MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Edge Selection Based on the Geometry of Dually Flat Spaces for Gaussian Graphical Models

Yoshihiro HIROSE and Fumiyasu KOMAKI

METR 2011–35 October 2011

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

Edge Selection Based on the Geometry of Dually Flat Spaces for Gaussian Graphical Models

Yoshihiro HIROSE* and Fumiyasu KOMAKI
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo

October 20, 2011

Abstract

We propose a method for selecting edges in Gaussian graphical models. Our algorithm takes after our previous work, an extension of Least Angle Regression (LARS), and it is based on the information geometry of dually flat spaces. Non-diagonal elements of the inverse of the covariance matrix, the concentration matrix, play an important role in edge selection. Our iterative method estimates these elements and selects covariance models simultaneously. A sequence of pairs of estimates of the concentration matrix and an independence graph is generated, whose length is the same as the number of non-diagonal elements of the matrix. In our algorithm, the next estimate of the graph is the nearest graph to the latest estimate of the concentration matrix. The next estimate of the concentration matrix is not just the projection of the latest estimate, and it is shrunk to the origin. We describe the algorithm and show results for some datasets.

Key words and phrases: divergence, dually flat space, edge selection, Gaussian graphical model, information geometry, shrinkage, update of estimator.

1 Introduction

In this paper, we propose a method for selecting edges in Gaussian graphical models. Edge selection is an important problem in statistical science, machine learning, and many other areas. Our method is based on the information geometry of dually flat spaces.

In 1972, Dempster proposed covariance selection [3]. The non-diagonal elements of the inverse of the covariance matrix, the concentration matrix,

*hirose@stat.t.u-tokyo.ac.jp

play an important role. For reducing the dimension of the parameter space, Maximum Likelihood Estimators (MLEs) are considered under constraints that some non-diagonal elements of the concentration matrix are 0. In [13], an algorithm was proposed for calculating the MLE for any covariance selection model, and Fortran code was given. In 1986, Speed and Kiiveri showed how 0s of the concentration matrix correspond to the conditional independence properties, thereby yielding a graph corresponding to a covariance selection model ([11]). The graph corresponding to a covariance selection model is called the independence graph of the model. In this paper, we propose an iterative algorithm for estimating non-diagonal elements of the concentration matrix and selecting covariance models, or independence graphs, simultaneously.

In the previous study [7], we extended the Least Angle Regression (LARS) algorithm [4] to generalized linear regression. We call this extension of LARS *bisector regression* because an estimator moves along bisectors of angles in the linear regression setting. LARS provides an efficient algorithm for computing LASSO, one of the most famous methods of regularization ([5, 6, 12]). LARS and LASSO estimate parameters and select explanatory variables simultaneously in the linear regression problem. A version of LARS is described in terms of Euclidean geometry, and it is easy to interpret in that correlations are used as angles of explanatory variables. For extending LARS, methods based on the information geometry of dually flat spaces was used ([1, 2, 8]). A new algorithm, bisector regression, was proposed which estimates parameters and selects explanatory variables simultaneously in the generalized linear regression problem. The important points of this extension are that exponential families of distributions form dually flat spaces, the algorithm is described as an estimator's move within a space, similar to LARS, and a sequence of pairs of a parameter estimate and a submodel is generated without the difficulty of combinations of variables. Our method of edge selection proposed in this paper follows the main idea of bisector regression. We consider the dually flat space of multivariate Gaussian distributions and propose the algorithm as updates of an estimator in the space. The algorithm is an iterative one. A sequence of graphs are generated and the number of graphs is the number of non-diagonal elements of the covariance matrix. Not all candidates of graphs are considered and we avoid the difficulty of combinations of edges. In terms of the graph, the next estimate of the graph is the nearest graph to the latest estimate of the concentration matrix. In terms of the concentration matrix, the next estimate of the matrix is not just the projection of the latest estimate of the matrix to the nearest graph. The next estimate of the matrix is shrunk to the origin.

In edge selection in Gaussian graphical models, the problem in this paper, we consider the covariance matrix and the inverse of distributions, while the mean is considered in the regression settings. In the case of the Gaussian

distribution, the space of the mean is Euclidean space. However, the space of the covariance matrix does not form Euclidean space. The dually flat space is introduced naturally for edge selection.

The remainder of this paper is organized as follows. In Section 2, we provide the settings and problem. The notations and parameters to be estimated are also introduced. A brief explanation of edge selection is given. The new algorithm for edge selection in Gaussian graphical models is proposed in Section 3. We provide an intuitive explanation and the detailed algorithm. Our iterative algorithm is based on the information geometry of dually flat spaces. A geometrical explanation is given. In Section 4, the result of our algorithm is shown for some datasets. The results are compared with those of the graphical LASSO [15]. We conclude the paper in Section 5.

2 Edge Selection in Gaussian Graphical Models

In this section, we consider the setting and problem, edge selection in Gaussian graphical models. We define the notations and parameters to be estimated. A brief explanation of edge selection is given for our algorithm.

2.1 Settings, Notations and Tools

We define the settings, notations, and tools used in the following. $X_1, \ldots, X_p$ are the random variables under consideration. $X = (X_1, \ldots, X_p)^\top$ has a multivariate Gaussian distribution with the mean vector $\mu = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix $\Sigma = (\sigma_{ab})$. The inverse of the covariance matrix Σ, the concentration matrix, is represented as $\Sigma^{-1} = (\sigma^{ab})$. We sometimes use the notation $(\Sigma)_{ab} = \sigma_{ab}$ and $(\Sigma^{-1})^{ab} = \sigma^{ab}$ for convenience.

The distribution of X is

$$f(x_1, \ldots, x_p | \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Let l denote the logarithm of f, that is,

$$\begin{aligned} l(x|\mu, \Sigma) &= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) \\ &= -\frac{1}{2}x^\top \Sigma^{-1} x + x^\top \left(\Sigma^{-1}\mu\right) \\ &\quad - \left\{-\frac{1}{2}\log\left|\Sigma^{-1}\right| + \frac{1}{2}\mu^\top \Sigma^{-1}\mu\right\} - \frac{p}{2}\log(2\pi). \end{aligned}$$

Defining $u = \Sigma^{-1}\mu$, we have

$$\begin{aligned} l(x|u,\Sigma^{-1}) &= -\frac{1}{2}x^\top\Sigma^{-1}x + x^\top u - \left\{-\frac{1}{2}\log\left|\Sigma^{-1}\right| + \frac{1}{2}u^\top\Sigma u\right\} \\ &= \sum_{1\le a<b\le p} -\sigma^{ab}x_a x_b + \sum_{a=1}^{p}\left(-\frac{1}{2}\sigma^{aa}\right)x_a^2 + \sum_{a=1}^{p} u^a x_a \\ &\qquad - \left\{-\frac{1}{2}\log\left|\Sigma^{-1}\right| + \frac{1}{2}u^\top\Sigma u\right\} - \frac{p}{2}\log(2\pi). \end{aligned}$$

Let $d := p(p-1)/2$, the number of non-diagonal elements of the covariance matrix. The natural parameter θ is defined by

$$\theta = \left(\left(-\sigma^{ab}\right)_{1\le a<b\le p}; \left(-\frac{1}{2}\sigma^{aa}\right)_{1\le a\le p}; (u^a)_{1\le a\le p}\right).$$

When $p = 4$, for example, it holds that $d = 6$ and

$$\begin{aligned} \theta = \Big(&-\sigma^{12}, -\sigma^{13}, -\sigma^{14}, -\sigma^{23}, -\sigma^{24}, -\sigma^{34}; \\ &-\frac{1}{2}\sigma^{11}, -\frac{1}{2}\sigma^{22}, -\frac{1}{2}\sigma^{33}, -\frac{1}{2}\sigma^{44}; u^1, u^2, u^3, u^4\Big). \end{aligned}$$

The expectation parameter η corresponding to the natural parameter θ is

$$\eta = \left(\left(\sigma_{ab} + \mu_a\mu_b\right)_{1\le a<b\le p}; \left(\sigma_{aa} + \mu_a^2\right)_{1\le a\le p}; (\mu_a)_{1\le a\le p}\right).$$

When $p = 4$, we have

$$\begin{aligned} \eta = \Big(&\sigma_{12} + \mu_1\mu_2, \sigma_{13} + \mu_1\mu_3, \sigma_{14} + \mu_1\mu_4, \sigma_{23} + \mu_2\mu_3, \sigma_{24} + \mu_2\mu_4, \sigma_{34} + \mu_3\mu_4; \\ &\sigma_{11} + \mu_1^2, \sigma_{22} + \mu_2^2, \sigma_{33} + \mu_3^2, \sigma_{44} + \mu_4^2; \mu_1, \mu_2, \mu_3, \mu_4\Big). \end{aligned}$$

Hereafter, we also call the pair (u, Σ^{-1}) the natural parameter. Similarly, the pair (μ, Σ) is called the expectation parameter. The two pairs (u, Σ^{-1}) and (μ, Σ) are just different names specifying a distribution.

We define two functions, called potential functions, with respect to the natural parameter (u, Σ^{-1}) and the expectation parameter (μ, Σ), respectively. Let ψ be the potential function of the natural parameter (u, Σ^{-1}) defined by

$$\psi(u, \Sigma^{-1}) = -\frac{1}{2}\log\left|\Sigma^{-1}\right| + \frac{1}{2}u^\top\Sigma u,$$

which is the last part of l. The potential function ϕ of the expectation parameter (μ, Σ) is defined by

$$\phi(\mu, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{p}{2}.$$

For the potential functions, it holds that

$$\phi(\mu, \Sigma) + \psi(u, \Sigma^{-1}) - \frac{1}{2} u^\top \Sigma u + \frac{p}{2} = 0.$$

Let $D(\cdot|\cdot)$ denote the Kullback-Leibler divergence, that is, the divergence between two Gaussian distributions (μ_1, Σ_1) and (μ_2, Σ_2), given by

$$\begin{aligned} D(\mu_1, \Sigma_1 \mid \mu_2, \Sigma_2) &= \mathrm{E}_{(\mu_1, \Sigma_1)} \left[\log \frac{\exp\left\{-\frac{1}{2}(X-\mu_1)^\top \Sigma_1^{-1}(X-\mu_1)\right\} / \left((2\pi)^{p/2}\sqrt{|\Sigma_1|}\right)}{\exp\left\{-\frac{1}{2}(X-\mu_2)^\top \Sigma_2^{-1}(X-\mu_2)\right\} / \left((2\pi)^{p/2}\sqrt{|\Sigma_2|}\right)} \right] \\ &= -\frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_2| + \frac{1}{2}\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1}\right) - \frac{p}{2} \\ &\qquad + \frac{1}{2}(\mu_2-\mu_1)^\top \Sigma_2^{-1}(\mu_2-\mu_1). \end{aligned}$$

In our method, the mean vector μ is fixed as explained later. The divergence from Σ_1 to Σ_2 is given as

$$D(\Sigma_1 \mid \Sigma_2) = -\frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_2| + \frac{1}{2}\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1}\right) - \frac{p}{2}.$$

In addition, we use the notation $D(\Sigma_1^{-1}|\,\Sigma_2^{-1}) = D(\Sigma_1|\,\Sigma_2)$.

2.2 Problem

We explain edge selection in Gaussian graphical models briefly. Only undirected graphs are considered and neither directed nor hybrid graphs are treated in this paper. Non-diagonal elements of the concentration matrix $\Sigma^{-1} = (\sigma^{ab})$ play an important role. For details, see references [9, 14].

We consider the graph $G = (E, V)$ corresponding to the random variables $X_1, \ldots, X_p$ with a multivariate Gaussian distribution. Each node $v_a \in V$ of the graph G corresponds to each random variable X_a and, for $a \neq b$, whether or not X_a and X_b have an edge $e_{ab} \in E$ between them depends on the value of σ^{ab}. There is no edge between X_a and X_b if $\sigma^{ab} = 0$. The condition $\sigma^{ab} = 0$ means that two random variables X_a and X_b are conditionally independent given that other random variables fixed. The graph G represents the conditional independence between random variables $X_1, \ldots, X_p$, and it is called the independence graph. In edge selection for the independence graph G, we must estimate the non-diagonal elements of the concentration matrix Σ^{-1}, $\sigma^{ab}\,(a \neq b)$. It is particularly important to decide which σ^{ab}s are 0.

In the following, we apply the main idea of bisector regression to edge selection in Gaussian graphical models. The mean vector μ is fixed and the space of Gaussian distributions is considered with two different names of the covariance matrix Σ and concentration matrix Σ^{-1}. Our main objectives are

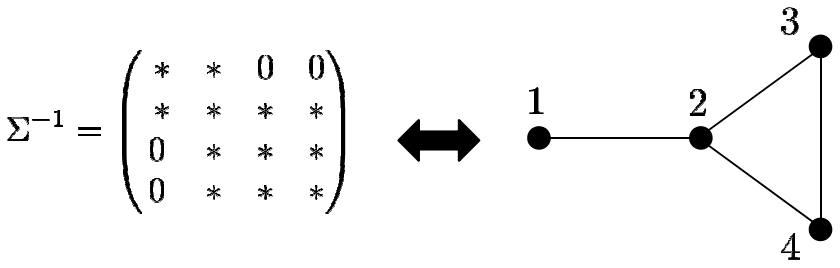

Figure 1: An example of the concentration matrix and independence graph for $p = 4$. The asterisk $*$ represents any value. In the concentration matrix Σ^{-1}, two elements are set to 0, that is, $\sigma^{13} = \sigma^{14} = 0$. In the independence graph corresponding to Σ^{-1}, the two edges e_{13} and e_{14} are deleted.

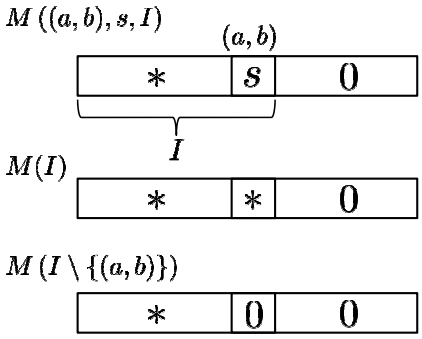

Figure 2: Elements of submodels. The non-diagonal elements of Σ^{-1}, σ^{ab}, are indicated for three submodels $M((a,b),s,I)$, $M(I)$, and $M(I \setminus \{(a,b)\})$. The asterisk $*$ represents any value.

to estimate the non-diagonal elements of Σ^{-1}, $\sigma^{ab}\,(a \neq b)$, and to estimate the independence graph G corresponding to Σ^{-1}. We consider submodels

$$M(I) = \left\{ \Sigma \mid \sigma^{a'b'} = 0 \ \left((a',b') \notin I \right) \right\},$$
$$M((a,b),s,I) = \left\{ \Sigma \mid \sigma^{ab} = s,\ \sigma^{a'b'} = 0 \ \left((a',b') \notin I \right) \right\}$$

for $(a,b) \in I \subseteq \{(a',b') | \, 1 \leq a' < b' \leq p\}$. The submodel $M(I)$ is the model where the edge $e_{a'b'}$ is deleted for $(a',b') \notin I$. The submodel $M((a,b),s,I)$ is the model where the edge $e_{a'b'}$ is deleted for $(a',b') \notin I$ and e_{ab} is deleted if $s = 0$. A graph corresponding to a distribution in $M(I)$ or $M((a,b),s,I)$ has edges in I at most.

3 Proposed Algorithm

In this section, the algorithm for our method is proposed. We give an intuitive explanation of the algorithm. The algorithm is illustrated geometrically and a detailed algorithm is proposed.

We describe the proposed algorithm for selecting edges in Gaussian graphical models. The main idea is based on our previous work [7] on bisector regression, which uses the information geometry of dually flat spaces. Our algorithm is an iterative one, generating a graph at each iteration. A sequence of graphs $\widehat{G}_{(0)}, \ldots, \widehat{G}_{(d)}$ is produced by the algorithm, where $d = p(p-1)/2$ is the number of non-diagonal elements of the covariance matrix Σ.

First, we illustrate our algorithm intuitively during the first iteration. A detailed explanation of the $(k+1)$th iteration is given later. The algorithm can be described as updates of one estimator in the dually flat space of multivariate Gaussian distributions with a fixed mean vector and fixed variances. The mean vector μ and variances σ_{aa}, the diagonal elements of the covariance matrix Σ, are fixed as $\mu = \mu_{\mathrm{MLE}}$ and $\sigma_{aa} = (\sigma_{\mathrm{MLE}})_{aa}\,(a = 1, \ldots, p)$, where μ_{MLE} and $(\sigma_{\mathrm{MLE}})_{aa}$ are the MLEs, the sample mean and sample variance, respectively. Both Σ and Σ^{-1} indicate the Gaussian distribution with the covariance matrix Σ. The subspace with fixed μ and fixed σ_{aa}s is a dually flat space, and our algorithm runs in this subspace. Our estimator starts at the MLE $\widehat{\Sigma}^{-1}_{(0)} = \widehat{\Sigma}^{-1}_{\mathrm{MLE}}$ of the model with no constraint, which corresponds to the complete graph denoted by $\widehat{G}_{(0)} = \widehat{G}_{\mathrm{MLE}}$. $\widehat{\Sigma}_{\mathrm{MLE}}$ is the sample covariance matrix. Note that the diagonal elements of $\widehat{\Sigma}_{\mathrm{MLE}}$ are the same as the $(\sigma_{\mathrm{MLE}})_{aa}$s. The terminal of the estimator is the MLE $\widehat{\Sigma}^{-1}_{(d)} = \widehat{\Sigma}^{-1}_{0}$ of the model with the condition that each pair of random variables $X_1, \ldots, X_p$ is conditionally independent, which corresponds to the graph with no edge denoted by $\widehat{G}_{(d)} = \widehat{G}_0$. The diagonal elements of $\widehat{\Sigma}_0$ are the same as the $(\sigma_{\mathrm{MLE}})_{aa}$s. For the graph $G^{-ab}_{(0)}\,(a \neq b)$, which has all edges except for the edge e_{ab} between X_a and X_b, we measure the distance from $\widehat{\Sigma}^{-1}_{(0)}$ to the graph $G^{-ab}_{(0)}$. The distance from $\widehat{\Sigma}^{-1}_{(0)}$ to $G^{-ab}_{(0)}$ is defined as the distance from $\widehat{\Sigma}^{-1}_{(0)}$ to the projection of $\widehat{\Sigma}^{-1}_{(0)}$ to the model corresponding to the graph $G^{-ab}_{(0)}$. This projection is known to be the MLE of the model of $G^{-ab}_{(0)}$. The nearest graph from $\widehat{\Sigma}^{-1}_{(0)}$ among all $a < b$ is the next estimate of the independence graph G, and it is denoted by $\widehat{G}_{(1)}$. $\widehat{G}_{(1)}$ has $d-1$ edges. In terms of Σ^{-1}, the next estimate of Σ^{-1}, $\widehat{\Sigma}^{-1}_{(1)}$, is not just the projection of $\widehat{\Sigma}^{-1}_{(0)}$ to the model of the graph $\widehat{G}_{(1)}$. $\widehat{\Sigma}^{-1}_{(1)}$ is shrunk to $\widehat{\Sigma}^{-1}_{0}$ depending on the distance between $\widehat{\Sigma}^{-1}_{(0)}$ and the MLE of the model of $\widehat{G}_{(1)}$. The details of this shrinkage are described later. Thus, we obtain the next estimate of the concentration matrix Σ^{-1} and independence graph G, $\widehat{\Sigma}^{-1}_{(1)}$ and $\widehat{G}_{(1)}$, respectively. The

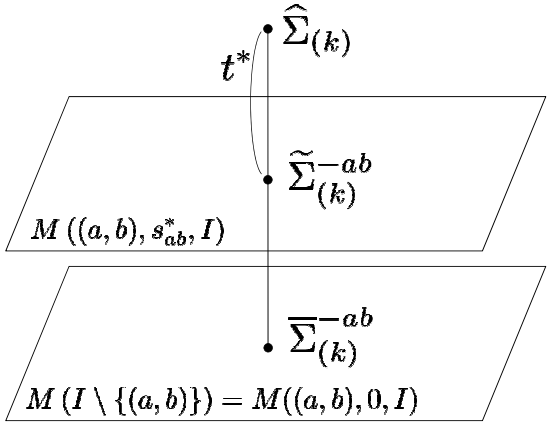

Figure 3: Submodels and projections. $\widehat{\Sigma}_{(k)}$: kth estimate of Σ, $I \subseteq \{(a',b') \mid 1 \leq a' < b' \leq p\}$: indices of non-diagonal elements to be estimated, $(a,b) \in I$: elements of I, $M((a,b),s,I)$: the submodel $\left\{\Sigma \,\middle|\, \sigma^{ab} = s,\, \sigma^{a'b'} = 0\ ((a',b') \notin I)\right\}$, $\overline{\Sigma}^{-ab}_{(k)}$: the projection of $\widehat{\Sigma}_{(k)}$ to the submodel $M\left(I \setminus \{(a,b)\}\right)$, $\widetilde{\Sigma}^{-ab}_{(k)}$: the projection of $\widehat{\Sigma}_{(k)}$ to the submodel $M((a,b), s^*_{ab}, I)$, s^*_{ab}: the value decided by the condition that $D\left(\widehat{\Sigma}_{(k)} \mid \widetilde{\Sigma}^{-ab}_{(k)}\right) = t^*$. The estimate and two projections are on a geodesic. $\widetilde{\Sigma}^{-ab}_{(k)}$ is on the geodesic connecting $\widehat{\Sigma}_{(k)}$ and $\overline{\Sigma}^{-ab}_{(k)}$.

first iteration of the algorithm is completed. In the second iteration, the algorithm proceeds in the same way as in the first, substituting $\widehat{\Sigma}^{-1}_{(1)}$ and $\widehat{G}_{(1)}$ for $\widehat{\Sigma}^{-1}_{(0)}$ and $\widehat{G}_{(0)}$, respectively.

After k iterations, we should have the kth estimate of the concentration matrix Σ^{-1} and independence graph G, $\widehat{\Sigma}^{-1}_{(k)}$ and $\widehat{G}_{(k)}$, respectively. $\widehat{\Sigma}^{-1}_{(k)}$ has k 0s and $\widehat{G}_{(k)} = (\widehat{E}_{(k)}, V)$ has $d-k$ edges, that is, $|\widehat{E}_{(k)}| = d-k$. The set of indices of the non-diagonal elements to be estimated, I, has $d-k$ components and $\widehat{E}_{(k)} = \{e_{ab} | \, (a,b) \in I\}$. In the following, we give an explanation mainly in terms of the covariance matrix Σ or concentration matrix Σ^{-1}, not in terms of the graph G. The graph G is determined by Σ^{-1}, depending on values of σ^{ab}. Strictly speaking, Σ^{-1} has more information than G because G does not know the values of σ^{ab}. Σ^{-1} knows the values of σ^{ab} while G knows only which σ^{ab} is 0, that is, G indicates a model, a covariance selection model. We use $M(I)$ and $M((a,b),s,I)$ instead of G. In the $(k+1)$th iteration, the algorithm proceeds as follows. Note that the mean vector μ and diagonal elements of the covariance matrix Σ, σ_{aa} for $1 \leq a \leq p$, are fixed at the values of the MLEs, and that $\hat{\sigma}^{a'b'} = 0$ for $(a',b') \notin I$. For $(a,b) \in I$, we measure the distance from $\widehat{\Sigma}^{-1}_{(k)}$ to the submodel $M(I \setminus \{(a,b)\})$, which is the same as $M\left((a,b), 0, I\right)$. The distance from $\widehat{\Sigma}^{-1}_{(k)}$ to $M(I \setminus \{(a,b)\})$

is the distance from $\widehat{\Sigma}_{(0)}^{-1}$ to the MLE $\overline{\Sigma}_{(k)}^{-ab}$ of the submodel $M(I \setminus \{(a,b)\})$. $\overline{\Sigma}_{(k)}^{-ab}$ is also the projection of $\widehat{\Sigma}_{(k)}$ to the submodel $M(I \setminus \{(a,b)\})$. Let $M(I \setminus \{(a^*,b^*)\})$ be the nearest submodel from $\widehat{\Sigma}_{(k)}^{-1}$ among all $(a,b) \in I$ and $t^* = \min_{(a,b)\in I} D\left(\widehat{\Sigma}_{(k)} \mid \overline{\Sigma}_{(k)}^{-ab}\right) = D\left(\widehat{\Sigma}_{(k)} \mid \overline{\Sigma}_{(k)}^{-a^*b^*}\right)$. For $(a,b) \in I$, the submodel $M(I \setminus \{(a,b)\})$ is translated so that the distance from $\widehat{\Sigma}_{(k)}$ to the translated model is equal to t^* as follows. The m-geodesic $l_{(k)}^{-ab}$ connecting $\widehat{\Sigma}_{(k)}$ and $\overline{\Sigma}_{(k)}^{-ab}$ is given by $l_{(k)}^{-ab} = \left\{\Sigma \middle| L \leq \sigma_{ab} \leq U, \sigma_{a'b'} = \left(\widehat{\Sigma}_{(k)}\right)_{a'b'} \text{ for } (a',b') \neq (a,b), (a',b') \in I, \sigma^{a'b'} = 0 \text{ for } (a',b') \notin I\right\}$, where $L = \min\left\{\left(\widehat{\Sigma}_{(k)}\right)_{ab}, \left(\overline{\Sigma}_{(k)}^{-ab}\right)_{ab}\right\}$ and $U = \max\left\{\left(\widehat{\Sigma}_{(k)}\right)_{ab}, \left(\overline{\Sigma}_{(k)}^{-ab}\right)_{ab}\right\}$. On the m-geodesic $l_{(k)}^{-ab}$, elements $\sigma_{a'b'}$ are fixed for $(a',b') \neq (a,b)$, $(a',b') \in I$ and $\sigma^{a'b'} = 0$ are fixed for $(a',b') \notin I$. We calculate s_{ab}^* and $\widetilde{\Sigma}_{(k)}^{-ab} \in l_{(k)}^{-ab}$ so that $D\left(\widehat{\Sigma}_{(k)} \mid \widetilde{\Sigma}_{(k)}^{-ab}\right) = t^*$ for $\widetilde{\Sigma}_{(k)}^{-ab} \in M\left((a,b), s_{ab}^*, I\right)$. This submodel $M\left((a,b), s_{ab}^*, I\right)$ is the translated model of $M(I\setminus\{(a,b)\})$. The next estimate of Σ^{-1}, $\widehat{\Sigma}_{(k+1)}^{-1}$, is defined as the intersection of the translated models. In detail, let $\hat{\sigma}_{(k+1)}^{ab} := \left((\widetilde{\Sigma}_{(k)}^{-ab})^{-1}\right)^{ab} = s_{ab}^*$ for $(a,b) \in I$, $\hat{\sigma}_{(k+1)}^{ab} := 0$ for $(a,b) \notin I$. Let $\widehat{E}_{k+1} := \widehat{E}_k \setminus \{e_{a^*b^*}\}$, and define $\widehat{G}_{k+1} = \left(\widehat{E}_{k+1}, V\right)$. Thus, we obtain the next estimate of the concentration matrix Σ^{-1} and independence graph G, $\widehat{\Sigma}_{(k+1)}^{-1}$ and $\widehat{G}_{(k+1)}$, respectively. The $(k+1)$th iteration of the algorithm is completed. In the $(k+2)$th iteration, the algorithm proceeds in the same way, substituting $\widehat{\Sigma}_{(k+1)}^{-1}$ and $\widehat{G}_{(k+1)}$ for $\widehat{\Sigma}_{(k)}^{-1}$ and $\widehat{G}_{(k)}$, respectively. After $d = p(p-1)/2$ iterations, our estimator comes to $\widehat{\Sigma}_0^{-1}$ and $\widehat{G}_0$, and the algorithm is completed.

Some remarks are necessary for describing the algorithm. In our algorithm, (a) the diagonal elements of the covariance matrix Σ are fixed, (b) the non-diagonal elements of the concentration matrix Σ^{-1} are estimated, (c) other elements of Σ and Σ^{-1} are decided by elements which are fixed or estimated, and (d) the mean vector is fixed. The algorithm is applied to the non-diagonal elements of Σ^{-1} and estimates them. In the following, we confine our attention to the non-diagonal elements of Σ^{-1}, and the description of other elements of Σ and Σ^{-1} is omitted.

The algorithm is given as follows. Steps 2 to 6 are iterated.

1. Let $\widehat{\Sigma}_{(0)} := \widehat{\Sigma}_{\mathrm{MLE}}$, $\widehat{G}_{(0)} := \widehat{G}_{\mathrm{MLE}}$, $I := \{(a,b) | \, 1 \leq a < b \leq p\}$, and $k := 0$.

2. Calculate the MLE $\overline{\Sigma}_{(k)}^{-ab}$ of the model $M\left(I \setminus \{(a,b)\}\right)$ for $(a,b) \in I$.

3. Find $t^* := \min_{(a,b)\in I} D\left(\widehat{\Sigma}_{(k)} \mid \overline{\Sigma}_{(k)}^{-ab}\right)$ and $(a^*,b^*) := \arg\min_{(a,b)\in I} D\left(\widehat{\Sigma}_{(k)} \mid \overline{\Sigma}_{(k)}^{-ab}\right)$.

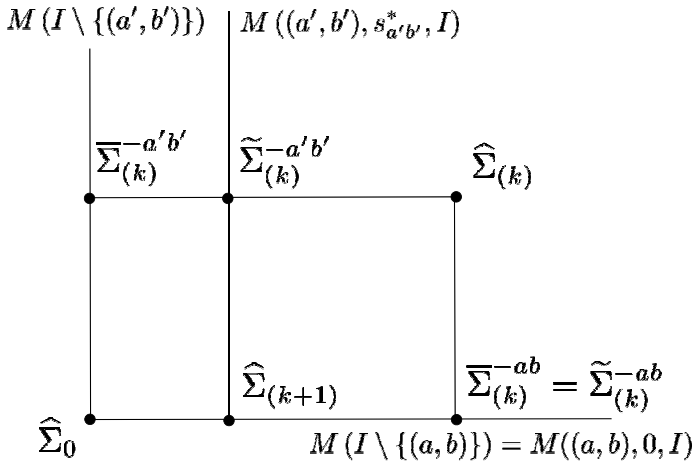

Figure 4: Update of the estimator. $\widehat{\Sigma}_{(k)}$: kth estimate of Σ, $\widehat{\Sigma}_{(k+1)}$: $(k+1)$th estimate of Σ, $\widehat{\Sigma}_0$: the MLE of the model with no edge, $M((a,b),s,I) = \left\{\Sigma \,\middle|\, \sigma^{ab} = s,\ \sigma^{a'b'} = 0\ ((a',b') \notin I)\right\}$, $\overline{\Sigma}^{-ab}_{(k)}$: the MLE of the submodel $M(I \setminus \{(a,b)\})$, $\widetilde{\Sigma}^{-ab}_{(k)}$: the projection of $\widehat{\Sigma}_{(k)}$ to the submodel $M((a,b), s^*_{ab}, I)$. $M(I \setminus \{(a,b)\})$ is nearer from $\widehat{\Sigma}_{(k)}$ than $M(I \setminus \{(a',b')\})$ is. The divergence to $\widetilde{\Sigma}^{-a'b'}_{(k)}$ from $\widehat{\Sigma}_{(k)}$ is the same as the divergence to $\overline{\Sigma}^{-ab}_{(k)}$. $\widehat{\Sigma}_{(k+1)}$ is defined as the intersection of $M((a,b),0,I)$ and $M((a',b'), s^*_{a'b'}, I)$.

4. For $(a,b) \in I$, calculate s^*_{ab} and $\widetilde{\Sigma}^{-ab}_{(k)} \in l^{-ab}_{(k)}$ satisfying $D(\widehat{\Sigma}_{(k)} \mid \widetilde{\Sigma}^{-ab}_{(k)}) = t^*$ and $\widetilde{\Sigma}^{-ab}_{(k)} \in M((a,b), s^*_{ab}, I)$.

5. Let $\hat{\sigma}^{ab}_{(k+1)} := s^*_{ab}$ for $(a,b) \in I$, $\hat{\sigma}^{ab}_{(k+1)} := 0$ for $(a,b) \notin I$, and $\widehat{E}_{(k+1)} := \widehat{E}_{(k)} \setminus \{e_{a^*b^*}\}$.

6. If $k+1 < d-1$, then go to step 2 with $k := k+1, I := I \setminus \{(a^*,b^*)\}$. If $k+1 = d-1$, then go to step 7.

7. Let $\widehat{\Sigma}_{(d)} := \widehat{\Sigma}_0$ and $\widehat{E}_{(d)} := \emptyset$. Stop the algorithm.

Note that in step 5, it holds that $\hat{\sigma}^{a^*b^*}_{(k+1)} = 0$ while $(a^*,b^*) \in I$. This fact indicates that one non-diagonal element of Σ^{-1} becomes 0 in an iteration and that our method selects covariance models sequentially.

4 Examples

We show the results of our method for some datasets and compare them with those of the graphical LASSO [15]. We used the software R [10] for computing the algorithm. The datasets are included in the `SIN` and `SMPracticals` packages of R.

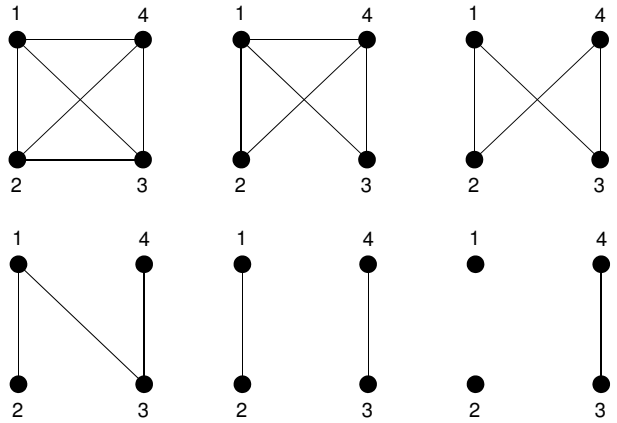

Figure 5: The result of our method for the frets' heads dataset. The sequence of independence graphs generated by our method is shown. 1: the head length of the first son, 2: the head breadth of the first son, 3: the head length of the second son, 4: the head breadth of the second son.

4.1 Frets' Heads Dataset

We show the result of our method for Frets' heads dataset. The data `frets` in the `SMPracticals` package was used. This dataset consists of head measurements of the first and the second adult son in a sample of 25 families. The dataset includes four variables, which are the head length of the first son, head breadth of the first son, the head length of the second son, and head breadth of the second son.

The result of our method is shown in Figure 5. The sequence of independence graphs is generated, the length of which is six, the total number of all edges. The graphical LASSO produces the sequence of graphs shown in Figure 6. Two sequences are almost the same but do not coincide strictly.

4.2 Mathematics Marks Dataset

The results for the mathematics marks dataset are shown. We used the `mathmarks` data in the `SIN` package. This dataset consists of the examination marks of 88 students in five subjects. The dataset includes five variables, which are mechanics, vectors, algebra, analysis, and statistics.

The result of our method is shown in Figure 7. The sequence of independence graphs is generated, the length of which is ten. The result of the graphical LASSO is shown in Figure 8. The two methods produce almost the same sequences of independence graphs, but they are not strictly the same.

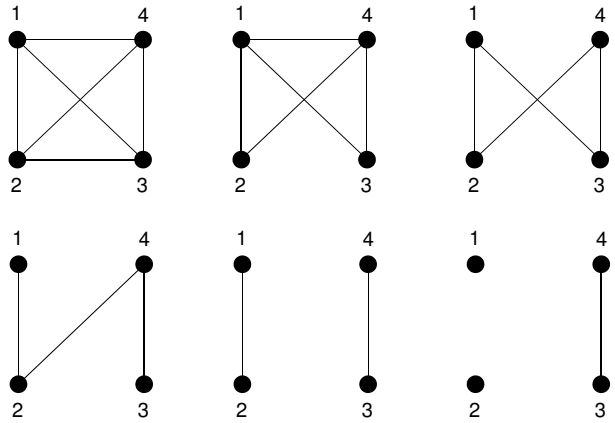

Figure 6: The result of the graphical LASSO for the frets' heads dataset. 1: the head length of the first son, 2: the head breadth of the first son, 3: the head length of the second son, 4: the head breadth of the second son.

5 Conclusion

We proposed a new method for selecting edges in Gaussian graphical models, where the main idea comes from our previous work on bisector regression. Our method is based on the information geometry of dually flat spaces, and it estimates the concentration matrix and selects edges of the graph simultaneously. A sequence of pairs of estimates of the concentration matrix and independence graph is generated, whose length is same as the number of non-diagonal elements of the matrix. The algorithm is efficient in that it avoids the difficulty of combinatorial choices on edges because all pairs of edges are not considered. Our algorithm is described as updates of an estimator in the dually flat space. Our estimator is updated with shrinkage concerning the intersection of submodels, and it goes into submodels in turn. This means that an estimate of the independence graph has one less edge than the previous estimate of the graph.

The results of our method were shown for some datasets and compared with those of the graphical LASSO. The two methods produce almost the same sequences of graphs, but the sequences do not necessarily coincide.

We are working on applying our main idea of bisector regression to other models and are making an effort to provide efficient and stable codes for our method. Sophisticated codes will enable readers to apply our method to suit their own purposes.

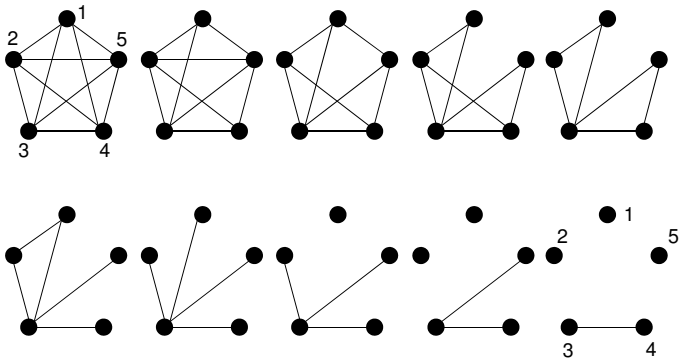

Figure 7: The result of our method for the mathematics marks dataset. 1: mechanics, 2: vectors, 3: algebra, 4: analysis, 5: statistics.

References

[1] S. Amari (1985). *Differential-Geometrical Methods in Statistics*. Springer Lecture Notes in Statistics, vol. 28, Springer.

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

[3] A. P. Dempster (1972). Covariance Selection. *Biometrics*, vol. 28, 157–175.

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

[5] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani (2007). Pathwise Coordinate Optimization. *The Annals of Applied Statistics*, vol. 1, 302–332.

[6] T. Hastie, R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer, New York.

[7] Y. Hirose and F. Komaki (2010). An Extension of Least Angle Regression Based on the Information Geometry of Dually Flat Spaces, *Journal of Computational and Graphical Statistics*, vol. 19, 1007–1023.

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

[9] S. L. Lauritzen (1996). *Graphical Models*. Clarendon Press, Oxford.

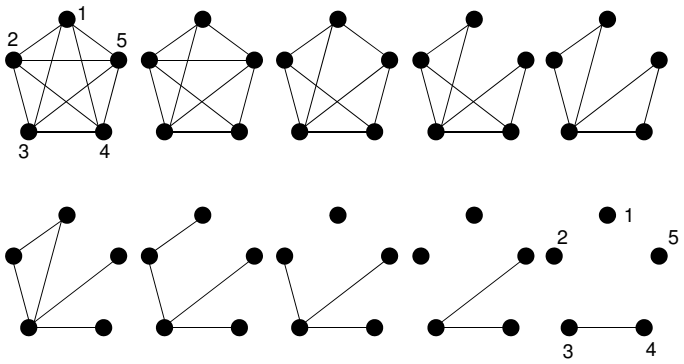

Figure 8: The result of the graphical LASSO for the mathematics marks dataset. 1: mechanics, 2: vectors, 3: algebra, 4: analysis, 5: statistics.

[10] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

[11] T. P. Speed and H. T. Kiiveri (1986). Gaussian Markov Distributions over Finite Graphs. *The Annals of Statistics*, vol. 14, 138–150.

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

[13] N. Wermuth and E. Scheidt (1977). Fitting a Covariance Selection Model to a Matrix. *Applied Statistics*, vol. 26, 88–92.

[14] J. Whittaker (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester.

[15] M. Yuan and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, vol. 94, 19–35.