MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Parametric modeling based on the gradient maps of convex functions

Tomonari SEI

METR 2006–51 October 2006

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.i.u-tokyo.ac.jp/mi/mi-e.htm

Parametric modeling based on the gradient maps of convex functions

Tomonari SEI *

October 11, 2006

Abstract

A unified framework of parametric statistical modeling based on the gradient maps of convex functions is presented. The treated data are assumed to be continuous. A class of statistical models called g-flat models is introduced as an affine subspace of the space of gradient maps. This model has many good properties including the concavity of the log-likelihood function. An application to detect the three-dimensional interaction of data is investigated.
Keywords: convex function, exact sampling, g-flat model, gradient representation, perturbation method, three-dimensional interaction.

1 Introduction

In this paper we propose a method of multivariate statistical modeling based on the gradient maps of convex functions.

For introduction let us consider a one-dimensional random variable Y with a cumulative distribution function $Q(y) = \mathrm{P}[Y \leq y]$ and consider a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$. Then the random variable X defined by $g(X) = Y$ has the distribution function

$$F(x) := \mathrm{P}[X \leq x] = \mathrm{P}[Y \leq g(x)] = Q(g(x)).$$

If Q is continuous and strictly increasing, then g is uniquely determined from F since $g(x) = Q^{-1}(F(x))$. Thus, if such a distribution function Q is fixed, a statistical model $\{F_\theta(\cdot) \mid \theta \in \Theta\}$ corresponds one-to-one with a set $\{g_\theta(\cdot) \mid \theta \in \Theta\}$ of strictly increasing functions on $\mathbb{R}$. Remark that any increasing function g on $\mathbb{R}$ is the gradient $\mathrm{d}\psi/\mathrm{d}x$ of a convex function ψ on $\mathbb{R}$, and vice versa.

*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Japan. sei@stat.t.u-tokyo.ac.jp

We generalize this argument to multivariate distributions. Let Y be a random vector subject to an m-dimensional probability distribution Q and let ψ be any convex function on $\mathbb{R}^m$. Then a random variable X is constructed by the unique solution to $(\nabla\psi)(X) = Y$, where $\nabla = (\partial/\partial x_i)_{i=1}^m$. The distribution of X is given by $P(A) = Q((\nabla\psi)(A))$ for any Borel set A. Here a question arises: for any continuous distribution P, is there a convex function ψ such that $P(A) = Q((\nabla\psi)(A))$ for any A? Several researchers showed that this question is positively answered. We state the fact in Section 2. We will call $\nabla\psi$ the gradient representation of P, and show that the gradient representation enables the 'exact sampling' from P as a natural extension of the inverse-function method.

We propose a useful class of statistical models called g-flat model. A g-flat model is a set of probability distributions P_θ whose gradient representation is affine with respect to the parameter θ. The g-flat model has good properties such as concavity of the log-likelihood function, nonnecessity of the normalization constant, description of independency, inclusion of all the multivariate normal distributions and possibility of the conic extension. We describe these properties and compare them with other statistical models in Section 3.

There is an important application of g-flat models. From the practical point of view, a model that describes the three-dimensional interaction of data is needed. However, this interaction is not described by any normal distribution and there is almost no tractable model to analyze it. We give an answer to this problem by using the g-flat model. An example of the distribution with the three-dimensional interaction is given in Section 2 and applied to a real data in Section 3.

We touch on some historical notes. Box & Cox (1964) introduced a class of coordinate-wise transformation like $Y_i = g_i(X_i)$ for $i = 1, \ldots, m$ to deal with non-normal data. This transformation is a special case of our gradient representation. De Oliveira (1997) used the coordinate-wise transformation to the prediction problem together with Bayesian inference. In non-parametric statistics, Easton & McCulloch (1990) generalized the quantile-quantile plot to multivariate data by means of the transportation problem. The transportation problem is closely connected with the gradient representation (see Remark 5).

This paper is organized as follows. We define the gradient representation of probability densities and give three examples in Section 2. The g-flat model is defined in Section 3. In this section we investigate the properties of g-flat models, give numerical results of an information quantity and apply a g-flat model to a real data. We give two addi-

tional theoretical results: one is the symmetry of distributions (Section 4), and another one is asymptotic analysis with respect to small perturbation of the potential function (Section 5). Finally we have some discussion in Section 6.

2 The gradient representation of multivariate distributions

2.1 The existence and uniqueness theorem

In McCann (1995) the following theorem was proved.

Theorem 1 (McCann 1995)**.** Let P and Q be any probability distributions on $\mathbb{R}^m$. Assume that $P(U) = 0$ for any Borel set $U \subset \mathbb{R}^m$ with Hausdorff dimension $m-1$. Then there exists a convex function ψ such that for any Borel set A,

$$Q(A) = P((\nabla\psi)^{-1}(A)). \tag{1}$$

The function ψ is P-a.e. unique up to arbitrary additive constant. □

From Theorem 1 the following definition is consistent.

Definition 2 (gradient representation)**.** Let p and q be the probability density functions of the distributions P and Q, respectively. We call the gradient map $\nabla\psi$ determined by (1) the *gradient representation* of the density p with respect to the *reference density* q. The convex function ψ is called the *potential function*. In this paper the reference density q is assumed to be the standard normal density $\phi(y) = (2\pi)^{-m/2}\exp(-y^\top y/2)$. □

In general the gradient representation of a given density p is not explicitly expressed (see Remark 5 below). Instead we construct a set of the transformations $y = g(x)$ to define a set of densities $p(x)$ (Figure 1). We denote the probability density having the gradient representation g by $p[g]$, which is uniquely determined by change of variables: $p[g](x) = \phi(g(x))\det[\nabla g^\top(x)]$. We recall that the reference density is the standard normal density $\phi(y)$.

Let $\mathrm{C}^2_\smile(\mathbb{R}^m)$ be the set of twice continuously differentiable and strictly convex functions. We define the set of gradient maps onto $\mathbb{R}^m$ by

$$\mathscr{G}_{\mathrm{all}} := \left\{ g : \mathbb{R}^m \to \mathbb{R}^m \mid \exists\psi \in \mathrm{C}^2_\smile(\mathbb{R}^m),\ \forall x \in \mathbb{R}^m,\ g(x) = \nabla\psi(x),\ g(\mathbb{R}^m) = \mathbb{R}^m \right\}.$$

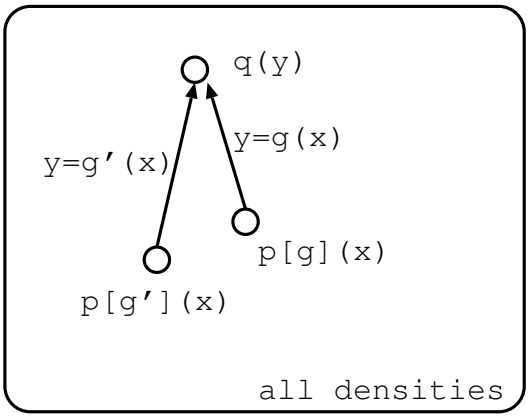


Figure 1: A reference density $q(y)$, gradient maps $g(x)$ and resultant densities $p[g](x)$.

The set $\mathscr{G}_{\mathrm{all}}$ is a subset of the linear space $\mathrm{C}^1(\mathbb{R}^m \to \mathbb{R}^m)$ that consists of all continuously differentiable maps from $\mathbb{R}^m$ to $\mathbb{R}^m$. All the gradient maps g considered in this paper are included in $\mathscr{G}_{\mathrm{all}}$. Although we are not aware of the characterization of the set $\mathcal{P} := \{p[g](x) \mid g \in \mathscr{G}_{\mathrm{all}}\}$ of densities, this class is sufficiently flexible as will be elucidated in the following sections.

Let us prove that $\mathscr{G}_{\mathrm{all}}$ is a convex cone in $\mathrm{C}^1(\mathbb{R}^m \to \mathbb{R}^m)$. We use the following lemma from convex analysis.

Lemma 3 (Rockafeller 1972, Corollary 13.3.1)**.** Let ψ be a differentiable convex function on $\mathbb{R}^m$. Then the range of $\nabla\psi$ is $\mathbb{R}^m$ if and only if ψ is co-finite, in that $\lim_{\lambda\to\infty} \psi(\lambda x)/\lambda = \infty$ for any $x \neq 0$. □

Theorem 4. The set $\mathscr{G}_{\mathrm{all}}$ is a convex cone.

Proof. We show $c_1 g_1 + c_2 g_2 \in \mathscr{G}_{\mathrm{all}}$ for any $g_1, g_2 \in \mathscr{G}_{\mathrm{all}}$ and $c_1, c_2 > 0$. Since there exist convex functions $\psi_1, \psi_2 \in \mathrm{C}^2_\smile(\mathbb{R}^m)$ such that $g_i = \nabla\psi_i$, we have $c_1 g_1 + c_2 g_2 = \nabla(c_1\psi_1 + c_2\psi_2)$. Since ψ_1 and ψ_2 are co-finite, $c_1\psi_1 + c_2\psi_2$ is also co-finite. □

Theorem 4 is used to construct the g-flat model in the next section.

Remark 5. Brenier (1991) showed the existence and uniqueness theorem (Theorem 1) under the condition where the support of P and Q is bounded. Rüchendorf & Rachev (1990) proved existence of ψ for any P and Q with finite second moments ($\int |x|^2 P(\mathrm{d}x) < \infty$ etc.). The latter paper generalized the existence theorem to the case of infinite-dimensional spaces.

Theorem 1 is closely related to the Monge-Kantorovich transportation problem (MKP). This problem is formulated as follows. For given two probability measures P and Q on $\mathbb{R}^m$, find a measure Γ on $\mathbb{R}^m \times \mathbb{R}^m$ that solves the following optimization problem.

$$W(P,Q) = \min \left\{ \int |x-y|^2 \Gamma(\mathrm{d}x, \mathrm{d}y) \mid \Gamma(\mathrm{d}x, \mathbb{R}^m) = P(\mathrm{d}x),\ \Gamma(\mathbb{R}^m, \mathrm{d}y) = Q(\mathrm{d}y) \right\}. \quad (2)$$

The MKP is an infinite-dimensional linear programming problem. The minimum value $W(P,Q)$ is called the Wasserstein metric between P and Q. Rüschendorf & Rachev (1990) proved that for any probability distributions P and Q with finite second moments, a measure Γ is a solution to (2) if and only if $y \in \partial\psi(x)$ Γ-a.e. for some closed convex function ψ. Here $\partial\psi(x)$ denotes the subgradient of ψ (see Rockafeller 1972, p. 214). A similar result is proved by Knott & Smith (1984). There are many references on the MKP. A consulting book is Rachev & Rüschendorf (1998).

One of the statistical methods related to the MKP is the multivariate generalization of the quantile-quantile plot by Easton & McCulloch (1990). Their method is essentially to solve the MKP when P and Q are the empirical measure of given data sets $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$. If the solution to the MKP is close to the identity map, it is concluded that the two data sets are similar.

To find the optimal transformation $g = \nabla\psi$ for given P and Q is difficult in general. Abdellaoui (1998) showed that if the support of Q is finite, the optimal transformation g is explicitly written. He also proposed an algorithm to give a sequence converging to g. For two-dimensional distributions Knott & Smith (1984) gave a procedure to find g that transports the uniform distribution over a bounded set A to the uniform distribution over another bounded set B, by using complex functions. □

2.2 Exact sampling

If we determine the gradient map $g \in \mathscr{G}_{\mathrm{all}}$, then samples drawn from $p[g](x)$ are exactly obtained on the basis of the inverse-function method. A sample X from $p[g](x)$ is obtained by solving $g(X) = Y$, where Y is a sample from the standard normal distribution. Since g is the gradient map of a convex function ψ, we can solve $g(X) = Y$ by solving the convex optimization problem

$$X = \operatorname*{argmin}_{x \in \mathbb{R}^m} \{\psi(x) - Y^\top x\}.$$

The solution exists uniquely for any Y because ψ is co-finite (see Lemma 3).

An independently and identically distributed (i.i.d.) sequence $\{X(i)\}_{i=1}^n$ is simultaneously obtained by solving

$$(X(1), \ldots, X(n)) = \operatorname*{argmin}_{(x(1),\ldots,x(n))} \sum_{t=1}^{n} \{\psi(x(t)) - Y(t)^\top x(t)\},$$

where $\{Y(t)\}_{t=1}^n$ is an i.i.d. sequence from the standard normal distribution. This procedure is valuable for the vector-oriented programming languages like R and MATLAB.

2.3 Examples

We give three examples having the explicit gradient representation. The first example is the distribution of three-dimensional interaction, the second one is the distribution quite different from the normal distribution and the third one is related to an electric circuit. To generate samples, we use the exact sampling described in the preceding subsection.

Example 1 (Distributions with three-dimensional interaction)**.** Let $m = 3$ and define

$$\psi(x) = \frac{x^\top x}{2} + \epsilon \sum_{\lambda=1}^{4} \arctan(e_\lambda^\top x),$$

where

$$(e_1, e_2, e_3, e_4) \;=\; \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \tag{3}$$

and $\epsilon \in \mathbb{R}$ is a small number such that convexity of ψ is assured. A result of numerical experiments when ϵ ranges over $\{0.00, 0.05, \ldots, 0.35\}$ is shown in Figure 2. The result shows that the third cumulant of $p[\nabla\psi](x)$ is unignorable. In Section 5 an asymptotic expression of the third cumulant is obtained under the limit as $\epsilon \to 0$. □

Example 2 (Curtain-type distribution)**.** Let $m = 2$ and define $\psi(x)$ by

$$\psi(x_1, x_2) = \begin{cases} (2t)^{-1} \left(-q|x_2| + \sqrt{(q^2+1)x_2^2 + x_1^2}\right)^2 & \text{if } |x_2| \geq |x_1|, \\ (2/t)^{-1} \left(q|x_1| + \sqrt{x_2^2 + (q^2+1)x_1^2}\right)^2 & \text{if } |x_2| < |x_1|, \end{cases}$$

where $t = 0.05$ and $q = (1-t)/\sqrt{2t}$. Samples drawn from the density $p[\nabla\psi](x)$ and the contour $\{x \mid \psi(x) = 1\}$ are shown in Figure 3. The curvature radius r of the contour is $(1+t^2)^{1/2}$ if $|x_2| > |x_1|$ and $(1+t^{-2})^{1/2}$ if $|x_2| < |x_1|$. □

Example 3 (Diode distribution)**.** Let $m = 2$ and define

$$\psi(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) + f(x_1 - x_2), \quad f(u) = \frac{u^2}{2}\mathbb{I}_{\{u \geq 0\}},$$

where $\mathbb{I}_{\{u>0\}}$ denotes the indicator function of the set $\{u \in \mathbb{R} \mid u > 0\}$. Scatter plot of samples are shown in Figure 4 (a). There is a boundary along the line $x_2 = x_1$ due to the irregularity of the function $f(u)$. We can easily resolve this unrealistic boundary by replacing $f(u)$ with a sufficiently smooth function, for example, $(\log(\exp(u) + 1))^2$. We call this distribution the diode distribution since there is an electric circuit with a diode that attains the characteristics of this transformation (see e.g. Murota 2003, Section 2.2 for the relation between electric circuits and convex functions). Let us consider the circuit given in Figure 4 (b). Assume that the diode is ideal: the diode connects its end-points if $V_1 > V_2$, and disconnects otherwise. Let the resistance values be $R_a = R_b = R_c = 1$. Then the circuit equation is

$$\begin{cases} I_1 &= V_1 + (V_1 - V_2)\mathbb{I}_{\{V_1 > V_2\}}, \\ I_2 &= -(V_1 - V_2)\mathbb{I}_{\{V_1 > V_2\}} + V_2. \end{cases}$$

The right hand side is the gradient of $\psi(V_1, V_2) = \frac{1}{2}(V_1^2 + V_2^2) + \frac{1}{2}(V_1 - V_2)^2\mathbb{I}_{\{V_1 > V_2\}}$. Hence, if (I_1, I_2) is normally-distributed, the samples of (V_1, V_2) are distorted as Figure 4 (a). □

3 The g-flat model

3.1 Definition and properties

We denote the probability density having the gradient representation g by $p[g]$. The Jacobian matrix of g is denoted by $G = \nabla g^\top$. The density $p[g]$ is explicitly expressed as $p[g](x) = \phi(g(x)) \det G(x)$, where $\phi(y) = (2\pi)^{-m/2} \exp(-y^\top y/2)$. Recall that $\mathscr{G}_{\text{all}}$ is the set of all the continuously differentiable gradient maps of convex functions.

Definition 6 (g-flat model)**.** A statistical model is called a *g-flat model* if it is represented as

$$\mathcal{M} = \{p[g](x) \mid g \in \mathscr{G}\}, \quad \mathscr{G} = \{g \mid g = \sum_{a=1}^{p} \theta^a g_a,\ (\theta_a) \in \Theta\} \subset \mathscr{G}_{\text{all}},$$

where $(g_a)_{a=1}^p$ is a set of gradient maps and Θ is a convex subset of $\mathbb{R}^p$. □

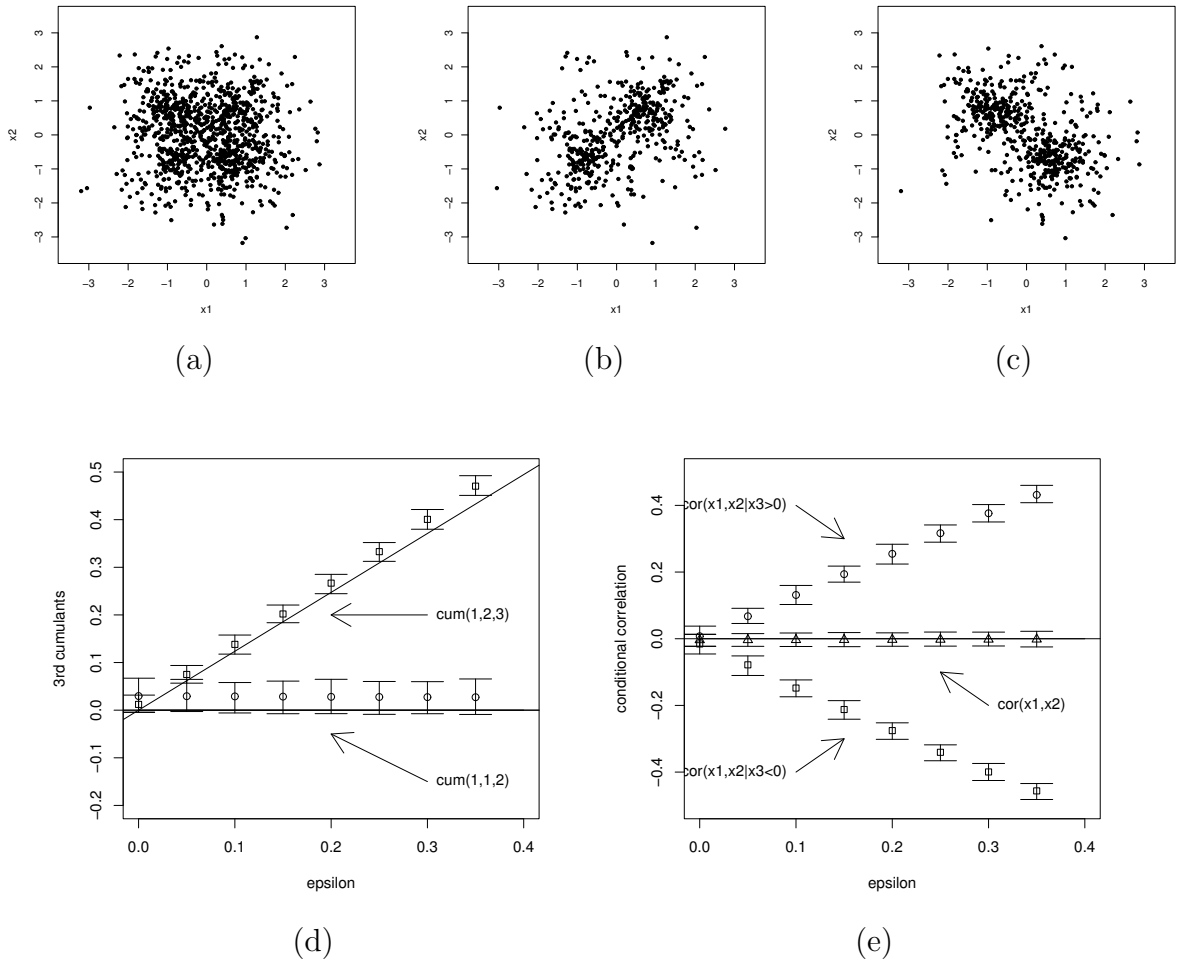
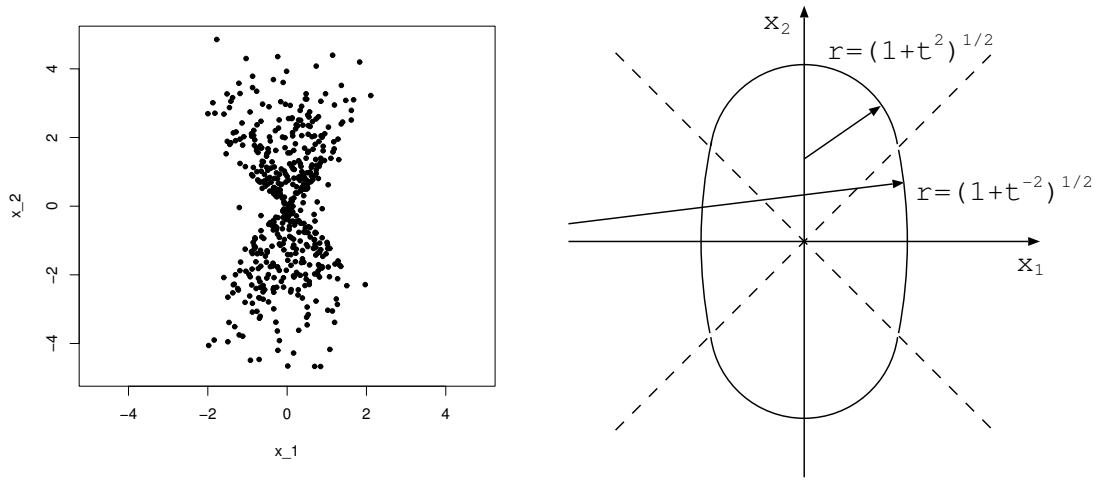
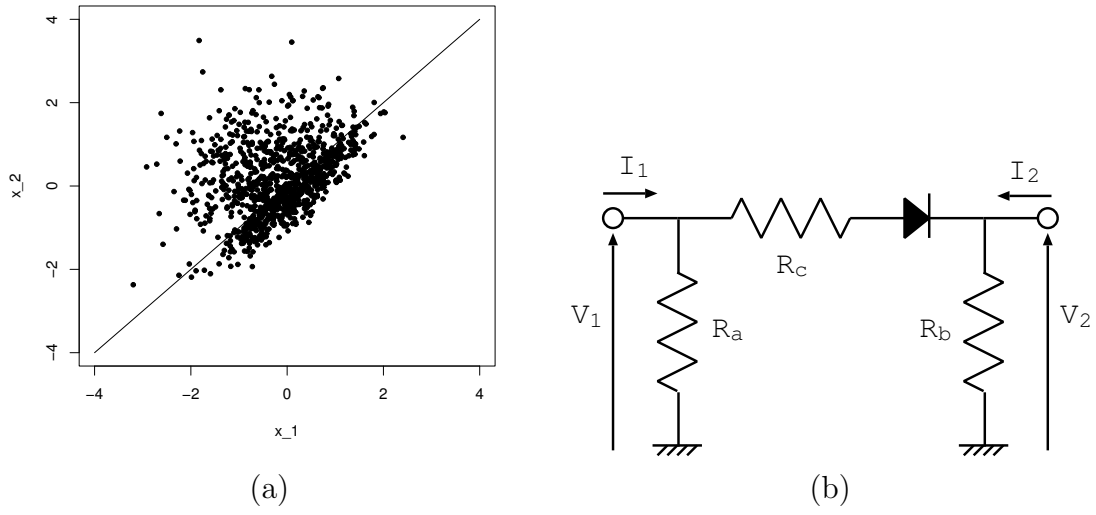


Figure 2: A simulation on the distribution with three-dimensional interaction. (a) Scatter plot of samples drawn from $p(x_1, x_2)$. The small parameter ϵ is 0.35. The sample size is 1000. (b) Scatter plot of $p(x_1, x_2|x_3 > 0)$. (c) Scatter plot of $p(x_1, x_2|x_3 < 0)$. (d) The third cumulants κ_{112} and κ_{123} against $\epsilon \in \{0.00, 0.05, \ldots, 0.35\}$. The sample size is 10000. The 95% confidence interval is based on the bootstrap method. The straight line (1.237ϵ) is obtained by asymptotic analysis in Subsection 5.2. (e) The marginal correlation $\mathrm{Cor}[x_1, x_2]$ and the conditional correlations $\mathrm{Cov}[x_1, x_2|x_3 \geq 0]$ and $\mathrm{Cov}[x_1, x_2|x_3 < 0]$ against $\epsilon \in \{0.00, 0.05 \ldots, 0.35\}$.



(a). (b).

Figure 3: A simulation on the curtain-type distribution. (a) Scatter plot of 500 samples. (b) The contour of the potential function.



(a) (b)

Figure 4: A simulation on the diode distribution. (a) Scatter plot of 500 samples. The solid line is $x_2 = x_1$. (b) An electric circuit satisfying $I = \nabla\psi(V)$.

Remark 7. The convex subset Θ is typically written as the first quadrant $\Theta = \mathbb{R}^m_{\geq 0}$ or the simplex $\Theta = \mathbb{R}^m_{\geq 0} \cap \{\sum_{a=1}^p \theta^a = 1\}$. But these sets are restrictive in some situations. For example, the normal model $\{N(\mu, \Sigma)\}$ is expressed in the gradient representation as

$$\mathcal{M} = \{p[g](x) \mid g(x) = Ax + b,\ A \in \mathrm{Sym}_+(\mathbb{R}^m),\ b \in \mathbb{R}^m\}$$

where $\mathrm{Sym}_+(\mathbb{R}^m)$ is the set of all positive definite matrices. □

The following theorem is one of the motivations to use the g-flat model.

Theorem 8. Let $\mathcal{M} = \{p[g](x) \mid g \in \mathscr{G}\}$ be a g-flat model. Then the log-likelihood function is concave with respect to g.

Proof. It is sufficient to prove that $\theta \in (0,1) \mapsto \log p[g_\theta](x)$ is concave, where $g_\theta(x) = (1-\theta)g_0(x) + \theta g_1(x)$ is the convex combination of arbitrary gradient functions g_0 and g_1 in $\mathscr{G}$. The logarithm of $p[g_\theta]$ is given by

$$\log p[g_\theta] = -g_\theta^\top g_\theta/2 + \log\det(G_\theta)$$

and therefore

$$\partial_\theta^2 \log p[g_\theta] = -(g_1 - g_0)^\top (g_1 - g_0) - \mathrm{tr}[G_\theta^{-1}(G_1 - G_0)G_\theta^{-1}(G_1 - G_0)] \leq 0.$$

The equality holds if and only if $g_1 = g_0$ and $G_1 = G_0$. □

From this theorem, if $\hat{\theta}$ is a local maximal point of $\log p[g_\theta]$, then it is actually the unique global maximal point. The numerical computation of the maximum likelihood estimator (MLE) is relatively simple. The penalized likelihood $\log p[g_\theta] - \lambda\,\mathrm{pen}(\theta)$ is also concave whenever the penalty term $\mathrm{pen}(\theta)$ is convex with respect to θ.

In the following, we enumerate the other properties of the g-flat models. We first remark that the multivariate normal model can be combined with any g-flat model. Consider any g-flat model $\mathcal{M} = \{p[g](x) \mid g(x) = \sum_{a=1}^p \theta^a g_a(x),\ \theta \in \Theta\}$. Then we can combine $\mathcal{M}$ with the normal model by putting

$$\begin{aligned}\mathcal{M}' \ = \ & \{p[g](x) \mid g(x) = Ax + b + \sum_{a=1}^p \theta^a g_a(x), \\ & \theta \in \Theta,\ A \in \mathrm{Sym}_+(\mathbb{R}^m),\ b \in \mathbb{R}^m\}\end{aligned}$$

This property of the g-flat model is particularly important for multivariate analysis because many statistical methods for multivariate data are based on the normal model and our g-flat model may enable to extend the methods.

Next we compare the g-flat model with the exponential family. Recall that a model $\mathcal{M}$ is called an exponential family (or *e-flat family*) if $\mathcal{M}$ is written as

$$\mathcal{M} = \{p_\theta(x) = c(x)\exp(\sum_{a=1}^{p}\theta^a t_a(x) - \psi(\theta)) \mid \theta \in \Theta\},$$

where $(t_a(x))$ is a set of functions and $\psi(\theta)$ is the normalizing constant that guarantees $\int p_\theta(x)\mathrm{d}x = 1$. For given $(t_a(x))$, to compute $\psi(\theta)$ is difficult in general and the Markov Chain Monte Carlo (MCMC) method is needed. Although the MCMC method is powerful especially in Bayesian analysis, it sometimes forces high-cost computations. On the other hand, the g-flat model is normalization-free so that the probability density is explicitly expressed.

Let us focus on the independence of two or more random variables. In the gradient representation, independence of variables is described by the (infinite-dimensional) affine subset in the set of all the gradient functions. Assume that $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ and denote the gradient representation of p, p_1 and p_2 by g, g_1 and g_2, respectively. Then we obtain $g(x_1, x_2) = g_1(x_1) + g_2(x_2)$. Remark that this property does not hold if a mixture family is considered. Recall that a model $\mathcal{M}$ is called a mixture (or *m-flat*) family if $\mathcal{M}$ is written as $\mathcal{M} = \{p \mid p(x) = \sum_{a=1}^{p}\theta^a p_a(x),\ \theta \in \Theta\}$.

We proceed to consider the conditional independence of two or more random variables. Unfortunately, the conditional independence is not described by affine subset in the gradient representation. For example, let $g(x) = Ax$ with a positive definite matrix A. Then x has the distribution $N(0, A^{-2})$. For the case where $x \sim N(0, \Sigma)$, it is widely known that x_1 and x_2 is conditionally independent given $(x_3, \cdots, x_m)$ if and only if $(\Sigma^{-1})_{12} = 0$. This condition is given by a non-affine relation

$$(A^2)_{12} = A_{11}A_{12} + A_{12}A_{22} + \cdots + A_{1m}A_{m2} = 0$$

in terms of A.

Finally we point out that any g-flat model has the conic extension. If the g-flat model is given by $\mathcal{M} = \{p[g](x) \mid g \in \mathscr{G}\}$, then the model $\widetilde{\mathcal{M}}$ defined by the following formula is also a g-flat model:

$$\widetilde{\mathcal{M}} := \{p[g](x) \mid g \in \widetilde{\mathscr{G}}\}, \quad \widetilde{\mathscr{G}} = \{g \mid g = a\bar{g} + b,\ \bar{g} \in \mathscr{G},\ a > 0,\ b \in \mathbb{R}^m\}.$$

We call $\widetilde{\mathcal{M}}$ the *conic extension* of $\mathcal{M}$. The conic extension makes the model flexible. For any $x_0 \in \mathbb{R}^m$ there exists a sequence $\{g_n\}_{n=1}^{\infty}$ in $\widetilde{\mathscr{G}}$ such that the distribution $p[g_n](x)\mathrm{d}x$

converges weakly to the Dirac distribution $\delta_{x_0}(\mathrm{d}x)$. In fact, take any gradient map $g \in \mathscr{G}$ and consider the sequence $g_n(x) = n(g(x) - g(x_0))$ in $\widetilde{\mathscr{G}}$. We remark that any e-flat models have a similar extension called the *exponential dispersion model* in that the dispersion parameter plays a role of our scaling parameter $a > 0$ (e.g. Jørgensen 1987). However the exponential dispersion model is not e-flat with respect to the dispersion parameter in general.

We summarize these properties in Table 1.

Table 1: Properties of models.

Advantages	g-flat model	m-flat model	e-flat model
Log-likelihood is concave	◯	◯	◯
All normal distributions can be included	◯	–	◯
Normalization is not needed	◯	◯	–
Independency can be described	◯	–	◯
Conditional independency can be described	–	–	◯
Conic extension is available	◯	–	–

3.2 Numerical evaluation of information loss

In this subsection, we only consider one-parameter model for simplicity. We evaluate the statistical curvature $\gamma(\theta)$ (Efron 1975) defined by

$$\begin{aligned} \gamma(\theta) &= \mathrm{E}[(\partial_\theta \partial_\theta \ell - \Gamma(\theta) J(\theta)^{-1} \partial_\theta \ell + J(\theta))^2]/(J(\theta)^2), \\ \Gamma(\theta) &= \mathrm{E}[(\partial_\theta \partial_\theta \ell)(\partial_\theta \ell)], \\ J(\theta) &= \mathrm{E}[(\partial_\theta \ell)^2], \end{aligned}$$

where ℓ is the log likelihood function $\ell = \log p_\theta(x)$. The quantity $\gamma(\theta)$ represents (asymptotically) the information loss when the given data is compressed into the maximum likelihood estimator. The e-flat models satisfy $\gamma(\theta) = 0$. We compare $\gamma(\theta)$ of g-flat models and that of m-flat models. Let g_0 and g_1 be two gradient maps. The g-flat model connecting the two maps is $p[(1-\theta)g_0 + \theta g_1](x)$ and the m-flat model is $(1-\theta)p[g_0](x) + \theta p[g_1](x)$, where $\theta \in [0, 1]$.

Example 4 (Normal and 3dimensional interaction distributions)**.** Let us consider $\psi_0(x) =$

$x^\top x/2$ and

$$\psi_1(x) = \frac{x^\top x}{2} + 0.2 \sum_{\lambda=1}^{4} \arctan(e_\lambda^\top x).$$

The vectors e_λ are defined in Example 1. We evaluate the curvature $\gamma(\theta)$ of the model connecting two of the three densities by the Monte-Carlo method. The result is shown in Figure 5. In the case, the g-flat model has less curvature than the m-flat model for all of $\theta \in [0, 1]$. □

Example 5 (Normal, curtain-type and diode distributions)**.** Let us consider three densities: the standard normal density, the curtain-type density (Example 2) and the diode density (Example 3). We evaluate the curvature $\gamma(\theta)$ of the model connecting two of the three densities by the Monte-Carlo method. The result is shown in Figure 6. In each case, the g-flat model has less curvature than the m-flat model for almost values of θ. There is an interesting property, in that the curvature of the g-flat model connecting the normal and diode densities is constant 1. In fact, by direct calculations, we can obtain $J(\theta) = 4(1 + 2\theta)^{-2}$, $\Gamma(\theta) = 8(1 + 2\theta)^{-3}$ and $\gamma(\theta) = 1$. □

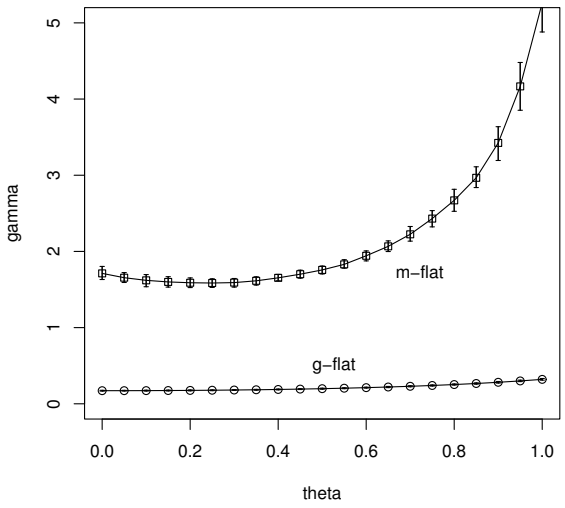
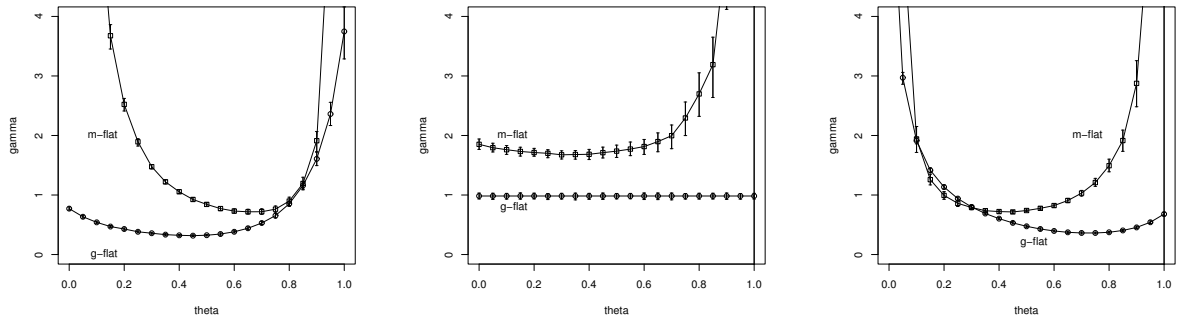


Figure 5: Statistical curvature of the g-flat and m-flat models connecting the normal and 3-dimensional interaction densities. The sample size is 10000 for each experiment and the 95% confidence interval is based on the bootstrap method.



(a) (b) (c)

Figure 6: Statistical curvature of the g-flat and m-flat models connecting (a) the normal and curtain-type densities, (b) the normal and diode densities, (c) the curtain-type and diode densities, respectively. The sample size is 10000 for each experiment and the 95% confidence interval is based on the bootstrap method.

3.3 Application to real data

We apply the g-flat model to detect three-dimensional interaction of a real data set. We use the data of decathlon (Miyakawa 1997). The data consist of 10 variables $\{X_i\}_{i=1}^{10}$ (100m, long-jump, shot-put, high-jump, 400m, 110m-hurdle, disc-throw, pole-vault, javelin-throw and 1500m) by 50 samples (50 athletes). The data are preprocessed before analysis such that the sample mean and variance of each variable are 0 and 1, respectively. We consider each marginal density $p(X_i, X_j, X_k)$ $(1 \leq i < j < k \leq 10)$, not the joint density $p(X_1, \ldots, X_{10})$, for simplicity. The empirical third cumulant is shown in Figure 7 (a). The model of three-dimensional interaction as Example 1 is used. The potential function is

$$\psi(x) = \frac{1}{2} x^\top A x + \theta \sum_{\lambda=1}^{4} \arctan(e_\lambda^\top x), \quad x = (x_i, x_j, x_k)^\top$$

where $\{e_\lambda\}$ is defined by (3), and $A \in \mathrm{Sym}_+(\mathbb{R}^m)$ and $\theta \in \mathbb{R}$ are unknown parameters.

We calculate Akaike's information criterion (AIC) of the two submodels $\theta = 0$ and $\theta \neq 0$ for all the triplets $\{(i, j, k)\}_{1 \leq i < j < k \leq 10}$ of the ten events. The result is shown in Figure 7 (b). The triplet having the most significant difference value is (X_4, X_5, X_6) (that denotes high-jump, 400m and 110m hurdle). The estimated potential is

$$\psi(x_4, x_5, x_6) = \frac{1}{2} x^\top \begin{pmatrix} 1.054 & -0.094 & 0.137 \\ -0.094 & 1.138 & -0.307 \\ 0.137 & -0.307 & 1.148 \end{pmatrix} x + 0.186 \sum_{\lambda=1}^{4} \arctan(e_\lambda^\top x).$$

Although the empirical third cumulant 0.204 of (X_4, X_5, X_6) is not so large, we find that two empirical conditional correlations are quite different: $\mathrm{Cor}[X_5, X_6|X_4 > 0] = 0.688$ and $\mathrm{Cor}[X_5, X_6|X_4 < 0] = -0.049$. The scatter plots shown in Figure 7 (b)–(d) also support the result. The consequence is that if an athlete is a good high-jumper, the scores of 400m and 110m hurdle are positively correlated, otherwise the two scores have almost no correlation. Remark that this result is never detected when only the normal distribution is used.

Of course there are naive non-parametric methods that detects the three-dimensional interaction by using some test statistics, e.g. the empirical third cumulant. However, if these naive methods are used, the predictive inference like the plug-in prediction $p_{\hat{\theta}}(\cdot)$, where $\hat{\theta}$ is the MLE, is not available.

4 Symmetric distributions

We discuss the gradient representation of symmetric distributions. In particular stationary and periodic time series is characterized by its gradient representation.

4.1 Invariance under orthogonal transformations

We first note that some transformation of the random variable is directly reflected in the transformation of the potential functions. Denote the orthogonal group on $\mathbb{R}^m$ by $O(\mathbb{R}^m)$.

Lemma 9. Let $U \in O(\mathbb{R}^m)$, $a > 0$ and $c \in \mathbb{R}^m$. Denote the potential of the density $p(x)$ by $\psi(x)$. Then the potential of the density $p(Ux)$, $p(ax)a^n$ and $p(x-c)$ is given by $\psi(Ux)$, $a^{-1}\psi(ax)$ and $\psi(x-c)$, respectively.

Proof. Let $g = \nabla\psi$ and $G = \nabla\nabla^\top\psi$. Then

$$\log p[g](Ux) = -\frac{1}{2}g(Ux)^\top g(Ux) + \log\det(G(Ux)).$$

Let $\tilde{\psi}(x) := \psi(Ux)$. Then $\nabla\tilde{\psi}(x) = U^\top g(Ux)$ and $\nabla\nabla^\top\tilde{\psi}(x) = U^\top G(Ux)U$. Hence

$$\log p[\nabla\tilde{\psi}](x) = -\frac{1}{2}g(Ux)^\top g(Ux) + \log\det(G(Ux)).$$

Thus $\tilde{\psi}$ is the potential function of $p(Ux)$. The other statements are similarly proved. □

Remark 10. Generalization of Lemma 9 to all the affine transformations or more general transformations is difficult. If g is the identity map and T is an invertible linear map, then

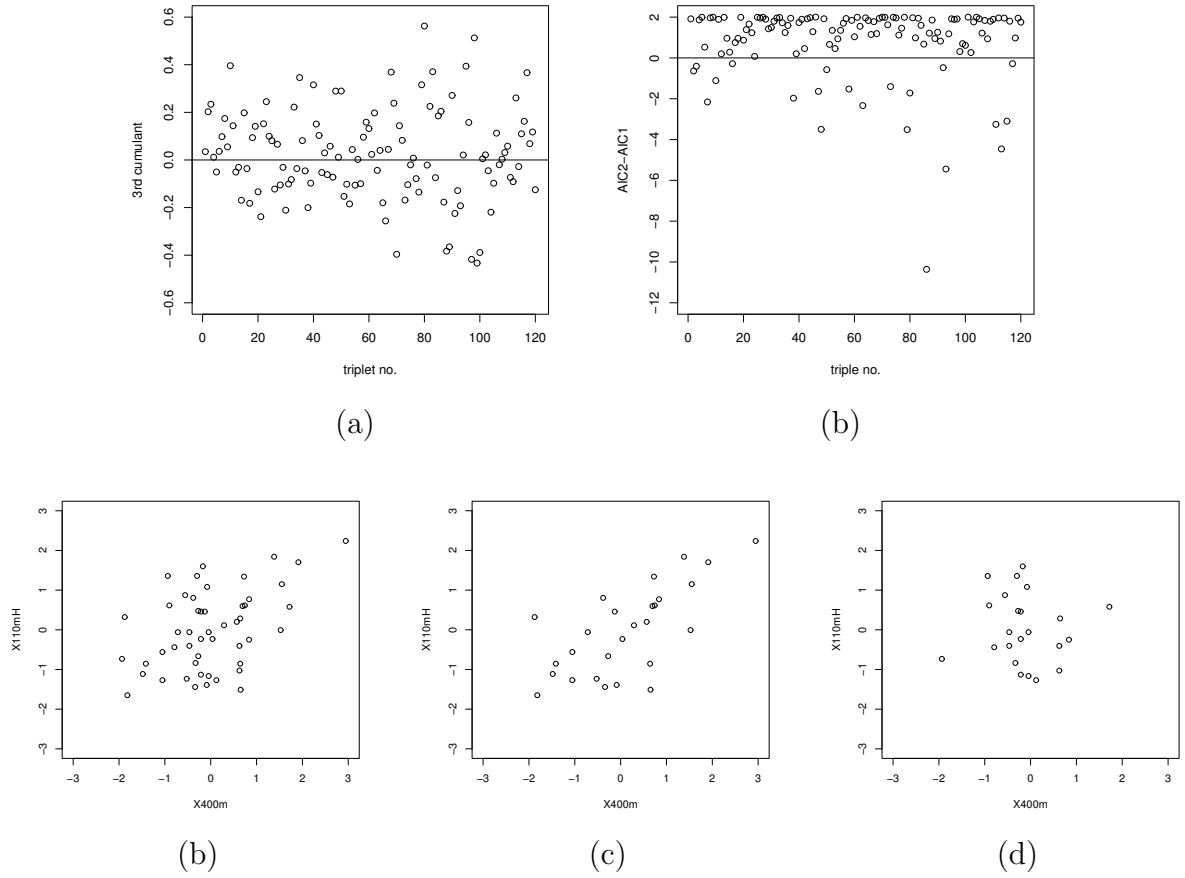


Figure 7: Detection of three-dimensional interaction on the decathlon data. (a) The empirical third cumulant for each triplet. The horizontal axis represents the 120 triplets arranged as $(1,2,3),(1,2,4),\ldots,(7,8,9)$. (b) Difference between AIC of the model $\theta \neq 0$ and $\theta = 0$. The point under the horizontal line implies that the model $\theta \neq 0$ is selected. The horizontal axis represents the 120 triplets. The 86th triplet $(4,5,6)$ has the most significant value. (b) Scatter plot of X_5 (horizontal) v.s. X_6 (vertical). The correlation is 0.465. (c) Scatter plot of X_5 v.s. X_6 conditioned by $X_4 > 0$ (the correlation is 0.688). (d) Scatter plot of X_5 v.s. X_6 conditioned by $X_4 < 0$ (the correlation is -0.049).

to find the potential function ψ corresponding to $\phi(Tx)\det(T)$ is equivalent to determine the polar factorization $T = UA$, where U is an orthogonal matrix and A is a positive definite matrix, because $\phi(Tx)\det(T) = \phi(Ax)\det(A)$. If g is not affine, the explicit form of ψ is not available. Brenier (1991) investigated the existence and uniqueness theorem (Theorem 1) from the view point of generalization of the polar factorization. □

The following theorem is immediately derived from Lemma 9.

Theorem 11. Let H be any subgroup of $O(\mathbb{R}^m)$. Denote the potential of the density $p(x)$ by $\psi(x)$. Then $p(x)$ is H-invariant if and only if $\psi(x)$ is H-invariant. □

Example 6 (Spherically symmetric distributions)**.** Consider a probability density function written as $p(x) = \pi(|x|)$, where $|x|$ denotes the Euclidean norm of x and π is a function on $\mathbb{R}_{\geq 0}$. From Theorem 11 the potential function is written as $\psi(x) = f(|x|)$ with some function f on $\mathbb{R}_{\geq 0}$. The gradient $\nabla\psi = (f'(|x|)/|x|)x$ is called a *radial transformation* (Rachev & Rüschendorf 1998, Example 3.2.16). We prove that ψ is convex if and only if $f'(r) > 0$ and $f''(r) > 0$ for all $r > 0$. In fact, the Hessian matrix of ψ is given by

$$\nabla\nabla^\top\psi(x) = \frac{f'(r)}{r}(I - ee^\top) + f''(r)ee^\top,$$

where $r = |x|$, $e = x/|x|$ and I is the identity matrix. The eigenvalues of $\nabla\nabla^\top\psi(x)$ are $f'(r)/r$ and $f''(r)$. Thus the Hessian matrix is positive definite if and only if $f'(r) > 0$ and $f''(r) > 0$.

We can generalize this result. Let us partition x into (x_A, x_B) and consider a probability density function written as $p(x) = \pi(|x_A|, |x_B|)$. The potential corresponding to the density $p(x)$ is written as $\psi(x) = f(|x_A|, |x_B|)$. The function ψ is convex if and only if $f(r_A, r_B)$ is a two-dimensional convex function and $\partial f/\partial r_A > 0$ and $\partial f/\partial r_B > 0$. □

Example 7 (Exchangeable distributions)**.** If a distribution is invariant under the symmetric group that consists of all permutations, then the distribution is called an *exchangeable distribution*. A distribution is exchangeable if and only if its potential function is invariant under the symmetric group. The density with 3-dimensional interaction discussed in Example 1 is an example of exchangeable densities. □

4.2 Stationary and periodic time series

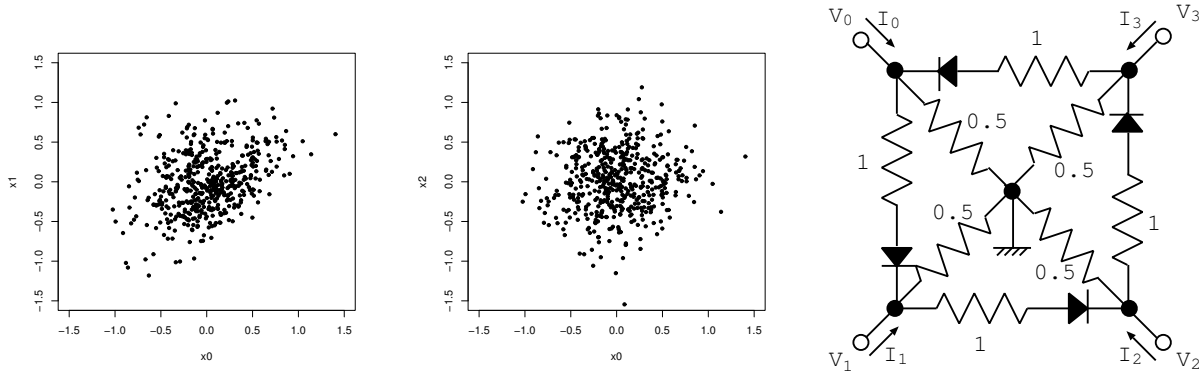
A random vector $X = (X_0, \cdots, X_{n-1})$ is called *stationary and periodic* if its distribution is invariant under the translation group $\tau^s : (X_t) \mapsto (X_{t-s})$, where $t - s$ is considered as

$t - s$ modulo n. Let ψ be the potential function that determines the density of X. Then the distribution of X is stationary and periodic if and only if ψ is invariant under the translation group. Let ψ be written as

$$\psi(x) \;=\; \sum_{s=0}^{n-1} f(\tau^s x), \tag{4}$$

where f is a convex function. This ψ is translation invariant since $\psi(\tau x) = \psi(x)$.

Example 8. Let $n = 4$. Put $f(x_0, x_1, x_2, x_3) = (x_0^2 + x_1^2)/2 + ((x_0 - x_1)^2/2)\mathbb{I}_{\{x_0 > x_1\}}$ like Example 3 and let ψ be given by (4). Samples drawn from the marginal densities $p(x_0, x_1)$ and $p(x_0, x_2)$ are drawn in Figure 8 (a) and (b). The potential is not invariant under inversion of time. In fact $\psi(0, 1, 2, 3) = 11.5$ and $\psi(3, 2, 1, 0) = 8.5$. The Kullback-Leibler (KL) divergence $\int p(x) \log(p(x)/p(\rho(x)))\mathrm{d}x$, where ρ denotes inversion of time, is estimated to $[0.88, 0.95]$ (95% interval). For comparison $\int p(x) \log(p(x)/\phi(x))\mathrm{d}x$ is estimated to $[2.57, 2.64]$ (95% interval). Note that any stationary and periodic Gaussian density is invariant under inversion of time. An electric circuit attaining this transformation is obtained by connecting the circuit of Example 3 as shown in Figure 8 (c). This example is clearly generalized to the case of arbitrary $n \geq 3$. □



(a) (b) (c)

Figure 8: A stationary and periodic distribution ($n = 4$). (a) Samples drawn from $p(X_0, X_1)$, (b) Samples drawn from $p(X_0, X_2)$ (c) An electric circuit satisfying $I = \nabla\psi(V)$.

A natural extension is to characterize the stationary process $X = (X_t \mid t \in \mathbb{Z})$, where $\mathbb{Z}$ stands for the set of all integers. The process X is called stationary if any finite-dimensional marginal distribution of X is translation invariant. However the gradient representation for the infinite-dimensional data needs further investigation and it is not discussed here.

5 Asymptotic analysis

We derive asymptotic expansion of the probability density functions and their cumulants under the limit where a positive parameter ϵ governing the potential ψ_ϵ converges to zero. In Subsection 5.1 we derive an asymptotic expression for general potential functions. In Subsection 5.2 the result is simplified for a restricted form of potential functions.

5.1 Expansion of cumulant generating functions

Denote the cumulant generating function of a probability density $p(x)$ by $\kappa[p](\eta) := \log \int p(x) \mathrm{e}^{\eta^\top x} \mathrm{d}x$. We first give a lemma on perturbed densities.

Lemma 12. Let $p_\epsilon(x) = p_0(x) + \epsilon r(x)$ be a density function. Then $\kappa[p_\epsilon](\eta)$is expanded as

$$\kappa[p_\epsilon](\eta) = \kappa[p_0](\eta) + \epsilon \int r(x) \mathrm{e}^{\eta^\top x - \kappa[p_0](\eta)} \mathrm{d}\eta + \mathrm{O}(\epsilon^2)$$

as $\epsilon \to 0$.

Proof. From the definition of $\kappa[p_\epsilon](\eta)$ we have

$$\kappa[p_\epsilon](\eta) \;=\; \log \left[\int p_0(x) \mathrm{e}^{\eta^\top x} \mathrm{d}x \right] + \epsilon \frac{\int r(x) \mathrm{e}^{\eta^\top x} \mathrm{d}x}{\int p_0(x) \mathrm{e}^{\eta^\top x} \mathrm{d}x} + \mathrm{O}(\epsilon^2)$$

and the result follows. □

We consider the perturbed potential

$$\psi_\epsilon(x) = \psi_0(x) + \epsilon \tau(x),$$

where ψ_0 is a convex function on $\mathbb{R}^m$ and τ is a function on $\mathbb{R}^m$. To assure the convexity of ψ_ϵ we suppose that the eigenvalues of $\nabla \nabla^\top \psi_0$ are bounded away from 0 and those of $\nabla \nabla^\top \tau$ are bounded. We will consider more specific $\psi_0(x)$ and $\tau(x)$ in the next subsection.

Lemma 13. Let $\psi_\epsilon(x) = \psi_0(x) + \epsilon \tau(x)$. Then the density function is expanded as

$$p_\epsilon(x) \;:=\; p[\nabla \psi_\epsilon](x) \;=\; p_0(x) + \epsilon \nabla^\top (p_0(x) v(x)) + \mathrm{O}(\epsilon^2)$$

as $\epsilon \to 0$, where $v(x) = (\nabla \nabla^\top \psi_0(x))^{-1} \nabla \tau(x)$.

Proof. The gradient $\nabla\psi_\epsilon(x) = \nabla\psi_0(x) + \epsilon\nabla\tau(x)$ is asymptotically rewritten as follows:

$$\nabla\psi_\epsilon(x) = \nabla\psi_0(x + \epsilon v(x)) + \mathrm{O}(\epsilon^2).$$

This map is the composition of the map $x \mapsto x + \epsilon v(x)$ and $\nabla\psi_0$. Therefore

$$\begin{aligned} p_\epsilon(x) &= p_0(x + \epsilon v(x)) \det(I + \epsilon\nabla v(x)^\top) + \mathrm{O}(\epsilon^2) \\ &= p_0(x) + \epsilon(\nabla^\top p_0(x))v(x) + \epsilon p_0(x)\nabla^\top v(x) + \mathrm{O}(\epsilon^2) \\ &= p_0(x) + \epsilon\nabla^\top(p_0(x)v(x)) + \mathrm{O}(\epsilon^2) \end{aligned}$$

and the desired formula is proved. □

Remark 14. Note that $\int \nabla^\top(p_0(x)v(x))\mathrm{d}x = 0$ by the integration-by-parts formula. This is directly deduced from measure-preserving property $\int p_\epsilon(x)\mathrm{d}x = \int p_0(x)\mathrm{d}x$. □

By applying Lemma 12 we obtain the following theorem.

Theorem 15. Let $\psi_\epsilon(x) = \psi_0(x) + \epsilon\tau(x)$ and $v(x) = (\nabla\nabla^\top\psi_0(x))^{-1}\nabla\tau(x)$. Then the cumulant generating function of p_ϵ is expanded as

$$\kappa[p_\epsilon](\eta) = \kappa[p_0](\eta) - \epsilon\int p_0(x)(\eta^\top v(x))\mathrm{e}^{\eta^\top x - \kappa[p_0](\eta)}\mathrm{d}x + \mathrm{O}(\epsilon^2).$$

Proof. From Lemma 12 and Lemma 13 we have

$$\begin{aligned} \kappa[p_\epsilon](\eta) &= \kappa[p_0](\eta) + \epsilon\int \nabla^\top\{p_0(x)v(x)\}\mathrm{e}^{\eta^\top x - \kappa[p_0](\eta)}\mathrm{d}x \\ &= \kappa[p_0](\eta) - \epsilon\int p_0(x)\eta^\top v(x)\mathrm{e}^{\eta^\top x - \kappa[p_0](\eta)}\mathrm{d}x, \end{aligned}$$

where we used the integration-by-parts formula. □

Corollary 16. Let $\psi_0(x) = x^\top Ax/2$. Then

$$\kappa[p_\epsilon](\eta) = \frac{1}{2}\eta^\top A^{-2}\eta - \epsilon\int(\eta^\top A^{-1}\nabla\tau(x))\phi(x|A^{-2}\eta, A^{-2})\mathrm{d}x,$$

where $\phi(x|\mu, \Sigma)$ is the normal density with the mean μ and covariance Σ.

Proof. By noting $p_0(x) = \phi(x|0, A^{-2})$ we obtain $p_0(x)\mathrm{e}^{\eta^\top x - \kappa[p_0](\eta)} = \phi(x|A^{-2}\eta, A^{-2})$. □

5.2 Basic perturbation around the identity map

As a special case of the perturbed potential we consider the following function

$$\psi_\epsilon(x) = \frac{x^\top x}{2} + \epsilon \sum_{\lambda \in \Lambda} f_\lambda(e_\lambda^\top x)$$

where Λ is a finite set, $\{f_\lambda\}$ is a set of functions on $\mathbb{R}$ and $\{e_\lambda = (e_{\lambda,i})\}$ is a set of vectors in $\mathbb{R}^m$. We call this form the *basic perturbation*. From Corollary 16 we have

$$\begin{aligned}
\kappa[p_\epsilon](\eta) &= \frac{1}{2}\eta^\top\eta - \epsilon \sum_\lambda (\eta^\top e_\lambda) \int f_\lambda^{(1)}(e_\lambda^\top x)\phi(x|\eta, I)\mathrm{d}x \\
&= \frac{1}{2}\eta^\top\eta - \epsilon \sum_\lambda (\eta^\top e_\lambda) \int f_\lambda^{(1)}(e_\lambda^\top x + e_\lambda^\top \eta)\phi(x|0, I)\mathrm{d}x \\
&= \frac{1}{2}\eta^\top\eta - \epsilon \sum_{k=1}^\infty \frac{1}{(k-1)!} \sum_\lambda (\eta^\top e_\lambda)^k \int f_\lambda^{(k)}(e_\lambda^\top x)\phi(x|0, I)\mathrm{d}x
\end{aligned}$$

up to $\mathrm{O}(\epsilon)$. The k-th cumulant is $\kappa_{i_1\cdots i_k}[p_\epsilon] = \kappa_{i_1\cdots i_k}[p_0](\eta) + \epsilon T_{i_1\cdots i_k} + \mathrm{O}(\epsilon^2)$, where

$$T_{i_1\cdots i_k} = -k \sum_\lambda F_\lambda^{(k)} e_{\lambda,i_1} \cdots e_{\lambda,i_k} \tag{5}$$

and $F_\lambda^{(k)} = \int f_\lambda^{(k)}(e_\lambda^\top x)\phi(x|0, I)\mathrm{d}x$. The basic perturbation has sufficient flexibility. In fact, for given finite number of coefficients $T_{i_1\cdots i_k}$, there exists a set of f_λ and e_λ such that (5) holds. This fact will be proved in Appendix.

For example let $f_\lambda(z) = \alpha_\lambda \arctan(z)$ and the set of e_λ be given in Table 2. Put $H_c^{(k)} = \mathrm{E}[(\arctan)^{(k)}(cZ)]$ for $c > 0$. From the formula (5) we have

$$T_i = 0, \quad T_{ij} = 0, \quad T_{ijk} = -3 \sum_\lambda \alpha_\lambda H_{|e_\lambda|}^{(3)} e_{\lambda,i} e_{\lambda,j} e_{\lambda,k}.$$

Conversely α_λ is written in terms of T_{ijk} as

$$\alpha_\lambda = -\frac{\beta_\lambda}{24 H_{|e_\lambda|}^{(3)}},$$

where the list of β_λ and $H_{|e_\lambda|}^{(3)}$ is given in Table 2. The quantities $H_c^{(k)}$ have a tractable form as follows. Denote the upper probability of χ^2-distribution (degree of freedom is 1) by $P_x = \int_x^\infty \mathrm{e}^{-u/2}/\sqrt{2\pi u}\,\mathrm{d}u$. Then

$$H_c^{(1)} = \mathrm{e}^{1/(2c^2)} \sqrt{\frac{\pi}{2c^2}} P_{1/c^2}, \quad H_c^{(2)} = 0, \quad H_c^{(3)} = c^{-4}(1 - (1 + c^2)H_c^{(1)}).$$

In fact,

$$H_c^{(1)} = \mathrm{E}\left[\frac{1}{1+c^2Z^2}\right] = \frac{1}{2}\int_0^\infty \mathrm{E}[\mathrm{e}^{-t(1+c^2Z^2)/2}]\mathrm{d}t = \frac{1}{2}\int_0^\infty \frac{\mathrm{e}^{-t/2}}{\sqrt{1+c^2t}}\mathrm{d}t = \mathrm{e}^{1/(2c^2)}\sqrt{\frac{\pi}{2c^2}}P_{1/c^2}.$$

The expression of $H_c^{(3)}$ follows from the integration-by-parts formula

$$\begin{aligned} H_c^{(3)} &= \int (\arctan)^{(3)}(cz)\phi(z)\mathrm{d}z \;=\; c^{-2}\int (\arctan)^{(1)}(cz)(z^2-1)\phi(z)\mathrm{d}z \\ &= c^{-4}\int (1-(1+c^2)(\arctan)^{(1)}(cz))\phi(z)\mathrm{d}z \;=\; c^{-4}(1-(1+c^2)H_c^{(1)}). \end{aligned}$$

It is easily derived that $H_c^{(k)} = 0$ for any even k.

In Example 1 we used $\{e_\lambda\}_{\lambda=19}^{22}$ of Table 2. Since $|e_\lambda| = \sqrt{3}$ and $e_{\lambda,1}e_{\lambda,2}e_{\lambda,3} = 1$, we obtain the straight line $\kappa_{123} = -3\epsilon \times 4H_{\sqrt{3}}^{(3)} = 1.237\epsilon$ in Figure 2 (d).

Table 2: The list of e_λ, β_λ and $H_{|e_\lambda|}^{(3)}$.

λ	$e_{\lambda,1}$	$e_{\lambda,2}$	$e_{\lambda,3}$	β_λ	$\lvert e_\lambda\rvert$	$H_{\lvert e_\lambda\rvert}^{(3)}$
1	3	0	0	T_{111}	3	$-.0279$
2	0	3	0	T_{222}		
3	0	0	3	T_{333}		
4	-1	0	0	$3T_{111}+6T_{122}+6T_{133}$	1	$-.3114$
5	0	-1	0	$6T_{112}+3T_{222}+6T_{233}$		
6	0	0	-1	$6T_{113}+6T_{223}+3T_{333}$		
7	2	1	0	$3T_{112}$	$\sqrt{5}$	$-.0573$
8	-2	1	0	$3T_{112}$		
9	2	0	1	$3T_{113}$		
10	-2	0	1	$3T_{113}$		
11	1	2	0	$3T_{122}$		
12	1	-2	0	$3T_{122}$		
13	1	0	2	$3T_{133}$		
14	1	0	-2	$3T_{133}$		
15	0	2	1	$3T_{223}$		
16	0	-2	1	$3T_{223}$		
17	0	1	2	$3T_{233}$		
18	0	1	-2	$3T_{233}$		
19	1	1	1	$6T_{123}$	$\sqrt{3}$	$-.1031$
20	-1	1	-1	$6T_{123}$		
21	-1	-1	1	$6T_{123}$		
22	1	-1	-1	$6T_{123}$		

6 Discussion

We defined the gradient representation of the probability densities and constructed the g-flat model by using it. The g-flat has many good properties including the concavity of the log-likelihood. A g-flat model was applied to detect the three-dimensional interaction of the decathlon data. From the theoretical viewpoint we discussed a class of symmetric potentials and studied asymptotic expansion with respect to small perturbation.

We have not stated about regression models. A generalization of g-flat models to this direction is available by adding the explanation variables in the potential function. More generally, graphical modeling corresponding to directed acyclic graphs will be described in terms of the gradient representation. These important generalizations are left on the future work.

A Flexibility of basic perturbations

We prove flexibility of the basic perturbations described in Subsection 5.2. Let $\mathrm{C}_{\mathrm{b}}^{\infty}(\mathbb{R})$ be the set of infinitely differentiable functions with the bounded derivatives. For vectors v and w, the tensors $(v_i w_i)$ and $(v_{i_1} \cdots v_{i_k})$ are denoted by $v \otimes w$ and $v^{\otimes k}$. The tensor $T_{i_1 \cdots i_k}$ in Eq. (5) is written as $(T_{i_1 \cdots i_k}) = -k \sum_{\lambda} F_{\lambda}^{(k)} (e_{\lambda})^{\otimes k}$. The direct sum of linear spaces V and W is denoted by $V \oplus W$.

Lemma 17. Fix a positive integer K. Let S^k be the set of all symmetric tensors of order k. Then for any element $(w^{(1)}, \ldots, w^{(K)}) \in S^1 \oplus \cdots \oplus S^K$ there exists a finite set $\{e_{\lambda}\}$ of vectors and a finite subset $\{f_{\lambda}\}$ of $\mathrm{C}_{\mathrm{b}}^{\infty}(\mathbb{R})$ such that

$$\sum_{\lambda} \left(F_{\lambda}^{(1)} e_{\lambda}^{\otimes 1}, \ldots, F_{\lambda}^{(K)} e_{\lambda}^{\otimes k} \right) = (w^{(1)}, \ldots, w^{(K)}),$$

where $F_{\lambda}^{(k)} = \mathrm{E}[f_{\lambda}^{(k)}(Z)]$ and $Z \in N(0,1)$.

Proof. Let $[K] = \{1, \ldots, K\}$. It is sufficient to prove the following two properties.

(i) For any $c = (c_1, \ldots, c_K) \in \mathbb{R}^K$ there exists a function $f \in \mathrm{C}_{\mathrm{b}}^{\infty}$ such that $F^{(k)} = c_k$ for all $k \in [K]$.

(ii) There exists a set $\{e_{\lambda}\}_{\lambda \in \Lambda}$ of vectors such that the space $S^1 \oplus \cdots \oplus S^K$ is spanned by $\{(e_{\lambda}^{\otimes 1}, \ldots, e_{\lambda}^{\otimes K})\}_{\lambda \in \Lambda}$.

We first prove (i) by using a topological technique. We assume that the norm $|c|$ of the vector c is less than 1 without loss of generality. For each $a \in \mathbb{R}^K$ let $f_a(z) := \sum_{k=1}^K a_k h_k(z)/k!$, where $h_k(z) = (-1)^k \phi^{(k)}(z)/\phi(z)$ is the Hermite polynomial. We have $\mathrm{E}[f_a^{(k)}(Z)] = \mathrm{E}[f_a(Z)h_k(Z)] = a_k$ for all $k \in [K]$. The function f_a is not in $\mathrm{C}_\mathrm{b}^\infty(\mathbb{R})$. Let $\eta(z)$ be a function in $\mathrm{C}_\mathrm{b}^\infty(\mathbb{R})$ such that $\eta(z) = 1$ if $|z| \le 1$ and 0 if $|z| \ge 2$. Define the function $f_{a,t} \in \mathrm{C}_\mathrm{b}^\infty(\mathbb{R})$ by $f_{a,t}(z) := (t \wedge 1)^K \eta(z/t) f_a(z)$ for $t \in (0, \infty)$. By Lebesgue's dominated convergence theorem we can prove that the function $(t, a) \mapsto \mathrm{E}[f_{a,t}^{(k)}(Z)]$ is continuous and

$$\mathrm{E}[f_{a,t}^{(k)}(Z)] \to \begin{cases} 0 & \text{as } t \to 0, \\ a_k & \text{as } t \to \infty. \end{cases}$$

Let $U_t := \{(\mathrm{E}[f_{t,a}^{(k)}(Z)])_{k=1}^K \mid |a| = 1\}$. Then U_∞ is the unit sphere and U_0 is the origin. Since U_t varies continuously from $t = \infty$ to 0, there exists some $t \ge 0$ such that $c \in U_t$. This implies that there exists a function $f := f_{t,a}$ such that $F^{(k)} = c_k$.

We next prove (ii). We construct the set $\{e_\lambda\}$ explicitly. Define the index set Λ by

$$\Lambda = \left\{(k, \tau, i) \mid k \in [K], \tau = (\tau_1, \dots, \tau_k) \in \{-1, 1\}^k, i = (i_1, \dots, i_k) \in [m]^k\right\}.$$

Let u_i be the i-th column vector of the $m \times m$ identity matrix. Then we define e_λ by

$$e_{(k,\tau,i)} := \sum_{j=1}^k \tau_j u_{i_j}, \quad (k, \tau, i) \in \Lambda.$$

For example $e_{(3,(1,-1,1),(1,2,3))} = u_1 - u_2 + u_3$ and $e_{(3,(1,1,-1),(1,1,2))} = 2u_1 - u_2$. For each $k \in [K]$ and $i \in [m]^k$, we define an element $w_{(k,i)} \in S^1 \oplus \cdots \oplus S^K$ by

$$w_{(k,i)} = (w_{(k,i)}^1, \dots, w_{(k,i)}^K) = \sum_{\tau \in \{-1,1\}^k} \tau_1 \cdots \tau_k \left(e_{(k,\tau,i)}^{\otimes 1}, \dots, e_{(k,\tau,i)}^{\otimes K}\right).$$

It is sufficient to prove that $\{w_{(k,i)}\}$ spans $S^1 \oplus \cdots \oplus S^K$. We can prove that

$$w_{(k,i)}^l = \begin{cases} 0 & \text{if } l < k, \\ 2^k \sum_{\sigma \in S(k)} u_{\sigma(i_1)} \otimes \cdots \otimes u_{\sigma(i_k)} & \text{if } l = k, \end{cases}$$

where $S(k)$ is the set of all permutations. Note that the set $\{\sum_\sigma u_{\sigma(i_1)} \otimes \cdots \otimes u_{\sigma(i_k)}\}_{i \in [m]^k}$ spans S^k. By induction from $k = K$ to $k = 1$, we can prove that the space $\{0\} \oplus \cdots \oplus \{0\} \oplus S^k \oplus \cdots \oplus S^K$ is spanned by $\{w_{(l,i)}\}_{l \ge k, i \in [m]^l}$. Thus $\{w_{(k,i)}\}_{k \in [K], i \in [m]^k}$ spans $S^1 \oplus \cdots \oplus S^K$. □

Acknowledgments

The author thanks Prof. Masami Miyakawa and Prof. Satoshi Aoki for their valuable comments and suggestions. He also thanks Prof. Akimichi Takemura and Prof. Fumiyasu Komaki for their encouragement to write this paper.

References

[1] Abdellaoui, T. (1998). Optimal solution of a Monge-Kantorovitch transportation problem, *J. Comp. Appl. Math.*, **96**, 149–161.

[2] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *J. Roy. Statist. Soc.*, **26**, 211–252.

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

[4] De Oliveira, V., Kedem, B. and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields, *J. Amer. Statist. Assoc.*, **92**, 1422–1433.

[5] Easton, G. S. and McCulloch, R. E. (1990). A multivariate generalization of quantile-quantile plots, *J. Amer. Statist. Assoc.*, **85**, 376–386.

[6] Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency), *Ann. Statist.*, **3**, 1189-1242.

[7] Jørgensen, B. (1987). Exponential dispersion models, *J. Roy. Statist. Soc.*, **49**, 127–162.

[8] Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions, *J. Optim. Theo. Appl.*, **43**, 39–49.

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

[10] Miyakawa, M. (1997). *Graphical modeling* (in Japanese), Asakura Shoten, Tokyo.

[11] Murota, K. (2003). *Discrete Convex Analysis*, SIAM, Philadelphia.

[12] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems – I: Theory and II: Applications*, Springer-Verlag, New York.

[13] Rockafellar, R. T. (1972). *Convex Analysis*, Princeton University Press.

[14] Rüschendorf, L. and Rachev, S. T. (1990). A characterization of random variables with minimum L^2-distance, *J. Multivariate Anal.*, **32**, 48–54.