MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

A Structural Model on a Hypercube Represented by Optimal Transport

Tomonari SEI

METR 2009–03 January 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

A structural model on a hypercube represented by optimal transport

Tomonari SEI

January 30, 2009

Abstract

We propose a flexible statistical model for high-dimensional quantitative data on a hypercube. Our model, called the structural gradient model (SGM), is based on a one-to-one map on the hypercube that is a solution for an optimal transport problem. As we show with many examples, SGM can describe various dependence structures including correlation and heteroscedasticity. The maximum likelihood estimation of SGM is effectively solved by the determinant-maximization programming. In particular, a lasso-type estimation is available by adding constraints. SGM is compared with graphical Gaussian models and mixture models.

Keywords: determinant maximization, Fourier series, graphical model, lasso, optimal transport, structural gradient model.

1 Introduction

In recent years, it becomes more important to treat high-dimensional quantitative data especially in biostatistics and spatial-temporal statistics. The graphical Gaussian model is one of the most important model. However, the Gaussian model represents only the second-order interaction without heteroscedasticity. In this paper, we introduce the structural gradient model (SGM) that represents both higher-order and heteroscedastic interactions of data. The model is defined by a transport map that pushes the target probability density forward to the uniform density. The data structure is described by the parameters in the transport map. This model is a practical specification of the gradient model defined in Sei [2006].

We consider probability density functions on the hypercube $[0,1]^m$ written as

$$p(x) \;=\; \det(D^2\psi(x)), \quad x \in [0,1]^m, \tag{1}$$

where ψ is a convex function and $D^2\psi(x)$ is the Hessian matrix of ψ at x. The function p is a probability density function if the gradient map $D\psi$ is a bijection on

$[0,1]^m$. In fact, by changing the variable from x to $y = D\psi(x)$, we obtain

$$\int_{[0,1]^m} \det(D^2\psi(x))\mathrm{d}x \;=\; \int_{[0,1]^m} \det\left(\frac{\partial y}{\partial x}\right)\mathrm{d}x \;=\; \int_{[0,1]^m} \mathrm{d}y \;=\; 1.$$

It is known that *any* probability density function on $[0,1]^m$ (actually on $\mathbb{R}^m$) is written as (1). This fact is deeply connected to the theory of optimal transport (see e.g. Villani [2003]). The bijective gradient map $D\psi$, called the Brenier map, is the optimal-transport plan from the density (1) to the uniform density. In this paper, we call ψ *the potential function*. Furthermore, as explained in Section 2, most density functions on $[0,1]^m$ are characterized by the Fourier series of ψ. When ψ is represented by the Fourier series, we will call the model (1) *the structural gradient model* and refer to it as SGM. Unknown parameters are the Fourier coefficients of the potential function ψ. SGM can describe not only two-dimensional correlations but also the three-dimensional interactions and heteroscedastic structures, unlike the graphical Gaussian model. We examine this flexibility by simulation and real-data analysis.

The maximum likelihood estimation of SGM is reduced to a determinant maximization problem with a robust convex feasible region. In practice, this region is not directly used because it is described by infinitely many constraints. We give two different approaches to overcome this difficulty. First we give a sequence converging to the feasible region from the inner side. Secondly we give a L^1-conservative region. These approaches enable us to calculate the estimator by the determinant maximization algorithm (Vandenberghe et al. [1998]). As a by-product of the second approach we have a lasso-type estimator for SGM. A related estimator is the lasso-type estimator for graphical Gaussian models (Meinshausen and Bühlmann [2006], Yuan and Lin [2007], Bunea et al. [2007], Banerjee et al. [2008]).

We consider only the case in which the sample space is a hypercube. However, this is not a strong assumption because we can transform any real-valued data into $[0,1]$-valued data by a fixed sigmoid function. Unlike the copula models (Nelsen [2006]), the marginal density of SGM does not need to be uniform. Our model can still adjust the marginal densities after the sigmoid transform. Another approach to deal with unbounded data is given by the author's past papers (Sei [2006], Sei [2007]), where optimal transport between the standard normal density and other densities is considered. In this paper, we use the uniform density instead of the normal density because the former is analytically simpler than the latter.

This paper is organized as follows. In Section 2, we define SGM and give various examples of it. In Section 3, we investigate the maximum likelihood estimation and propose a lasso-type estimator. In Section 4, we compare SGM with graphical

Gaussian models and mixture models by numerical experiments. Finally we have some discussions in Section 5. All mathematical proofs are given in Appendix.

2 The structural gradient model (SGM)

In this section, we first give the formal definition and some theoretical properties of SGM. Then various examples follow.

2.1 Definition and basic facts

Let m be a fixed positive integer. Denote the gradient operator on $[0,1]^m$ by $D = (\partial/\partial x_i)_{i=1}^m$ and the Hessian operator by $D^2 = (\partial^2/\partial x_i \partial x_j)_{i,j=1}^m$. The determinant of a matrix A is denoted by $\det A$. The notation $A \succ B$ (resp. $A \succeq B$) means that $A - B$ is positive definite (resp. positive semi-definite). Let $\mathbb{Z}_{\geq 0}$ be the set of all non-negative integers.

Definition 1 (SGM)**.** Let $\mathcal{U}$ be a finite subset of $\mathbb{Z}_{\geq 0}^m$. We define *the structural gradient model* (abbreviated as *SGM*) by Eq. (1) with the potential function

$$\psi(x|\theta) \;=\; \frac{1}{2}x^\top x - \sum_{u \in \mathcal{U}} \frac{\theta_u}{\pi^2} \prod_{j=1}^m \cos(\pi u_j x_j), \tag{2}$$

where $x = (x_j) \in [0,1]^m$ and $\theta = (\theta_u) \in \mathbb{R}^{\mathcal{U}}$. We call $\mathcal{U}$ *the frequency set.* The parameter space of SGM is

$$\Theta \;=\; \left\{\theta \in \mathbb{R}^{\mathcal{U}} \mid D^2\psi(x|\theta) \succeq 0, \quad \forall x \in [0,1]^m\right\}. \tag{3}$$

A vector $\theta \in \mathbb{R}^{\mathcal{U}}$ is called *feasible* if $\theta \in \Theta$. We also call Θ *the feasible region.* □

The following lemma is fundamental.

Lemma 1. If θ is feasible, then $p(x|\theta)$ is a probability density function on $[0,1]^m$.

SGM has sufficient flexibility for multivariate modeling because the following theorem by Caffarelli [2000] holds. To state the theorem, we prepare some notations. Denote the $2m$ faces of $[0,1]^m$ by $F_j^b = \{x \in [0,1]^m \mid x_j = b\}$ for $j \in \{1,\ldots,m\}$ and $b \in \{0,1\}$. For a smooth function ψ on $[0,1]^m$, we consider a Neumann condition

$$\frac{\partial \psi(x)}{\partial x_j} \;=\; b \;\text{ for any }\; x \in F_j^b. \tag{4}$$

It is easily confirmed that the function ψ defined by (2) satisfies the Neumann condition (4). Conversely, if $\psi(x)$ satisfies the Neumann condition (4), then it is expanded by an infinite cosine series in L^2 sense (see e.g. page 300 of Zygmund [2002]). In other words, the function (2) approximates any potential function satisfying (4) if we make the frequency set $\mathcal{U}$ large. Now we describe the Caffarelli's theorem. Here we put a slightly stronger assumption than his.

Theorem 1 (Theorem 5 of Caffarelli [2000])**.** Let $p(x)$ be a strictly positive and continuously differentiable function on $[0,1]^m$. Assume that $p(x)$ satisfies a Neumann condition $\partial p(x)/\partial x_j = 0$ for any $x \in F_j^b$. Then there exists a twice-differentiable convex function $\psi(x)$ such that (1) and (4) hold.

Since the conditions for $p(x)$ in the above theorem are differentiability and a boundary condition, we can construct sufficiently many statistical models by SGM. In the following subsection, we enumerate various examples of SGM. In Section 5, we discuss removal of the boundary condition for $p(x)$ by removing the twice-differentiability condition for $\psi(x)$.

For the one-dimensional case ($m = 1$), SGM becomes a mixture model as will be explained in the following subsection. For the multi-dimensional case ($m > 1$), SGM is not a mixture model except for essentially one-dimensional case.

Lemma 2. SGM is not a mixture model unless there exists some $i \in \{1, \ldots, m\}$ such that $\mathcal{U} \subset \mathbb{Z}_i$, where $\mathbb{Z}_i = \{u \in \mathbb{Z}_{\geq 0}^m \mid u_j = 0 \ \forall j \neq i\}$.

We use the following mixture model as a reference.

Definition 2 (MixM)**.** Let $\mathcal{U}$ be a finite subset of $\mathbb{Z}_{\geq 0}^m$. We define a structural mixture model (referred to as *MixM*) by

$$\tilde{p}(x|\theta) \ = \ 1 + \sum_{u \in \mathcal{U}} \theta_u \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j), \tag{5}$$

where $x = (x_j) \in [0,1]^m$, $\theta = (\theta_u) \in \mathbb{R}^{\mathcal{U}}$ and $\|u\|^2 = \sum_{j=1}^m u_j^2$. The feasible region is $\tilde{\Theta} := \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \tilde{p}(x|\theta) \geq 0 \ \forall x \in [0,1]^m\}$. □

In the following lemma, we prove that SGM and MixM have a common score function at the origin $\theta = 0$ of the parameter space. The Fisher information matrix at the origin is also calculated.

Lemma 3. The score vector at the origin $\theta = 0$ of both SGM and MixM is equal to $(\|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j))_{u \in \mathcal{U}}$. The Fisher information matrix $(J_{uv})_{u,v \in \mathcal{U}}$ at the origin $\theta = 0$ of both the models is given by

$$J_{uv} \ = \ \frac{\|u\|^4 1_{\{u=v\}}}{2^{|\sigma(u)|}},$$

where $\sigma(u) = \{j \in \{1, \ldots, m\} \mid u_j > 0\}$. In particular, J_{uv} is diagonal.

The Fisher information matrix J at the origin is useful if we deal with the testing of hypothesis $\theta = 0$. Under this hypothesis, the maximum likelihood estimator $\hat{\theta}$ is approximated by a Gaussian random vector with mean 0 and variance $(nJ)^{-1}$. In Section 4, we will use the scaled maximum likelihood estimator $J^{1/2}\hat{\theta}$ to detect which components of $\hat{\theta}$ are significant. A method of computation for the maximum likelihood estimator is given in Section 3. In general, it seems difficult to calculate the Fisher information at the other points $\theta \neq 0$. Exceptional cases will be stated in the following examples.

2.2 Examples

We enumerate examples of SGM. We mainly compare SGM with MixM defined in Definition 2. For SGM, the following sufficient condition for feasibility of θ is useful to deal with the examples. In Theorem 3, we will show that θ is feasible if

$$1 - \sum_{u \in \mathcal{U}} |\theta_u| u_j^2 \ \geq \ 0 \tag{6}$$

for any $j = 1, \ldots, m$. This condition is also necessary if, for example, $\mathcal{U}$ is a one-element set (see Theorem 3 for details).

Example 1 (1-dimensional case)**.** If $m = 1$, then the probability density of SGM is given by the Fourier series

$$p(x_1|\theta) \ = \ 1 + \sum_{u \in \mathcal{U}} \theta_u u^2 \cos(\pi u x_1).$$

This coincides with MixM (Definition 2). The model is considered as a particular case of the circular model proposed by Fernández-Durán [2004]. If $\mathcal{U} = \{u\}$ with some $u \in \mathbb{Z}_{>0}$, then the Fisher information $J_{uu}(\theta)$ is explicitly expressed for any feasible $\theta = \theta_u$. In fact,

$$J_{uu}(\theta) \ = \ \frac{1 - \sqrt{1 - \theta^2 u^4}}{\theta^2 \sqrt{1 - \theta^2 u^4}}. \tag{7}$$

The proof is given in Appendix. □

Example 2 (Independence)**.** Let $m = 2$ and

$$\mathcal{U} = \{(u_1, 0) \mid u_1 \in \mathcal{U}_1\} \cup \{(0, u_2) \mid u_2 \in \mathcal{U}_2\},$$

where $\mathcal{U}_i$ $(i = 1, 2)$ is a finite subset of $\mathbb{Z}_{\geq 0}$. Then SGM becomes an independent model

$$p(x_1, x_2|\theta) \ = \ \left(1 + \sum_{u_1 \in \mathcal{U}_1} \theta_{(u_1,0)} u_1^2 \cos(\pi u_1 x_1)\right) \left(1 + \sum_{u_2 \in \mathcal{U}_2} \theta_{(0,u_2)} u_2^2 \cos(\pi u_2 x_2)\right).$$

Independence of higher-dimensional variables is similarly described. On the other hand, if we consider MixM

$$\tilde{p}(x_1, x_2|\theta) \;=\; 1 + \sum_{u_1 \in \mathcal{U}_1} \theta_{(u_1,0)} u_1^2 \cos(\pi u_1 x_1) + \sum_{u_2 \in \mathcal{U}_2} \theta_{(0,u_2)} u_2^2 \cos(\pi u_2 x_2),$$

then x_1 and x_2 are not independent except for trivial cases. □

Example 3 (Correlation)**.** Let $m = 2$ and $\mathcal{U} = \{(1,1)\}$. Then a pair (X_1, X_2) drawn from $p(x_1, x_2|\theta)$ has positive or negative correlation if $\theta_{(1,1)} > 0$ or < 0, respectively (see Figure 1). We confirm this observation by explicit calculation. We denote $\theta = \theta_{(1,1)}$, $c(\xi) = \cos(\pi\xi)$ and $s(\xi) = \sin(\pi\xi)$ for simplicity. The density is

$$\begin{aligned} p(x_1, x_2|\theta) &= \det \begin{pmatrix} 1 + \theta c(x_1)c(x_2) & -\theta s(x_1)s(x_2) \\ -\theta s(x_1)s(x_2) & 1 + \theta c(x_1)c(x_2) \end{pmatrix} \\ &= 1 + 2\theta c(x_1)c(x_2) + \frac{\theta^2}{2}(c(2x_1) + c(2x_2)). \end{aligned}$$

By the condition (6), the feasible region for θ is $[-1, 1]$. The marginal density of X_i $(i = 1, 2)$ is exactly calculated as

$$p(x_i|\theta) \;=\; 1 + \frac{\theta^2}{2} c(2x_i).$$

The mean and variance of X_i $(i = 1, 2)$ are $1/2$ and $(1/12) + \theta^2/(4\pi^2)$, respectively. The correlation is

$$\frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{V}[X_1]\mathrm{V}[X_2]}} \;=\; \frac{8\theta/\pi^4}{(1/12) + \theta^2/(4\pi^2)} \;=\; \frac{96\theta/\pi^4}{1 + 3\theta^2/\pi^2}.$$

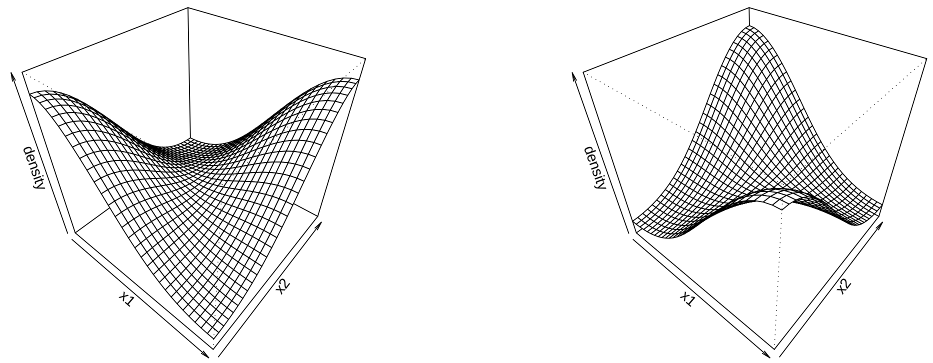
The maximum correlation over $\theta \in [-1, 1]$ is $96/(\pi^4 + 3\pi^2) \simeq 0.7558$ at $\theta = 1$. In contrast, if we consider MixM

$$\tilde{p}(x_1, x_2|\theta) \;=\; 1 + 2\theta c(x_1)c(x_2),$$

then the feasible region (i.e. the set of θ that assures $\tilde{p}(x_1, x_2|\theta) \geq 0$) is $|\theta| \leq 1/2$. The correlation is $96\theta/\pi^4$ and its maximum value is $48/\pi^4 \simeq 0.4928$ at $\theta = 1/2$. Thus SGM can describe a distribution with higher correlation than MixM. The Fisher information $J_{uu}(\theta)$ is explicitly expressed for any feasible θ, where $u = (1, 1)$. The formula is

$$J_{uu}(\theta) \;=\; \frac{2(1 - \sqrt{1 - \theta^2})}{\theta^2\sqrt{1 - \theta^2}}. \tag{8}$$

The proof is given in Appendix. □



(a) $\theta = 0.5$.

(b) $\theta = -0.5$.

Figure 1: The probability density $p(x|\theta)$ for $\mathcal{U} = \{(1,1)\}$ and $\theta = \theta_{(1,1)} = \pm 0.5$. The correlation coefficient is about ± 0.458 for $\theta = \pm 0.5$, respectively.

Example 4 (Heteroscedasticity)**.** Let $m = 2$ and $\mathcal{U} = \{(1,2)\}$. Then a pair (X_1, X_2) drawn from $p(x_1, x_2|\theta)$ has the following property: the conditional mean of X_2 given X_1 does not depend on X_1 but the conditional variance does (see Figure 2). In other words, X_2 has heteroscedasticity in terms of regression analysis. We confirm this fact. The joint density is

$$\begin{aligned} p(x_1, x_2|\theta) &= \det \begin{pmatrix} 1 + \theta c(x_1)c(2x_2) & -2\theta s(x_1)s(2x_2) \\ -2\theta s(x_1)s(2x_2) & 1 + 4\theta c(x_1)c(2x_2) \end{pmatrix} \\ &= 1 + 5\theta c(x_1)c(2x_2) + 2\theta^2 c(2x_1) + 2\theta^2 c(4x_2) \end{aligned}$$

where we put $c(\xi) = \cos(\pi\xi)$, $s(\xi) = \sin(\pi\xi)$, and $\theta = \theta_{(1,2)}$. The marginal density of X_1 is $p(x_1) = 1 + 2\theta^2 c(2x_1)$. The conditional density of X_2 given X_1 is

$$p(x_2|x_1, \theta) = 1 + \frac{5\theta c(x_1)c(2x_2) + 2\theta^2 c(4x_2)}{1 + 2\theta^2 c(2x_1)}$$

The conditional mean of X_2 given X_1 is exactly $1/2$, and therefore the correlation between X_1 and X_2 is zero. However, the conditional variance of X_2 given X_1 is not constant:

$$\int_0^1 (x_2 - 1/2)^2 p(x_2|x_1, \theta)\mathrm{d}x_2 = \frac{1}{12} + \frac{10\theta c(x_1) + \theta^2}{4\pi^2\{1 + 2\theta^2 c(2x_1)\}}.$$

In order to measure the dependency of X_1, let us consider the quantity

$$\begin{aligned} \beta_{122}(\theta) &= \frac{\mathrm{E}[(X_1 - 1/2)(X_2 - 1/2)^2]}{\{\mathrm{V}[X_1]\}^{1/2}\mathrm{V}[X_2]}. \\ &= \frac{-5\theta/\pi^4}{\{(1/12) + \theta^2/\pi^2\}^{1/2}\{(1/12) + \theta^2/(4\pi^2)\}} \end{aligned}$$

The maximum value of $\beta_{122}(\theta)$ over the feasible region $\theta \in [-1/4, 1/4]$ is $\beta_{122}(-1/4) \simeq 0.5047$. In contrast, for MixM $\tilde{p}(x_1, x_2|\theta) = 1 + 5\theta c(x_1)c(2x_2)$, the maximum of $\beta_{122}(\theta)$ over the feasible region $\theta \in [-1/5, 1/5]$ is $\tilde{\beta}_{122}(-1/5) \simeq 0.4267$. Thus SGM can describe more heteroscedastic distributions than MixM. The heteroscedasticity appears in regression analysis, where explanatory and response variables are *a priori* selected. Remark that our model does not need a priori selection of variables. □

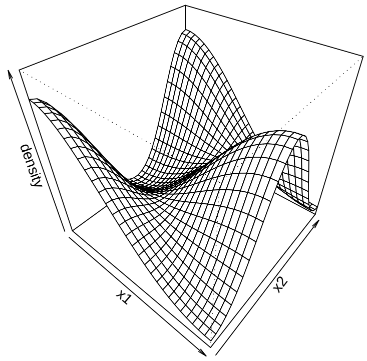


Figure 2: The probability density for $\mathcal{U} = \{(1, 2)\}$ and $\theta = 0.2$. The conditional density $p(x_2|x_1)$ is unimodal if x_1 is close to 1, and bimodal if x_1 is close to 0.

Example 5 (three-dimensional interaction)**.** Let $m = 3$ and $\mathcal{U} = \{(1, 1, 1)\}$. Then the triplet (X_1, X_2, X_3) has the three-dimensional interaction although the marginal two-dimensional correlation for any pair vanishes. We confirm this. The joint probability density is

$$\begin{aligned} p(x_1, x_2, x_3|\theta) &= 1 + 3\theta c_1c_2c_3 + 3\theta^2 c_1^2c_2^2c_3^2 + \theta^3 c_1^3c_2^3c_3^3 \\ &\quad -2\theta^3 c_1s_1^2c_2s_2^2c_3s_3^2 - (1 + \theta c_1c_2c_3)\theta^2(c_1^2s_2^2s_3^2 + s_1^2c_2^2s_3^2 + s_1^2s_2^2c_3^2), \end{aligned}$$

where $c_i = \cos(\pi x_i)$ and $s_i = \sin(\pi x_i)$ for $i = 1, 2, 3$. The density is symmetric with respect to permutation of axes. The feasible region is $|\theta| \leq 1$ by (6). The 2-dimensional and 1-dimensional marginal densities are $p(x_1, x_2|\theta) = 1+\theta^2(4c_1^2c_2^2-1)/2$ and $p(x_1|\theta) = 1+\theta^2(2c_1^2-1)/2$, respectively. In particular, the mean of X_i is 1/2 and the correlation of X_i and X_j $(i \neq j)$ is zero. However, there exists three-dimensional interaction between (X_1, X_2, X_3). We calculate

$$\beta_{123}(\theta) := \frac{\mathrm{E}[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)(X_3 - \mathrm{E}X_3)]}{\sqrt{\mathrm{V}[X_1]\mathrm{V}[X_2]\mathrm{V}[X_3]}}.$$

The result is

$$\beta_{123}(\theta) \;=\; \frac{-24\theta/\pi^6 - 1944\theta^3/729\pi^6}{(1/12+\theta^2/(4\pi^2))^{3/2}}.$$

The maximum value of $\beta_{123}(\theta)$ over the feasible region $|\theta| \le 1$ is $\beta_{123}(-1) \simeq 0.7743$. In contrast, for MixM $\tilde{p}(x_1,x_2,x_3|\theta) = 1+3\theta c_1c_2c_3$, we have $\beta_{123}(\theta) = -288\sqrt{12}\theta/\pi^6$. Its maximum value over the feasible region $|\theta| \le 1/3$ is about 0.3459 at $\theta = -1/3$. □

Example 6 (Approximately conditional independence)**.** Let $m = 3$ and (X_1, X_2, X_3) be drawn from a probability density $p(x_1, x_2, x_3)$. In general, conditional independence of X_1 and X_2 given X_3 is described by $p(x_1, x_2, x_3) = p(x_3)p(x_1|x_3)p(x_2|x_3)$ or, equivalently, the conditional mutual information

$$I_{12|3} \;=\; \int p(x_1,x_2,x_3)\log\frac{p(x_1,x_2|x_3)}{p(x_1|x_3)p(x_2|x_3)}\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3$$

vanishes. A log-linear model $\exp(f(x_1, x_3) + g(x_2, x_3))$ satisfies this condition. Although SGM does not represent any conditional-independence model, we can construct an approximately conditional-independence model. Let $m = 3$ and $\mathcal{U} = \{(1, 0, 1), (0, 1, 1)\}$. Then, by putting $c_i = \cos(\pi x_i)$, $s_i = \sin(\pi x_i)$, $\theta = \theta_{(1,0,1)}$ and $\phi = \theta_{(0,1,1)}$, we have

$$\begin{aligned}
&p(x_1,x_2,x_3|\theta,\phi)\\
&\quad= \det\begin{pmatrix} 1+\theta c_1c_3 & 0 & -\theta s_1s_3 \\ 0 & 1+\phi c_2c_3 & -\phi s_2s_3 \\ -\theta s_1s_3 & -\phi s_2s_3 & 1+\theta c_1c_3+\phi c_2c_3 \end{pmatrix}\\
&\quad= 1+2\theta c_1c_3+2\phi c_2c_3+3\theta\phi c_1c_2c_3^2+\theta^2(c_1^2c_3^2-s_1^2s_3^2)+\phi^2(c_2^2c_3^2-s_2^2s_3^2)\\
&\qquad +\theta^2\phi(c_1^2c_3^2-s_1^2s_3^2)c_2c_3+\theta\phi^2(c_2^2c_3^2-s_2^2s_3^2)c_1c_3
\end{aligned}$$

Now assume that $\epsilon := \max(|\theta|, |\phi|)$ is close to zero. Then the conditional mutual information is, after tedious calculations,

$$I_{12|3} \;=\; \frac{3}{16}\theta^2\phi^2 + \mathrm{O}(\epsilon^5).$$

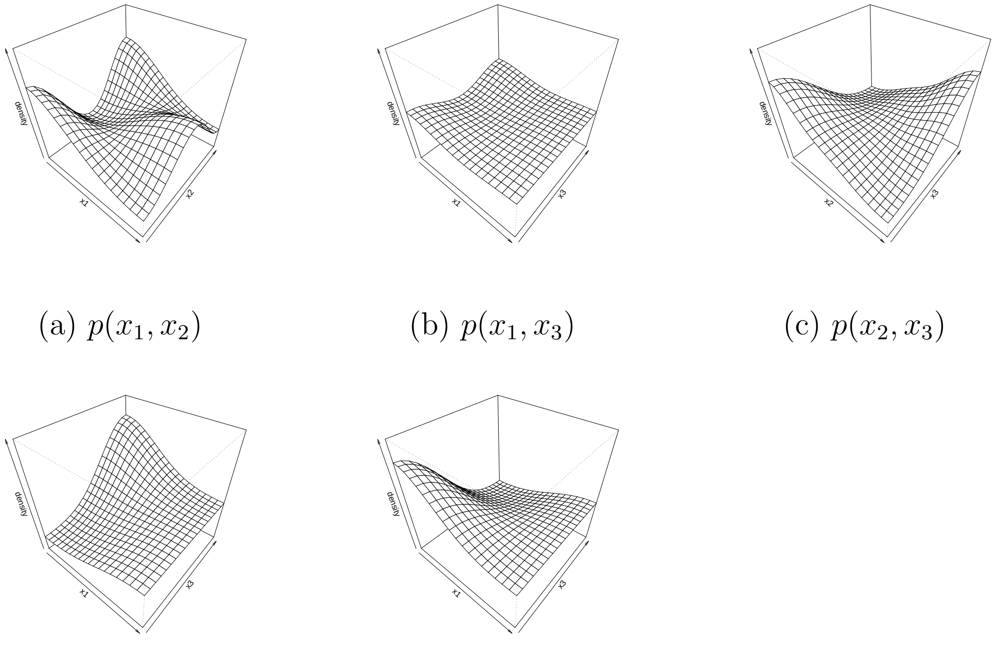
On the other hand, MixM $\tilde{p}(x_1, x_2, x_3|\theta, \phi) = 1+2\theta c_1c_3+2\phi c_2c_3$ has the conditional mutual information $I_{12|3} = (3/4)\theta^2\phi^2 + \mathrm{O}(\epsilon^5)$. The leading term is 4 times larger than that of SGM. □

We summarize the above examples in Table 1.

Example 7. We can construct more complicated densities by combining the preceding ones. For example, let $m = 3$ and $\mathcal{U} = \{(1, 2, 0), (0, 1, 1), (1, 1, 1)\}$. Let the corresponding parameter vector be $\theta = (0.1, 0.3, 0.2)$. The vector θ is feasible since (6) is satisfied. The marginal and conditional 2-dimensional densities are illustrated in Figure 3. □

Table 1: Summary of the examples. For each example, the characteristics of SGM and MixM are compared.

#	Model name	m	Characteristic	SGM	MixM
1	1-dim.	1	(SGM=MixM)	—	—
2	independence	2	'is independent'	TRUE	FALSE
3	correlation	2	maximum correlation	0.7558	0.4928
4	heteroscedasticity	2	maximum β_{122}	0.5047	0.4267
5	3-dim. interaction	3	maximum β_{123}	0.7743	0.3459
6	conditional independence	3	leading coefficient of $I_{12\|3}$	3/16	3/4



(a) $p(x_1, x_2)$ (b) $p(x_1, x_3)$ (c) $p(x_2, x_3)$

(d) $p(x_1, x_3 | x_2 = 3/4)$ (e) $p(x_1, x_3 | x_2 = 1/4)$

Figure 3: The marginal and conditional densities for $\mathcal{U} = \{(1, 2, 0), (0, 1, 1), (1, 1, 1)\}$. The figures (a), (b) and (c) are the marginal density $p(x_i, x_j)$ for each pair (i, j). The figures (d) and (e) are the conditional density $p(x_1, x_3 | x_2)$ for specific values of x_2.

3 Maximum likelihood estimation of SGM

Let $x(1),\ldots,x(n)$ be independent samples drawn from the true density $p_0(x)$ whose support is $[0,1]^m$. From the definition of SGM, the maximum likelihood estimation of SGM is formulated as a convex optimization program:

$$\begin{aligned} &\text{maximize} && \sum_{t=1}^{n} \log\det\left(I+\sum_{u\in\mathcal{U}}\theta_u H_u(x(t))\right), \\ &\text{subject to} && \theta\in\Theta=\left\{\theta\in\mathbb{R}^{\mathcal{U}} \;\middle|\; I+\sum_{u\in\mathcal{U}}\theta_u H_u(\xi)\succeq 0 \quad \forall\xi\in[0,1]^m\right\}, \end{aligned}$$

where we put $H_u(x)=D^2(-\pi^{-2}\prod_{\rho=1}^{m}\cos(\pi u_\rho x_\rho))$. Recall that D^2 is the Hessian operator and $\mathcal{U}$ is a finite subset of $\mathbb{Z}_{\geq 0}^m$.

It is hard to write down Θ explicitly. The difficulty follows from the statement "for any $\xi\in[0,1]^m$" in the definition of Θ. In general, for a set of feasible regions Θ_α indexed by α, the region $\cap_\alpha\Theta_\alpha$ is called a robust feasible region (see Ben-tal and Nemirovski [1998]).

We consider two approaches to solve this problem. We will first give a sequence Θ_M° of regions converging to Θ°, the interior of Θ, as $M\to\infty$. Hence the maximum likelihood estimator is calculated with arbitrary accuracy in principle. However, Θ_M° has about M^m constraints on θ and therefore it is usually expensive if $m\geq 3$. For the second approach, we give a proper subset Θ^{lit} of Θ, which consists of only m constraints. As a by-product of the second approach, we obtain a lasso-type estimator because Θ^{lit} is compatible with L^1-constraints. We call the maximizer of the log-likelihood over these constrained regions *the constrained maximum likelihood estimator*. The constrained maximum likelihood estimator is calculated via the determinant maximization algorithm (Vandenberghe et al. [1998]).

If $m=1$, the feasible region is the set of Fourier coefficients of non-negative functions. To deal with the feasible region, Fernández-Durán [2004] used Fejér's characterization: the Fourier series of any non-negative function is written as the square of a Fourier series. More specifically, for any $r(x)=\sum_{u=0}^{\infty} r_u\cos(\pi ux)$, its square $r(x)^2$ is of course non-negative and written by a Fourier series. The Fourier coefficients of $r(x)^2$ are written by quadratic polynomials of $(r_u)_{u=0}^{\infty}$. However, it is hard to use this representation for our problem because we assume $\theta_u=0$ for $u\notin\mathcal{U}$ and this restriction is not affine in r_u.

3.1 Inner approximation of feasible region

Let Θ° be the interior of Θ. We give a sequence of tractable sets Θ°_M that converges to Θ° from inside as $M \to \infty$. We first remark the following lemma.

Lemma 4. The set Θ° is equal to $\left\{\theta \in \mathbb{R}^{\mathcal{U}} \mid D^2\psi(x|\theta) \succ 0 \ \ \forall x \in [0,1]^m\right\}$.

We prepare some notations for constructing Θ°_M. We consider the lattice points L_M^m, where $L_M = \{\frac{0}{M}, \frac{1}{M}, \cdots, \frac{M}{M}\}$. Let $\|u\|_\infty = \max_j |u_j|$ and $U_{\max} = \max_{u \in \mathcal{U}} \|u\|_\infty$. Define a linear operator K_M on $\mathbb{R}^{\mathcal{U}}$ by $(K_M\theta)_u = \theta_u / \prod_{j=1}^m (1 - u_j/M)$ for $\theta \in \mathbb{R}^{\mathcal{U}}$. Finally, we define Θ°_M for each $M \geq U_{\max} + 1$ by

$$\Theta^\circ_M \ = \ \left\{\theta \in \mathbb{R}^{\mathcal{U}} \ \middle|\ D^2\psi(\xi|K_M\theta) \succ 0, \ \forall \xi \in L_M^m\right\}.$$

Remark that Θ°_M is written in a finite number of constraints, in contrast to Θ° and Θ. We have the following theorem.

Theorem 2. For any $M \geq U_{\max} + 1$, we have $\Theta^\circ_M \subset \Theta^\circ$ and

$$\Theta^\circ \ = \ \limsup_{M \to \infty} \Theta^\circ_M,$$

where $\limsup_{M\to\infty} \Theta^\circ_M$ is defined by $\cap_{M' \geq 1} \cup_{M \geq M'} \Theta^\circ_M$.

The constrained maximum likelihood estimator of θ over Θ°_M is calculated via the determinant maximization algorithm (Vandenberghe et al. [1998]). Hence, in principle, we can calculate the maximum likelihood estimator with arbitrary accuracy. However, the region Θ°_M consists of $|L_M^m| = (M+1)^m$ constraints. This number is usually expensive if $m \geq 3$. In the following subsection, we give a proper subset of Θ which consists of only m constraints.

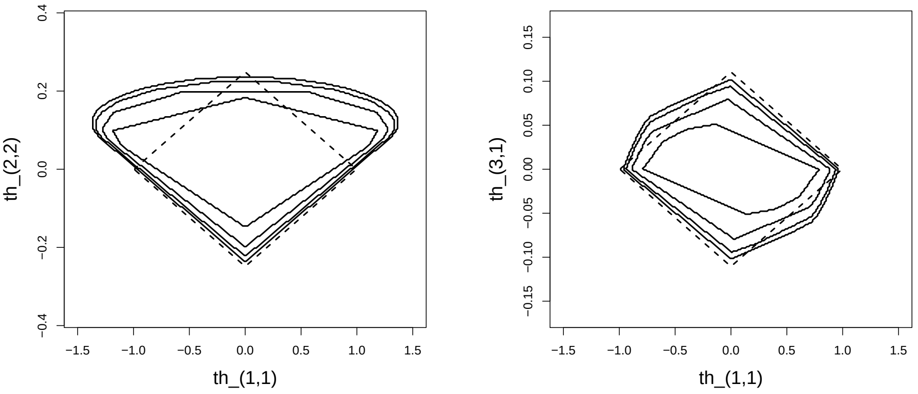
Example 8. Let $m = 2$ and $\mathcal{U} = \{(1,1),(2,2)\}$. The approximated regions Θ°_M ($M = 5, 10, 20, 40$) are illustrated in Figure 4 (a). For this case, we can give a precise expression of Θ. The two eigenvalues of the Hessian matrix $D^2\psi(x|\theta)$ are given by

$$\lambda_\pm \ = \ 1 + \theta_{(1,1)}\cos(\pi(x_1 \pm x_2)) + 4\theta_{(2,2)}\cos(2\pi(x_1 \pm x_2)).$$

In the theory of time-series analysis, the function $f(z) := 1 + \sum_k \rho_k \cos(kz)$ of z is the spectral density of a MA(k) process with the autocorrelation coefficients $(\rho_j)_{j=1}^k$. In particular, for MA(2), it is known that $f(z)$ is non-negative for any z if and only if $|\rho_1| + |\rho_2| \leq 1$ or $\rho_1^2 \leq 4\rho_2(1-\rho_2)$ holds (see Box and Jenkins [1976], Section 3.4). Therefore the feasible region for $\mathcal{U} = \{(1,1),(2,2)\}$ is given by

$$|\theta_{(1,1)}| + |4\theta_{(2,2)}| \ \leq \ 1 \quad \text{or} \quad (\theta_{(1,1)})^2 \leq 4(4\theta_{(2,2)})(1 - 4\theta_{(2,2)}).$$

The region Θ°_M shown in Figure 4 (a) is close to this region. We also illustrate the approximated regions for another example $\mathcal{U} = \{(1,1),(3,1)\}$ in Figure 4 (b). □



(a) $\mathcal{U} = \{(1,1),(2,2)\}$. (b) $\mathcal{U} = \{(1,1),(3,1)\}$.

Figure 4: The approximated region Θ_M° (solid line; $M = 5, 10, 20, 40$ from inner side) and the little parameter space $\Theta^{\rm lit}$ (dashed line) defined in Subsection 3.2.

We remark that the feasible region for MixM (Definition 2) is approximated from the inner side by

$$\tilde{\Theta}_M^\circ \ := \ \left\{\theta \in \mathbb{R}^{\mathcal{U}} \mid \tilde{p}(\xi|K_M\theta) > 0, \ \text{for } \xi \in L_M^m\right\}$$

The proof is similar to that of Theorem 2 and omitted here.

3.2 A conservative region and Lasso-type estimation

We give a sufficient condition such that $\theta \in \Theta$. Define a set $\Theta^{\rm lit}$ by

$$\Theta^{\rm lit} \ = \ \left\{\theta \in \mathbb{R}^{\mathcal{U}} \ \middle| \ 1 - \sum_{u \in \mathcal{U}} |\theta_u| u_j^2 \geq 0 \quad (\forall j = 1, \ldots, m)\right\}.$$

We call $\Theta^{\rm lit}$ *the little parameter space.* It is an intersection of m constraints. In the following theorem, we show that the little parameter space $\Theta^{\rm lit}$ is a subset of the feasible region Θ. In other words, $\Theta^{\rm lit}$ is more conservative than Θ in the sense of robustness. We say that a subset $\mathcal{V}$ of $\mathcal{U}$ is linearly independent modulo 2 if a linear map $\ell : \{0,1\}^{\mathcal{V}} \mapsto \{0,1\}^m$ defined by $\ell(\epsilon) = \sum_{u \in \mathcal{V}} \epsilon_u u \pmod 2$ has the kernel $\{0\}$. For each $\mathcal{V} \subset \mathcal{U}$, the set of vectors that have only $\mathcal{V}$-components is denoted by $\mathbb{R}_{\mathcal{V}} = \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \theta_u = 0 \ \forall u \notin \mathcal{V}\}$.

Theorem 3. For any $\mathcal{U}$, $\Theta^{\rm lit} \subset \Theta$. Furthermore, if a subset $\mathcal{V}$ of $\mathcal{U}$ is linearly independent modulo 2, then we have $\Theta^{\rm lit} \cap \mathbb{R}_{\mathcal{V}} = \Theta \cap \mathbb{R}_{\mathcal{V}}$. In particular, if $\mathcal{U}$ itself is linearly independent modulo 2, then $\Theta^{\rm lit} = \Theta$.

By letting $\mathcal{V}$ be a one-element set $\{u\}$, we have the relation $\Theta^{\mathrm{lit}} \cap \mathbb{R}_{\{u\}} = \Theta \cap \mathbb{R}_{\{u\}}$. This shows that Θ^{lit} contains at leat $2|\mathcal{U}|$ boundary points of Θ. The little parameter space for $\mathcal{U} = \{(1,1),(2,2)\}$ and $\mathcal{U} = \{(1,1),(3,1)\}$ is indicated in Figure 4 (a) and (b), respectively.

The constrained maximum likelihood estimator of θ over Θ^{lit} is computed via the determinant maximization algorithm by introducing non-negative slack variables θ_u^+ and θ_u^- such that $\theta_u = \theta_u^+ - \theta_u^-$ and $|\theta_u| = \theta_u^+ + \theta_u^-$. The estimator is usually sparse. This sparsity is closely related to the lasso estimator Tibshirani [1996] in that the regression method is executed with L^1-constraints. Our little parameter space Θ^{lit} is also represented by L^1-constraints. Hence we call the constrained maximum likelihood estimator of θ over Θ^{lit} *the lasso-type estimator for SGM*. Furthermore, we will use an indexed set $\Theta_\tau^{\mathrm{lit}}$ with a tuning parameter $\tau \in [0,1]$ by

$$\Theta_\tau^{\mathrm{lit}} = \left\{ \theta \in \mathbb{R}^{\mathcal{U}} \;\middle|\; \tau - \sum_{u \in \mathcal{U}} |\theta_u| u_j^2 \geq 0 \quad (\forall j = 1, \ldots, m) \right\}.$$

In particular, $\Theta_0^{\mathrm{lit}} = \{0\}$ and $\Theta_1^{\mathrm{lit}} = \Theta^{\mathrm{lit}}$. The tuning parameter τ can be selected by cross validation.

We remark that the feasible region for MixM (Definition 2) has the following conservative region

$$\tilde{\Theta}^{\mathrm{lit}} := \left\{ \theta \in \mathbb{R}^{\mathcal{U}} \;\middle|\; 1 - \sum_{u \in \mathcal{U}} |\theta_u| \|u\|^2 \geq 0 \right\}.$$

Furthermore, if a subset $\mathcal{V}$ of $\mathcal{U}$ is linearly independent modulo 2, then we have $\tilde{\Theta}^{\mathrm{lit}} \cap \mathbb{R}_{\mathcal{V}} = \tilde{\Theta} \cap \mathbb{R}_{\mathcal{V}}$. The proof is similar to that of Theorem 3 and is omitted here.

Recently, lasso-type estimators for graphical Gaussian models are proposed by several authors: Yuan and Lin [2007], Banerjee et al. [2008] and Friedmann et al. [2008]. On the other hand, a sparse density estimation (SPADES) for mixture models is considered in Bunea et al. [2007]. Our MixM is considered as a version of SPADES although the estimation procedure is different. In Section 4, we compare SGM with MixM and the graphical Gaussian model by numerical examples.

4 Numerical examples

We give numerical examples on simulated and real datasets. We calculate the constrained maximum likelihood estimator and study its predictive performance. We compare SGM with the graphical Gaussian model (with lasso) and MixM (Definition 2).

We describe some notations and assumptions. We use the following frequency set for SGM throughout this section:

$$\mathcal{U} \;=\; \left\{u \in \mathbb{Z}_{\geq 0}^m \mid \|u\|_\infty \leq 2,\ \|u\|_1 \leq 3\right\}, \tag{9}$$

where $\|u\|_\infty = \max_j |u_j|$ and $\|u\|_1 = \sum_j |u_j|$. The elements of $\mathcal{U}$ are given by $(1,0,\ldots,0)$, $(2,0,\ldots,0)$, $(1,1,0,\ldots,0)$, $(2,1,0,\ldots,0)$, $(1,1,1,0,\ldots,0)$ and their permutations of the components. The cardinality of $\mathcal{U}$ is $m(m+1)(m+5)/6$. Let $\hat{\theta}_M^\circ = (\hat{\theta}_{M,u}^\circ)_{u\in\mathcal{U}}$ and $\hat{\theta}_\tau^{\mathrm{lit}} = (\hat{\theta}_{\tau,u}^{\mathrm{lit}})_{u\in\mathcal{U}}$ denote the constrained maximum likelihood estimators of θ over the regions Θ_M° and $\Theta_\tau^{\mathrm{lit}}$, respectively (see Section 3 for the definition of Θ_M° and $\Theta_\tau^{\mathrm{lit}}$). We call $\hat{\theta}_\tau^{\mathrm{lit}}$ the lasso-type estimator of SGM. The same notations on the estimators are used also for MixM.

The graphical Gaussian lasso estimator $\hat{C} = \hat{C}(\tau)$ of the concentration matrix (Yuan and Lin [2007]) is formulated as follows

$$\text{min.} \quad \{\log\det(C) + \mathrm{tr}(\hat{\Sigma}C)\} \quad \text{s.t.} \quad \sum_{i<j} |C_{ij}| \leq \tau \sum_{i<j} |(\hat{\Sigma}^{-1})_{ij}|,$$

where $\hat{\Sigma}$ is the sample correlation and the tuning parameter τ ranges over $[0,1]$. If $\tau = 1$, the graphical Gaussian lasso estimator coincides with the maximum likelihood estimator (this is not the case for the lasso-type estimators of SGM and MixM). The partial correlation coefficient of x_i and x_j is estimated by $\hat{\rho}_{ij} = -\hat{C}_{ij}/\sqrt{\hat{C}_{ii}\hat{C}_{jj}}$.

For given raw data $(D_{ti})_{1\leq t\leq n, 1\leq i\leq m}$, we preprocess it before estimation. For Gaussian models, we use the data $\tilde{D}_{ti}$ scaled by the standard way:

$$\tilde{D}_{ti} \;=\; \frac{D_{ti} - \bar{D}_{\cdot i}}{\mathrm{sd}(D_{\cdot i})}, \quad \bar{D}_{\cdot i} \;=\; \frac{1}{n}\sum_{t=1}^n D_{ti}, \quad \mathrm{sd}(D_{\cdot i}) \;=\; \sqrt{\frac{1}{n}\sum_{t=1}^n (D_{ti} - \bar{D}_{\cdot i})^2}.$$

For SGM and MixM, the data is further transformed into $X_{ti} = \Phi(\tilde{D}_{ti})$, where Φ is the standard normal cumulative distribution function, in order that X_{ti} ranges over $[0,1]$. By the transform Φ, the standard normal density as the null Gaussian model is transformed into the uniform density as the null SGM and the null MixM.

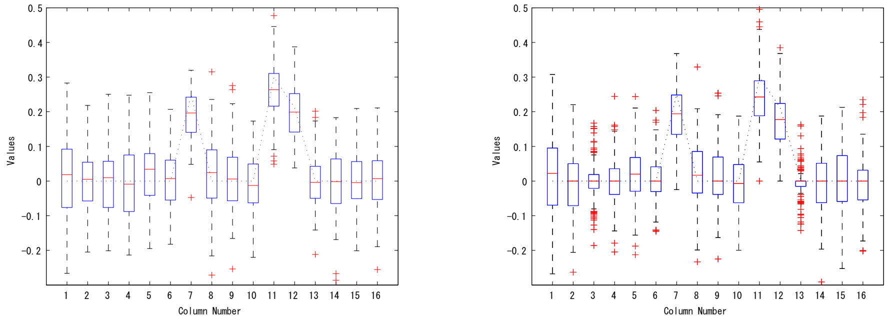
We used the package SDPT3 for solving the determinant-maximization problem on MATLAB (Toh et al. [2006]).

4.1 Simulation

We first confirm that the maximum likelihood estimator is actually computed by the method described in Section 3. Consider Example 7 of Subsection 2.2. The true parameter is $\theta_{(1,2,0)} = 0.1$, $\theta_{(0,1,1)} = 0.3$ and $\theta_{(1,1,1)} = 0.2$ with the true frequency set $\mathcal{U}_0 = \{(1,2,0),(0,1,1),(1,1,1)\}$. The frequency set (9) we use for estimation is written in a matrix form

$$\mathcal{U} \;=\; \begin{pmatrix} 1 & 2 & 0 & 1 & 2 & 0 & 1 & 0 & 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 \end{pmatrix}. \qquad (10)$$

The columns are arranged according to the lexicographic order. A result of estimation is given in Figure 5. The sample size is $n = 100$ and the number of experiments is 100. The samples were generated by the exact method of Sei [2006]. Both estimators actually distribute around the true parameter.



(a) $\sqrt{J_{uu}}\hat{\theta}^{\circ}_{M,u}$ $(M = 5)$.

(b) $\sqrt{J_{uu}}\hat{\theta}^{\mathrm{lit}}_{\tau,u}$ $(\tau = 1)$.

Figure 5: A simulation of estimation of SGM. The box-plot shows each component of the constrained maximum likelihood estimators (a) $\hat{\theta}^{\circ}_{M}$ for $M = 5$ and (b) $\hat{\theta}^{\mathrm{lit}}_{\tau}$ for $\tau = 1$. The values are normalized by the square root $\sqrt{J_{uu}}$ of the Fisher information. The horizontal axis denotes $u \in \mathcal{U}$ arranged according to (10). The dashed line denotes the true parameter. The sample size is $n = 100$ and the number of experiments is 100.

We next compare SGM with MixM and Gaussian models. We consider a five-dimensional example. Let $\phi(x|\mu,\Sigma)$ denote the normal density with mean μ and covariance Σ. Let $m = 5$ and define the true density $p_0(x)$ by

$$p_0(x) \;=\; \phi(x_1|0,1)\phi(x_2|x_1,1)\phi(x_3|0,\sigma_3^2(x_2))\phi(x_4,x_5|0,\Sigma_{45}(x_3)), \qquad (11)$$

where

$$\sigma_3^2(x_2) \;=\; 1 + \tanh(x_2) \quad \text{and} \quad \Sigma_{45}(x_3) \;=\; \begin{pmatrix} 1 & \tanh(x_3) \\ \tanh(x_3) & 1 \end{pmatrix}.$$

By the definition, the set of variables (x_1, x_2) has positive correlation, the variable x_3 has heteroscedasticity against x_2, and the set of variables (x_3, x_4, x_5) has three-dimensional interaction. Remark that the density does not belong to SGM. A numerical result is shown in Table 2. The sample size is $n = 40$ and the number of experiments is 200. All of the three models detected the correlation of the pair (x_1, x_2). However, only SGM effectively detected the heteroscedasticity of (x_2, x_3) and the three-dimensional interaction (x_3, x_4, x_5). The estimator of MixM was too sparse, and did not effectively detect them.

For the same true density, we also computed the predictive performance of the estimators of SGM, MixM and Gaussian. We use the expected predictive log-likelihood as the index of the predictive performance. The arbitrary constant of the log-likelihood is determined in such a way that the log-likelihood of the null model is zero. The sample size is $n = 40$ for observation and 10 for prediction. The number of experiments is 200. The maximum mean predictive log-likelihood of SGM is estimated as $3.37(\pm 0.33)$ at $\tau = 1.0$, where the confidence interval is based on the 95% interval with the normal approximation. For MixM and Gaussian, the maximum value is estimated as $1.99(\pm 0.15)$ at $\tau = 1.0$ and $2.72(\pm 0.26)$ at $\tau = 0.32$, respectively. Hence SGM has better predictive performance than MixM and Gaussian.

4.2 Real dataset

We consider the digoxin clearance data reported in Halkin et al. [1975] (see also Edwards [2000]). The data consists of creatinine clearance (x_1), digoxin clearance (x_2) and urine flow (x_3) of 35 patients. In Table 3, we compare the lasso-type estimators of SGM, MixM and the Gaussian model. The result shows that for the data our SGM gives slightly better predictive performance than MixM and the Gaussian models. As stated in Edwards [2000], partial correlation of (x_1, x_3) is not significant. However, our model suggests a heteroscedastic effect of x_1 (creatinine clearance) against x_3 (urine flow).

5 Discussion

We defined SGM as a set of the potential functions ψ and studied its feasible region to calculate the constrained maximum likelihood estimator. SGM was applied to both simulated and real dataset. We discuss remaining mathematical and practical problems.

We used the finite Fourier expansion to define the potential function ψ as Eq. (2). It is sometimes hard to describe local behavior of the density function if we use this expansion. For such purposes, we can use wavelets instead of the cosine functions as long as the resultant potential function satisfies the Neumann condition (4). For example, assume that we want to describe tail behavior of two-dimensional data

Table 2: Mean value of the lasso-type estimators for the five-dimensional data. The tuning parameter τ is set to 1. The sample size is $n = 40$ and the number of experiments is 200. The confidence interval is based on the 95% interval with the normal approximation. For SGM and MixM, only top ten values of $\sqrt{J_{uu}}\hat{\theta}^{\mathrm{lit}}_{\tau,u}$ are shown. For the Gaussian model, u is the indicator vector of a pair (i, j).

SGM		MixM		Gaussian	
u	$\mathrm{E}[\sqrt{J_{uu}}\hat{\theta}^{\mathrm{lit}}_{\tau,u}]$	u	$\mathrm{E}[\sqrt{J_{uu}}\hat{\theta}^{\mathrm{lit}}_{\tau,u}]$	u	$\mathrm{E}[\hat{\rho_{ij}}(\tau)]$
$(1,1,0,0,0)$	0.510 (±0.013)	$(1,1,0,0,0)$	0.123 (±0.006)	$(1,1,0,0,0)$	0.706 (±0.011)
$(0,0,1,1,1)$	-0.297 (±0.017)	$(0,1,2,0,0)$	-0.031 (±0.005)	$(1,0,0,0,1)$	-0.023 (±0.023)
$(0,1,2,0,0)$	-0.232 (±0.015)	$(0,0,1,1,1)$	-0.007 (±0.003)	$(0,1,1,0,0)$	0.014 (±0.023)
$(0,0,2,0,0)$	-0.106 (±0.014)	$(0,0,2,0,0)$	-0.006 (±0.002)	$(1,0,0,1,0)$	-0.010 (±0.022)
$(2,0,0,0,0)$	-0.095 (±0.011)	$(0,2,0,0,0)$	-0.002 (±0.001)	$(0,1,0,0,1)$	0.008 (±0.024)
$(0,2,0,0,0)$	-0.084 (±0.010)	$(1,0,2,0,0)$	-0.002 (±0.001)	$(0,0,0,1,1)$	-0.007 (±0.028)
$(0,0,0,0,2)$	-0.043 (±0.013)	$(2,0,0,0,0)$	-0.001 (±0.001)	$(0,1,0,1,0)$	0.007 (±0.024)
$(0,0,0,2,0)$	-0.043 (±0.010)	$(0,2,0,1,0)$	-0.000 (±0.001)	$(0,0,1,1,0)$	-0.006 (±0.023)
$(1,0,2,0,0)$	-0.036 (±0.009)	$(0,0,1,0,2)$	-0.000 (±0.001)	$(1,0,1,0,0)$	-0.004 (±0.021)
$(0,0,0,2,1)$	-0.015 (±0.015)	$(0,0,0,0,2)$	-0.000 (±0.001)	$(0,0,1,0,1)$	0.004 (±0.023)

Table 3: A result for the digoxin data. The lasso-type estimators of SGM, MixM and the graphical Gaussian model are shown. Only non-zero values are displayed. For the Gaussian model, the estimated partial correlation of the pairs $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$ is displayed on the row $u = (1, 1, 0), (1, 0, 1), (0, 1, 1)$, respectively. The cross-validated predictive log-likelihood (referred to as CV prediction) is put on the bottom. For each model, the asterisk '$*$' indicates the optimal tuning parameter selected by CV prediction.

		SGM		MixM		Gaussian	
		$\tau = 0.5$	$\tau = 1.0^*$	$\tau = 0.5$	$\tau = 1.0^*$	$\tau = 0.25^*$	$\tau = 1.0$
	$(1,1,0)$	0.351	0.558	0.177	0.354	0.480	0.758
	$(0,1,1)$	0.149	0.301			0.217	0.485
	$(2,0,1)$		-0.166				
	$(1,0,1)$	0.149	0.148				-0.191
u	$(0,0,2)$	-0.070	-0.147				
	$(0,2,0)$		-0.088				
	$(1,0,2)$		0.072				
	$(0,0,1)$	0.073	0.050				
	$(0,1,2)$		-0.039				
CV prediction		11.19	$\underline{14.54}$	6.95	12.26	14.49	-0.92

around $x = (1,1)$. Then we can use a function

$$\psi(x|\theta,a) = (x_1^2 + x_2^2)/2 + \pi^{-2}\theta(2 + \cos(\pi x_1) + \cos(\pi x_2))^a,$$

where $a > 1/2$. A typical shape of the density function $p(x|\theta,a) = \det(D^2\psi(x|\theta,a))$ is given in Figure 6. One can confirm that the gradient map $D\psi$ is continuous on $[0,1]^2$ and satisfies the Neumann condition (4). A sufficient condition for convexity of ψ is $0 \leq \theta \leq 2^{1-2a}/a$. If $a < 1$, then the tail behavior of $p(x|\theta,a)$ is

$$p(x|\theta,a) \;\simeq\; \theta^2 a^2 (2a-1)\left(\frac{\pi^2}{2}\{(1-x_1)^2 + (1-x_2)^2\}\right)^{2(a-1)}$$

as $(x_1, x_2) \to (1,1)$. The proofs of these facts are omitted. Although estimation of θ is described by the determinant maximization, that of a is not. Further investigation is needed.

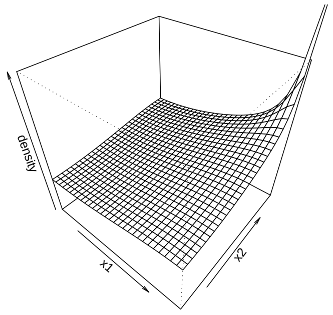


Figure 6: The density function $p(x|\theta,a)$ for $a = 0.75$ and $\theta = 2^{1-2a}/a$.

If any covariates are available together with given data, we can include the covariates in the parameter θ of SGM. However, since the parameter space Θ of SGM is not the whole Euclidean space, its use is restricted.

The author recently proved an inequality on Efron's statistical curvature, in that the curvature of SGM at the origin $\theta = 0$ is always smaller than that of MixM (5). This fact is not so practical but it supports SGM. Since the statement and the proof of this inequality are rather complicated, we will present them in a forthcoming paper.

We constructed a lasso-type estimator on SGM as a byproduct of the conservative feasible region in Section 3. Performance of the estimator is numerically studied in Section 4. For the existing lasso estimators, some asymptotic results are known when the sample size n and/or the number m of variates increase (Knight and Fu [2000], Meinshausen and Bühlmann [2006], Yuan and Lin [2007], Bunea et al. [2007], Banerjee et al. [2008]). We think it is important to compare our SGM with the Gaussian, mixture and exponential models on the asymptotic argument.

A Proofs

A.1 Proof of Lemma 1

Let ψ have the form (2) and choose any θ such that $D^2\psi(x|\theta) \succeq 0$ for every $x \in [0,1]^m$. We prove that the gradient map $D\psi(\cdot|\theta)$ is a bijection on $[0,1]^m$. If $\theta = 0$, then the bijectivity of $D\psi(x|\theta) = x$ is clear. Therefore we assume $\theta \neq 0$. We can extend the domain of $\psi(\cdot|\theta)$ from $[0,1]^m$ to whole $\mathbb{R}^m$ by using Eq. (2), and denote the extended function by $\tilde{\psi}(x) = \tilde{\psi}(x|\theta)$ for $x \in \mathbb{R}^m$. Since $\tilde{\psi}(x)$ is a periodic and even function along each axis, the convexity condition $D^2\tilde{\psi} \succeq 0$ holds over $x \in \mathbb{R}^m$. We will prove that (i) $D\tilde{\psi}$ is a bijection on $\mathbb{R}^m$ and (ii) $D\tilde{\psi}$ is a bijection on each hyperplane $\{x \mid x_j = b\}$, where $j \in \{1,\ldots,m\}$ and $b \in \{0,1\}$. We first show that the bijectivity on $[0,1]^m$ follows from the conditions (i) and (ii). Indeed, if (i) and (ii) are fulfilled, then for each $j \in \{1,\ldots,m\}$ the sandwiched region $\{x \in \mathbb{R}^m \mid 0 \leq x_j \leq 1\}$ between two hyperplanes is mapped onto itself because $D\tilde{\psi}$ is continuous. Therefore $[0,1]^m$ is injectively mapped onto itself. To prove (i), it is sufficient to show that $\tilde{\psi}$ is strictly convex and co-finite: $\lim_{\lambda\to\infty} \tilde{\psi}(\lambda x)/\|x\| = 0$ whenever $x \neq 0$ (see Theorem 26.6 of Rockafeller [1970]). We define a function $f(z)$ of $z \in \mathbb{R}$ by $f(z) = \tilde{\psi}(x_0 + ze)$, where $x_0 \in \mathbb{R}^m$ and $e \in \mathbb{R}^m \setminus \{0\}$ are arbitrary. Then $f''(z) \geq 0$ for any z since $D^2\tilde{\psi}(x) \succeq 0$ for any $x \in \mathbb{R}^m$. However, since $f''(z)$ is a non-constant analyitc function (recall that $\theta \neq 0$), $f''(z)$ must be positive except for a finite number of z for each bounded interval. Hence f, and therefore $\tilde{\psi}$, is strictly convex. The co-finiteness of $\tilde{\psi}$ is immediate because $\tilde{\psi}$ is sum of $x^\top x/2$ and a bounded function. Hence (i) was proved. Next we prove the condition (ii). We consider the hyperplane $\{x \mid x_m = b\}$, where $b \in \{0,1\}$, without loss of generality. Denote the restriction of $\tilde{\psi}$ to $\{x \mid x_m = b\}$ by $\tilde{\psi}_{m-1}$. Then $\tilde{\psi}_{m-1}$ has the following expression

$$\tilde{\psi}_{m-1}(x_1,\ldots,x_{m-1}) \;=\; \frac{b^2}{2} + \frac{1}{2}\sum_{i=1}^{m-1} x_i^2 - \sum_{u\in\mathcal{U}} \pi^{-2}\theta_u(-1)^{u_j b} \prod_{i=1}^{m-1} \cos(\pi u_j x_j).$$

This function is the same form as Eq. (2) with the dimension $m-1$. The convexity condition $(\partial^2\tilde{\psi}_{m-1}/\partial x_i \partial x_j) \succeq 0$ is also satisfied because $\tilde{\psi}_{m-1}$ is a restriction of $\tilde{\psi}$. Thus (ii) is proved in the same manner as the proof of (i).

A.2 Proof of Lemma 2

A statistical model is a mixture model if and only if all the second derivatives of the density function with respect to the parameter vanish. Hence we calculate the second derivative of the density function of SGM. Put $\mathbb{Z}_i := \{u \in \mathbb{Z}^m_{\geq 0} \mid u_j = 0 \ \forall j \neq i\}$. If $\mathcal{U} \subset \mathbb{Z}_i$ for some i, then it is easy to confirm that SGM becomes a mixture model

$$p(x|\theta) \ = \ 1 + \sum_{u \in \mathcal{U}} \theta_u u_i^2 \cos(\pi u_i x_i).$$

Hence we assume that $\mathcal{U} \not\subset \mathbb{Z}_i$ for any i. Then there exist $u, v \in \mathcal{U}$ (the case $u = v$ is available) such that $|\sigma(u) \cup \sigma(v)| \geq 2$, where $\sigma(u) = \{j \mid u_j > 0\}$. Putting $A_u = \{D^2\psi(x|\theta)\}^{-1}\{\partial/\partial\theta_u(D^2\psi(x|\theta))\}$ we have

$$\frac{\partial^2 p(x|\theta)}{\partial\theta_u \partial\theta_v} \ = \ \operatorname{tr} A_u \operatorname{tr} A_v - \operatorname{tr}[A_u A_v].$$

Since $A_u|_{\theta=0,x=0} = \operatorname{diag}(u_1^2, \ldots, u_m^2)$, we have

$$\left.\frac{\partial^2 p(x|\theta)}{\partial\theta_u \partial\theta_v}\right|_{\theta=0,x=0} \ = \ \|u\|^2\|v\|^2 - \sum_i u_i^2 v_i^2 \ = \ \sum_i \sum_{j \neq i} u_i^2 v_j^2 \ > \ 0,$$

where the last inequality follows from $|\sigma(u) \cup \sigma(v)| \geq 2$. Thus SGM is not a mixture model as long as $\mathcal{U} \not\subset \mathbb{Z}_i$ for any i.

A.3 Proof of Lemma3

The score function of SGM at $\theta = 0$ is directly calculated as

$$L_u \ := \ \left.\frac{\partial}{\partial\theta_u} \log p(x|\theta)\right|_{\theta=0} \ = \ \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j).$$

The score function of MixM is also easily proved to be L_u. Then the Fisher information matrix of both the models is

$$J_{uv} \ = \ \int p(x|0) L_u L_v \mathrm{d}x \ = \ \|u\|^2\|v\|^2 \prod_{j=1}^m \int_0^1 \cos(\pi u_j x_j)\cos(\pi v_j x_j)\mathrm{d}x_j.$$

Here the integral is calculated by the following formula

$$\int_0^1 \cos(\pi u_j x_j)\cos(\pi v_j x_j)\mathrm{d}x_j \ = \ \begin{cases} 1 & \text{if } u_j = v_j = 0, \\ 1/2 & \text{if } u_j = v_j > 0, \\ 0 & \text{if } u_j \neq v_j. \end{cases}$$

A.4 Proof of Equations (7) and (8)

We first prove Eq. (7). Let $m = 1$ and $\mathcal{U} = \{u\}$. We only consider the case $u = 1$. The other cases are similarly proved. Put $\theta = \theta_1$. Since $p(x_1|\theta) = 1 + \theta\cos(\pi x_1)$, we have

$$J_{uu}(\theta) \;=\; \int_0^1 \frac{\cos^2(\pi x_1)}{1+\theta\cos(\pi x_1)}\mathrm{d}x_1.$$

By putting $z = \exp(\mathrm{i}\pi u x_1)$, we obtain

$$J_{uu}(\theta) \;=\; \frac{1}{2\pi\mathrm{i}}\oint_{|z|=1} \frac{(z+z^{-1})^2/4}{1+\theta(z+z^{-1})/2}\frac{\mathrm{d}z}{z} \;=\; \frac{1}{4\pi\mathrm{i}}\oint_{|z|=1}\frac{(z^2+1)^2}{z^2(\theta z^2+2z+\theta)}\mathrm{d}z.$$

The poles of the integrand inside the unit circle are 0 and z_+, where $z_\pm := (-1 \pm \sqrt{1-\theta^2})/\theta$. By the residue theorem, we obtain

$$J_{uu}(\theta) \;=\; \frac{1}{2}\left(\frac{-2}{\theta^2}\right) + \frac{1}{2}\frac{(z_+^2+1)^2}{z_+^2\theta(z_+-z_-)} \;=\; \frac{1-\sqrt{1-\theta^2}}{\theta^2\sqrt{1-\theta^2}}.$$

This proves Eq. (7).

We next prove Eq. (8). Put $u = (1,1)$ and $\theta = \theta_u$. We use the following identity

$$\begin{aligned} p(x|\theta) \;&=\; \det\begin{pmatrix} 1+\theta\cos(x_1)\cos(x_2) & -\theta\sin(x_1)\sin(x_2) \\ -\theta\sin(x_1)\sin(x_2) & 1+\theta\cos(x_1)\cos(x_2) \end{pmatrix} \\ &=\; (1+\theta\cos(\pi(x_1-x_2)))(1+\theta\cos(\pi(x_1+x_2))). \end{aligned}$$

The Fisher information is

$$\begin{aligned} J_{uu}(\theta) \;&=\; \int_{[0,1]^2}\left(\frac{\cos^2(\pi(x_1-x_2))}{1+\theta\cos(\pi(x_1-x_2))} + \frac{\cos^2(\pi(x_1+x_2))}{1+\theta\cos(\pi(x_1+x_2))}\right)\mathrm{d}x_1\mathrm{d}x_2 \\ &=\; \frac{1}{4}\int_{[-1,1]^2}\left(\frac{\cos^2(\pi(x_1-x_2))}{1+\theta\cos(\pi(x_1-x_2))} + \frac{\cos^2(\pi(x_1+x_2))}{1+\theta\cos(\pi(x_1+x_2))}\right)\mathrm{d}x_1\mathrm{d}x_2 \\ &=\; \frac{1}{4}\int_{[-1,1]^2}\left(\frac{\cos^2(\pi y_1)}{1+\theta\cos(\pi y_1)} + \frac{\cos^2(\pi y_2)}{1+\theta\cos(\pi y_2)}\right)\mathrm{d}y_1\mathrm{d}y_2 \end{aligned}$$

where the last equality follows from the transformation $y_1 = x_1 - x_2$ and $y_2 = x_1 + x_2$, and from the periodicity of the integrand. Then (8) is proved in the same manner as the proof of (7).

A.5 Proof of Lemma 4

We use the following elementary lemma. Put $\mathcal{S} = \{A \succeq 0 \mid \mathrm{tr}\,A = 1\}$. Note that $\mathcal{S}$ is compact.

Lemma 5. *Let X be a real symmetric matrix. Then the minimum eigenvalue of X is given by $\min_{A\in\mathcal{S}}\mathrm{tr}(AX)$.*

Proof. Let $X = \sum_i \xi_i e(i)e(i)^\top$ be the spectral decomposition of X, where $\xi_1 \le \cdots \le \xi_m$ and $e(i)^\top e(i) = 1$. For any $A \in \mathcal{S}$,

$$\operatorname{tr}(AX) \;=\; \sum_i \xi_i (e(i)^\top A e(i)) \;\ge\; \xi_1 \sum_j (e(j)^\top A e(j)) \;=\; \xi_1.$$

The equality is attained at $A = e(1)e(1)^\top$. □

Let $H_u(x) = D^2(-\pi^{-2}\prod_{j=1}^m \cos(\pi u_j x_j))$. Then

$$D^2\psi(x|\theta) \;=\; I + \sum_{u\in\mathcal{U}} \theta_u H_u(x).$$

The minimum eigenvalue $\rho_{\min}(\theta)$ of $D^2\psi(x|\theta)$ minimized over $x \in [0,1]^m$ is

$$\rho_{\min}(\theta) \;=\; 1 + \min_{x\in[0,1]^m, A\in\mathcal{S}} \sum_{u\in\mathcal{U}} \theta_u \operatorname{tr}(AH_u(x)).$$

Recall that the parameter space Θ is expressed as $\Theta = \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \rho_{\min}(\theta) \ge 0\}$. We prove that the interior of Θ is $\Theta^\circ = \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \rho_{\min}(\theta) > 0\}$. Put

$$\mu \;=\; \max_{u\in\mathcal{U}} \max_{x\in[0,1]^m} \max_{A\in\mathcal{S}} |\operatorname{tr}(AH_u(x))| \;<\; \infty.$$

We first prove that if $\rho_{\min}(\theta) > 0$, then $\theta \in \Theta^\circ$. Indeed, if $\eta \in \mathbb{R}^{\mathcal{U}}$ is sufficiently small, then

$$\rho_{\min}(\theta + \eta) \;\ge\; \rho_{\min}(\theta) - \mu \sum_{u\in\mathcal{U}} |\eta_u| \;\ge\; 0.$$

We next prove that if $\rho_{\min}(\theta) = 0$, then $\theta \in \Theta \setminus \Theta^\circ$. Since $\rho_{\min}(\theta) = 0$, there exist some $A \in \mathcal{S}$ and some $x \in [0,1]^m$ such that $\operatorname{tr}(AD^2\psi(x|\theta)) = 0$. For such an x, there exists some $v \in \mathcal{U}$ such that $\theta_v \operatorname{tr}(AH_v(x)) < 0$. Define a vector $\eta \in \mathbb{R}^{\mathcal{U}}$ by $\eta_u = \theta_v 1_{\{u=v\}}$. Then, for any $\epsilon > 0$, we have

$$\rho_{\min}(\theta + \epsilon\eta) \;\le\; \operatorname{tr}(AD^2\psi(x|\theta + \epsilon\eta)) \;=\; \epsilon\theta_v \operatorname{tr}(AH_v(x)) \;<\; 0.$$

This implies that θ is a boundary point of Θ. Hence Lemma 4 was proved.

A.6 Proof of Theorem 2

We first recall some notations. We use $[m] = \{1, \ldots, m\}$ and $L_M = \{\frac{0}{M}, \frac{1}{M}, \cdots, \frac{M}{M}\}$. The supremum norm of $s \in \mathbb{Z}^m$ is defined by $\|s\|_\infty := \max_j |s_j|$. Recall that $U_{\max} = \max_{u\in\mathcal{U}} \|u\|_\infty$. We denote $U = U_{\max}$ for simplicity. Recall that K_M is a linear map on $\mathbb{R}^{\mathcal{U}}$ defined by $K_M\theta = (\theta_u / \prod_{j=1}^m (1 - u_j/M))_{u\in\mathcal{U}}$.

Define a set $K_M^{-1}\Theta^\circ$ by

$$K_M^{-1}\Theta^\circ \;:=\; \{K_M^{-1}\theta \mid \theta \in \Theta^\circ\} \;=\; \{\theta \mid D^2\psi(x|K_M\theta) \succ 0 \quad \forall x \in [0,1]^m\}.$$

Then we have $K_M^{-1}\Theta^\circ \;\subset\; \Theta_M^\circ$ by the definition of Θ_M°. Hence, the theorem follows from the following two claims.

(i) $\limsup_{M\to\infty} K_M^{-1}\Theta^\circ = \Theta^\circ$.

(ii) $\Theta_M^\circ \subset \Theta^\circ$ for any M.

We first prove (i). Put $\mathcal{S} = \{A \succeq 0 \mid \operatorname{tr} A = 1\}$ and $f(x|\theta, A) = \operatorname{tr}[AD^2\psi(x|\theta)]$. By Lemma5 and compactness of $[0,1]^m \times \mathcal{S}$, a vector θ belongs to Θ° if and only if

$$\min_{x\in[0,1]^m, A\in\mathcal{S}} f(x|\theta, A) \ > \ 0.$$

Now it is sufficient to prove that, for any $\theta \in \mathbb{R}^{\mathcal{U}}$, $f(x|K_M\theta, A)$ converges to $f(x|\theta, A)$ uniformly in $x \in [0,1]^m$ and $A \in \mathcal{S}$. Let $H_u(x) := D^2(-\pi^{-2}\prod_{j=1}^m \cos(\pi u_j x_j))$. Then we have $f(x|\theta, A) = 1 + \sum_{u\in\mathcal{U}} \theta_u \operatorname{tr}[AH_u(x)]$ and therefore

$$|f(x|K_M\theta, A) - f(x|\theta, A)| \ \leq \ \sum_{u\in\mathcal{U}} |\{(K_M\theta)_u - \theta_u\} \operatorname{tr}[AH_u(x)]| \,. \qquad (12)$$

Since the function $\operatorname{tr}[AH_u(x)]$ of $(x, A) \in [0,1]^m \times \mathcal{S}$ is bounded and since $(K_M\theta)_u$ converges to θ_u for each $u \in \mathcal{U}$ as $M \to \infty$, the right hand side of (12) converges to 0 uniformly in x and A.

Next we prove (ii). Let $R_M = \{-\frac{M-1}{M}, \ldots, \frac{M-1}{M}, \frac{M}{M}\}$. We extend the domain of ψ from $[0,1]^m$ to $\mathbb{R}^m$ as done in the proof of Lemma 1, and denote it again by ψ. If $\theta \in \Theta_M^\circ$, then $D^2\psi(\xi|K_M\theta)$ is positive definite for any $\xi \in R_M^m$ because $\psi(x|\theta)$ is an even function with respect to each coordinate x_j. Then it is sufficient to prove that $D^2\psi(x|\theta)$ for any x is written as a convex combination of $\{D^2\psi(\xi|K_M\theta)\}_{\xi\in R_M^m}$. Define a Fejér-type kernel Q_M by

$$Q_M(z) \ = \ \frac{1}{2M^2} \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \mathrm{e}^{\mathrm{i}\pi(a-b)z} \ = \ \frac{1}{2M^2} \left(\frac{\sin(\pi M z/2)}{\sin(\pi z/2)} \right)^2 .$$

Then the following lemma holds.

Lemma 6. For any $M \geq U + 1$, we have

$$D^2\psi(x|\theta) \ = \ \sum_{\xi\in R_M^m} D^2\psi(\xi|K_M\theta) \prod_{j=1}^m Q_M(x_j - \xi_j).$$

The right hand side is a convex combination of $\{D^2\psi(\xi|K_M\theta)\}_{\xi\in R_M^m}$.

Proof. For each $j \in \{1, \ldots, m\}$, define an operator $K_{M,j}$ on $\mathbb{R}^{\mathcal{U}}$ by

$$(K_{M,j}\theta)_u = \frac{\theta_u}{1 - u_j/M}.$$

Then we have $K_M = \prod_{j=1}^m K_{M,j}$ from the definition. It is sufficient to show that

$$D^2\psi(x|\theta) \ = \ \sum_{\xi_j\in R_M} D^2\psi(\xi_j, x_{\setminus j}|K_{M,j}\theta) Q_M(x_j - \xi_j), \qquad (13)$$

where $x_{\setminus j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m)$. In fact, if (13) is proved, then

$$\begin{aligned}
D^2\psi(x|\theta) &= \sum_{\xi_1 \in R_M} D^2\psi(\xi_1, x_2, \ldots, x_m|K_{M,1}\theta)Q_M(x_1 - \xi_1) \\
&= \sum_{\xi_1 \in R_M} \sum_{\xi_2 \in R_M} D^2\psi(\xi_1, \xi_2, \ldots, x_M|K_{M,1}K_{M,2}\theta) \prod_{j=1}^{2} Q_M(x_j - \xi_j) \\
&= \cdots \\
&= \sum_{\xi \in R_M^m} D^2\psi(\xi|K_M\theta) \prod_{j=1}^{m} Q_M(x_j - \xi_j).
\end{aligned}$$

We prove (13) for $j = 1$ without loss of generality. We first describe $D^2\psi(x|\theta)$ in terms of $\{\mathrm{e}^{\mathrm{i}\pi s^\top x}\}_{s \in \mathbb{Z}^m}$. For each $s \in \mathbb{Z}^m$, we define a $m \times m$ matrix

$$F_s = \begin{cases} I & \text{if } s = 0, \\ \theta_u 2^{-|\sigma(u)|} s s^\top & \text{if } |s_j| = u_j \text{ for all } j \in [m] \text{ for some } u \in \mathcal{U}, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $\sigma(u) = \{j \in [m] \mid u_j > 0\}$. Then, by applying the Euler's formula $\cos(\pi u_j x_j) = (\mathrm{e}^{\mathrm{i}\pi u_j x_j} - \mathrm{e}^{-\mathrm{i}\pi u_j x_j})/2$ to Eq. (2), we can show that

$$D^2\psi(x|\theta) = \sum_{\|s\|_\infty \le U} F_s \mathrm{e}^{\mathrm{i}\pi s^\top x}.$$

Recall that $U = \max_{u \in \mathcal{U}} \|u\|_\infty$. The right hand side of (13) with $j = 1$ is

$$\begin{aligned}
&\sum_{\xi_1 \in R_M} D^2\psi(\xi_1, x_{\setminus 1}|K_{M,1}\theta)Q_M(x_1 - \xi_1) \\
&= \sum_{\xi_1 \in R_M} \left(\sum_{\|s\|_\infty \le U} \frac{F_s \mathrm{e}^{\mathrm{i}\pi(s_1\xi_1 + s_{\setminus 1}^\top x_{\setminus 1})}}{1 - |s_1|/M} \right) \left(\frac{1}{2M^2} \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \mathrm{e}^{\mathrm{i}\pi(a-b)(x_1 - \xi_1)} \right) \\
&= \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \sum_{\|s\|_\infty \le U} \frac{F_s \mathrm{e}^{\mathrm{i}\pi((a-b)x_1 + s_{\setminus 1}^\top x_{\setminus 1})}}{M - |s_1|} \frac{1}{2M} \sum_{\xi_1 \in R_M} \mathrm{e}^{\mathrm{i}\pi(s_1 - a + b)\xi_1} \\
&= \sum_{a=0}^{M-1} \sum_{b=0}^{M-1} \sum_{\|s\|_\infty \le U} \frac{F_s \mathrm{e}^{\mathrm{i}\pi s^\top x}}{M - |s_1|} 1_{\{s_1 \equiv a - b \bmod 2M\}}.
\end{aligned}$$

For any s_1 with $|s_1| \le U < M$, the cardinality of the set

$$\{(a, b) \in \{0, \ldots, M-1\}^2 \mid s_1 = a - b\}$$

is $M - |s_1|$. Hence we have

$$\sum_{\xi_1 \in R_M} D^2\psi(\xi_1, x_{\setminus 1}|K_{M,1}\theta)Q_M(x_1 - \xi_1) = \sum_{\|s\|_\infty \le U} F_s \mathrm{e}^{\mathrm{i}\pi s^\top x} = D^2\psi(x|\theta).$$

Therefore (13) was proved.

Now we prove that $\{\prod_{j=1}^m Q_M(x_j-\xi_j)\}_{\xi\in R_M^m}$ becomes a probability vector. In fact, non-negativity follows from the definition of Q_M and the total mass is 1 because

$$\sum_{\xi_1\in R_M} Q_M(x_1-\xi_1) \;=\; \frac{1}{2M^2}\sum_{a=0}^{M-1}\sum_{b=0}^{M-1}\sum_{\xi_1\in R_M} \mathrm{e}^{\mathrm{i}\pi(a-b)(x_1-\xi_1)} \;=\; \frac{1}{M}\sum_{a=0}^{M-1}\sum_{b=0}^{M-1} 1_{\{a=b\}} \;=\; 1.$$

Therefore the lemma and Theorem 2 are proved. □

A.7 Proof of Theorem 3

Let $\theta\in\Theta^{\mathrm{lit}}$. We show that $D^2\psi(x|\theta)\succeq 0$ for all $x\in[0,1]^m$. By Euler's formula, we obtain

$$\prod_{j=1}^m \cos(\pi u_j x_j) \;=\; 2^{-m}\sum_{\alpha\in\{-1,1\}^m}\cos(\pi\alpha^\top d(u)x),$$

where $d(u)$ is the $m\times m$ diagonal matrix with the diagonal vector u. Note that $2^{-m}\sum_{\alpha\in\{-1,1\}^m}\alpha\alpha^\top = I$. Then

$$\begin{aligned}
D^2\psi(x|\theta) \;&=\; I+\sum_{u\in\mathcal{U}}\frac{\theta_u}{2^m}\sum_{\alpha\in\{-1,1\}^m}\cos(\pi\alpha^\top d(u)x)d(u)\alpha\alpha^\top d(u)\\
&\succeq\; I-\sum_{u\in\mathcal{U}}\frac{|\theta_u|}{2^m}\sum_{\alpha\in\{-1,1\}^m} d(u)\alpha\alpha^\top d(u)\\
&=\; I-\sum_{u\in\mathcal{U}}|\theta_u|d(u)^2\\
&\succeq\; 0.
\end{aligned}$$

This implies that $\theta\in\Theta$.

Next we assume that $\mathcal{V}\subset\mathcal{U}$ is linearly independent modulo 2. Since $\Theta^{\mathrm{lit}}\subset\Theta$, it is sufficient to prove that $\Theta\cap\mathbb{R}_{\mathcal{V}}\subset\Theta^{\mathrm{lit}}\cap\mathbb{R}_{\mathcal{V}}$. Let $\theta\in\Theta\cap\mathbb{R}_{\mathcal{V}}$. We evaluate $D^2\psi(x|\theta)$ at lattice points $\xi\in\{0,1\}^m$. For any $\xi\in\{0,1\}^m$ and any $v\in\mathbb{Z}^m$, we have

$$D^2\left(-\pi^{-2}\prod_{j=1}^m\cos(\pi v_j x_j)\right)\Bigg|_{x=\xi} = (-1)^{v^\top\xi}d(v)^2.$$

Since $\mathcal{V}$ is linearly independent modulo 2, we can choose $\xi\in\{0,1\}^m$ such that $v^\top\xi = 1_{\{\theta_v>0\}}$ (mod 2) for all $v\in\mathcal{V}$. Then

$$0 \;\preceq\; D^2\psi(x|\theta)\big|_{x=\xi} \;=\; 1+\sum_{v\in\mathcal{V}}\theta_v(-1)^{v^\top\xi}d(v)^2 \;=\; 1-\sum_{v\in\mathcal{U}}|\theta_v|d(v)^2.$$

This means $\theta\in\Theta^{\mathrm{lit}}\cap\mathbb{R}_{\mathcal{V}}$.

Acknowledgements

This study was partially supported by the Global Center of Excellence "The research and training center for new development in mathematics" and by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), No. 19700258.

References

O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J Machine Lear. Res.*, 9:485–516, 2008.

A. Ben-tal and A. Nemirovski. Robust convex optimization. *Math. Oper. Res.*, 23 (4):769–805, 1998.

G. E. P. Box and G. M. Jenkins. *Time series analysis – forecasting and control.* Holden-Day Inc., San Francisco, 1976.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparse density estimation with l1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007, Lecture Notes in Artificial Intelligence*, pages 530–544. Springer-Verlag, Heidelberg, 2007.

L. A. Caffarelli. Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.*, 214:547–563, 2000.

D. Edwards. *Introduction to Graphical Modeling.* Springer-Verlag, New York, second edition, 2000.

J. J. Fernández-Durán. Circular distributions based on nonnegative trigonometric sums. *Biometrics*, 60(JUNE):499–503, 2004.

J. Friedmann, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

H. Halkin, L. B. Sheiner, C. C. Peck, and K. L. Melmon. Determinants of the renal clearance of digoxin. *Clin. Pharmacol. Ther.*, 17(4):385–394, 1975.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5): 1356–1378, 2000.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

R. B. Nelsen. *An Introduction to Copulas.* Springer-Verlag, New York, second edition, 2006.

R. T. Rockafeller. *Convex analysis.* Princeton University Press, 1970.

T. Sei. Parametric modeling based on the gradient maps of convex functions. Technical report, METR2006-51, Department of Mathematical Engineering, University of Tokyo, 2006.

T. Sei. Gradient modeling for multivariate analysis. In *The Pyrenees International Workshop on Statistics, Probability and Operations Research (SPO 2007)*, Jaca, Spain, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc., B*, 58(1):267–288, 1996.

K. C. Toh, R. H. Tütüncü, and M. J. Todd. *On the implementation and usage of SDPT3 — a MATLAB software package for semidefinite-quadratic-linear programming, version 4.0*, 2006.

L. Vandenberghe, S. Boyd, and S. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl*, 19(2):499–533, 1998.

C. Villani. *Topics in Optimal Transportation.* AMS, Providence, 2003.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

A. Zygmund. *Trigonometric Series*, volume 2. Cambridge Mathematical Library, third edition, 2002.