案している．人工的に共変量シフトの仮定に従う学習及びテストデータ集合を作成し，各種の重み付き経験リスク最小化手法による予測実験により提案手法の優位性を示している．

第 4 章では，第 3 章で議論した重み付き経験リスク最小化手法における重みのパラメトライズ方法が，f-平均と呼ばれる測度同士の特殊な平均操作に対応していることに注目し，

f-平均から導かれる divergence として $\alpha$-geodesical skew-divergence を提案し，その性質を考察している．この divergence がパラメータの選択により既存の様々な divergence を表現できることを始めとして，パラメータに関する単調性や連続性，劣加法性などの性質を明らかにしている．

第５章では，本論文の寄与である重み付き経験リスク最小化手法と $\alpha$-geodesical skew-divergence の位置付けを述べた上で，今後の課題として高次元データへの対応や深層学習モデルとの組み合わせといった発展と，より現実的なデータへの応用の展望を述べている．

[論文の評価]

機械学習において転移学習の問題は盛んに研究されており，共変量シフトはその中でも最もシンプルな代表的問題として様々なアプローチから研究がなされている．しかしながら，情報幾何学の観点から共変量シフトの問題に取り組む研究は少ない．既存の複数の手法を包含する統一的フレームワークを構築し，幾何学的観点からそれらの特徴付けを行い，いくつかの理論的性質を明らかにしたこと，並びに新たな divergence の提案とその性質に関する考察を与えたことは，統計的機械学習の発展に寄与する研究成果であるといえる．以上の理由により，本論文は統計科学の博士論文として十分な意義を持つと考える．

なお，第３章の内容は，査読付き国際論文誌 Neural Computation(vol 34(9), 1944-1977, 2022)に，第４章の内容は，査読付き国際論文誌 Entropy(vol.23 (5), 2021)に掲載されている．