MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

An Extension of Least Angle Regression Based on the Information Geometry of Dually Flat Spaces

Yoshihiro HIROSE and Fumiyasu KOMAKI

METR 2009–09 March 2009

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

An Extension of Least Angle Regression Based on the Information Geometry of Dually Flat Spaces

Yoshihiro HIROSE* and Fumiyasu KOMAKI*†
*Graduate School of Information Science and Technology
University of Tokyo, Tokyo, Japan
†RIKEN Brain Science Institute, Wako-shi, Japan

March 31st, 2009

Abstract

We extend the least angle regression algorithm using the information geometry of dually flat spaces. The least angle regression algorithm is based on a bisector in Euclidean space, and it is used for estimating parameters and selecting explaining variables for linear regression. The extended least angle regression algorithm is used for estimating parameters in generalized linear regression, and it has a function of selecting explaining variables. We use curves corresponding to bisectors in Euclidean space for this purpose.

1 Introduction

We consider parametric regressions, i.e., linear regression and generalized linear regression. We extend the least angle regression (LARS) algorithm [4] using the information geometry of dually flat spaces. LARS is used for estimating parameters and selecting explaining variables for linear regression [4]. The extended LARS algorithm can be used for estimating parameters and selecting explaining variables for generalized linear regression.

In the iterative LARS algorithm, we use the geometry of the Euclidean space spanned by explaining variable vectors. The algorithm selects one explaining variable in each iteration for constructing the estimators. In this procedure, a bisector or its extension to higher dimensional spaces were used. The estimator moves along the bisector or its extension. In the LARS algorithm, the bisectors and distance in Euclidean space play an important role in estimating parameters.

One of the main advantages of the LARS algorithm is its efficiency. In fact, the number of iterations is the same as the number of explaining variables. Furtheremore, LARS is associated with the lasso proposed by Tibshirani [10]. Lasso minimizes the L_2-norm of the residual of the estimated response and observed response subject to a constraint on the L_1-norm of the estimator. Lasso has been studied extensively, and [8] can be referred to. A slightly modified LARS algorithm yields the lasso estimator. This implies that we can obtain the lasso estimator with lesser computational effort.

In this study, we extend the LARS algorithm using the information geometry of dually flat spaces. A dually flat space is a generalization of the Euclidean space [1, 2]. The model manifold of an exponential family of distributions is a dually flat space. The exponential family of distributions appears in generalized linear regression. We estimate the parameters of the exponential family in generalized linear regression using the information geometry of a dually flat space. In a dually flat space, geodesics and divergence correspond to straight lines and distance in the Euclidean space, respectively. In order to obtain the estimator, we consider a curve corresponding to a bisector in a Euclidean space.

In section 2, we propose the extended LARS algorithm. We describe the information geometry of dually flat spaces. Then, we describe the extended LARS algorithm, and show the geometrical aspect of this algorithm. In section 3, we show the results of the extended LARS algorithm for two types of databases. In section 4, we present the conclusions.

2 Extended least angle regression algorithm

2.1 Settings

We consider a generalized linear regression model. For observed data

$$\{y_a, x^a = (x^a_1, x^a_2, \dots, x^a_d)\}_{a=1,2,\dots,n},$$

the design matrix X is defined by

$$X = (x^a_i)_{1 \le a \le n,\, 1 \le i \le d} = (x_1, x_2, \dots, x_d),$$

where $x_i = (x^1_i, x^2_i, \dots, x^n_i)^\top$ $(i = 1, 2, \dots, d)$ and X is a $(n \times d)$ matrix. Let $\mathbf{1}$ be the vector with n 1s, i.e., $(1, 1, \dots, 1)^\top$. The matrix $\tilde{X}$ can be defined as

$$\tilde{X} = (\mathbf{1}|X)\,.$$

The model that we consider is an exponential family

$$p(y|\xi) = \exp\left(\sum_{a=1}^{n} y_a \xi^a + \sum_{b=1}^{r} u_b(y)\xi^{b+n} - \psi(\xi)\right), \tag{1}$$

$$\xi' = \tilde{X}\theta',$$

$$\xi'' = \theta'',$$

where $\xi := (\xi^1, \xi^2, \dots, \xi^{n+r})^\top$, $\xi' := (\xi^1, \xi^2, \dots, \xi^n)^\top$, $\xi'' := (\xi^{n+1}, \xi^{n+2}, \dots, \xi^{r+n})^\top$, $\theta := (\theta^0, \theta^1, \dots, \theta^{d+r})^\top$, $\theta' := (\theta^0, \theta^1, \dots, \theta^d)^\top$, and $\theta'' := (\theta^{d+1}, \theta^{d+2}, \dots, \theta^{r+d})^\top$. Here, $u(y) := \{y_1, \dots, y_n, u_1(y), u_2(y), \dots, u_r(y)\}$ is a sufficient statistic of ξ, and $\psi(\cdot)$ is a convex function corresponding to the normalizing constant. The parameter ξ is the natural parameter and $\psi(\cdot)$ is the potential function of ξ. The expectation $\mu = (\mu_1, \dots, \mu_{n+r})^\top$, where $\mu_a = \mathrm{E}[y_a]\,(a = 1, \dots, n)$ and $\mu_{b+n} = \mathrm{E}[u_b(y)]\,(b = 1, \dots, r)$, is called the expectation parameter.

We define the simplest model as

$$\left\{\theta \middle|\, \theta^1 = \theta^2 = \cdots = \theta^d = 0\right\}. \tag{2}$$

In the case of the simplest model, $\xi^1 = \xi^2 = \cdots = \xi^n$.

Example 1 (Normal regression)**.** In normal regression, an exponential family of normal distributions is given as

$$p(y|m, \sigma^2) = \prod_{a=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_a - m_a)^2}{2\sigma^2}\right\},$$

where $m = (m_1, m_2, \dots, m_n)^\top$ is the mean vector and σ^2 is the unknown variance. The natural parameter is

$$\begin{aligned} \xi^a &= \frac{m_a}{\sigma^2} \quad (a = 1, 2, \dots, n), \\ \xi^{n+1} &= -\frac{1}{2\sigma^2}, \end{aligned}$$

and $r = 1$ in model (1). The distribution is given by

$$p(y|\xi) = \exp\left(\sum_{a=1}^{n} y_a \xi^a + \left(\sum_{a=1}^{n} y_a^2\right) \xi^{n+1} - \psi(\xi)\right),$$

where

$$\psi(\xi) = -\frac{\sum_{a=1}^{n} (\xi^a)^2}{4\xi^{n+1}} - \frac{n}{2}\log(-\xi^{n+1}) + \frac{n}{2}\log\pi$$

is the potential function of ξ. The expectation parameter μ is given by

$$\begin{aligned} \mu_a &= m_a = -\frac{\xi^a}{2\xi^{n+1}} \quad (a = 1, 2, \dots, n), \\ \mu_{n+1} &= \sum_{a=1}^{n} (m_a^2 + \sigma^2) = \sum_{a=1}^{n} \left(\frac{\xi^a}{2\xi^{n+1}}\right)^2 - \frac{n}{2\xi^{n+1}}. \end{aligned}$$

The model we consider is

$$\begin{aligned} \xi' &= \tilde{X}\theta', \\ \xi^{n+1} &= \theta^{d+1}. \end{aligned}$$

In the simplest model that we assume, $\xi^1 = \xi^2 = \cdots = \xi^n$. This implies that $\mu_1 = \mu_2 = \cdots = \mu_n$. □

Example 2 (Logistic regression)**.** In logistic regression, we consider the following exponential family.

$$\begin{aligned} p(y|\xi) &= \prod_{a=1}^{n} \frac{\exp(y_a \xi^a)}{1 + \exp \xi^a} \\ &= \exp\left(\sum_{a=1}^{n} y_a \xi^a - \psi(\xi) \right), \\ \xi &= \tilde{X}\theta, \end{aligned}$$

where $\xi = (\xi^1, \xi^2, \dots, \xi^n)^\top$ is the natural parameter, $\theta = (\theta^0, \theta^1, \dots, \theta^d)^\top$, $y \in \{0,1\}^n$, $r = 0$, and

$$\psi(\xi) = \sum_{a=1}^{n} \log(1 + \exp \xi^a)$$

is the potential function of ξ. The expectation parameter μ is given by

$$\mu_a = \mathrm{E}[y_a] = \frac{\exp \xi^a}{1 + \exp \xi^a} \quad (a = 1, 2, \dots, n).$$

In the simplest model that we assume, $\xi^1 = \xi^2 = \cdots = \xi^n$. This implies that $\mu_1 = \mu_2 = \cdots = \mu_n$, i.e., $\mathrm{Prob}(y_1 = 1) = \mathrm{Prob}(y_2 = 1) = \cdots = \mathrm{Prob}(y_n = 1)$. □

2.2 Information geometry for the algorithm

Before we describe the extended LARS algorithm, we summarize some notions of the information geometry of dually flat spaces that are used in this study. For details, refer to [1], [2], and [6]. Based on the information geometry, the model manifold of the exponential family is known to be a dually flat space. In a dually flat space, there exists a pair of a coordinate system ξ, called the e-affine coordinate, and a convex function ψ, i.e., the potential function of ξ. Similarly, it also contains a pair of a coordinate system μ, called the m-affine coordinate, and a convex function ϕ, i.e., the potential function of μ. In the model manifold of the exponential family, the natural parameter ξ is an e-affine coordinate system, and the expectation parameter μ is an m-affine coordinate system. The ξ and μ coordinate systems are called mutually dual because of the following relations:

$$\mu_a = \frac{\partial}{\partial \xi^a}\psi(\xi), \tag{3}$$

$$\xi^a = \frac{\partial}{\partial \mu_a}\phi(\mu), \tag{4}$$

$$\phi(\mu) + \psi(\xi) - \mu \cdot \xi = 0. \tag{5}$$

Then, the relations

$$g_{ab} = \frac{\partial^2}{\partial \xi^a \partial \xi^b}\psi(\xi), \tag{6}$$

$$g^{ab} = \frac{\partial^2}{\partial \mu_a \partial \mu_b}\phi(\mu) \tag{7}$$

also hold, where (g_{ab}) denotes the Fisher information matrix and (g^{ab}), the inverse of (g_{ab}). In the dually flat space of the exponential family, the potential function $\phi(\mu)$ is given as

$$\phi(\mu) = \xi \cdot \mu - \psi(\xi) = -\int p(y|\mu)\log p(y|\mu)\mathrm{d}y = -H(p(y|\mu)),$$

where $p(y|\mu)$ is the density function of y given the parameter μ and $H(\cdot)$ is the entropy in information theory.

Let $\xi(P)$ and $\mu(P)$ denote the e-affine coordinate and m-affine coordinate of point P, respectively. For a dually flat space, two different geodesics, an e-geodesic and an m-geodesic, are defined. The e-geodesic $\xi(t)$ connecting two points, P and Q, is represented by

$$\xi(t) = (1-t)\xi(P) + t\xi(Q), \quad t \in [0,1]$$

in the e-affine coordinate system ξ. The m-geodesic $\mu(t)$ connecting two points, P and Q, is represented by

$$\mu(t) = (1-t)\mu(P) + t\mu(Q), \quad t \in [0,1]$$

in the m-affine coordinate system μ. The geodesics correspond to a straight line in Euclidean space.

We define the orthogonality of an e-geodesic and an m-geodesic (Figure 1). Let P, Q, and R denote different points in a dually flat space. We represent the m-geodesic connecting P and Q as l_{m}, and the e-geodesic connecting R and Q as l_{e}. The two geodesics l_{m} and l_{e} intersect at point Q. The e-geodesic l_{e} and m-geodesic l_{m} are orthogonal if the equation

$$\begin{aligned}(\mu(P) - \mu(Q)) \cdot (\xi(R) - \xi(Q)) &:= \sum_a (\mu_a(P) - \mu_a(Q))(\xi^a(R) - \xi^a(Q)) \\ &= 0\end{aligned}$$

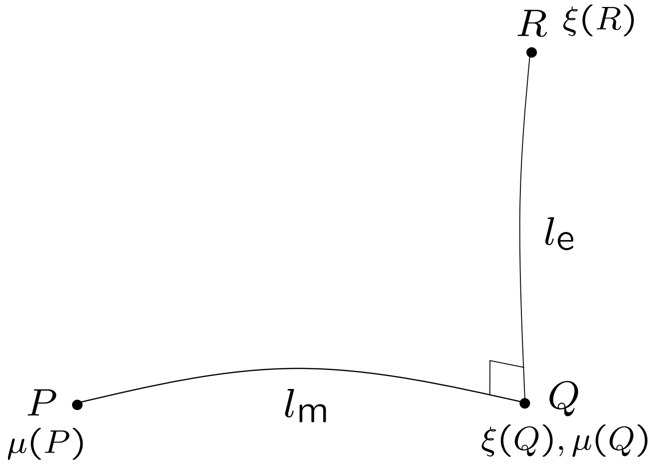

Figure 1: The orthogonality, divergence, extended Pythagorean theorem, and m-projection. $\mu(w)$: m-affine coordinate of the point w ($w \in \{P, Q\}$). $\xi(w)$: e-affine coordinate of the point w ($w \in \{Q, R\}$). l_{m}: m-geodesic connecting P and Q. l_{e}: e-geodesic connecting R and Q. l_{e} and l_{m} are orthogonal, and therefore, the equation $(\mu(P) - \mu(Q)) \cdot (\xi(R) - \xi(Q)) = 0$ is satisfied. Because of the extended Pythagorean theorem, the equality $D(P, R) = D(P, Q) + D(Q, R)$ holds. The e-geodesic l_{e} is e-flat in the dually flat space. The m-projection of R on l_{e} is Q.

is satisfied.

The divergence $D(P, Q)$ from P to Q is defined by

$$D(P, Q) = \phi(\mu(P)) + \psi(\xi(Q)) - \mu(P) \cdot \xi(Q).$$

The divergence is given by the square of the distance in Euclidean space. If l_{m} and l_{e} are orthogonal, the equation

$$D(P, R) = D(P, Q) + D(Q, R) \tag{8}$$

holds. Relation (8) is called the extended Pythagorean theorem.

Finally, we define the m-projection in a dually flat space F. Let S be an e-flat subspace in the dually flat space F. We call S an e-flat subspace in the dually flat space F if, for arbitrary points $P, Q \in S$, the e-geodesic connecting P and Q in F lies in the subspace S. The m-projection $\bar{P} \in S$ of point P on the subspace S in F satisfies

$$(\mu(P) - \mu(\bar{P})) \cdot (\xi - \xi(\bar{P})) = 0 \quad (\forall \xi \in S),$$

where $\mu(P)$ is the m-affine coordinate of P, ξ is the e-affine coordinate of the point in S, and $\xi(\bar{P})$ and $\mu(\bar{P})$ are the e-affine and m-affine coordinates of

$\bar{P}$, respectively. That is, the m-geodesic connecting P and $\bar{P}$ is orthogonal with an arbitraty e-geodesic in S which intersects the m-geodesic.

Example 1 (Normal regression, continued)**.** In Example 1, the expectation parameter μ and the potential function ψ of ξ are

$$\begin{aligned}
\mu_a &= -\frac{\xi^a}{2\xi^{n+1}} \quad (a = 1, 2, \ldots, n), \\
\mu_{n+1} &= \sum_{a=1}^{n} \left(\frac{\xi^a}{2\xi^{n+1}} \right)^2 - \frac{n}{2\xi^{n+1}}, \\
\psi(\xi) &= -\frac{\sum_{a=1}^{n} (\xi^a)^2}{4\xi^{n+1}} - \frac{n}{2} \log(-\xi^{n+1}) + \frac{n}{2} \log \pi,
\end{aligned}$$

respectively. The natural parameter is represented by the expectation parameter as

$$\begin{aligned}
\xi^a &= \frac{m_a}{\sigma^2} = \frac{n\mu_a}{\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2} \quad (a = 1, 2, \ldots, n), \\
\xi^{n+1} &= -\frac{1}{2\sigma^2} = -\frac{n}{2(\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2)}.
\end{aligned}$$

The potential function ϕ of μ is given by

$$\begin{aligned}
\phi(\mu) &= \xi \cdot \mu - \psi(\xi) \\
&= \sum_{a=1}^{n} \frac{n\mu_a^2}{\mu_{n+1} - \sum_{b=1}^{n} \mu_b^2} - \frac{n\mu_{n+1}}{2(\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2)} \\
&\quad - \left(\sum_{a=1}^{n} \frac{n\mu_a^2}{2(\mu_{n+1} - \sum_{b=1}^{n} \mu_b^2)} + \frac{n}{2} \log \left(\frac{2(\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2)}{n} \right) \right) \\
&= -\frac{n}{2} \log \left(\frac{2(\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2)}{n} \right) - \frac{n}{2}(1 + \log \pi).
\end{aligned}$$

The natural parameter ξ and the expectation parameter μ are mutually dual coordinate systems. Relations

$$\begin{aligned}
\frac{\partial\psi(\xi)}{\partial\xi^a} &= -\frac{\xi^a}{2\xi^{n+1}} \\
&= \mu_a \quad (a = 1, 2, \dots, n), \\
\frac{\partial\psi(\xi)}{\partial\xi^{n+1}} &= \sum_{a=1}^{n} \left(\frac{\xi^a}{2\xi^{n+1}}\right)^2 - \frac{n}{2\xi^{n+1}} \\
&= \mu_{n+1}, \\
\frac{\partial\phi(\mu)}{\partial\mu_a} &= \frac{n\mu_a}{\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2} \\
&= \xi^a \quad (a = 1, 2, \dots, n), \\
\frac{\partial\phi(\mu)}{\partial\mu_{n+1}} &= -\frac{n}{2(\mu_{n+1} - \sum_{a=1}^{n} \mu_a^2)} \\
&= \xi^{n+1}
\end{aligned}$$

hold. The divergence from a point P to another point Q is given by

$$\begin{aligned}
D(P, Q) &= \phi(\mu(P)) + \psi(\xi(Q)) - \mu(P) \cdot \xi(Q) \\
&= -\frac{n}{2} \log\left(\frac{2(\mu_{n+1}(P) - \sum_{a=1}^{n} \mu_a^2(P))}{n}\right) - \frac{n}{2} \\
&\quad - \frac{\sum_{a=1}^{n} (\xi^a(Q))^2}{4\xi^{n+1}(Q)} + \frac{n}{2} \log\pi - \mu(P) \cdot \xi(Q)
\end{aligned}$$

□

Example 2 (Logistic regression, continued)**.** In Example 2, the expectation parameter μ and the potential function ψ of ξ are

$$\begin{aligned}
\mu_a &= \frac{\exp\xi^a}{1 + \exp\xi^a} \quad (a = 1, 2, \dots, n), \\
\psi(\xi) &= \sum_{a=1}^{n} \log\left(1 + \exp\xi^a\right),
\end{aligned}$$

respectively. The potential function ϕ of μ is given by

$$\begin{aligned}
\phi(\mu) &= \xi \cdot \mu - \psi(\xi) \\
&= \xi \cdot \mu - \sum_{a=1}^{n} \log\left(1 + \exp\xi^a\right) \\
&= \sum_{a=1}^{n} \left\{\mu_a \log\mu_a + (1 - \mu_a) \log\left(1 - \mu_a\right)\right\}.
\end{aligned}$$

The natural parameter ξ and the expectation parameter μ are mutually dual coordinate systems. Two relations

$$\begin{aligned}\frac{\partial\psi(\xi)}{\partial\xi^a} &= \frac{\exp\xi^a}{1+\exp\xi^a}\\ &= \mu_a \quad (a=1,2,\dots,n),\\ \frac{\partial\phi(\mu)}{\partial\mu_a} &= \log\frac{\mu_a}{1-\mu_a}\\ &= \xi^a \quad (a=1,2,\dots,n)\end{aligned}$$

hold. The divergence from a point P to another point Q is given by

$$\begin{aligned}D(P,Q) &= \phi(\mu(P)) + \psi(\xi(Q)) - \mu(P)\cdot\xi(Q)\\ &= \sum_{a=1}^{n}\{\mu_a(P)\log\mu_a(P) + (1-\mu_a(P))\log(1-\mu_a(P))\}\\ &\quad + \sum_{a=1}^{n}\log(1+\exp\xi^a(Q)) - \mu(P)\cdot\xi(Q).\end{aligned}$$

□

In the generalized linear regression analysis, we need to choose one distribution from the exponential family of distributions. We propose an algorithm that is applicable in the dually flat space of the exponential family.

We introduce a new dually flat space S (Figure 2). The model manifold F of the exponential family forms a dually flat space. The natural parameter ξ is the e-affine coordinate system, and the expectation parameter μ is the m-affine coordinate system in F. Let ψ^* denote the potential function of ξ and ϕ^* denote the potential function of μ. We consider the subspace S of model (1) in F. The subspace S is an e-flat subspace in F. Since the space F is a dually flat space and the transform between ξ and θ, i.e., $\xi' = \tilde{X}\theta'$ and $\xi'' = \theta''$, is an affine transform, S is also a dually flat space. We define the function $\psi(\theta)$ as $\psi(\theta) = \psi^*(\xi', \xi'') = \psi^*(\tilde{X}\theta', \theta'')$. Since the potential function $\psi^*(\xi)$ is convex, $\psi(\theta)$ is convex. We introduce a new coordinate system $\eta = (\eta_0, \eta_1, \dots, \eta_{d+r})^\top$ defined by

$$\eta_i = \frac{\partial}{\partial\theta^i}\psi(\theta) \quad (i=0,1,\dots,d+r),$$

and define the function $\phi(\eta)$ as

$$\phi(\eta) = \eta\cdot\theta - \psi(\theta).$$

For $i = 1,2,\dots,d$ and $j = d+1,\dots,d+r$, we have

$$\begin{aligned}\eta_i &= \frac{\partial}{\partial\theta^i}\psi(\theta) = \sum_{a=1}^{n}\frac{\partial}{\partial\xi^a}\psi^*(\xi)\cdot x_i^a = \sum_{a=1}^{n}\mu_a x_i^a,\\ \eta_j &= \frac{\partial}{\partial\theta^j}\psi(\theta) = \frac{\partial}{\partial\xi^j}\psi^*(\xi) = \mu_j,\end{aligned}$$

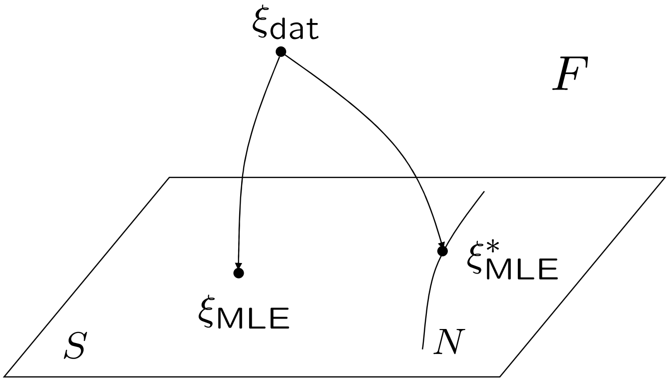

Figure 2: Model subspaces. F: dually flat space of the exponential family. S: e-flat subspace of model (1). N: e-flat subspace of the simplest model (2). ξ_{dat}: point corresponding to data. ξ_{MLE}: MLE for model (1). ξ^*_{MLE}: MLE for the simplest model (2). ξ_{MLE} is the m-projection of ξ_{dat} on S. ξ^*_{MLE} is the m-projection of ξ_{dat} on N.

implying that

$$\begin{aligned}\eta' &= \tilde{X}^\top \mu',\\ \eta'' &= \mu'',\end{aligned}$$

where $\mu' = (\mu_1, \mu_2, \ldots, \mu_n)^\top$, $\mu'' = (\mu_{n+1}, \mu_{n+2}, \ldots, \mu_{r+n})^\top$, $\eta' = (\eta_0, \eta_1, \ldots, \eta_d)^\top$, and $\eta'' = (\eta_{d+1}, \eta_{d+2}, \ldots, \eta_{r+d})^\top$. For $i = 0, 1, \ldots, d+r$, the relation

$$\begin{aligned}\frac{\partial}{\partial \eta_i}\phi(\eta) &= \frac{\partial}{\partial \eta_i}\left(\eta\cdot\theta - \psi(\theta)\right)\\ &= \theta^i + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \sum_{j=1}^{d}\frac{\partial\theta^j}{\partial\eta_i}\frac{\partial\psi(\theta)}{\partial\theta^j}\\ &= \theta^i + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \sum_{j=1}^{d}\frac{\partial\theta^j}{\partial\eta_i}\eta_j\\ &= \theta^i\end{aligned}$$

holds. Therefore, θ is an e-affine coordinate system in S and η is an m-affine coordinate system in S. Coordinate systems θ and η are mutually dual. Convex functions ψ and ϕ are the potential functions of θ and η, respectively.

We describe the relationship between two MLEs for the two different

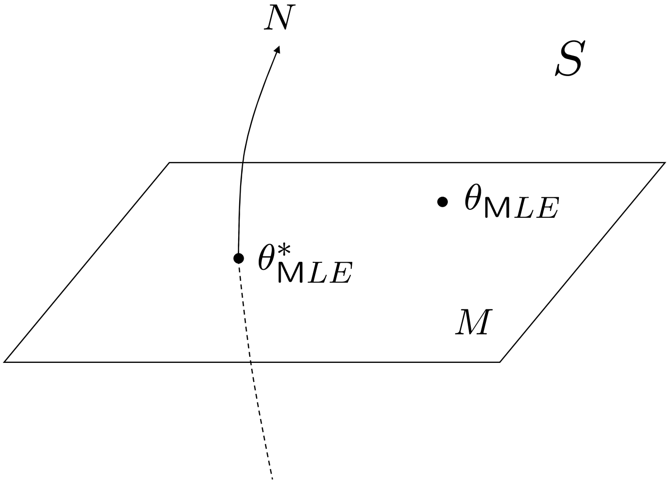

Figure 3: The subspace M that the algorithm works with. M: m-flat subspace in S that is orthogonal to N at θ^*_{MLE}. $N := \{\theta | \theta^1 = \theta^2 = \cdots = \theta^d = 0\}$: e-flat subspace. θ^*_{MLE}: MLE for the simplest model (2). θ_{MLE}: MLE for the model (1). M contains both θ_{MLE} and θ^*_{MLE}.

models (1) and (2). Let

$$\theta_{\mathrm{MLE}} = (\theta^0_{\mathrm{MLE}}, \theta^1_{\mathrm{MLE}}, \ldots, \theta^{d+r}_{\mathrm{MLE}})^\top$$

be the MLE for model (1), and let

$$\theta^*_{\mathrm{MLE}} = (\theta^{*0}_{\mathrm{MLE}}, 0, \ldots, 0, \theta^{*d+1}_{\mathrm{MLE}}, \ldots, \theta^{*d+r}_{\mathrm{MLE}})^\top$$

be the MLE for the simplest model (2). The point θ^*_{MLE} lies in the e-flat subspace $N := \{\theta | \theta^1 = \theta^2 = \cdots = \theta^d = 0\}$. We define the m-flat subspace M as a subspace M containing θ^*_{MLE} and it is orthogonal to N at θ^*_{MLE} (Figure 3). Thus, the point θ_{MLE} lies in M. The m-flat subspace M is represented by

$$M = \{\eta | \, \eta_0 = (\eta^*_{\mathrm{MLE}})_0, \eta_{d+1} = (\eta^*_{\mathrm{MLE}})_{d+1}, \ldots, \eta_{d+r} = (\eta^*_{\mathrm{MLE}})_{d+r}\}.$$

Since M is an m-flat subspace, M is also a dually flat space and $(\eta_1, \ldots, \eta_d)$ is an m-affine coordinate in M. The e-affine coordinate that is dual with $(\eta_1, \ldots, \eta_d)$ is $(\theta^1, \ldots, \theta^d)$. A point lying in M is specified by the mixture coordinate $((\eta^*_{\mathrm{MLE}})_0, \theta^1, \ldots, \theta^d, (\eta^*_{\mathrm{MLE}})_{d+1}, \ldots, (\eta^*_{\mathrm{MLE}})_{d+r})$ in S. The algorithm we propose works with the m-flat subspace M. We omit the condition $\eta_0 = (\eta^*_{\mathrm{MLE}})_0, \eta_{d+1} = (\eta^*_{\mathrm{MLE}})_{d+1}, \ldots, \eta_{d+r} = (\eta^*_{\mathrm{MLE}})_{d+r}$. We use the coordinates $(\theta^1, \ldots, \theta^d)$ in M.

Example 1 (Normal regression, continued)**.** We consider the e-affine coordinate θ and the m-affine coordinate η in the subspace S. The m-affine coordinate η, which is dual with θ, is given by $\eta = (\eta', \eta'')$, where $\eta' = \tilde{X}^\top \mu'$ and $\eta'' = \mu''$. The m-affine coordinate system η is given by

$$\begin{aligned}
\eta_0 &= -\sum_{a=1}^{n} \frac{\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j}{2\theta^{d+1}} \\
\eta_i &= -\sum_{a=1}^{n} \left(\frac{\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j}{2\theta^{d+1}} \cdot x_i^a \right) \qquad (i = 1, 2, \ldots, d), \\
\eta_{d+1} &= \sum_{a=1}^{n} \left(\frac{\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j}{2\theta^{d+1}} \right)^2 - \frac{n}{2\theta^{d+1}}.
\end{aligned}$$

The potential function of θ is

$$\begin{aligned}
\psi(\theta) &= \psi^*(\tilde{X}\theta', \theta'') \\
&= -\frac{\sum_{a=1}^{n} \left(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j\right)^2}{4\theta^{d+1}} - \frac{n}{2}\log(-\theta^{d+1}) + \frac{n}{2}\log \pi.
\end{aligned}$$

□

Example 2 (Logistic regression, continued)**.** We consider the e-affine coordinate θ and the m-affine coordinate η in the subspace S. The m-affine coordinate η that is dual with θ is given by $\eta = \tilde{X}^\top \mu$. The m-affine coordinate system η is given by

$$\begin{aligned}
\eta_0 &= \sum_{a=1}^{n} \frac{\exp(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j)}{1 + \exp(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j)}, \\
\eta_i &= \sum_{a=1}^{n} \left(\frac{\exp(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j)}{1 + \exp(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j)} \cdot x_i^a \right) \qquad (i = 1, 2, \ldots, d).
\end{aligned}$$

The potential function of θ is

$$\begin{aligned}
\psi(\theta) &= \psi^*(\tilde{X}\theta) \\
&= \sum_{a=1}^{n} \log \left(1 + \exp\left(\theta^0 + \sum_{j=1}^{d} x_j^a \theta^j \right) \right).
\end{aligned}$$

□

We consider subspaces $M(I) := \{\theta | \, \theta^j = 0 \, (j \notin I)\}$ in M for $I \subset \{1, 2, \ldots, d\}$ (Figure 4). We define $\theta^{[I]} := (\theta^i)_{i \in I}$ and $\eta_{[I]} := (\eta_i)_{i \in I}$, where θ^i is an e-affine coordinate system and η_i is an m-affine coordinate system of S. Let $\psi^{[I]}(\theta^{[I]}) = \psi(\theta) \, (\theta^j = 0, \, j \notin I)$ and $\phi^{[I]}(\eta^{[I]}) = \eta^{[I]} \cdot \theta^{[I]} - \psi^{[I]}(\theta^{[I]})$,

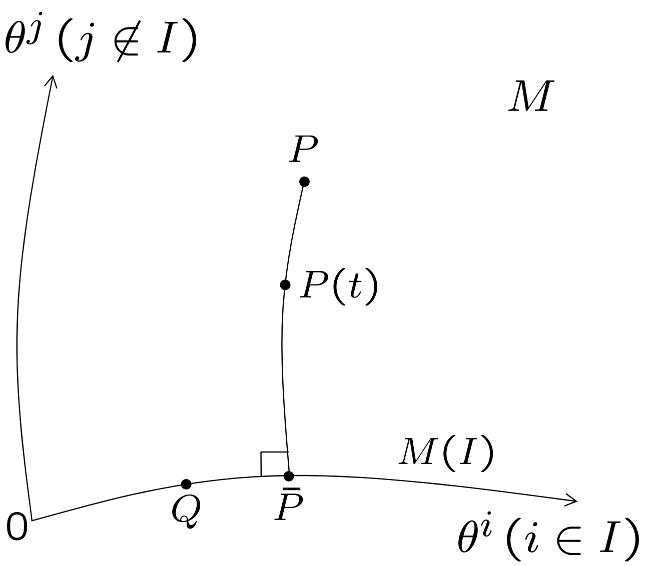

Figure 4: The m-projection and e-affine coordinate system of the subspace M. M: dually flat space. θ: e-affine coordinate system of M. $M(I) = \{\theta | \, \theta^j = 0, \, j \notin I\}$: an e-flat subspace in M and a dually flat space. $\bar{P}$: m-projection of P on $M(I)$. $P(t)\,(0 \le t \le 1)$: point on the m-geodesic connecting P and $\bar{P}$. Q: arbitrary point on $M(I)$. Since the m-geodesic and $M(I)$ are orthogonal, $(\eta(P) - \eta(\bar{P})) \cdot (\theta(Q) - \theta(\bar{P})) = 0$ holds. For $i \in I$, the i-th component $\eta_i(P(t))$ is a constant. If $\theta^{[I]} = (\theta^i)_{i \in I}$, then $\theta^{[I]}$ is an e-affine coordinate system of $M(I)$.

where ψ is the potential function of θ and ϕ is the potential function of η in S. We have the following lemma. The proof of this lemma is given in the appendix.

Lemma 1. For an arbitrary set $I \subset \{1, 2, \ldots, d\}$, the space $M(I)$ is dually flat. The coordinate system $\theta^{[I]}$ is an e-affine coordinate system in $M(I)$. The coordinate system $\eta^{[I]}$ is the m-affine coordinate system that is dual with $\theta^{[I]}$ in $M(I)$. The functions $\psi^{[I]}$ and $\phi^{[I]}$ are the potential functions of $\theta^{[I]}$ and $\eta^{[I]}$ in $M(I)$, respectively. The divergence from P to Q in $M(I)$ is given by $D^{[I]}(\eta^{[I]}(P), \theta^{[I]}(Q)) = \phi^{[I]}(\eta^{[I]}(P)) + \psi^{[I]}(\theta^{[I]}(Q)) - \eta^{[I]}(P) \cdot \theta^{[I]}(Q)$. □

For $I \subset \{1, 2, \ldots, p\}$ and $i \in I$, we consider the m-projection $\bar{P}$ of P on $M(i, \alpha, I) := \{\theta | \, \theta^i = \alpha, \theta^j = 0 \, (j \notin I)\}$ in $M(I)$. The subspace $M(i, \alpha, I)$ is an e-flat subspace in $M(I)$. Let $l(i, I)$ be the m-geodesic connecting P and $\bar{P}$. Then, the m-coordinate of every point on $l(i, I)$ is given by $u\eta^{[I]}(P) + (1 - u)\eta^{[I]}(\bar{P})$ $(u \in [0, 1])$, where $\eta^{[I]}(P)$ and $\eta^{[I]}(\bar{P})$ are the $\eta^{[I]}$-coordinates of P and $\bar{P}$, respectively. Since $l(i, I)$ and $M(i, \alpha, I)$ are orthogonal, we obtain

$$(\eta^{[I]}(P) - \eta^{[I]}(\bar{P})) \cdot (\theta^{[I]}(Q) - \theta^{[I]}(\bar{P})) = 0 \quad (\forall Q \in M(i, \alpha, I)), \tag{9}$$

where $\theta^{[I]}(\bar{P})$ and $\theta^{[I]}(Q)$ represent $\theta^{[I]}$-coordinates of $\bar{P}$ and Q, respectively. The i-th coordinates $\theta^i(Q)$ and $\theta^i(\bar{P})$ are constantly equal to α while $\theta^q(Q)$ and $\theta^q(\bar{P})$ $(q \in I\backslash\{i\})$ are not constants. Therefore, we obtain

$$\eta_q(\bar{P}) = \eta_q(P) \quad (q \in I\backslash\{i\})$$

as the condition to satisfy condition (9). On the m-geodesic $l(i, I)$, all the components of the coordinates except for the i-th one are constants in the $\eta^{[I]}$-coordinate system in $M(I)$.

2.3 Extended LARS algorithm

In this subsection, we describe the extended LARS algorithm. Let $\hat{\theta}_{(k)}$ be the estimator attained in the k-th iteration and let $I \subseteq \{1, 2, \ldots, d\}$ be the set such that $\theta^j = 0\,(j \notin I)$ and $\theta^i \neq 0\,(i \in I)$. First, we define $I = \{1, 2, \ldots, d\}$ and $k = 0$. We define the first estimator $\hat{\theta}_{(0)} = \theta_{\mathrm{MLE}}$, the MLE for model (1).

In this algorithm, we consider the estimator in the dually flat space $M(I)$ (Figure 5). Let $\bar{\theta}(i, I)$ denote the m-projection of $\hat{\theta}_{(k)}$ on $M(i, 0, I) = M(I \setminus \{i\})$ and let $l(i, I)$ denote the m-geodesic from $\hat{\theta}_{(k)}$ to $M(i, 0, I)$. Let the point $\theta(t, i, I) \in l(i, I)$ be a point such that the divergence from $\hat{\theta}_{(k)}$ is equal to $t > 0$. Using $\theta(t, i, I)$, we define $\theta^*(t, I)$ as $(\theta^*(t, I))^i = (\theta(t, i, I))^i\,(i \in I)$, $(\theta^*(t, I))^j = 0\,(j \notin I)$. The point $\theta^*(t, I)$ is the intersection point of $(|I|-1)$-dimensional e-flat spaces $M(i, (\theta(t, i, I))^i, I) = \{\theta |\, \theta^i = (\theta(t, i, I))^i, \theta^j = 0\,(j \notin I)\}\,(i \in I)$. The space $M(i, (\theta(t, i, I))^i, I)$ is orthogonal to the m-geodesic $l(i, I)$ at ponit $\theta(t, i, I)$. By the extended Pythagorean theorem, we obtain

$$D^{[I]}(\theta(t, i_1, I), \theta^*(t, I)) = D^{[I]}(\theta(t, i_2, I), \theta^*(t, I)) \quad (\forall i_1, i_2 \in I,\, \forall t > 0).$$

This implies that $\{\theta^*(t, I)|\, t > 0\}$ is the curve corresponding to a bisector in Euclidean space.

We use the curve $\{\theta^*(t, I)|\, t > 0\}$ in estimating the parameters. As $t > 0$ increases, the curve $\{\theta^*(t, I)|\, t > 0\}$ intersects some spaces $M(i, 0, I)\,(i \in I)$ one-by-one. Let $M(i^*, 0, I)$ denote the first space and let $t^* = D(\hat{\theta}_{(k)}, \bar{\theta}(i^*, I))$. We define the next estimator $\hat{\theta}_{(k+1)}$ as $\hat{\theta}_{(k+1)} = \theta^*(t^*, I)$. Therefore, the estimator $\hat{\theta}_{(k+1)}$ is in the space $M(i^*, 0, I) = M(I\backslash\{i^*\})$.

If $k < d - 1$, then let $I := I\backslash\{i^*\}$, $k := k + 1$ and repeat the above mentioned process. If $k = d - 1$, then let $\hat{\theta}_{(d)} = 0$ and terminate the algorithm.

The extended LARS algorithm is described below.

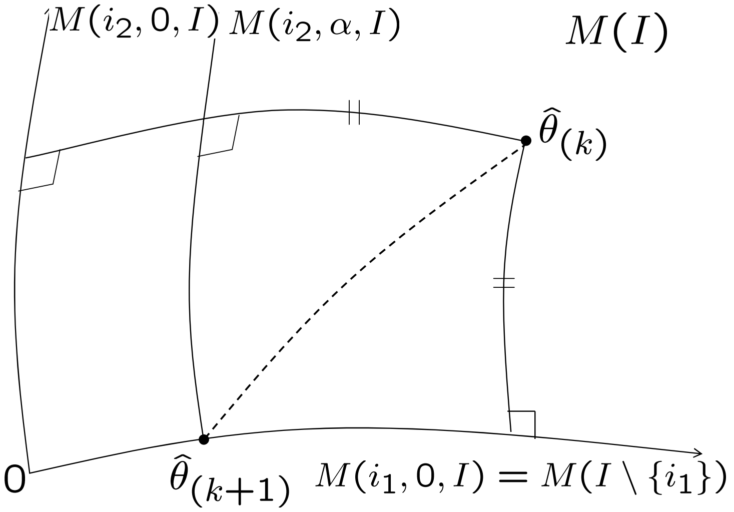

Figure 5: The extended LARS algorithm. $M(I) = \{\theta | \, \theta^j = 0 \, (j \notin I)\}$: dually flat space corresponding to the index set $I \subset \{1, 2, \ldots, d\}$. $M(i, 0, I) = \{\theta | \, \theta^i = 0, \theta^j = 0 \, (j \notin I)\} = M(I \setminus \{i\})$. $M(i, \alpha_i, I) = \{\theta | \, \theta^i = \alpha_i, \theta^j = 0 \, (j \notin I)\}$. $\hat{\theta}_{(k)} \in M(I)$: the k-th estimator. The dotted curve corresponds to a bisector in Euclidean space. The estimator moves along the curve in $M(I)$ from $\hat{\theta}_{(k)}$ until it intersects the first hyperplane $M(i_1, 0, I)$ $(i_1 \in I)$. Then we define the crossing point as the new estimator $\hat{\theta}_{(k+1)}$. We have $M(i_1, 0, I) = M(I \backslash \{i_1\})$, and therefore, we iterate the above process in $M(I \backslash \{i_1\})$.

1. Let $I = \{1, 2, \ldots, d\}$, $\hat{\theta}_{(0)} := \hat{\theta}_{\mathrm{MLE}}$, and $k = 0$.
2. Let $M(i, 0, I) = \{\theta | \, \theta^i = 0, \theta^j = 0 \, (j \notin I)\} = M(I \setminus \{i\}) \, (i \in I)$ and calculate the m-projection $\bar{\theta}(i, I)$ of $\hat{\theta}_{(k)}$ on $M(i, 0, I)$.
3. Let $t^* = \min_{i \in I} D^{[I]}(\hat{\theta}_{(k)}, \bar{\theta}(i, I))$ and $i^* = \arg\min_{i \in I} D^{[I]}(\hat{\theta}_{(k)}, \bar{\theta}(i, I))$.
4. For $\alpha^i \in \mathrm{R}, i \in I$, let $M(i, \alpha^i, I) = \{\theta | \, \theta^i = \alpha^i, \theta^j = 0 \, (j \notin I)\}$. For every $i \in I$, compute α^i such that the m-projection $\bar{\theta}'(i, \alpha^i, I)$ of $\hat{\theta}_{(k)}$ on $M(i, \alpha^i, I)$ satisfies $t^* = D^{[I]}(\hat{\theta}_{(k)}, \bar{\theta}'(i, \alpha^i, I))$.
5. Let $\hat{\theta}^i_{(k)} = \alpha^i \, (i \in I)$ and $\hat{\theta}^j_{(k)} = 0 \, (j \notin I)$.
6. If $k = d - 1$, then go to step 7. If $k < d - 1$, then go to step 2 with $k := k + 1, I := I \backslash \{i^*\}$.
7. Let $\hat{\theta}_{(d)} = 0$ and quit the algorithm.

3 Examples

In this section, we apply the extended LARS algorithm to two types of databases. We used normalized design matrices, i.e., each column vector of design matrices has mean 0 and variance 1. We used the free software R for this purpose [7].

3.1 Normal regression

3.1.1 Data of diabetes

We applied the extended LARS algorithm to the data of diabetes in Efron et al. [4]. The data consists of ten explaining variables $x_1, x_2, \ldots, x_{10}$ and one response variable y of $n = 443$ people.

The explaining variables are x_1: age, x_2: sex, x_3: BMI, x_4: blood pressure, $x_5, \ldots, x_{10}$: serum measurements.

The first estimator, the MLE for model (1), is represented at the right-hand side of Figure 6. The algorithm starts from the right-hand side and proceeds to the left-hand side in this figure. The algorithm ends when the estimator reaches the origin.

According to the extended LARS algorithm, all the components of the estimator $\hat{\theta}$ become 0 in the sequence of $\theta_1, \theta_7, \theta_{10}, \theta_8, \theta_6, \theta_2, \theta_4, \theta_5, \theta_3, \theta_9$ (Figure 6).

3.2 Logistic regression

3.2.1 Data of heart disease

We applied the extended LARS algorithm to South African Heart Disease (SAHD) data [5] that was originally reported in [9]. We used the data included in the ElemStatLearn package in R. This data consists of nine explaining variables $x_1, x_2, \ldots, x_9$ and one response variable y of $n = 462$ people.

Response y is chd. Explaining variables are x_1: sbp, x_2: tobacco, x_3: ldl, x_4: adiposity, x_5: famhist, x_6: typea, x_7: obesity, x_8: alcohol, and x_9: age.

The result of the extended LARS algorithm for this data shows that all the components of the estimator $\hat{\theta}$ become 0 in the sequence of $\theta_8, \theta_4, \theta_1, \theta_7, \theta_3, \theta_2, \theta_6, \theta_5, \theta_9$ (Figure 7). The earlier the coefficient of an explaining variable becomes 0, the weaker is its influence.

4 Conclusion

In this study, we extended the least angle regression (LARS) algorithm. LARS is described in terms of Euclidean geometry. However, we extended

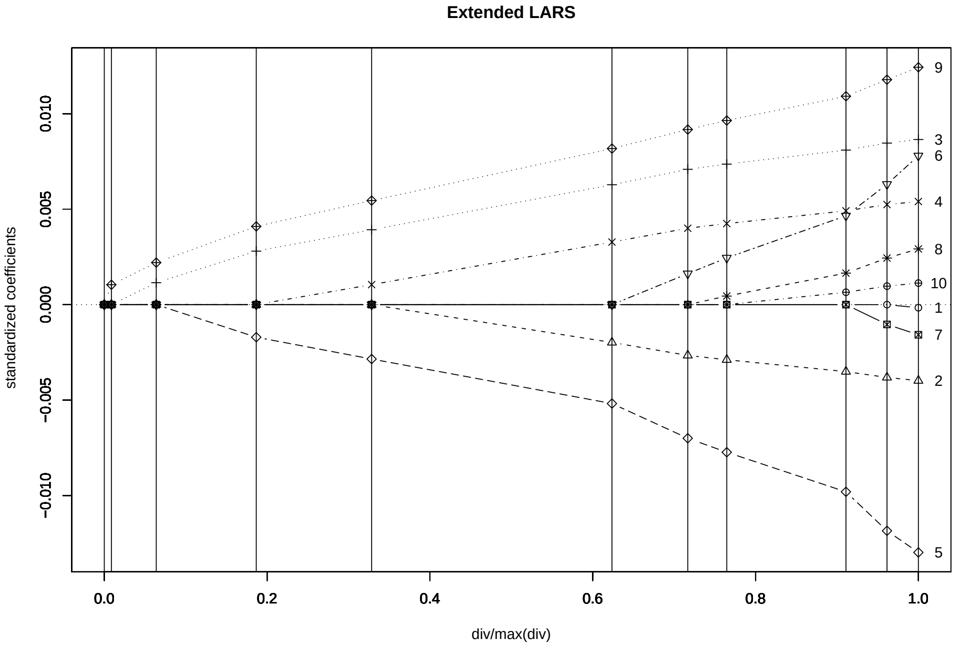

Figure 6: Result of the extended LARS algorithm for the normal regression of the diabetes data. The horizontal axis indicates the normalized divergence from the estimator to the origin. The vertical axis indicates $\theta_i \, (i = 1, 2, \ldots, p)$, i.e., the coefficients of explaining variables $x_i \, (i = 1, 2, \ldots, p)$. The right-hand side of the graph corresponds to the first estimator, that is, the MLE for model (1). Components of the estimator $\hat{\theta}$ become 0 in the sequence of $\theta_1, \theta_7, \theta_{10}, \theta_8, \theta_6, \theta_2, \theta_4, \theta_5, \theta_3, \theta_9$.

LARS using the information geometry of dually flat spaces. The extended LARS algorithm is used for estimating parameters and selecting variables in generalized linear regression. Since dually flat spaces and their information geometry are more general notions than Euclidean spaces and their geometry, the extended LARS algorithm also works in Euclidean spaces. However, the extended and original LARS algorithms differ significantly. The former reduces one explaining variable in each iteration, while the latter increases one explaining variable in each iteration. Thus, the extended LARS algorithm works in dually flat spaces.

It should be noted that the extended and original LARS algorithms are similar in several aspects. First, the extended LARS algorithm can select explaining variables because it reduces one explaining variable in each iteration. Second, the extended LARS algorithm is efficient because the number of iterations is equal to the number of explaining variables. These

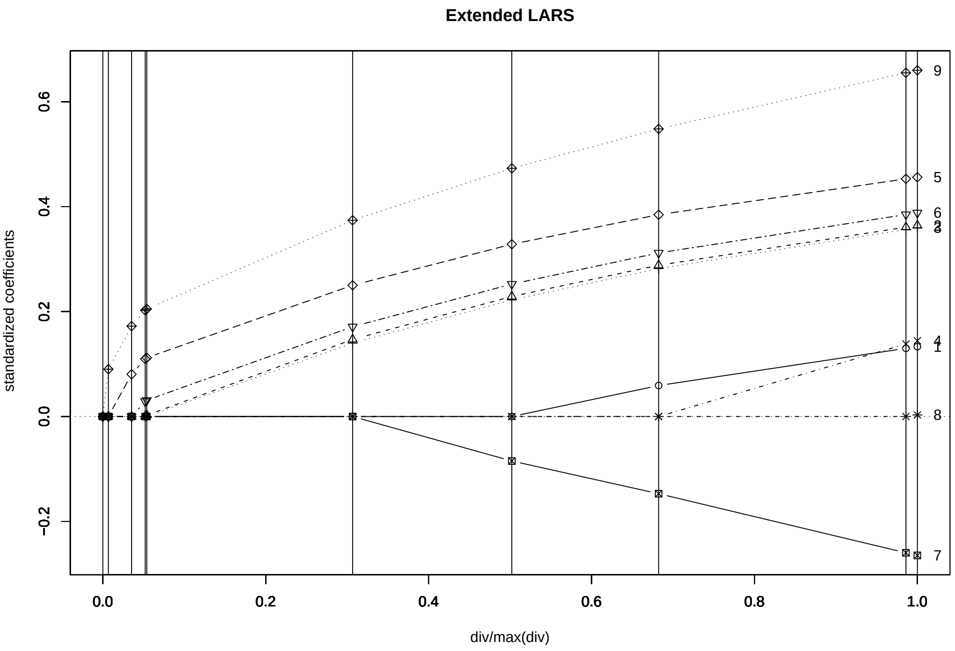

Figure 7: Result of the extended LARS algorithm for the logistic regression of the SAHD data. Components of the estimator become 0 in the sequence of $\theta_8, \theta_4, \theta_1, \theta_7, \theta_3, \theta_2, \theta_6, \theta_5, \theta_9$.

two properties are advantages of the original LARS algorithm. Therefore, it is evident that the extended LARS algorithm has the advantages of the original LARS algorithm.

Moreover, we applied the extended LARS algorithm to two types of databases. The behavior of this algorithm can be observed from the figures.

Finally, we list several interesting problems we are interested in. First, the original and extended LARS algorithms give more than one pair of explaining variables and estimate of the parameter. Therefore, we need to set a criterion to select one estimate. Second, the extended LARS algorithm is based on the assumption that columns of the design matrix are linear independent. However, this assumption is not necessarily valid. For example, a necessary condition for this assumption is that $n \geq p$ in the $n \times p$ design matrix. However, we know cases in which $n < p$; for example, the analysis of gene expressions. Therefore, the problem of how parameters can be estimated without the assumption of linear independence remains unresolved [3].

Acknowledgment

This work was supported in part by Global COE Program "The research and training center for new development in mathematics", MEXT, Japan.

References

[1] S. Amari: *Differential-Geomethical Methods in Statistics.* Springer Lecture Notes in Statistics 28, 1985.

[2] S. Amari and H. Nagaoka: *Methods of Information Geometry.* Translations of Mathematical Monographs, Vol. 191, Oxford University Press, 2000.

[3] E. Candes and T. Tao: The Dantzig Selector: Statistical Estimation When p is Much Larger Than n (with discussion). *The Annals of Statistics*, vol. 35 (2007), pp. 2313–2351.

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani: Least Angle Regression (with discussion). *The Annals of Statistics*, vol. 32 (2004), pp. 407–499.

[5] T. Hastie, R. Tibshirani, and J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[6] R. Kass and P. Vos: *Geometrical Foundations of Asymptotic Inference.* John Wiley, New York, 1997.

[7] J. Maindonald and J. Braun: *Data Analysis and Graphics Using R - an Example-based Approach.* Cambridge University Press, 2003.

[8] M. Osborne, B. Presnell, and B. Turlach: A New Approach to Variable Selection in Least Squares Problems. *IMA Journal of Numerical Analysis*, vol. 20 (2000), pp. 389–403.

[9] J. Rousseauw, J. du Plessis, A. Benade, P. Jordan, J. Kotze, P. Jooste, and J. Ferreira: Coronary Risk Factor Screening in Three Rural Communities. *South African Medical Journal*, vol. 64 (1983), pp. 430–436.

[10] R. Tibshirani: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, vol. 58 (1996), pp. 267–288.

A The proof of lemma 1

Without loss of generality, we consider the case that $I = \{1, 2, \dots, d-1\}$. It is sufficient to prove that relations (3) and (4) hold when ξ and μ are substituted by θ and η, respectively. Since $\theta^d = 0$, for $i = 1, 2, \dots, d-1$, the relations

$$\begin{aligned}
\frac{\partial}{\partial(\theta^{[I]})^i}\psi^{[I]}(\theta^{[I]}) &= \frac{\partial}{\partial\theta^i}\psi(\theta) \\
&= \eta_i \\
&= \eta_i^{[I]}, \\
\frac{\partial}{\partial\eta_i^{[I]}}\phi^{[I]}(\eta^{[I]}) &= \frac{\partial}{\partial\eta_i^{[I]}}\left\{\eta^{[I]}\cdot\theta^{[I]} - \psi^{[I]}(\theta^{[I]})\right\} \\
&= \frac{\partial}{\partial\eta_i}(\eta\cdot\theta) - \frac{\partial}{\partial\eta_i}\psi(\theta) \\
&= \left(\frac{\partial\eta}{\partial\eta_i}\cdot\theta\right) + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \frac{\partial}{\partial\eta_i}\psi(\theta) \\
&= \theta^i + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \sum_{k=1}^{d-1}\frac{\partial\theta^k}{\partial\eta_i}\frac{\partial\psi(\theta)}{\partial\theta^k} \\
&= \theta^i + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \sum_{k=1}^{d-1}\frac{\partial\theta^k}{\partial\eta_i}\eta_k \\
&= \theta^i + \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) - \left(\eta\cdot\frac{\partial\theta}{\partial\eta_i}\right) \\
&= \theta^i \\
&= (\theta^{[I]})^i
\end{aligned}$$

hold. Therefore, the subspace $M(I) = \{\theta |\, \theta^d = 0\}$, i.e., $M(I) = \{\theta |\, \theta^d = 0, \eta_0 = (\eta^*_{\mathrm{MLE}})_0, \eta_{d+1} = (\eta^*_{\mathrm{MLE}})_{d+1}, \dots, \eta_{d+r} = (\eta^*_{\mathrm{MLE}})_{d+r}\}$, is a dually flat space. Coordinate systems $\theta^{[I]}$ and $\eta^{[I]}$ are an e-affine coordinate system and an m-affine coordinate system, respectively. The two coordinate systems $\theta^{[I]}$ and $\eta^{[I]}$ are mutually dual in $M(I)$. The functions $\psi^{[I]}$ and $\phi^{[I]}$ are the potential functions of $\theta^{[I]}$ and $\eta^{[I]}$, respectively.