Statistica Sinica **21** (2011), 1291-1314

A STRUCTURAL MODEL ON A HYPERCUBE REPRESENTED BY OPTIMAL TRANSPORT

Tomonari Sei

Keio University

Abstract: We propose a flexible statistical model for high-dimensional quantitative data on a hypercube. Our model, the structural gradient model (SGM), is based on a one-to-one map on the hypercube that is a solution to an optimal transport problem. As we show with many examples, SGM can describe various dependence structures including correlation and heteroscedasticity. The likelihood function is explicitly expressed without any normalizing constant. Simulation of SGM is achieved through a direct extension of the inverse function method. The maximum likelihood estimation of SGM is reduced to the determinant-maximization known as a convex optimization problem. In particular, a lasso-type estimation is available by adding constraints. SGM is compared with graphical Gaussian models and mixture models.

Key words and phrases: Determinant maximization, Fourier series, graphical model, lasso, optimal transport, structural gradient model.

1. Introduction

In recent years, it has become important to treat high-dimensional quantitative data, especially in biostatistics and spatial-temporal statistics. The graphical Gaussian model is an important one. However, the Gaussian model represents only second-order interactions 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 (2009).

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

$$p(x) = \det(D^2\psi(x)), \quad x \in [0,1]^m, \tag{1.1}$$

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 $[0,1]^m$ can be written as (1.1). This fact is deeply connected to the theory of optimal transport (see e.g. Villani (2003)). A precise statement is the following.

Theorem 1.(Brenier (1991), McCann (1995)) *Let p and q be any two probability densities with respect to the Lebesgue measure on $\mathbb{R}^m$. Then there exists a convex function ψ satisfying $p(x) = q(D\psi(x))\det(D^2\psi(x))$. The function ψ is p-a.s. unique up to an arbitrary additive constant.*

For any density $p(x)$ on $[0,1]^m$, the representation (1.1) is obtained if the density $q(x)$ in Theorem 1 is set to the uniform density on $[0,1]^m$. The bijective gradient map $T := D\psi$ is the optimal-transport plan that minimizes the cost functional $\mathrm{E}[\|X - T(X)\|^2]$ subject to the density of X and $T(X)$ being p and q, respectively (e.g., Villani (2003)). A sample X from the density (1.1) is obtained by an extension of the inverse function method:

$$X = (D\psi)^{-1}(Y) = \operatorname*{argmin}_{x\in[0,1]^m}\{\psi(x) - x^\top Y\},$$

where Y is a sample drawn from the uniform density on $[0,1]^m$ (see also Sei (2009)).

In this paper we call ψ *the potential function*. 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 call the model (1.1) *the structural gradient model* and refer to it as SGM. Unknown parameters are the Fourier coefficients of 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 through simulation and 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. To overcome this, we give a L^1-conservative region that enables us to calculate the estimator by the determinant maximization algorithm (Vandenberghe, Boyd, and Wu (1998)). As a by-product of the 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, Tsybakov, and Wegkamp (2007), Banerjee, Ghaoui, and d'Aspremont (2008)).

We consider only the case in which the sample space is a hypercube. This is not a strong assumption since we can transform any real-valued data into $[0,1]$-valued data with a fixed sigmoid function. Another approach to deal with unbounded data within the framework of gradient maps is given in Sei (2007),

Sei (2009), where optimal transport between the standard normal density and other densities is considered. Here we use the uniform density instead of the normal density because the former is analytically simpler.

A tractable statistical model on the hypercube $[0,1]^m$ is the copula model (see e.g. Nelsen (2006)), in which every marginal density is uniform on $[0,1]$. The copula model is a useful tool to see dependency of multivariate data independently of the marginal density. Unfortunately, as is shown in Section 2, SGM is not a copula model because the one-dimensional marginal density functions of SGM are not uniform except for special cases. In other words, SGM adjusts the marginal densities simultaneously with dependency.

The paper is organized as follows. In Section 2, we define SGM and give various examples of it. In Section 3, we investigate maximum likelihood estimation and propose a lasso-type estimator. In Section 4, we compare SGM with graphical Gaussian models and mixture models using numerical experiments. Finally we give some discussion in Section 5. All proofs are given in the 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 and let $(\mathbb{Z}_{\geq 0}^m)^+ = \mathbb{Z}_{\geq 0}^m \setminus \{0\}$ be the set of non-zero vectors with non-negative integer components.

Definition 1.([SGM]) Let $\mathcal{U}$ be a finite subset of $(\mathbb{Z}_{\geq 0}^m)^+$. *The structural gradient model* (abbreviated as *SGM*) is given by (1.1) with the potential function

$$\psi(x|\theta) = \frac{1}{2} x^\top x - \sum_{u \in \mathcal{U}} \frac{\theta_u}{\pi^2} \prod_{j=1}^m \cos(\pi u_j x_j), \tag{2.1}$$

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

$$\Theta = \left\{ \theta \in \mathbb{R}^{\mathcal{U}} \mid D^2 \psi(x|\theta) \succeq 0 \ \text{ for all } \ x \in [0,1]^m \right\}. \tag{2.2}$$

A vector $\theta \in \mathbb{R}^{\mathcal{U}}$ is called *feasible* if $\theta \in \Theta$, *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 of a theorem of Caffarelli (2000). To state the theorem, we need some notation. Denote the $2m$ faces of $[0,1]^m$ by $F_j^b = \{x \in [0,1]^m \mid x_j = b\}$ for $j \in \{1,\dots,m\}$ and $b \in \{0,1\}$. For a smooth function ψ on $[0,1]^m$, consider the Neumann condition

$$\frac{\partial \psi(x)}{\partial x_j} = b \quad \text{for any} \quad x \in F_j^b. \tag{2.3}$$

It is easily confirmed that ψ at (2.1) satisfies (2.3). Conversely, if $\psi(x)$ satisfies (2.3), then it can be expanded by an infinite cosine series in L^2 sense (see e.g. page 300 of Zygmund (2002)). In other words, the function (2.1) approximates any potential function satisfying (2.3) if we make the frequency set $\mathcal{U}$ large. Now we give Caffarelli's theorem, but with slightly stronger assumption.

Theorem 2.(Caffarelli (2000)) *Let $p(x)$ be a strictly positive and continuously differentiable function on $[0,1]^m$. Assume that $p(x)$ satisfies $\partial p(x)/\partial x_j = 0$ for any $x \in F_j^b$. Then there exists a twice-differentiable convex function $\psi(x)$ such that* (1.1) *and* (2.3) *hold.*

Since the conditions for $p(x)$ here are differentiability and a boundary condition, we can construct sufficiently many statistical models by SGM. In the next subsection, we give various examples. 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 next subsection. For the multi-dimensional case ($m > 1$), SGM is not a mixture model except for essentially one-dimensional cases.

Lemma 2. *SGM is not a mixture model with respect to θ unless there exists some $i \in \{1,\dots,m\}$ such that $\mathcal{U} \subset \mathbb{Z}_i$, where $\mathbb{Z}_i = \{u \in (\mathbb{Z}_{\geq 0}^m)^+ \mid u_j = 0 \text{ if } j \neq i\}$.*

We use the following mixture model as a reference.

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

$$\tilde{p}(x|\theta) \;=\; 1 + \sum_{u \in \mathcal{U}} \theta_u \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j), \tag{2.4}$$

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 $\tilde{\Theta} = \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \tilde{p}(x|\theta) \geq 0 \text{ for all } x \in [0,1]^m\}$.

Lemma 3. *The score vector at $\theta = 0$ of both SGM and MixM is $(\|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j))_{u\in\mathcal{U}}$. The Fisher information matrix $J = (J_{uv})_{u,v\in\mathcal{U}}$ at $\theta = 0$ of both the models is $J_{uv} = \|u\|^4 2^{-|\sigma(u)|} 1_{\{u=v\}}$, where $\sigma(u) = \{j \in \{1,\dots,m\} \mid u_j > 0\}$ and $|\sigma(u)|$ denotes the cardinality of $\sigma(u)$. In particular, J_{uv} is diagonal.*

The Fisher information matrix J at the origin is useful if we deal with testing $\theta = 0$. 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 use the scaled maximum likelihood estimator $J^{1/2}\hat{\theta}$ to detect significant components of $\hat{\theta}$. 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 points $\theta \neq 0$. Exceptional cases are in the next subsection.

We describe a relation between SGM and copula models. A probability density $p(x)$ on $[0,1]^m$ is *a copula density* if every one-dimensional marginal density of $p(x)$ is uniform on $[0,1]$. *A copula model* is a statistical model that consists of copula densities; Nelsen (2006) for a comprehensive review. A referee has pointed out that densities of SGM cannot be copula densities except for special cases. A precise statement follows.

Lemma 4. *Let $\mathcal{U} \subset (\mathbb{Z}_{\geq 0}^m)^+$ be non-empty and Θ be the feasible region of SGM. Then for almost all $\theta \in \Theta$, the density $p(x|\theta)$ is not a copula density.*

An example of the exceptional subset of Θ is given in Example 8. We also remark that MixM becomes a copula model if $|\sigma(u)| \geq 2$ for all $u \in \mathcal{U}$, where $\sigma(u) = \{j \mid u_j > 0\}$.

2.2. Examples

We give examples of SGM. We mainly compare SGM with MixM of Definition 2. For SGM, a sufficient condition for feasibility of θ is useful in dealing with examples. Theorem 3 has it that θ is feasible if

$$1 - \sum_{u\in\mathcal{U}} |\theta_u| u_j^2 \;\geq\; 0 \tag{2.5}$$

for any $j = 1,\dots,m$. 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 $m = 1$, 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. 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, for any feasible $\theta = \theta_u$,

$$J_{uu}(\theta) = \frac{1 - \sqrt{1 - \theta^2 u^4}}{\theta^2 \sqrt{1 - \theta^2 u^4}}. \tag{2.6}$$

The proof is given in the 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. Let $\theta = \theta_{(1,1)}$, $c(\xi) = \cos(\pi\xi)$, 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 (2.5), the feasible region for θ is $[-1, 1]$. The marginal density of X_i $(i = 1, 2)$ is $p(x_i|\theta) = 1 + \frac{\theta^2}{2} c(2x_i)$. The mean and variance of X_i $(i = 1, 2)$ are $1/2$ and $(1/12) + \theta^2/(4\pi^2)$, respectively. The correlation coefficient 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 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 $J_{uu}(\theta)$ of SGM is, for any feasible θ and $u = (1, 1)$,

$$J_{uu}(\theta) = \frac{2(1 - \sqrt{1 - \theta^2})}{\theta^2\sqrt{1 - \theta^2}}. \tag{2.7}$$

The proof is given in the Appendix.

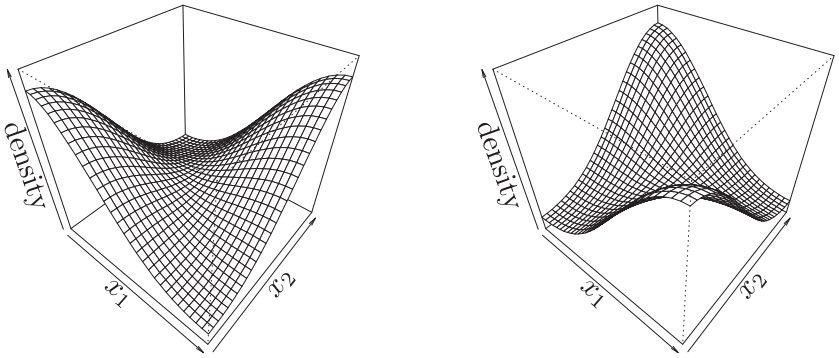


Figure 1. The probability density $p(x|\theta)$ for $\mathcal{U} = \{(1,1)\}$ and $\theta = \theta_{(1,1)} = \pm 0.5$. The correlation coefficient is about ± 0.458 for $\theta = \pm 0.5$, respectively.

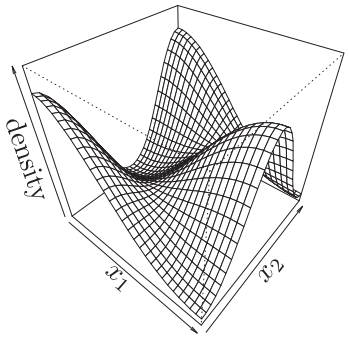


Figure 2. The probability density for $\mathcal{U} = \{(1,2)\}$ and $\theta = 0.2$. The conditional density $p(x_2|x_1)$ is unimodal if x_1 is close to 1, and bimodal if x_1 is close to 0.

Example 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 $1/2$, and therefore the correlation coefficient between X_1 and X_2 is zero. However, the conditional variance of X_2 given X_1 is

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

In order to measure the dependency of X_1, consider

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

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 about 0.4267 at $\theta = -1/5$. Thus SGM can describe more heteroscedastic distributions than MixM. The heteroscedasticity appears in regression analysis, where explanatory and response variables are *a priori* selected.

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

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

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 (2.5). 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):

$$\begin{aligned}\beta_{123}(\theta) &= \frac{\mathrm{E}[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)(X_3 - \mathrm{E}X_3)]}{\sqrt{\mathrm{V}[X_1]\mathrm{V}[X_2]\mathrm{V}[X_3]}} \\ &= \frac{-24\theta/\pi^6 - 1944\theta^3/729\pi^6}{(1/12 + \theta^2/(4\pi^2))^{3/2}}.\end{aligned}$$

The maximum value of $\beta_{123}(\theta)$ over the feasible region $|\theta| \leq 1$ 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 $\beta_{123}(\theta) = -288\sqrt{12}\theta/\pi^6$ with a maximum value over the feasible region $|\theta| \leq 1/3$ of about 0.3459 at $\theta = -1/3$.

Example 6.(Approximate 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 approximate conditional-independence model. Let $m = 3$ and $\mathcal{U} = \{(1, 0, 1), (0, 1, 1)\}$. Then, 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

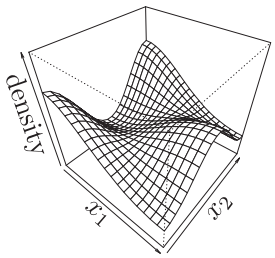
$$\begin{aligned}
&p(x_1, x_2, x_3|\theta, \phi) \\
&= \det \begin{pmatrix} 1 + \theta c_1 c_3 & 0 & -\theta s_1 s_3 \\ 0 & 1 + \phi c_2 c_3 & -\phi s_2 s_3 \\ -\theta s_1 s_3 & -\phi s_2 s_3 & 1 + \theta c_1 c_3 + \phi c_2 c_3 \end{pmatrix} \\
&= 1 + 2\theta c_1 c_3 + 2\phi c_2 c_3 + 3\theta\phi c_1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

Now assume that $\epsilon = \max(|\theta|, |\phi|)$ is close to zero. Then the conditional mutual information is $I_{12|3} = (3/16)\theta^2\phi^2 + \mathrm{O}(\epsilon^5)$. On the other hand MixM, $\tilde{p}(x_1, x_2, x_3|\theta, \phi) = 1 + 2\theta c_1 c_3 + 2\phi c_2 c_3$, has the conditional mutual information $I_{12|3} = (3/4)\theta^2\phi^2 + \mathrm{O}(\epsilon^5)$. The leading term is four times larger than that of SGM.

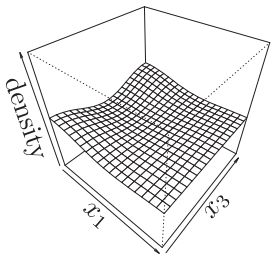
Example 7. We can construct more complicated densities by combining the preceding ones. For example, let $m = 3$, $\mathcal{U} = \{(1, 2, 0), (0, 1, 1), (1, 1, 1)\}$, and $\theta = (0.1, 0.3, 0.2)$. The vector θ is feasible since (2.5) is satisfied. The marginal and conditional 2-dimensional densities are illustrated in Figure 3.

Example 8. From Lemma 4, it is impossible to let every one-dimensional marginal density of $p(x|\theta)$ for any $\theta \in \Theta$ be uniform. Here we construct an example of subsets $\Theta_0 \subset \Theta$ such that $p(x|\theta)$ is a copula for any $\theta \in \Theta_0$. Let $m = 2$ and $\mathcal{U} = \{(2, 0), (1, 1), (0, 2)\}$. Let

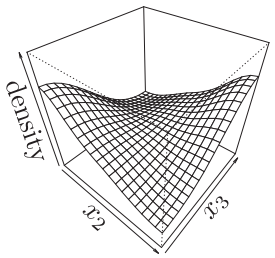
$$\Theta_0 = \{(\theta_{(2,0)}, \theta_{(1,1)}, \theta_{(0,2)}) \in \Theta \mid \theta_{(1,1)} = \gamma, \theta_{(2,0)} = \theta_{(0,2)} = -\frac{\gamma^2}{8} \text{ for some } \gamma \in \mathbb{R}\}.$$



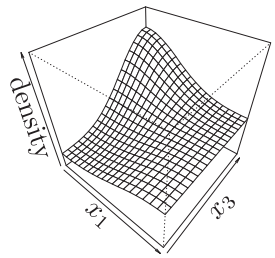
(a) $p(x_1, x_2)$



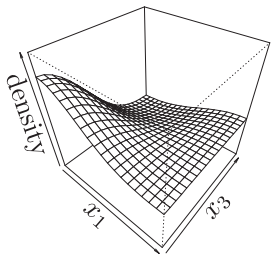
(b) $p(x_1, x_3)$



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

Put $c(\xi) = \cos(\pi\xi)$ and $s(\xi) = \sin(\pi\xi)$. Then for any $\theta = (-\gamma^2/8, \gamma, -\gamma^2/8) \in \Theta_0$, we have

$$\begin{aligned} &p(x_1, x_2|\theta) \\ &= \det \begin{pmatrix} 1 + \gamma c(x_1)c(x_2) + (-\frac{\gamma^2}{2})c(2x_1) & -\gamma s(x_1)s(x_2) \\ -\gamma s(x_1)s(x_2) & 1 + \gamma c(x_1)c(x_2) + (-\frac{\gamma^2}{2})c(2x_2) \end{pmatrix} \\ &= 1 + 2\gamma c(x_1)c(x_2) - \frac{\gamma^3}{2}\{c(2x_1)c(x_1)c(x_2) + c(x_1)c(2x_2)c(x_2))\} \\ &\quad + \frac{\gamma^4}{4}c(2x_1)c(2x_2). \end{aligned}$$

One can confirm that the marginal densities $p(x_1|\theta)$ and $p(x_2|\theta)$ are uniform.

We summarize the examples in Table 1.

3. Maximum Likelihood Estimation of SGM

In this section, we consider the maximum likelihood estimation of SGM. We first formulate it as a robust convex optimization problem.

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

Let $x(1), \dots, x(n)$ 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 problem:

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

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

It is hard to write Θ down explicitly; the difficulty follows from the "for any $\xi \in [0,1]^m$" in its definition. 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)). Hence our problem is a robust convex optimization problem.

In the next subsection, we give a tractable subset Θ^{lit} of Θ that consists of only m constraints. The constrained maximum likelihood estimator over Θ^{lit} is calculated via the determinant-maximization algorithm (Vandenberghe, Boyd, and Wu (1998)). As a by-product of the approach, we obtain a lasso-type estimator since Θ^{lit} is compatible with L^1-constraints.

If $m = 1$, the feasible region Θ 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 $r(x) = \sum_{u=0}^{\infty} r_u \cos(\pi u x)$, its square $r(x)^2$ is non-negative and a Fourier series with Fourier coefficients that are quadratic polynomials in $(r_u)_{u=0}^{\infty}$. However, it is hard to use this representation here since we assume $\theta_u = 0$ for $u \notin \mathcal{U}$, and this restriction is not affine in r_u.

3.1. A conservative region and Lasso-type estimation

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

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

We call Θ^{lit} *the little parameter space*, it is an intersection of m constraints. In the next theorem, we show that Θ^{lit} is a subset of the feasible region Θ. In other words, Θ^{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 $\mathbb{R}_{\mathcal{V}} = \{\theta \in \mathbb{R}^{\mathcal{U}} \mid \theta_u = 0 \text{ if } u \notin \mathcal{V}\}$.

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

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 least $2|\mathcal{U}|$ boundary points of Θ.

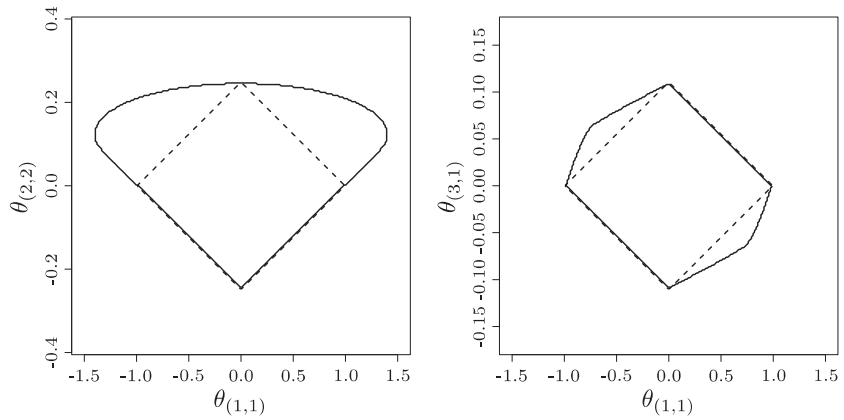
Example 9. Let $m = 2$. The little parameter space for $\mathcal{U} = \{(1,1),(2,2)\}$ is indicated in Figure 4(a), here $\mathcal{U}$ is not linearly independent modulo 2. For this case, we can write

$$\Theta = \left\{\theta \mid |\theta_{(1,1)}| + 4|\theta_{(2,2)}| \leq 1\right\} \cup \left\{\theta \mid (\theta_{(1,1)})^2 \leq 16\theta_{(2,2)}(1 - 4\theta_{(2,2)})\right\}.$$

The expression is the same as the feasible region of autocorrelation parameters of the MA(2) model in time series analysis (see Box and Jenkins (1976), Section 3.4). We also illustrate the regions for another example $\mathcal{U} = \{(1,1),(3,1)\}$ in Figure 4(b).

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 in Tibshirani (1996), in that the regression method is executed with L^1-constraints. Note that Θ^{lit} is also represented by L^1-constraints. Furthermore, we use an indexed set $\Theta_\tau^{\mathrm{lit}}$ with a tuning parameter, $\tau \in [0,1]$,

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



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

Figure 4. The contour of the parameter space Θ (solid line) and the little parameter space Θ^{lit} (dashed line). Here Θ is calculated by a brute-force method.

In particular, $\Theta_0^{\mathrm{lit}} = \{0\}$ and $\Theta_1^{\mathrm{lit}} = \Theta^{\mathrm{lit}}$. We call the constrained maximum likelihood estimator $\hat{\theta}_\tau^{\mathrm{lit}}$ over $\Theta_\tau^{\mathrm{lit}}$ *the lasso-type estimator for SGM*. The tuning parameter τ can be selected by cross validation.

We remark that the feasible region for MixM has the 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 $\tilde{\Theta}^{\mathrm{lit}} \cap \mathbb{R}_{\mathcal{V}} = \tilde{\Theta} \cap \mathbb{R}_{\mathcal{V}}$. The proof is similar to that of Theorem 3 and is omitted here.

Recently, lasso-type estimators for graphical Gaussian models have been proposed by several authors: Yuan and Lin (2007), Banerjee, Ghaoui, and d'Aspremont (2008), and Friedmann, Hastie, and Tibshirani (2008). A sparse density estimation (SPADES) for mixture models is considered in Bunea, Tsybakov, and Wegkamp (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 through numerical examples.

4. Numerical Examples

We give examples of simulations and 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.

We describe some notations and assumptions. We use the following frequency

set for SGM throughout this section:

$$\mathcal{U} = \left\{ u \in (\mathbb{Z}^m_{\geq 0})^+ \mid \|u\|_\infty \leq 2,\ \|u\|_1 \leq 3 \right\}, \tag{4.1}$$

where $\|u\|_\infty = \max_j |u_j|$ and $\|u\|_1 = \sum_j |u_j|$. The elements of $\mathcal{U}$ are $(1,0,\dots,0)$, $(2,0,\dots,0)$, $(1,1,0,\dots,0)$, $(2,1,0,\dots,0)$, $(1,1,1,0,\dots,0)$, and their permutations of the components. The cardinality of $\mathcal{U}$ is $m(m+1)(m+5)/6$. Let $\hat{\theta}^{\mathrm{lit}}_\tau = (\hat{\theta}^{\mathrm{lit}}_{\tau,u})_{u\in\mathcal{U}}$ denote the lasso-type estimator of θ over the region $\Theta^{\mathrm{lit}}_\tau$. The notation on the estimators is 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

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

where $\hat{\Sigma}$ is the sample correlation matrix 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 data $(D_{ti})_{1\leq t\leq n, 1\leq i\leq m}$ valued in $\mathbb{R}^{n\times m}$, we preprocess it before estimation. For Gaussian models, we use the data $\tilde{D}_{ti}$ scaled in 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, Tütüncü, and Todd (2006)).

4.1. Simulation

We first confirm that the lasso-type estimator for SGM described in Section 3 actually works. 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 (4.1) we use for estimation is

$$\mathcal{U} = \begin{pmatrix} 1 & 2 & 0 & 1 & 2 & 0 & 1 & 0 & 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 2 & 2 & 0 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 \end{pmatrix} \tag{4.2}$$

in matrix form, with columns arranged according to lexicographic order. A result of estimation is given in Figure 5, where Figure 5(a) and (b) are the results under the tuning parameter $\tau = 1.0$ and $\tau = 0.5$, respectively. The sample size is $n = 100$ and the number of experiments is 100. The samples were generated by the exact method described in Section 1. For $\tau = 1.0$, the estimator actually distributes around the true parameter. The estimated components $\hat{\theta}_u$ for $u \in \mathcal{U} \setminus \mathcal{U}_0$ become zero at a considerable rate. For $\tau = 0.5$, the shrinkage effect is stronger while the estimated components $\hat{\theta}_u$ for $u \in \mathcal{U}_0$ are more biased. Here the true parameter θ belongs to Θ_1^{lit} but not to $\Theta_{0.5}^{\mathrm{lit}}$.

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

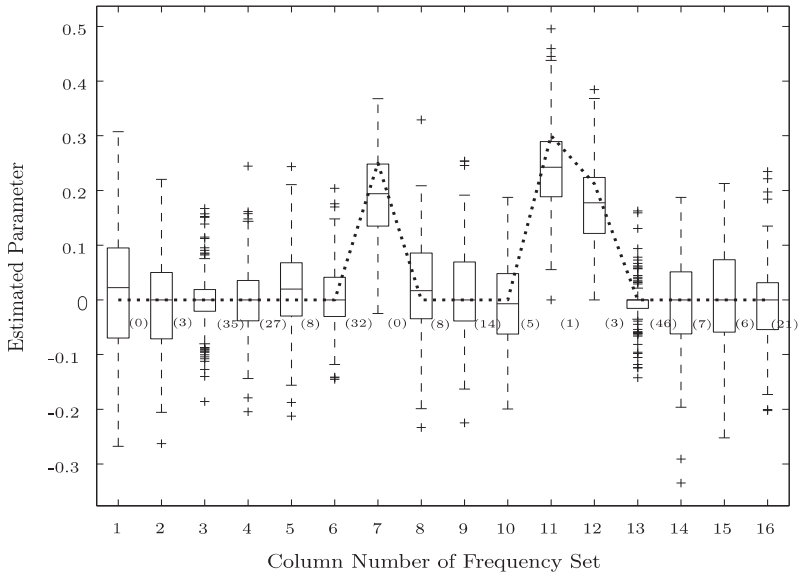
$$p_0(x) = \phi(x_1|0,1)\phi(x_2|x_1,1)\phi(x_3|0,\sigma_3^2(x_2))\phi(x_4,x_5|0,\Sigma_{45}(x_3)), \tag{4.3}$$

where

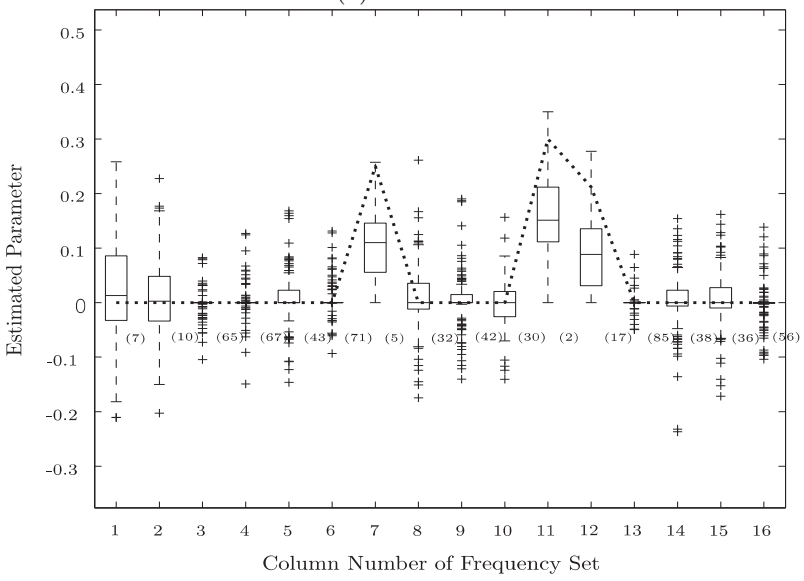
$$\sigma_3^2(x_2) = 1 + \tanh(x_2) \quad \text{and} \quad \Sigma_{45}(x_3) = \begin{pmatrix} 1 & \tanh(x_3) \\ \tanh(x_3) & 1 \end{pmatrix}.$$

By 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. The sampled data is preprocessed before estimation as described above. Remark that the true 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 took the tuning parameter $\tau \in \{i/10\}_{i=0}^{10}$ for SGM and MixM, and $\tau \in \{i/100\}_{i=0}^{100}$ for Gaussian. We used the expected predictive log-likelihood as the index of the predictive performance. The arbitrary constant of the log-likelihood was determined in such a way that the log-likelihood of the null model was zero. The sample size was $n = 40$ for observation and 10 for prediction, the number of experiments was 200. Then the maximum mean predictive log-likelihood of SGM was estimated as $3.37(\pm 0.33)$ at $\tau = 1.0$, where the 95% confidence interval is based on the normal approximation. For MixM and Gaussian, the maximum value was estimated as $1.99(\pm 0.15)$ at $\tau = 1.0$ and $2.72(\pm 0.26)$ at $\tau = 0.32$, respectively. Hence SGM had better predictive performance than MixM or Gaussian.



(a) $\tau = 1.0$.



(b) $\tau = 0.5$.

Figure 5. A simulation of estimation of SGM. The box-plot shows the (normalized) lasso-type estimator $\sqrt{J_{uu}}\hat{\theta}^{\mathrm{lit}}_{\tau,u}$ against $u \in \mathcal{U}$, where (a) $\tau = 1$ and (b) $\tau = 0.5$. The normalizing constant $\sqrt{J_{uu}}$ is the square root of the Fisher information at $\theta = 0$; the horizontal axis denotes $u \in \mathcal{U}$ arranged according to (4.2); the dashed line denotes the true parameter; the sample size is $n = 100$ and the number of experiments is 100. The data set is common to (a) and (b). The number in parentheses is the count of exact-zero estimations out of the 100 experiments.

Table 2. Mean value of the estimators for the five-dimensional data. The tuning parameter τ for each model 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 the 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 ($\pm$0.013)	$(1,1,0,0,0)$	0.123 ($\pm$0.006)	$(1,1,0,0,0)$	0.706 ($\pm$0.011)
$(0,0,1,1,1)$	-0.297 ($\pm$0.017)	$(0,1,2,0,0)$	-0.031 ($\pm$0.005)	$(1,0,0,0,1)$	-0.023 ($\pm$0.023)
$(0,1,2,0,0)$	-0.232 ($\pm$0.015)	$(0,0,1,1,1)$	-0.007 ($\pm$0.003)	$(0,1,1,0,0)$	0.014 ($\pm$0.023)
$(0,0,2,0,0)$	-0.106 ($\pm$0.014)	$(0,0,2,0,0)$	-0.006 ($\pm$0.002)	$(1,0,0,1,0)$	-0.010 ($\pm$0.022)
$(2,0,0,0,0)$	-0.095 ($\pm$0.011)	$(0,2,0,0,0)$	-0.002 ($\pm$0.001)	$(0,1,0,0,1)$	0.008 ($\pm$0.024)
$(0,2,0,0,0)$	-0.084 ($\pm$0.010)	$(1,0,2,0,0)$	-0.002 ($\pm$0.001)	$(0,0,0,1,1)$	-0.007 ($\pm$0.028)
$(0,0,0,0,2)$	-0.043 ($\pm$0.013)	$(2,0,0,0,0)$	-0.001 ($\pm$0.001)	$(0,1,0,1,0)$	0.007 ($\pm$0.024)
$(0,0,0,2,0)$	-0.043 ($\pm$0.010)	$(0,2,0,1,0)$	-0.000 ($\pm$0.001)	$(0,0,1,1,0)$	-0.006 ($\pm$0.023)
$(1,0,2,0,0)$	-0.036 ($\pm$0.009)	$(0,0,1,0,2)$	-0.000 ($\pm$0.001)	$(1,0,1,0,0)$	-0.004 ($\pm$0.021)
$(0,0,0,2,1)$	-0.015 ($\pm$0.015)	$(0,0,0,0,2)$	-0.000 ($\pm$0.001)	$(0,0,1,0,1)$	0.004 ($\pm$0.023)

4.2. A 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 predictive performance was estimated by the 5-fold cross-validated predictive log-likelihood. The tuning parameter examined was $\tau \in \{i/10\}_{i=0}^{10}$ for SGM and MixM, and $\tau \in \{i/100\}_{i=0}^{100}$ for Gaussian. The result shows that for the data our SGM gives slightly better predictive performance than MixM or the Gaussian models. As stated in Edwards (2000), the partial correlation of (x_1, x_3) is not significant. SGM suggests a heteroscedastic effect of x_1 (creatinine clearance) against x_3 (urine flow).

5. Discussion

We defined SGM as a set of potential functions ψ and studied the lasso-type estimator. SGM was applied in both simulations and data. We discuss remaining mathematical and practical problems.

We used the finite Fourier expansion to define the potential function ψ at (2.1). It is sometimes hard to describe the 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 (2.3). For example, assume that we want to describe tail behavior of two-dimensional data around $x = (1, 1)$. Then we can use $\psi(x|\theta, a) = (x_1^2 + x_2^2)/2 + \pi^{-2}\theta(2 +$

Table 3. A result for the digoxin data. The lasso-type estimators of SGM, MixM and the graphical Gaussian model for highlighted values of τ are shown. The exact-zero estimation is not 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	<u>14.54</u>	6.95	12.26	14.49	-0.92

$\cos(\pi x_1)+\cos(\pi x_2))^a$, 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 (2.3). A sufficient condition for convexity of ψ is $0\le\theta\le 2^{1-2a}/a$. If $a<1$, then the tail behavior of $p(x|\theta,a)$ is

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

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.

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: the curvature of SGM at $\theta=0$ is always smaller than that of MixM (2.4). This fact is not so practical but it supports SGM. Since the statement and the proof of this inequality are rather complicated, we present them in a forthcoming paper.

We constructed a lasso-type estimator for SGM as a byproduct of the conservative feasible region in Section 3. Performance of the estimator was numerically studied in Section 4. For the existing lasso estimators, some asymptotic

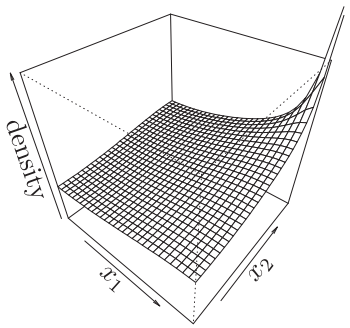


Figure 6. The density function $p(x|\theta, a)$ for $a = 0.75$ and $\theta = 2^{1-2a}/a$.

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, Tsybakov, and Wegkamp (2007), Banerjee, Ghaoui, and d'Aspremont (2008)). We think it important to compare our SGM with the Gaussian, mixture, and exponential models on asymptotics.

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.

Appendix: Proofs

A.1. Proof of Lemma 1

Let ψ have the form (2.1) 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, so we take $\theta \neq 0$. We can extend the domain of $\psi(\cdot|\theta)$ from $[0,1]^m$ to $\mathbb{R}^m$ using (2.1), 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 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 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 set $f(z) = \tilde{\psi}(x_0 + ze)$,

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 analytic 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 the sum of $x^\top x/2$ and a bounded function. Hence (i) is proved. For (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

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

This function is the same form as (2.1) 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 (i).

A.2. Proof of Lemma 2

A statistical model is a mixture model with respect to a given parameter 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 $\mathbb{Z}_i = \{u \in (\mathbb{Z}^m_{\geq 0})^+ \mid u_j = 0 \text{ if } j \neq i\}$. If $\mathcal{U} \subset \mathbb{Z}_i$ for some i, then it is easy to confirm that SGM coincides with MixM. 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 Lemma 3

The score function L_u of SGM at $\theta = 0$ is directly calculated as $L_u = (\partial/\partial\theta_u)\log p(x|\theta)|_{\theta=0} = \|u\|^2 \prod_{j=1}^m \cos(\pi u_j x_j)$. 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 easily calculated.

A.4. Proof of Lemma 4

Let $p(x|\theta)$ be the density function of SGM. For each $i \in \{1, \dots, m\}$, let $x_{-i} = (x_j)_{j \in \{1,\dots,m\} \setminus \{i\}}$ and denote the marginal density of x_i by $r_i(x_i|\theta) = \int_{[0,1]^{m-1}} p(x|\theta) \, \mathrm{d}x_{-i}$. Then $r_i(x_i|\theta)$ is a polynomial with respect to $\theta = (\theta_u)_{u \in \mathcal{U}}$ by the definition of SGM. Note that $r_i(x_i|\theta = 0) = 1$ for any i and x_i because $p(x|\theta = 0) = 1$. We prove by contradiction that there exist i and x_i such that $r_i(x_i|\theta)$ is a non-constant polynomial of θ. Assume that $r_i(x_i|\theta)$ is the constant polynomial 1 for any i and x_i. Fix $u \in \mathcal{U}$. Put $c_j = \cos(\pi u_j x_j)$ and $s_j = \sin(\pi u_j x_j)$ for each j. Put $A_u = \{D^2\psi(x|\theta)\}^{-1}\{\partial/\partial\theta_u (D^2\psi(x|\theta))\}$. Then, for each i, the second derivative of r_i with respect to θ_u at $\theta = 0$ is

$$\begin{aligned}
0 = \left. \frac{\partial^2 r_i(x_i|\theta)}{\partial \theta_u^2} \right|_{\theta=0} &= \int_{[0,1]^{m-1}} \left. \frac{\partial^2 p(x|\theta)}{\partial \theta_u^2} \right|_{\theta=0} \mathrm{d}x_{-i} \\
&= \int_{[0,1]^{m-1}} \left\{ (\mathrm{tr} A_u)^2 - \mathrm{tr}(A_u^2) \right\} \big|_{\theta=0} \, \mathrm{d}x_{-i} \\
&= \int_{[0,1]^{m-1}} \left\{ \sum_{j,k=1}^{m} 1_{\{j \neq k\}} u_j^2 u_k^2 (c_j^2 c_k^2 - s_j^2 s_k^2) \prod_{l \neq j,k} c_l^2 \right\} \mathrm{d}x_{-i} \\
&= 2u_i^2 \Big(\sum_{j \neq i} u_j^2 \Big) (c_i^2 - s_i^2) 2^{-|\sigma(u) \setminus \{i\}|},
\end{aligned}$$

where $\sigma(u) = \{j \mid u_j > 0\}$. Thus we have $u_i^2 = 0$ or $\sum_{j \neq i} u_j^2 = 0$ for each i. This implies $u = (0, \dots, 0, u_\rho, 0, \dots, 0)$ for some $\rho \in \{1, \dots, m\}$. Then we have, however,

$$0 = \left. \frac{\partial r_\rho(x_\rho|\theta)}{\partial \theta_u} \right|_{\theta=0} = \int_{[0,1]^{m-1}} \mathrm{tr}(A_u)|_{\theta=0} \, \mathrm{d}x_{-\rho} = u_\rho^2 c_\rho$$

and $u_\rho = 0$. This contradicts $(0, \dots, 0) \notin \mathcal{U}$. Hence there exist i and x_i such that $r_i(x_i|\theta)$ is a non-constant polynomial of θ. Since the set of zero-points of a given non-zero polynomial has zero Lebesgue measure, we obtain the result.

A.5. Proofs of Equations (2.6) and (2.7)

We first prove (2.6). Let $m = 1$ and $\mathcal{U} = \{u\}$. We only consider the case $u = 1$, the other cases are proved similarly. Put $\theta = \theta_1$. Since $p(x_1|\theta) = 1+\theta\cos(\pi x_1)$, we have $J_{uu}(\theta) = \int_0^1 \cos^2(\pi x_1)/(1+\theta\cos(\pi x_1))\mathrm{d}x_1$. By putting $z = \exp(\mathrm{i}\pi x_1)$, 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 $J_{uu}(\theta) = (1-\sqrt{1-\theta^2})/(\theta^2\sqrt{1-\theta^2})$. This proves (2.6).

We next prove (2.7). Put $u = (1,1)$ and $\theta = \theta_u$. We use the identity

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

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 (2.7) is proved in the same manner as (2.6).

A.6. 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

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

Since the last formula is non-negative definite, we have $\theta \in \Theta$.

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 $D^2(-\pi^{-2}\prod_{j=1}^m \cos(\pi v_j x_j))|_{x=\xi} = (-1)^{v^\top \xi} d(v)^2$. Since $\mathcal{V}$ is linearly independent modulo 2, we can choose $\xi \in \{0,1\}^m$ such that $v^\top \xi = 1_{\{\theta_v > 0\}} \pmod 2$ for all $v \in \mathcal{V}$. Then

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

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

References

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

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

Box G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis – Forecasting and Control*. Holden-Day Inc., San Francisco.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44**, 375-417.

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

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

Edwards, D. (2000). *Introduction to Graphical Modeling*, 2nd edition, Springer-Verlag, New York.

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

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

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

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

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80**, 309-323.

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

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd edition, Springer-Verlag, New York.

Rockafeller, R. T. (1970). *Convex Analysis*. Princeton University Press.

Sei, T. (2007). Gradient modeling for multivariate analysis. In *The Pyrenees International Workshop on Statistics, Probability and Operations Research* (SPO 2007), `http://metodosestadisticos.unizar.es/~jaca2007/Ficheros/actasSPO07web.pdf`. Jaca, Spain.

Sei, T. (2009). Gradient modeling for multivariate quantitative data. *Ann. Inst. Statist. Math.*, to appear.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

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

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

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

Zygmund, A. (2002). *Trigonometric Series*, Volume 2, 3rd edition, Cambridge Mathematical Library.

Department of Mathematics, Faculty of Science and Technology, Keio University, Yagami Campus, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan.

E-mail: sei@math.keio.ac.jp

(Received February 2009; accepted January 2010)