論文の内容の要旨

# Dimension Reduction Based on the Geometry of Dually Flat Spaces
# （双対平坦空間の幾何学に基づいた次元削減）

廣瀬 善大

We propose methods for estimating parameters and selecting models for generalized linear regression problems, for contingency tables, and for edge selection in Gaussian graphical models. Our method for generalized linear regression can be interpreted as an extension of Least Angle Regression (LARS). Our purpose is to provide methods which help us to estimate parameters efficiently and to narrow down candidate models simultaneously. Our methods can be described in a simple way, and they are easy to interpret. Tools which we use are popular and/or natural in statistics. We take advantage of natural structures of probability distributions to build our algorithms. Our methods are based on the information geometry of dually flat spaces. We give a detailed explanation on our algorithms. Examples of results are shown for some datasets.

Linear regression is one of the most basic problems in statistics. A response variable is observed with explanatory variables for some samples. Ideally, the response variable is represented as a linear combination of the explanatory variables, but there is an error concerning observation. Our interest is to estimate regression coefficient, parameter of a linear model, based on observed data. One of the most famous estimators is the least squares estimator (LSE), which is also the maximum likelihood estimator (MLE) under the condition that the distribution of errors is a normal distribution. It is known that the MLE is not the best for prediction and that shrinkage is effective. This disadvantage of the MLE is called overfitting. One method for avoiding overfitting is regularization. For example, ridge regression is used for estimating parameters in linear regression. Ridge regression is defined by an optimization problem which minimizes the mean squared error with a penalty term of 2-norm of regression coefficient. In the decades, many methods of regularization were proposed, and recently sparsity of parameters is given much attention to. Sparsity means some zeros of a parameter. A representative

example is Least Absolute Shrinkage and Selection Operator (LASSO) from the view point of the sparsity. LASSO is defined by an optimization problem which minimizes the mean squared error with a penalty term of 1-norm of a parameter. The LASSO solution is sparse because of the term of 1-norm.

LARS is an algorithm for estimating parameters and selecting variables simultaneously in linear regression. The LARS algorithm is described in terms of Euclidean geometry. A version of the LARS algorithm is known to compute the LASSO solution. Another version computes the forward stagewise regression which is a method for estimating parameters. LARS has three versions for LARS itself, LASSO, and forward stagewise regression. A framework of LARS provides a unified treatment of LARS, LASSO, and forward stagewise regression and gives a geometrical view of parameter estimation. This fact is a motivation and an advantage of LARS. Another motivation or advantage of LARS is about computation. For example, we need a computation method if we want the LASSO solution, because LASSO is just defined as an optimization problem. The cost for computing LARS is known to be the same as that of the MLE and LARS gives us an efficient algorithm for LASSO. The LARS algorithm is originally defined as an estimator's move in Euclidean space spanned by explanatory variables. The algorithm is described simply in terms of Euclidean geometry, and it is easy to give a statistical interpretation. Correlations represent inner products or angles between explanatory variables and response variable and bisectors of angles play an important role in estimation. LARS outputs model candidates, the number of which is much less than the total number of all possible models. LARS helps us to avoid the difficulty of combinations.

Our method for generalized linear regression, named *bisector regression*, is an extension of LARS based on the information geometry of dually flat spaces. Bisector regression is an algorithm to generate candidates of parameter estimates and submodels. The bisector regression algorithm is described as an estimator's move in a dually flat space. The reason why we use the word extension is that, similar to LARS, bisector of an angle and its extension to high dimensions play an important role in bisector regression. However, bisector regression is essentially different from LARS. In LARS, an estimator starts at the origin, and it moves along bisectors of angles. The LARS estimator finally reaches the MLE of the full model, generating candidates of regression coefficients and submodels. In bisector regression, our estimator starts at the MLE of the full model, and it moves along curves corresponding to bisectors. Our estimator finally reaches the origin, generating candidates. We obtain a sequence of parameter estimates and submodels by bisector regression and the number of candidates is much smaller than the total number of all possible submodels, which means that bisector regression narrows down candidate submodels considerably. In the information geometry of dually flat spaces, we do not use angles, but it is possible to define curves corresponding to bisectors of angles with Kullback-Leibler divergence. This idea is based on the property that a point on a bisector of an angle has the same distance from two projections on two straight lines forming the angle. The extended Pythagorean theorem helps this idea too which is an information-geometrical version of Pythagorean theorem in Euclidean space. In the bisector regression algorithm, we measure divergences from the current estimate to submodels in which one more elements of parameters become 0. Our estimator moves along a curve corresponding to a bisector, and it comes into the nearest submodel. We iterate this process, yielding an estimate, until our estimator comes to the origin. The new estimate has one more 0s than the previous estimate. The length of a sequence of estimates generated by bisector regression is the number of parameters to be estimated. We do not need to consider all candidates of combinations of explanatory variables and the number of the candidates is much smaller than total number of all possible models. Note that the bisector regression has some iterations, but it is different from an iterative method like Newton method. All estimates and submodels generated by bisector regression are candidates and this is not the case that we want just a converged value. Iterations are not for convergence of the algorithm but for making various candidates. Our extension of LARS is based on the original LARS algorithm, not other versions. We pay attention to simpleness of the LARS algorithm and let our extension take after it.

We apply the main idea of bisector regression to contingency tables. We consider multinomial distributions and introduce a parametrization. Parameters represent main effects and interactions, and parameters corresponding to interactions are estimated by our method. Our method generates a sequence of parameter estimates and models, the length of which is the number of parameters to be estimated. An estimate in the sequence has one more zero than the previous one. For contingency tables, the bisector regression algorithm works in a dually flat space of multinomial distributions.

The main idea of bisector regression is also applied to edge selection in Gaussian graphical models. Our interest is to estimate non-diagonal elements of the concentration matrix, the inverse of the covariance matrix, and to generate a independence graph. Our method generates a sequence of estimates of the concentration matrix, the length of which is the total number of all edges, or the number of the non-diagonal elements of the concentration matrix. The concentration matrices generated by our method yield independence graphs. The number of all possible graphs is much bigger than the number of our candidates. Our method helps us to estimate an independence graph efficiently. For edge selection in Gaussian graphical models, the bisector regression algorithm works in the dually flat space of multivariate normal distributions.

Note that, for example, edge selection in Gaussian graphical models is an example of problems which cannot be solved directly by LARS. Of course, we can apply LARS and LASSO to a problem if the problem under consideration can be recast in a linear regression problem. However, it is not always possible to make a problem a linear regression and the recast is not necessarily natural from the point of view of the problem itself. Edge selection in Gaussian graphical models is equivalent to estimation of the concentration matrix. A family of positive definite matrices forms a dually flat space and it is not Euclidean space. This means that this edge selection problem should not be treated in terms of Euclidean geometry even if it is possible to recast the problem in a linear regression problem. The information geometry provides us useful and natural tools for this edge selection problem. Bisector regression is not just a geometrical interpretation of existing methods, but it is a new method for statistical estimation which is based on a natural structure of a problem.

Three algorithms of bisector regression for three problems are based on the same idea. However, they have differences depending on problems which they are applied to. For generalized linear regression, we remark on the fact that submodels depend on the design matrix, or an observation of each explanatory variable. A submodel is embedded in the space of all distributions corresponding to a regression problem. We have different alignments of submodels for different observations. In problems of contingency tables and edge selection, submodels are decided without observations. For contingency tables, parameters are not dealt with equally while bisector regression treats parameters equally in generalized linear regression. Parameters are separated and bisector regression is applied to each group of parameters. We delete higher-order interactions first and leave lower-order interactions. In edge selection for Gaussian graphical models, parameters to be estimated are non-diagonal elements of the concentration matrix. Our aim is to estimate the concentration matrix. We consider the mean of distributions in regression, but we need to consider the matrix in edge selection. A dually flat space is introduced naturally even in the case of normal distributions while only Euclidean geometry are necessary for normal linear regression. We propose an algorithm for edge selection based on a natural structure of probability distributions. This also means that edge selection in Gaussian graphical models is a problem to which LARS and LASSO cannot be applied directly.