Ann Inst Stat Math (2011) 63:675–688
DOI 10.1007/s10463-009-0261-1

Gradient modeling for multivariate quantitative data

Tomonari Sei

Received: 7 October 2008 / Revised: 17 March 2009 / Published online: 11 November 2009

Abstract We propose a new parametric model for continuous data, a "g-model", on the basis of gradient maps of convex functions. It is known that any multivariate probability density on the Euclidean space is uniquely transformed to any other density by using the gradient map of a convex function. Therefore the statistical modeling for quantitative data is equivalent to design of the gradient maps. The explicit expression for the gradient map enables us the exact sampling from the corresponding probability distribution. We define the g-model as a convex subset of the space of all gradient maps. It is shown that the g-model has many desirable properties such as 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-Model · Gradient representation · Three-dimensional interaction

1 Introduction

In this paper we propose a method of gradient-based modeling for multivariate quantitative data. An essential fact we use is that any probability density function on the Euclidean space is transformed to any other density function by using the gradient map of some convex function. We will call the map the gradient representation of the probability density and show that the statistical modeling on the basis of the gradient representation has various advantages.

T. Sei (✉)
Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: sei@stat.t.u-tokyo.ac.jp

Let us consider the one-dimensional case to clarify the idea. Let Q be a cumulative distribution function on $\mathbb{R}$. Assume that Q is continuous and strictly increasing. For any increasing function g from $\mathbb{R}$ onto $\mathbb{R}$, the function F defined by $F(x) = Q(g(x))$ is a cumulative distribution function. If Q is fixed, there is a one-to-one correspondence between g and F. Therefore the statistical model can be described in terms of the increasing function g. Remark that any increasing function on $\mathbb{R}$ is the gradient map of a convex function on $\mathbb{R}$, and vice versa.

The idea can be generalized to the multivariate distribution. In any dimension there exists a one-to-one correspondence between probability distributions and gradient maps of convex functions. This remarkable fact was proven by Brenier (1991) for a restricted case and McCann (1995) for the general case. The details are stated in Sect. 2. We will show that the exact sampling is available by the inverse-function method once the gradient representation is obtained.

A statistical model whose gradient representation is linear with respect to the parameter is called a g-model in this paper. This model has many desirable properties such as concavity of the log-likelihood function, nonnecessity of normalization, description of independency, inclusion of all the multivariate normal distributions and possibility of the conic extension. We will describe these properties and compare the g-model with other types in Sect. 3.

There is an important application of g-models. From the practical point of view, a model describing the three-dimensional interaction of data is needed, where we say that the set of variables (x_1, x_2, x_3) has the three-dimensional interaction if the third-order derivative of the joint log-probability density with respect to x_1, x_2 and x_3 does not vanish (see Chapter 2 of Whittaker 1990). However, this interaction is not described by any normal distribution. There is almost no tractable parametric model to analyze it. We will propose such a model by using the g-model and apply it to a real data. Roughly speaking, the density function has three-dimensional interaction if and only if the corresponding gradient map has three-dimensional interaction.

We make some historical notes on use of the gradient representation in statistics. Box and Cox (1964) introduced a class of coordinate-wise transformation to deal with non-normal data. This transformation is a special case of our gradient representation. De Oliveira et al. (1997) used the coordinate-wise transformation to the prediction problem together with Bayesian inference. In non-parametric statistics, Easton and McCulloch (1990) proposed the quantile–quantile plot for multivariate data by means of the transportation problem. The transportation problem is closely connected with the gradient representation (see Sect. 2).

This paper is organized as follows. In Sect. 2, we give the precise definition of the gradient representation and its examples. In Sect. 3, we define the g-models and investigate their properties. We apply the g-model to a real data in Sect. 4. Finally we have some discussions in Sect. 5. A technical lemma is given in Appendix.

2 The gradient representation of multivariate distributions

We define the gradient representation of multivariate continuous distributions and discuss its properties.

2.1 The existence and uniqueness theorem

Let Y be a random vector subject to a probability density q with respect to the Lebesgue measure on $\mathbb{R}^m$ and let ψ be any convex function on $\mathbb{R}^m$. Denote the gradient map of ψ by $\nabla\psi(x) = (\partial\psi(x)/\partial x_i)_{i=1}^m$. If the gradient map $\nabla\psi(x)$ is one-to-one and onto the support of q, then a random vector X is defined by the unique solution of $(\nabla\psi)(X) = Y$. By change of variables, the probability density of X is

$$p(x) \;=\; q(\nabla\psi(x))\det(\nabla\nabla^\top\psi(x)), \tag{1}$$

where $\nabla\nabla^\top\psi(x) = (\partial^2\psi(x)/\partial x_i\partial x_j)_{i,j=1}^m$ is the Hessian matrix of $\psi(x)$. Thus a convex function induces a probability distribution.

The following theorem shows that the converse is also true.

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

From Theorem 1 the following definition is consistent.

Definition 2 (gradient representation) Let p and q be probability densities with respect to the Lebesgue measure on $\mathbb{R}^m$. We call the gradient map $\nabla\psi$ satisfying (1) the *gradient representation* of the density p with respect to the *reference density* q. The convex function ψ is called the *potential function*.

We mainly use the standard normal density as the reference density in this paper. In general the gradient representation of a given density is not explicitly expressed. Instead we first determine the gradient map $\nabla\psi$ and obtain the density $p(x)$ as a result. The probability density having the gradient representation $\nabla\psi$ is denoted by $p[\nabla\psi](x) := q(\nabla\psi(x))\det(\nabla\nabla^\top\psi(x))$.

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

$$\mathscr{G}_{\mathrm{all}} := \left\{\nabla\psi \mid \psi \in \mathrm{C}^2(\mathbb{R}^m),\ \nabla\psi(\mathbb{R}^m) = \mathbb{R}^m,\ \nabla\nabla^\top\psi(x) \succ 0\right\},$$

where $\succ$ denotes the positive definiteness. 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 considered in this paper except for Example 2 are included in $\mathscr{G}_{\mathrm{all}}$. Although we are not aware of the characterization of the set $\{p[\nabla\psi](x) \mid \nabla\psi \in \mathscr{G}_{\mathrm{all}}\}$ of densities, the set contains all Hölder-continuous positive densities on $\mathbb{R}^m$ (see, e.g. Villani 2003, Theorem 4.14).

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 (Rockafellar 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 $x \neq 0$.*

Proposition 4 *The set $\mathcal{G}_{\mathrm{all}}$ is a convex cone.*

Proof We prove $c_1\nabla\psi_1 + c_2\nabla\psi_2 \in \mathcal{G}_{\mathrm{all}}$ for any $\nabla\psi_i \in \mathcal{G}_{\mathrm{all}}$ and $c_i > 0$ $(i = 1, 2)$. In fact, if ψ_1 and ψ_2 are co-finite convex functions, $c_1\psi_1 + c_2\psi_2$ is a co-finite convex function. □

Proposition 4 is used to construct the g-model in the next section. For the rest of this subsection we briefly review known facts on the gradient representation.

McCann (1995) showed the existence and uniqueness theorem (i.e. Theorem 1) under a more general condition: if two probability measures P and Q on $\mathbb{R}^m$ have no mass on any set with Hausdorff dimension $m-1$, then there exists the gradient representation $\nabla\psi$. In general it is difficult to find $\nabla\psi$ for given P and Q both analytically and numerically. We refer to Abdellaoui (1998), Knott and Smith (1984) and Haker et al. (2004) in this direction.

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 "coupling measure" Γ on $\mathbb{R}^m \times \mathbb{R}^m$ that solves a minimization problem

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

The MKP is an infinite-dimensional linear programming problem. It is known that if P and Q have the densities p and q, respectively, then the "deterministic coupling" $y = \nabla\psi(x)$ is the solution of the MKP, where ψ is the potential function of Theorem 1 (see Knott and Smith 1984; Rüschendorf and Rachev 1990; Rachev and Rüschendorf 1998; Villani 2003). One of the statistical methods related to the MKP is the multivariate generalization of the quantile–quantile plot by Easton and 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 $\{y_i\}_{i=1}^n$.

2.2 Exact sampling

Assume that samples from the reference density q are easily obtained, as in the case of the standard normal density. Once we determine the gradient map $\nabla\psi$, we can sample from $p[\nabla\psi](x)$ exactly on the basis of the inverse-function method. A sample X from $p[\nabla\psi](x)$ is obtained by solving $\nabla\psi(X) = Y$, where Y is a sample from q. The equation is equivalent to the following 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). Newton's method efficiently solves this optimization problem.

An independently and identically distributed (i.i.d.) sequence $\{X(t)\}_{t=1}^n$ 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 reference density q. This procedure is convenient for the vector-oriented programming languages like R and MATLAB.

2.3 Examples of gradient representation

We give two examples having the explicit gradient representation. We assume that the reference density $q(y)$ is standard normal in these examples. The first example is a distribution with a three-dimensional interaction. The second one is a distribution quite different from the normal distribution. To generate samples, we use the exact sampling described in the preceding subsection.

We say that a set of random variables (X_1, X_2, X_3) has the three-dimensional interaction if the joint density function $p(x_1, x_2, x_3)$ has the following property (see Sect. 2 of Whittaker 1990):

$$\frac{\partial^3}{\partial x_1 \partial x_2 \partial x_3} \log p(x_1, x_2, x_3) \neq 0 \quad \text{for some} \quad (x_1, x_2, x_3).$$

We use the third-order cumulant $\kappa_{123} = \mathrm{E}[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)(X_3 - \mathrm{E}X_3)]$ of (X_1, X_2, X_3) as a quantity representing the three-dimensional interaction because this quantity vanishes if (but not only if) there is no three-dimensional interaction. Indeed, if $(\partial^3/\partial x_1 \partial x_2 \partial x_3) \log p(x_1, x_2, x_3) = 0$ for any (x_1, x_2, x_3), then $p(x_1, x_2|x_3)$ does not depend on x_3 and therefore

$$\kappa_{123} = \mathrm{E}[\mathrm{E}[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)|X_3](X_3 - \mathrm{E}X_3)] = 0.$$

Example 1 (Three-dimensional interaction) Let $m = 3$ and define

$$\psi(x) = \frac{x^\top x}{2} + \epsilon \sum_{\lambda=1}^{4} \arctan(e_\lambda^\top x), \quad x = (x_1, x_2, x_3)^\top,$$

where

$$(e_1, e_2, e_3, e_4) = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \tag{2}$$

and $\epsilon \in \mathbb{R}$ is a small number such that convexity of ψ is assured. In Appendix A, we will prove that ψ is strictly convex if and only if $|\epsilon| < 2 \times 3^{-3/2} \approx 0.3849$. A result of numerical experiments when ϵ ranges over $\{0.00, 0.05, \ldots, 0.35\}$ is shown in Fig. 1. The result shows that the third-order cumulant κ_{123} of $p[\nabla\psi](x)$ is nonnegligible. In Sect. 4 we will use this density in order to detect the three-dimensional interaction of real data. One can show that the third-order cumulant of (X_1, X_2, X_3) is approximated as $\kappa_{123} \approx 1.237\epsilon$ when ϵ is small (see Sei 2006 for details).

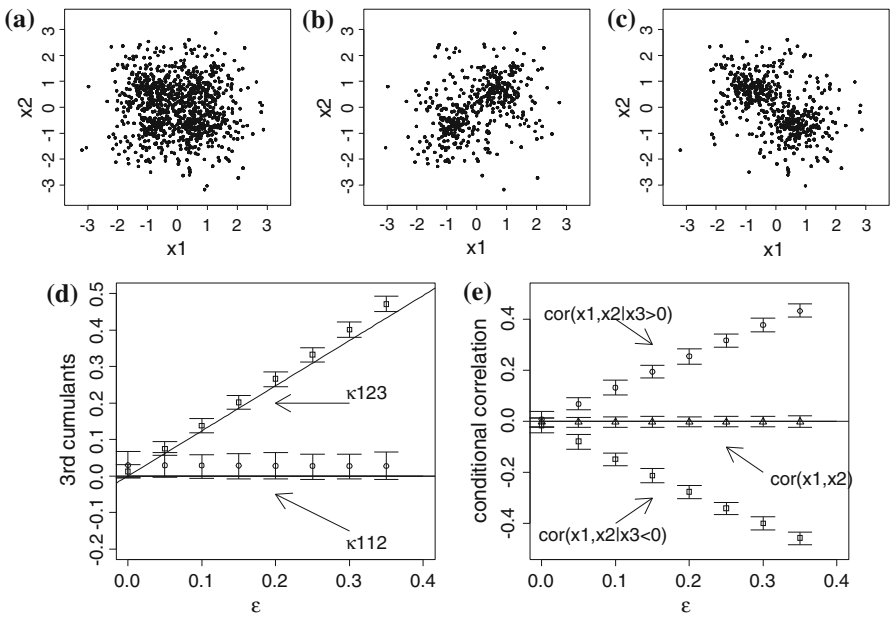

Fig. 1 A simulation on the distribution with three-dimensional interaction. **a** Scatter plot of samples drawn from $p(x_1, x_2)$. The small parameter ϵ is 0.35. The sample size is 1,000. **b** Scatter plot of $p(x_1, x_2|x_3 > 0)$. **c** Scatter plot of $p(x_1, x_2|x_3 < 0)$. **d** The third cumulants κ_{112} and κ_{123} against $\epsilon \in \{0.00, 0.05, \ldots, 0.35\}$. The sample size is 10,000. The 95% confidence interval is based on the normal approximation. The *diagonal line* indicates 1.237ϵ. **e** The marginal correlation $\mathrm{cor}(x_1, x_2)$ and the conditional correlations $\mathrm{cor}(x_1, x_2|x_3 \geq 0)$ and $\mathrm{cor}(x_1, x_2|x_3 < 0)$ against $\epsilon \in \{0.00, 0.05, \ldots, 0.35\}$

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

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

where t is a positive constant and $\rho = (1-t)/(2t)^{1/2}$. Samples drawn from the density $p[\nabla\psi](x)$ for $t = 0.05$ and the contour $\{x \mid \psi(x) = 1\}$ are shown in Fig. 2. The contour consists of four arcs. The radius r of each arc is $(1+t^2)^{1/2}$ if $|x_2| > |x_1|$ and $(1+t^{-2})^{1/2}$ if $|x_2| < |x_1|$.

3 The g-model

3.1 Definition and properties of the g-model

We define the g-models and discuss their properties. For a given convex function ψ, we use the symbols $g = \nabla\psi$ and $G = \nabla\nabla^\top\psi$ for the gradient vector and the Hessian matrix. We denote the probability density having the gradient representation g by $p[g]$. The density is explicitly expressed as $p[g](x) = q(g(x)) \det G(x)$,

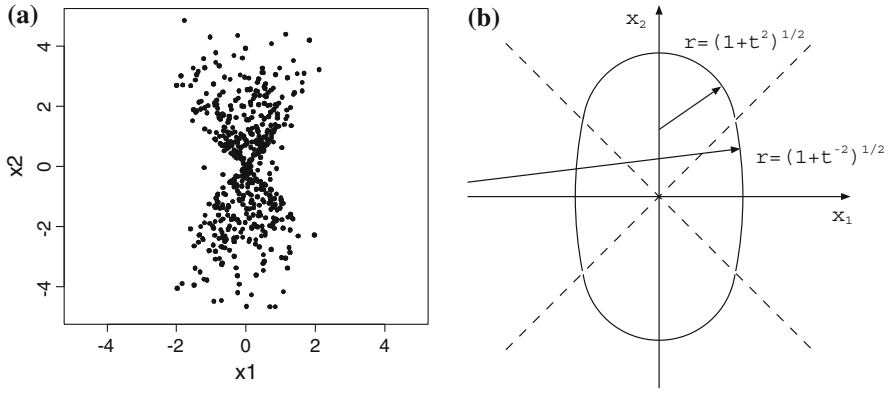

Fig. 2 A simulation on the curtain-type distribution. **a** Scatter plot of 500 samples. **b** The contour of the potential function

where q is the reference density (see Definition 2). Recall that $\mathscr{G}_{\mathrm{all}}$ is the set of all the continuously differentiable gradient maps onto $\mathbb{R}^m$.

Definition 5 (g-model) A statistical model is called a gradient model or *g-model* if it is written as

$$\mathcal{M} = \left\{ p[g](x) \mid g = \sum_{a=1}^{p} \theta^a g_a,\ \theta = (\theta^a)_{a=1}^{p} \in \Theta \right\},$$

where $(g_a)_{a=1}^{p}$ are fixed gradient maps in $\mathscr{G}_{\mathrm{all}}$ and Θ is a convex subset of $\mathbb{R}^p$.

Remark 6 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, if the reference density q is the standard normal $N(0, I)$, then the normal model $\{N(\mu, \Sigma)\}$ is expressed in the gradient representation as

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

where $\mathbb{S}_+(m)$ is the set of all positive definite $m \times m$ matrices. The density $p[g](x)$ is the normal density with the mean $-A^{-1}b$ and the variance A^{-2}.

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

Theorem 7 *Assume that the reference density q is log-concave. Then the log-likelihood function of any g-model is concave.*

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 $\mathscr{G}_{\mathrm{all}}$. The logarithm of $p[g_\theta](x)$ is given by

$$\log p[g_\theta](x) = \log q(g_\theta(x)) + \log \det(G_\theta(x)),$$

where $G_\theta(x) := \nabla(g_\theta(x)^\top)$ is positive definite for any $\theta \in (0,1)$ and $x \in \mathbb{R}^m$. Let $H(y) := (-(\partial^2/\partial y_i \partial y_j) \log q(y))_{i,j=1}^m$. Then we have

$$\frac{\partial^2}{\partial \theta^2} \log p[g_\theta] = -(g_1 - g_0)^\top H(g_\theta)(g_1 - g_0) - \mathrm{tr}\{G_\theta^{-1}(G_1 - G_0)G_\theta^{-1}(G_1 - G_0)\},$$

where we omitted the argument x for simplicity. Since $H(g_\theta(x))$ is positive semi-definite and $G_\theta(x)$ is positive definite, the first and second terms are non-positive, respectively. Thus $\log p[g_\theta]$ is concave. □

From this theorem, any local maximal point of the log-likelihood function is actually the global maximal point when we use a g-model. The numerical computation of the maximum likelihood estimator (MLE) is relatively simple. A penalized log-likelihood function $\log p[g_\theta](x) - \mathrm{pen}(\theta)$ is also concave whenever the penalty function $\mathrm{pen}(\theta)$ is convex.

3.2 Comparison to other models

We enumerate the properties of g-model and compare it with other models. The properties depend on the structure of the reference density $q(y)$. Thus we specify the reference density for each case.

In Theorem 7, we proved that the log-likelihood function of any g-model is concave. There are two other classes of statistical models whose log-likelihood function is concave. One is the exponential model (or *e-model*)

$$p_\theta(x) = p_0(x) \exp\left(\sum_{a=1}^p \theta^a t_a(x) - C(\theta) \right), \tag{3}$$

where $p_0(x)$ is a probability density, $(t_a(x))_{a=1}^p$ is a set of functions, $(\theta^a)_{a=1}^p$ is the set of parameters and $C(\theta)$ is the normalizing constant that guarantees $\int p_\theta(x)\mathrm{d}x = 1$. The other one is the mixture model (or *m-model*)

$$p_\theta(x) = p_0(x) + \sum_{a=1}^p \theta^a t_a(x),$$

where $p_0(x)$ is a probability density, $(t_a(x))_{a=1}^p$ is a set of functions and $(\theta^a)_{a=1}^p$ is the set of parameters. In the following, we enumerate various properties of the g-model and compare it with the e-model and m-model.

We first remark that the multivariate normal model can be combined with any g-model if the reference density q is normal. Consider any g-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

$$\mathcal{M}' = \left\{ p[g](x) \mid g(x) = Ax + b + \sum_{a=1}^{p} \theta^a g_a(x),\ \theta \in \Theta,\ A \in \mathbb{S}_+(m),\ b \in \mathbb{R}^m \right\}.$$

Recall that $\mathbb{S}_+(m)$ is the set of all $m \times m$ positive definite matrices. The potential function is $\psi(x) = \frac{1}{2} x^\top Ax + b^\top x + \sum_{a=1}^{p} \theta^a \psi_a(x)$ with $g_a = \nabla \psi_a$. This property of the g-model is particularly important for multivariate analysis because many statistical methods for multivariate data are based on the normal model and our g-model extends the methods. The e-model also has this property but the m-model does not. More generally, if the reference density is symmetric with respect to the orthogonal transformation, then the elliptical distribution is obtained in a similar manner.

Let us focus on the independence of two or more random variables. Independence of variables is reflected in decomposition of the potential function. Let $p_1(x_1)$ and $p_2(x_2)$ be two densities, where x_1 and x_2 do not necessary have the same dimension. Assume that the reference density $q(y_1, y_2)$ is independent and written as $q(y_1, y_2) = q_1(y_1)q_2(y_2)$. Denote the potential function of $p_i(x_i)$ with respect to the reference density $q_i(y_i)$ by $\psi_i(x_i)$ $(i = 1, 2)$. Then the potential function of $p_1(x_1)p_2(x_2)$ is $\psi(x_1, x_2) := \psi_1(x_1) + \psi_2(x_2)$. In fact, we can easily prove $p[\nabla\psi](x) = p_1(x_1)$ $p_2(x_2)$. The e-model also has such property but the m-model does not.

We proceed to consider the conditional independence of two or more random variables. Unfortunately, the conditional independence is not described by any affine subset in the gradient representation. For example, let the reference density $q(y)$ be the standard normal density and let the gradient map $g(x)$ be a linear map Kx with a positive definite matrix K. Then x has the distribution $N(0, K^{-2})$. Let $\Sigma = K^{-2}$. It is well known that x_1 and x_2 is conditionally independent given $(x_3, \ldots, x_m)$ if and only if $(\Sigma^{-1})_{12} = 0$. This condition is given by a non-affine relation

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

in terms of K. The e-model can exactly describe the conditional independence.

Next we point out that the g-model does not have any normalizing constant to obtain the likelihood function. This property is quite different from the e-model. For the e-model, $C(\theta)$ in (3) is difficult to compute in general and the Markov Chain Monte Carlo method is often needed.

Finally we remark that any g-model can be extended such that any point-mass distribution is included in the extended model as an extreme point. Consider a g-model $\mathcal{M} = \{p[g](x) \mid g(x) = \sum_{a=1}^{p} \theta^a g_a(x),\ \theta \in \Theta\}$. Then the model $\widetilde{\mathcal{M}}$ defined by the following formula is also a g-model:

$$\widetilde{\mathcal{M}} := \left\{ p[g](x) \mid g(x) = b + \tau \sum_{a=1}^{p} \theta^a g_a(x),\ \theta \in \Theta,\ b \in \mathbb{R}^m,\ \tau > 0 \right\}.$$

The potential function is $\psi(x) = b^\top x + \tau \sum_{a=1}^{p} \theta^a \psi_a(x)$ with $g_a = \nabla \psi_a$. We call $\widetilde{\mathcal{M}}$ the *conic extension* of $\mathcal{M}$. The conic extension makes the model flexible as the following proposition.

Table 1 Properties of models

Advantages	g-Model (condition of q)	m-Model	e-Model
Log-likelihood is concave	○ (log-concave)	○	○
All normal distributions can be included	○ (normal)	–	○
Independency can be described	○ (independent)	–	○
Conditional independency can be described	– (–)	–	○
There is no normalizing constant	○ (any)	○	–
Conic extension is available	○ (any)	–	–

Proposition 8 *Let $\mathcal{M}$ be a g-model and $\widetilde{\mathcal{M}}$ be its conic extension. Then, for any $x_0 \in \mathbb{R}^m$ there exists a sequence $\{p[g^{(n)}]\}_{n=1}^{\infty}$ in $\widetilde{\mathcal{M}}$ such that the distribution $p[g^{(n)}](x)\mathrm{d}x$ converges weakly to the Dirac distribution $\delta_{x_0}(\mathrm{d}x)$.*

Proof Let $p[g](x)$ be an element of $\mathcal{M}$ and consider a sequence $g^{(n)}(x) = n(g(x) - g(x_0))$. Then $p[g^{(n)}]$ is an element of $\widetilde{\mathcal{M}}$ from the definition. We prove that $p[g^{(n)}](x)\mathrm{d}x$ converges weakly to $\delta_{x_0}(\mathrm{d}x)$. Let f be any bounded continuous function on $\mathbb{R}^m$. Then

$$\begin{aligned}\int f(x)\, p[g^{(n)}](x)\mathrm{d}x &= \int f(x) q(n(g(x) - g(x_0)))\det(nG(x))\mathrm{d}x \\ &= \int f(g^{-1}(g(x_0) + y/n))q(y)\mathrm{d}y.\end{aligned}$$

By the dominated convergence theorem, the right-hand side converges to $f(x_0)$ as $n \to \infty$. This completes the proof. □

We remark that any e-models have a similar extension called the *exponential dispersion model* in that the dispersion parameter plays a role of our scaling parameter $\tau > 0$ (e.g. Jørgensen 1987). However the exponential dispersion model is not an e-model with respect to the dispersion parameter in general. We summarize these properties in Table 1.

4 Application

We apply the g-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 $\{X_i\}_{i=1}^{10}$ (100 m, long-jump, shot-put, high-jump, 400 m, 110 m-hurdle, disc-throw, pole-vault,

javelin-throw and 1,500 m) by 50 athletes. We first normalize the data such that the sample mean and variance of each variable are 0 and 1, respectively. We consider each marginal density $p(X_i, X_j, X_k)$ $(1 \le i < j < k \le 10)$, not the joint density $p(X_1, \ldots, X_{10})$, for simplicity. The empirical third cumulant is shown in Fig. 3a. The model of three-dimensional interaction as Example 1 is used. The potential function is

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

where $\{e_\lambda\}$ is defined by (2), and $K \in \mathbb{S}_+(m)$ and $\theta \in \mathbb{R}$ are unknown parameters. A conservative region that assures the convexity of ψ is given by $|\theta| < (2 \times 3^{-3/2})\rho_{\min}(K)$, where $\rho_{\min}(K)$ is the minimum eigenvalue of K, as will be shown in Appendix A.

We calculate Akaike's information criterion (AIC) of the two submodels $\theta = 0$ and $\theta \neq 0$ for all triplets $\{(i, j, k)\}_{1 \le i < j < k \le 10}$ of the ten variables. The result is shown in Fig. 3b. The triplet having the most significant difference of AIC is (X_4, X_5, X_6) (that denotes high-jump, 400 m and 110 m 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 Fig. 3c–e 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.

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.

5 Discussion

We defined the g-model by using the gradient representation of the probability densities. The g-model has many good properties including the concavity of the log-likelihood. A g-model was applied to detect the three-dimensional interaction of the decathlon data.

We have not discussed regression models. A generalization of g-models to this direction will be available by adding the explanatory variables in the potential function. More generally, graphical modeling will be described in terms of the gradient representation. These important generalizations are left on the future work.

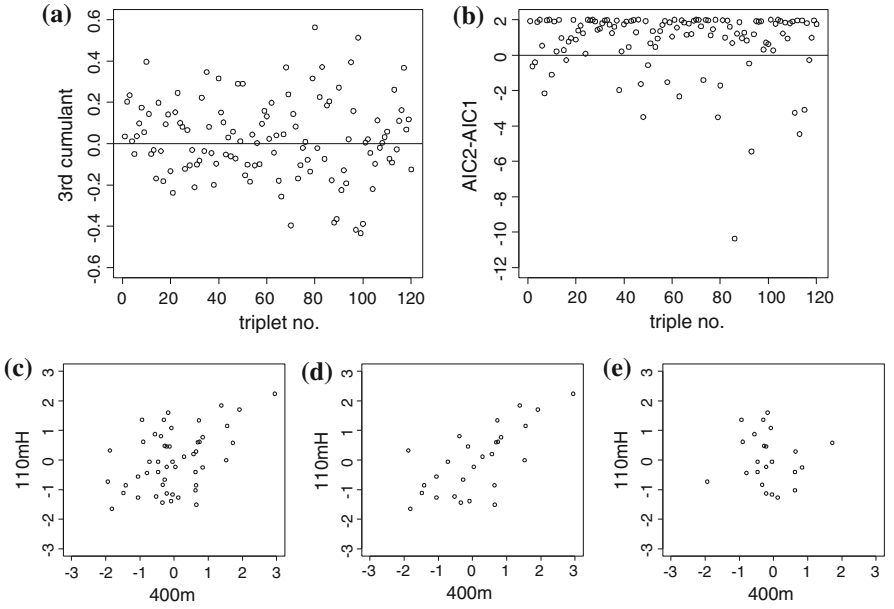

Fig. 3 Detection of three-dimensional interaction on the decathlon data. **a** The empirical third cumulant for each triplet. The *horizontal axis* represents the 120 triplets arranged as $(1, 2, 3), (1, 2, 4), \ldots, (7, 8, 9)$. **b** Difference between AIC of the model $\theta \neq 0$ and $\theta = 0$. 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. **c** Scatter plot of X_5 (horizontal) versus X_6 (vertical). The correlation is 0.465. **d** Scatter plot of X_5 versus X_6 conditioned by $X_4 > 0$ (the correlation is 0.688). **e** Scatter plot of X_5 versus X_6 conditioned by $X_4 < 0$ (the correlation is -0.049)

Appendix A: The feasible region of the three-dimensional interaction model

Consider the potential function

$$\psi(x) = \frac{1}{2}x^\top K x + \epsilon \sum_{\lambda=1}^{4} f(e_\lambda^\top x), \quad x \in \mathbb{R}^3, \tag{4}$$

where K is a positive definite matrix, $\{e_\lambda\}_{\lambda=1}^4$ is given by (2), and f is a one-dimensional odd function (e.g. $f(z) = \arctan(z)$). We give a lemma on convexity of ψ. Denote the minimum and maximum eigenvalue of K by $\rho_{\min}(K)$ and $\rho_{\max}(K)$, respectively. Let μ be the maximum value of $f''(z)$ over $z \in \mathbb{R}$. For example, $\mu = 3^{3/2}/8$ if $f(z) = \arctan(z)$.

Lemma 9 *If $|\epsilon| < (4\mu)^{-1}\rho_{\min}(K)$, then ψ in (4) is strictly convex. Conversely, if ψ is strictly convex, then $|\epsilon| < (4\mu)^{-1}\rho_{\max}(K)$.*

Proof For any vector x and any unit vector u, we have

$$u^\top(\nabla\nabla^\top\psi(x))u = u^\top K u + \epsilon \sum_{\lambda=1}^{4} f''(e_\lambda^\top x)(e_\lambda^\top u)^2.$$

Since the minimum value of $\epsilon f''(z)$ over $z \in \mathbb{R}$ is $-\mu|\epsilon|$ and $\sum_{\lambda=1}^{4} e_\lambda e_\lambda^\top$ is $4I$, we have

$$u^\top(\nabla\nabla^\top\psi(x))u \geq \rho_{\min}(K) - \mu|\epsilon| \sum_{\lambda=1}^{4}(e_\lambda^\top u)^2 \;=\; \rho_{\min}(K) - 4\mu|\epsilon|.$$

Thus the first part of the lemma is proved.

We now prove the second part. Assume that ϵ is positive. The negative case is similar. Let z_* be the maximal point of $f''(z)$. Let $x = (z_*, z_*, z_*)^\top$ and $u = (1, -1, 0)^\top/\sqrt{2}$. We can check that $e_\lambda^\top u = 0$ for $\lambda = 1$ and 3, and that $f''(e_\lambda^\top x) = -\mu$ for $\lambda = 2$ and 4. Therefore

$$\begin{aligned} u^\top(\nabla\nabla^\top\psi(x))u &\leq \rho_{\max}(K) + \epsilon \sum_{\lambda=1}^{4} f''(e_\lambda^\top x)(e_\lambda^\top u)^2 \\ &= \rho_{\max}(K) - \epsilon \sum_{\lambda=1}^{4} \mu(e_\lambda^\top u)^2 \\ &= \rho_{\max}(K) - 4\mu\epsilon. \end{aligned}$$

The left-hand side is positive if ψ is strictly convex. □

Corollary 10 *Let $K = I$. Then ψ in (4) is strictly convex if and only if $|\epsilon| < (4\mu)^{-1}$.*

Acknowledgments The author thanks professors M. Miyakawa, S. Aoki, A. Takemura and F. Komaki for their helpful comments and encouragement to write this paper. This work is supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Abdellaoui, T. (1998). Optimal solution of a Monge–Kantorovitch transportation problem. *Journal of Computational and Applied Mathematics, 96*, 149–161.

Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B* (*Methodological*), *26*, 211–252.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics, 44*, 375–417.

De Oliveira, V., Kedema, B., Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association, 92*(440), 1422–1433.

Easton, G. S., McCulloch, R. E. (1990). A multivariate generalization of quantile–quantile plots. *Journal of the American Statistical Association, 85*(410), 376–386.

Haker, S., Zhu, L., Tannenbaum, A., Angenent, S. (2004). Optimal mass transport for registration and warping. *International Journal of Computer Vision, 60*(3), 225–240.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society, Series B* (*Methodological*), *49*(2), 127–162.

Knott, M., Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications, 43*(1), 39–49.

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal, 80*(2), 309–323.

Miyakawa, M. (1997). *Graphical modeling* (*in Japanese*). Tokyo: Asakura Shoten.

Rachev, S. T., Rüschendorf, L. (1998). *Mass transportation problems—I: Theory and II: Applications*. New York: Springer.

Rockafellar, R. T. (1972). *Convex analysis*. Princeton: Princeton University Press.
Rüschendorf, L., Rachev, S. T. (1990). A characterization of random variables with minimum L^2-distance. *Journal of the Multivariate Analysis, 32*, 48–54.
Sei, T. (2006). Parametric modeling based on the gradient maps of convex functions. Technical report, METR2006-51, Department of Mathematical Engineering, University of Tokyo.
Villani, C. (2003). *Topics in optimal transportation*. Providence: AMS.
Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: Wiley.