Information geometry of modal linear regression

Keishi Sando¹ · Shotaro Akaho^{2,3} · Noboru Murata^{3,4} · Hideitsu Hino^{3,5}

Received: 29 August 2018 / Revised: 12 June 2019 / Published online: 1 July 2019
© The Author(s) 2019

Abstract

Modal linear regression (MLR) is used for modeling the conditional mode of a response as a linear predictor of explanatory variables. It is an effective approach to dealing with response variables having a multimodal distribution or those contaminated by outliers. Because of the semiparametric nature of MLR, constructing a statistical model manifold is difficult with the conventional approach. To overcome this difficulty, we first consider the information geometric perspective of the modal expectation–maximization (EM) algorithm. Based on this perspective, model manifolds for MLR are constructed according to observations, and a data manifold is constructed based on the empirical distribution. In this paper, the *em* algorithm, which is a geometric formulation of the EM algorithm, of MLR is shown to be equivalent to the conventional EM algorithm of MLR. The robustness of the MLR model is also discussed in terms of the influence function and information geometry.

Keywords Modal linear regression · Information geometry · Kernel density estimation

Part of this work is supported by JST KAKENHI 16K16108, 16H02842, 17H01793 and JST CREST JPMJCR1761.

✉ Hideitsu Hino
hino@ism.ac.jp

¹ Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan

² National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki 305-8568, Japan

³ RIKEN AIP, Tokyo 103-0027, Japan

⁴ Faculty of Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

⁵ The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8565, Japan

1 Introduction

Linear regression is used to model the conditional mean of a response variable y given the predictor variable x . A well-known least-squares estimator for linear regression coefficients is highly sensitive to outliers. To alleviate this problem, many estimators have been developed, such as robust M-estimators [12,13]. However, the consistency of the robust M-estimators requires the conditional error distribution to have homoscedasticity and symmetry given a predictor. In reality, various types of data do not follow these assumptions (e.g., wages, prices, and expenditures). Baldauf and Silva [3] pointed out that the estimation cannot be consistent unless data follow proper assumptions, which is also the case in some real-world applications [18].

Modal linear regression (MLR) is used to model the conditional mode of y given x by using a linear predictor function of x . MLR relaxes the distribution assumptions for the M-estimators of linear regression and is robust against outliers compared to the least-squares estimation of linear regression coefficients. It is also robust against violations of standard assumptions on the usual mean regression, such as heavy-tailed noise and skewed conditional and noise distributions. Kemp and Silva [14] and Yao and Li [23] proved that their estimators for the MLR model are consistent even when the error distribution is asymmetric. For the above reasons, improving methods for mode estimation has been an important topic of research for many years.

In information geometry [2], a manifold that consists of statistical models is called a model manifold. Information geometric formulation of estimating algorithm is useful for understanding the behavior and characteristic of the algorithm. For example, the procedure of parameter estimation is regarded as a projection from a point in the statistical manifold to a point in the model manifold. Modal linear regression is known to be robust to outliers, and we aim to elucidate the source of robustness by formulating the estimation procedure as geometric operations. Identifying and understanding the source of robustness of the estimation algorithm and statistical model associated with the modal linear regression would be helpful for developing other algorithms and models robust to outliers. In information geometry, a model manifold is often constructed by using a parametric distribution. Because of the lack of a parametric distribution, constructing a manifold that corresponds to the MLR model is difficult with conventional approaches. The contribution of this paper is being the first to obtain an information geometric perspective on MLR.

1.1 Related works

The modal regression model is related to kernel density estimation (KDE) [22], which is a nonparametric method for estimating the probability density function of observed data. Parzen [19] revealed the sufficient conditions for the L^2 -consistency and asymptotic normality of KDE. They also derived the conditions for the consistency and asymptotic normality of a mode estimation constructed on the basis of KDE. Epanechnikov [7] found the optimal kernel function for KDE under some conditions. In general, if the probability density function of a random variable X has a unique mode and is symmetric with respect to the mode, $\Pr(p - w \leq X \leq p + w)$ with fixed w is max-

imized when p is the mode. Based on this property, Lee [15] proposed an estimator for the coefficients of MLR. The MLR model is formulated as follows:

$$y = x^{\top} \beta + \epsilon, \quad \text{where } \text{Mode}[\epsilon; x] = 0.$$

The estimator of Lee [15] is consistent when there exists $w > 0$ and the probability density function of ϵ is symmetric in the range of $0 \pm w$. Lee [16] improved the objective function for the MLR model to become more tractable by using a quadratic kernel. Lee [15,16] required the conditional probability density function of ϵ given x to be symmetric around 0 in the range of $\pm w$. Kemp and Silva [14] proved that the mode estimator for the coefficients of MLR is consistent even if symmetry is not satisfied. Yao and Li [23] proposed an expectation–maximization (EM) algorithm to estimate the coefficients of MLR.

Besides linear modeling, other studies have used semiparametric or nonparametric approaches to modal regression. Gannoun et al. [9] proposed a semiparametric modal regression model that assumes a linear relation between the mean, median, and mode. Yao et al. [24] developed a local modal regression that estimates the conditional mode of response as a polynomial function of predictors. Chen et al. [4] defined the modal manifold as a union of sets in which the first derivative of the conditional density is zero and the second derivative is negative. Then, the modal manifold is estimated by kernel estimates of derivatives of densities.

In information geometry, a model manifold is often constructed by using a parametric distribution. Estimates are regarded as the projection of an empirical distribution onto the model manifold. In the case of linear regression, we construct a model manifold based on the assumption that an error variable has a normal distribution. Because of the lack of a parametric distribution, constructing a model manifold that corresponds to the MLR model is difficult with conventional approaches. Some studies have considered nonparametric models for information geometry. Pistone and Sempì [20] showed a well-defined Banach manifold for probability measures. Grasselli [10] addressed the Fisher information and α -connections for the Banach manifold. Zhang [25] discussed the relationship between divergence functions, the Fisher information, α -connections, and fundamental issues in information geometry. Takano et al. [21] proposed a framework for a nonparametric e -mixture estimation. In contrast to these nonparametric approaches to information geometry, in this paper we consider information geometry associated with a semiparametric MLR model. We propose to construct a model manifold by using observations, as is done when constructing an empirical distribution with conventional approaches. Our proposal gives a geometric viewpoint to the MLR model.

2 Modal linear regression

Let $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$ be a set of predictor variables and a response variable, respectively. The original least-squares for linear regression estimates a conditional mean of y given x , while MLR estimates a conditional mode of y given x . In this section, we briefly explain the EM algorithm of MLR introduced by Yao and Li [23].

2.1 Formulation

Suppose that $\{x_i, y_i\}_{i=1}^N$ are i.i.d. observations, where the i -th predictor variable is denoted by $x_i \in \mathbb{R}^p$ and the corresponding response is denoted by $y_i \in \mathbb{R}$. With MLR, we model a conditional mode of y given x by a linear function of x :

$$\text{Mode}[y; x] = x^\top \beta,$$

where $\text{Mode}[y; x] = \underset{y}{\operatorname{argmax}} f(y|x)$ for the conditional density function $f(y|x)$. Namely, y and x are related as

$$y = x^\top \beta + \epsilon, \quad \text{where} \quad \text{Mode}[\epsilon; x] = 0. \quad (1)$$

To estimate β , Lee [15] introduced a loss function with the form

$$l(\beta; y, x) = -\phi_h \left(y - x^\top \beta \right), \quad (2)$$

where $\phi_h(x) = \frac{1}{h} \phi \left(\frac{x}{h} \right)$, $\phi(\cdot)$ is a kernel function, and h is a bandwidth parameter. Minimizing the empirical loss leads to the estimate $\hat{\beta}$ of the linear coefficient:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^N \phi_h(y_i - x_i^\top \beta). \quad (3)$$

In this paper, $\phi(\cdot)$ denotes a standard normal density function. The consistency and asymptotic normality of the estimate $\hat{\beta}$ obtained by (3) have been established under certain regularity conditions on the samples, kernel function, parameter space, and vanishing rate of the bandwidth parameter [14].

2.2 EM algorithm for MLR

Here, we introduce the EM algorithm for MLR parameter estimation proposed by Yao and Li [23]. The algorithm consists of two steps starting from an initial estimate $\beta^{(1)}$:
E-Step

Consider the surrogate function

$$\gamma(\beta; \beta^{(k)}) = \sum_{i=1}^N \pi_i^{(k)} \log \left[\frac{\frac{1}{N} \phi_h(y_i - x_i^\top \beta)}{\pi_i^{(k)}} \right], \quad (4)$$

where

$$\pi_i^{(k)} = \frac{\phi_h(y_i - x_i^\top \beta^{(k)})}{\sum_{j=1}^N \phi_h(y_j - x_j^\top \beta^{(k)})}, \quad i = 1 \dots N. \quad (5)$$

This function satisfies

$$\gamma(\beta^{(k)}; \beta^{(k)}) = \log \left[\frac{1}{N} \sum_{i=1}^N \phi_h \left(y_i - x_i^\top \beta^{(k)} \right) \right] \quad (6)$$

and

$$\begin{aligned} & \log \left[\frac{1}{N} \sum_{i=1}^N \phi_h \left(y_i - x_i^\top \beta \right) \right] \\ &= \log \left[\sum_{i=1}^N \pi_i^{(k)} \frac{\frac{1}{N} \phi_h \left(y_i - x_i^\top \beta \right)}{\pi_i^{(k)}} \right], \quad \text{by Jensen's inequality} \\ &\geq \sum_{i=1}^N \pi_i^{(k)} \log \left[\frac{\frac{1}{N} \phi_h \left(y_i - x_i^\top \beta \right)}{\pi_i^{(k)}} \right] = \gamma(\beta; \beta^{(k)}). \end{aligned} \quad (7)$$

M-Step

In this step, the parameter β is updated to increase the value of $\frac{1}{N} \sum_{i=1}^N \phi_h \left(y_i - x_i^\top \beta \right)$. The updated parameter $\beta^{(k+1)}$ is given as

$$\beta^{(k+1)} = \underset{\beta}{\operatorname{argmax}} \gamma(\beta; \beta^{(k)}). \quad (8)$$

The following inequality holds:

$$\begin{aligned} \log \left[\frac{1}{N} \sum_{i=1}^N \phi_h \left(y_i - x_i^\top \beta^{(k+1)} \right) \right] &\geq \gamma(\beta^{(k+1)}; \beta^{(k)}) \\ &\geq \gamma(\beta^{(k)}; \beta^{(k)}) = \log \left[\frac{1}{N} \sum_{i=1}^N \phi_h \left(y_i - x_i^\top \beta^{(k)} \right) \right]. \end{aligned}$$

Equation (8) is equivalent to (9).

$$\beta^{(k+1)} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^N \pi_i^{(k)} \log \phi_h(y_i - x_i^\top \beta). \quad (9)$$

When $\phi(\cdot)$ is a standard normal density function,

$$\beta^{(k+1)} = \left(X^\top W_k X \right)^{-1} X^\top W_k y, \quad W_k = \operatorname{diag} \left(\pi_1^{(k)} \dots \pi_N^{(k)} \right).$$

The property of the estimate $\hat{\beta}$ was discussed in [23].

3 Information geometry

In this section, we briefly explain information geometry, statistical inference, and the *em* algorithm. Information geometry [2] is a framework for formulating spaces consisting of probability density functions by means of differential geometry. Let S be a statistical manifold which is composed of probability distributions.

A parametric family of probability distributions of the form

$$f(x; \theta) = \exp \left\{ C(x) + \sum_{i=1}^n \theta_i F_i(x) - \psi(\theta) \right\},$$

$$\text{where } \psi(\theta) = \log \int \exp \left\{ C(x) + \sum_{i=1}^n \theta_i F_i(x) \right\} dx \quad (10)$$

is called the exponential family and it has a critical role in information geometry. When $\{F_1 \dots F_n, 1\}$ is linearly independent, θ and $f(x; \theta)$ have a one-to-one correspondence, and $(\theta_i)_{i=1}^n$ is an affine coordinate system of the statistical manifold. There is another useful coordinate system:

$$\eta_i = E_{f(x; \theta)} [F_i(x)] = \int f(x; \theta) F_i(x) dx, \quad i = 1 \dots n. \quad (11)$$

$(\theta_i)_{i=1}^n$ and $(\eta_i)_{i=1}^n$ are called the natural parameters and expectation parameters, respectively, of the exponential family.

For an exponential family equipped with coordinate systems $\{(\theta_i)_{i=1}^n, (\eta_i)_{i=1}^n\}$, there exist the potential functions $\psi(\theta)$ and $\phi(\eta)$ satisfying the following relation [2]:

$$\forall i = 1 \dots n, \quad \eta_i = \frac{\partial \psi}{\partial \theta_i}, \quad \theta_i = \frac{\partial \phi}{\partial \eta_i}, \quad \psi(\theta) + \phi(\eta) - \sum_{i=1}^n \theta_i \eta_i = 0. \quad (12)$$

In information geometry, a statistical inference is often regarded as the projection of an empirical distribution onto a model manifold, which is a submanifold of S . Projection is defined as a procedure to find a point which minimizes the discrepancy between a point and a submanifold, hence it is important to measure the discrepancy between two probability distributions. We can define *divergence* from a point (probability distribution) $p \in S$ to $q \in S$ by using the coordinate systems and potential functions as follows:

$$D(p||q) = \psi(\theta(p)) + \phi(\eta(q)) - \sum_{i=1}^n \theta_i(p) \eta_i(q), \quad \forall p, q \in S. \quad (13)$$

For an exponential family, there are two natural definitions of divergences, the *e*- and *m*-divergence. The *m*-divergence is defined as

$$\begin{aligned}
 D^{(m)}(p||q) &= \varphi(\eta(p)) + \psi(\theta(q)) - \sum_{i=1}^n \eta_i(p)\theta_i(q), \quad \text{from Eq. (12)} \\
 &= \sum_{i=1}^n \theta_i(p)\eta_i(p) - \psi(\theta(p)) \\
 &\quad - \left\{ \sum_{i=1}^n \theta_i(q)\eta_i(p) - \psi(\theta(q)) \right\}, \quad \text{from Eq. (11)} \\
 &= E_{\theta(p)} \left[\sum_{i=1}^n \theta_i(p)F_i(x) - \psi(\theta(p)) \right] - E_{\theta(p)} \left[\sum_{i=1}^n \theta_i(q)F_i(x) - \psi(\theta(q)) \right] \\
 &= E_{\theta(p)} [\log p(x)] - E_{\theta(p)} [\log q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx,
 \end{aligned}$$

which is equal to the KL-divergence. The e -divergence is defined as follows:

$$\begin{aligned}
 D^{(e)}(p||q) &= \psi(\theta(p)) + \varphi(\eta(q)) - \sum_{i=1}^n \theta_i(p)\eta_i(q) \\
 &= \varphi(\eta(q)) + \psi(\theta(p)) - \sum_{i=1}^n \eta_i(q)\theta_i(p) \\
 &= D^{(m)}(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx.
 \end{aligned}$$

The EM algorithm [5] is a method for estimating the maximum likelihood parameters in a latent variable model. In information geometry, the exponential - mixture (em) algorithm [1] corresponds to the EM algorithm. For a latent variable model, an empirical distribution based on observations is not unique. A manifold that consists of empirical joint probability distributions of observable variables and latent variables is called a data manifold $\mathcal{D}$. The em algorithm finds the points $p^* \in \mathcal{M}$ and $q^* \in \mathcal{D}$ that minimize the KL-divergence from q^* to p^* by iterating the following two steps starting from an initial guess $p^{(1)}$. Figure 1 is a conceptual diagram of the em algorithm.

e -step

e -projection of $p^{(k)} \in \mathcal{M}$ onto the data manifold $\mathcal{D}$.

$$q^{(k)} = \operatorname{argmin}_{q \in \mathcal{D}} D^{(e)}(p^{(k)}||q).$$

m -step

m -projection of $q^{(k)} \in \mathcal{D}$ onto the model manifold $\mathcal{M}$.

$$p^{(k+1)} = \operatorname{argmin}_{p \in \mathcal{M}} D^{(m)}(q^{(k)}||p).$$

Fig. 1 Update process of the *em* algorithm

4 MEM algorithm and its information geometry

In this section, we introduce the modal EM (MEM) algorithm [17] for estimating the mode given a probability density function and provide an information geometric perspective. The MEM algorithm is an iterative procedure similar to the EM algorithm, but there are no explicit *observations*, hence the construction of an empirical density function is nontrivial. In this paper, we propose to construct an empirical density function based on the assumption that one *pseudo-observation* is obtained. This assumption has a critical role for constructing an empirical density function of the MEM algorithm and the MLR model, in which the same difficulty exists because of the absence of the explicit observation.

4.1 MEM algorithm

Consider the Gaussian mixture model, whose parameters are known. In general, it is difficult to express the mode of the Gaussian mixture model in a closed form, even if the parameters are known. In order to obtain the mode, we need to execute a numerical optimization. The MEM algorithm [17] is an iterative method for finding a local mode of a mixture distribution with the form

$$f(x) = \sum_{i=1}^K \pi_i f_i(x), \quad x \in \mathbb{R}^p, \quad \text{where} \quad \begin{cases} \pi_i \geq 0, & \sum_{i=1}^K \pi_i = 1, \\ f_i : \mathbb{R}^p \rightarrow \mathbb{R} \text{ is a probability density function,} \end{cases}$$

where all of the parameters in this model are known. The purpose of the MEM algorithm is to find the mode of $f(x)$: $x^* = \underset{x}{\operatorname{argmax}} f(x)$. Li et al [17] proposed to iterate the following two steps starting from the initial estimate $x^{(1)}$:

E-step

$$p_i^{(k)} = \frac{\pi_i f_i(x^{(k)})}{f(x^{(k)})}, \quad i = 1 \dots K. \quad (14)$$

M-step

$$x^{(k+1)} = \operatorname{argmax}_x \sum_{i=1}^K p_i^{(k)} \log f_i(x). \quad (15)$$

4.2 Pseudo-observation

The purpose of the MEM algorithm is to find an optimal point that maximizes a probability density function $f(x)$. In information geometry, projection is a procedure of finding the probability density function in a set of density functions that minimizes the discrepancy from the empirical density function. In this section, we show how we can interpret the MEM algorithm as a problem of finding an optimal function from a set of probability density functions, which leads us to the information geometric perspective of the MEM algorithm.

Let a probability density function $s(x ; m)$ be

$$s(x ; m) = f(x + m),$$

where $m \in \mathbb{R}^p$ is a parameter for $s(x ; m)$. From this definition, the mode of $s(x ; m^*)$ is 0 when the mode of $f(x)$ is m^* . On the other hand, if m^* is the optimal solution of $\max_m s(0 ; m)$, then the mode of $f(x)$ is m^* . Two problems $\max_x f(x)$ and $\max_m s(0 ; m)$ have the same solution, but the idea behind these problems are different. The former is the problem that finds an optimal point from a domain of $f(x)$, while the latter considers a set of probability density functions shifted by m and finds an optimal function that maximizes the value at $x = 0$ from the set. Search space of these two problems are different, and that of the latter interpretation is convenient when we formulate the MEM algorithm from an information geometric perspective because its search space is a set of function as in the procedure of projection.

To derive the information geometric formulation of the MEM algorithm, we introduce the latent variable $Z \in \{1 \dots K\}$ to the mixture model $f(x) = \sum_{i=1}^K \pi_i f_i(x)$. The latent variable specifies a mixture component from which an observation x is obtained. The joint probability density function $g(x, z)$ is expressed as follows [1]:

$$g(x, z) = \prod_{i=1}^K [\pi_i f_i(x)]^{\delta_i(z)}, \quad \text{where} \quad \begin{cases} \pi_i \geq 0, \quad \sum_{i=1}^K \pi_i = 1, \\ f_i \text{ is a probability density function,} \\ \delta_i(z) = \begin{cases} 1 & i = z, \\ 0 & i \neq z. \end{cases} \end{cases} \quad (16)$$

In the above discussion, we considered finding $\operatorname{argmax}_m s(0 ; m)$ from a set of probability density function parameterized by m . In the same way, we define a joint probability density function $l(x, z; m)$ and a corresponding model manifold $\mathcal{M}$ as follows:

$$l(x, z; m) = g(x + m, z),$$

$$\mathcal{M} = \{l(x, z; m) \mid m \in \mathbb{R}^p\}.$$

We then consider an empirical density function and data manifold. In general, an empirical density function is constructed based on observations. For example, when observations $\{x_i \in \mathbb{R}^p\}_{i=1}^N$ are i.i.d., the empirical density function is defined as $\frac{1}{N} \sum_{i=1}^N \delta(x - x_i)$, where $\delta(\cdot)$ denotes the Dirac delta function. In the formulation of the MEM algorithm, the construction of the empirical density function is nontrivial because there are no explicit *observations*. In the interpretation of $\operatorname{argmax}_m s(0 ; m)$, we considered the case of $x = 0$ and $s(0 ; m)$ is treated as a function of m . It is equivalent to assuming that “one observation $x = 0$ is obtained”, and we interpret the procedure of the MEM algorithm as the maximum likelihood estimation, namely, finding the parameter values $m \in \mathbb{R}^p$ that maximizes the likelihood of $s(0 ; m)$. This assumption leads to the following definition of the empirical density function $p(x)$:

$$p(x) = \delta(x - 0) = \delta(x).$$

By introducing the latent variable $Z \in \{1 \dots K\}$, we extend $p(x)$ to an empirical joint density function of X and Z :

$$h(x, z) = p(x)q(z \mid x),$$

and by introducing the parameters $\{q_i\}_{i=1}^K$, we model the conditional density function $q(z \mid x)$ as

$$q(z \mid x) = \sum_{i=1}^K q_i \delta_i(z), \quad \text{where } q_i \geq 0, \sum_{i=1}^K q_i = 1.$$

Then, the empirical joint density function $h(x, z ; q_1 \dots q_K)$ is expressed as

$$h(x, z ; q_1 \dots q_K) = \sum_{i=1}^K q_i \delta(x) \delta_i(z), \quad \text{where } q_i \geq 0, \sum_{i=1}^K q_i = 1. \quad (17)$$

The data manifold $\mathcal{D}$ is defined as follows:

$$\mathcal{D} = \left\{ h(x, z ; q_1 \dots q_K) \mid q_i \geq 0, \sum_{i=1}^K q_i = 1 \right\}. \quad (18)$$

It is shown in Appendix A that $\mathcal{D}$ is in a mixture family.

We consider the MEM algorithm to be a maximum likelihood estimation problem. Thus, we can consider the e - and m -projection between the model manifold $\mathcal{M}$ and the data manifold $\mathcal{D}$ as follows:

$$h(x, z; q_1^{(k)} \dots q_K^{(k)}) = \operatorname{argmin}_{h \in \mathcal{D}} D^{(e)}(l(\cdot, \cdot; m^{(k)}) || h), \quad (19)$$

$$g(x, z; m^{(k+1)}) = \operatorname{argmin}_{g \in \mathcal{M}} D^{(m)}(h(\cdot, \cdot; q_1^{(k)} \dots q_K^{(k)}) || g). \quad (20)$$

The detailed calculation of the e -projection in (19) is provided in Appendix B, and the optimal parameter for $h \in \mathcal{D}$ is given as

$$q_i^{(k)} = \frac{\pi_i f_i(m^{(k)})}{f(m^{(k)})}, \quad i = 1 \dots K. \quad (21)$$

The m -projection in (20) is equivalent to

$$\max_{m \in \mathbb{R}^p} \sum_{i=1}^K q_i^{(k)} \log f_i(m). \quad (22)$$

The detailed derivation is shown in Appendix C.

4.3 Summary: information geometry of MEM

To provide an information geometric perspective of the MEM algorithm, we interpret the algorithm as a problem of finding an optimal probability density function that maximizes the value at $x = 0$. This enables us to cast the MEM algorithm as the maximum likelihood estimation. The e -projection of the model distribution $l(x, z; m^{(k)})$ onto $\mathcal{D}$ gives the optimal $q_i^{(k)}$, $i = 1 \dots K$, which is equivalent to (14) in the original MEM algorithm. The m -projection of $h(x, z; q_1^k \dots q_K^{(k)})$ onto $\mathcal{M}$ derives the optimal $m^{(k+1)}$, which is consistent with (15) in the original MEM algorithm.

5 Information geometry of MLR

In this section, we analyze MLR from the viewpoint of information geometry. We elucidate the source of the difficulty with constructing a model manifold and data manifold for the MLR model and propose a framework for geometrically formulating the MLR model.

5.1 Problem of constructing manifolds

In order to elucidate the source of the difficulty with constructing manifolds for the MLR model, we consider the parameter estimation of a Gaussian mixture model as a

specific example of statistical inferences in information geometry. Suppose that observations $\{x_i \in \mathbb{R}^p\}_{i=1}^N$ are i.i.d. subject to a Gaussian mixture distribution expressed as

$$f(x; \mu, \Sigma) = \sum_{i=1}^K \pi_i g(x; \mu_i, \Sigma_i),$$

$$\text{where } \begin{cases} \pi_i \geq 0, \sum_{i=1}^K \pi_i = 1, \\ g(x; \mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^p} \sqrt{\det(\Sigma_i)}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i) \right\}. \end{cases}$$

Then, the model manifold consists of Gaussian mixture density functions whose parameters are the means and covariance matrices. The data manifold is constructed based on the empirical density function $\frac{1}{N} \sum_{i=1}^N \delta(x - x_i)$.

In the parameter estimation of the Gaussian mixture model, the model manifold is constructed based on the parametric distribution. On the other hand, there is no assumption of parametric distributions in MLR. This makes it nontrivial to construct a model manifold and data manifold.

To construct the model manifold for the MLR model, we consider (i) the assumption that $\text{Mode}[\epsilon; x] = 0$ and (ii) the form of the objective function of β for the MLR model: $\frac{1}{N} \sum_{i=1}^N \phi_h(y_i - x_i^\top \beta)$. From this assumption and fact, the optimization problem expressed in (3) is regarded as a maximization problem of KDE at $\epsilon = 0$ for the probability density function of ϵ . Based on the given observations, we propose constructing the following model for MLR:

$$f(\epsilon; \beta) = \frac{1}{N} \sum_{i=1}^N \phi_h(\epsilon - \epsilon_i(\beta)), \quad (23)$$

where $\epsilon_i(\beta) = y_i - x_i^\top \beta$, $i = 1 \dots N$ and the variable ϵ denotes an error variable. We introduce the latent variable $Z \in \{1 \dots N\}$, which specifies a mixture component from which an observation is obtained. The joint density function of ϵ and Z is

$$g(\epsilon, z; \beta) = \prod_{i=1}^N \left[\frac{1}{N} \phi_h(\epsilon - \epsilon_i(\beta)) \right]^{\delta_i(z)}. \quad (24)$$

The model manifold $\mathcal{M}$ is denoted by

$$\mathcal{M} = \{g(\epsilon, z; \beta) \mid \beta \in \mathbb{R}^p\}. \quad (25)$$

It is shown in Appendix D that $\mathcal{M}$ is in a curved exponential family.

We next propose constructing a data manifold for the MLR model. The empirical density function is often constructed based on observations. Consider (i) the construction proposed in Sect. 4.2 and (ii) the assumption that $\text{Mode}[\epsilon; x] = 0$. We propose constructing the empirical density function as follows:

$$p(\epsilon) = \delta(\epsilon - 0) = \delta(\epsilon). \quad (26)$$

By introducing the latent variable $Z \in \{1 \dots N\}$ to (26), we extend $p(\epsilon)$ to the empirical joint density function of ϵ and Z :

$$h(\epsilon, z) = p(\epsilon)q(z | \epsilon).$$

By introducing the parameters $\{q_i\}_{i=1}^N$, we model the conditional density function $q(z | \epsilon)$ as

$$q(z | \epsilon) = \sum_{i=1}^N q_i \delta_i(z), \quad \text{where } q_i \geq 0, \sum_{i=1}^N q_i = 1$$

Then, the empirical joint density function $h(\epsilon, z, q_1 \dots q_N)$ is expressed as

$$h(\epsilon, z; q_1 \dots q_N) = \sum_{i=1}^N q_i \delta(\epsilon) \delta_i(z), \quad \text{where } q_i \geq 0, \sum_{i=1}^N q_i = 1 \quad (27)$$

The data manifold $\mathcal{D}$ is defined as follows:

$$\mathcal{D} = \left\{ h(\epsilon, z; q_1 \dots q_N) \mid q_i \geq 0, \sum_{i=1}^N q_i = 1 \right\}. \quad (28)$$

It is shown in Appendix E that $\mathcal{D}$ is in a mixture family.

Here, we consider the e -projection of a model with the parameters $\beta^{(k)}$ onto the data manifold:

$$\begin{aligned} & \min_{h \in \mathcal{D}} D^{(e)}(g(\cdot, \cdot; \beta^{(k)}) || h). \\ & \rightarrow \left\{ \begin{array}{l} \min_{q_1 \dots q_N} D^{(e)}(g(\cdot, \cdot; \beta^{(k)}) || h(\cdot, \cdot; q_1 \dots q_N)), \\ \text{s.t. } q_i \geq 0, \sum_{i=1}^N q_i = 1 \end{array} \right. \quad (29) \end{aligned}$$

The detailed calculation is shown in Appendix F. An optimal solution for (29) is

$$q_i^{(k)} = \frac{\phi_h(y_i - x_i^\top \beta^{(k)})}{\sum_{j=1}^N \phi_h(y_j - x_j^\top \beta^{(k)})}, \quad i = 1 \dots N, \quad (30)$$

which is equivalent to the E-step in (5). Then, we consider the m -projection of the empirical joint density function with the parameters $q_i = q_i^{(k)}$, $i = 1 \dots N$ onto the model manifold:

$$\min_{g \in \mathcal{M}} D^{(m)}(h(\cdot, \cdot; q_1 = q_1^{(k)} \dots q_N = q_N^{(k)}) || g). \quad (31)$$

Fig. 2 Conceptual diagram of the *em* algorithm corresponding to the MLR model

The detailed calculation is shown in Appendix G. The optimization problem expressed as (31) is equivalent to

$$\max_{\beta} \sum_{i=1}^N q_i^{(k)} \log \phi_h \left(y_i - x_i^{\top} \beta \right), \quad (32)$$

which is equivalent to the M-step (9).

Figure 2 shows the update process of the *em* algorithm corresponding to the MLR model parameter estimation.

In this section, we propose constructing a model manifold and data manifold for the MLR model. Although a model manifold is often constructed based on a parametric distribution assumption, we construct it based on observations. Concerning the empirical distribution, our construction is based on (i) the assumption that $\text{Mode}[\epsilon; x] = 0$ and (ii) the proposed construction in Sect. 4.2. We apply the framework of the *em* algorithm to the proposed manifolds. We demonstrate that the *e*-projection of the model onto the data manifold derives (5), and the *m*-projection of the empirical distribution onto the model manifold leads to (9).

6 On the influence function for MLR

An influence function quantifies the influence of one observation on the estimate. Let T be the functional defined for a set of probability measures and F be the probability measure. The influence function of T at F is expressed by $\text{IF}(x; T, F)$ and is defined as follows:

$$\text{IF}(x; T, F) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)F + \epsilon \Delta_x) - T(F)}{\epsilon}. \quad (33)$$

Yao and Li [23] argued that MLR is robust against outliers after investigating its breakdown point. Elucidating the reason for the robustness is important because it can be a clue to developing a novel robust estimator. There are various approaches in the literature for making an estimation robust. One representative approach is to change the manifold for the m -projection. For example, the median is often used as a robust estimate of the mean. This approach corresponds to adopting the model manifold composed of Laplacian distributions instead of Gaussian distributions. On the other hand, a robust estimator can also be obtained by changing the projection method, such as using robust divergences instead of the KL-divergence [8].

So far, we have established the equivalence of the em and EM algorithms for MLR. The em algorithm is based on the KL-divergence, which is sensitive to outliers. Thus, we conjecture that the m -projection onto the model manifold, which is composed of the error distributions given by KDE, is the source of the robustness of MLR. Here, we focus on the influence function [11] of the estimator for the MLR model.

Unfortunately, it is difficult to derive the influence functions correspond to the update of the estimate of coefficient by the e - and m -steps for MLR problem because it is not trivial how to deal with the effect of an outlier on the projection to the data-dependent manifold.

In the finite-sample robustness analysis, the rescaled version of the influence function [11] is used to evaluate the effect of an outlier $(u, v) \in \mathbb{R}^p \times \mathbb{R}$ on each projection. Suppose that the current estimate is $\beta^{(k)}$. Without outlier case, the joint density function in Eq. (24) is projected onto the data manifold $\mathcal{D}$. When the dataset is contaminated with an outlier (u, v) , the joint density is modified as

$$g_{N+1}(\epsilon, z; \beta^{(k)}) = \prod_{i=1}^{N+1} \left[\frac{1}{N+1} \phi_h(\epsilon - \epsilon_i(\beta^{(k)})) \right]^{\delta_i(z)} \quad (34)$$

where $\epsilon_{N+1}(\beta^{(k)}) = v - u^\top \beta^{(k)}$, and it is projected onto the outlier-contaminated data manifold $\mathcal{D}_{N+1}$, which is composed of $q_1 \dots q_{N+1}$. Let us express the result of the usual e -projection as $q^{(k)} = (q_1^{(k)} \dots q_N^{(k)})$ and the result of the e -projection with the outlier as $q'^{(k)} = (q_1'^{(k)} \dots q_{N+1}'^{(k)})$. Then the rescaled version of the influence function on the e -projection is expressed as follows

$$\begin{aligned} & \frac{q_i'^{(k)} - q_i^{(k)}}{\frac{1}{N+1}} \\ &= - \frac{(N+1)\phi_h(v - u^\top \beta^{(k)})\phi_h(\epsilon_i(\beta^{(k)}))}{\left(\sum_{j=1}^N \phi_h(\epsilon_j(\beta^{(k)})) \right) \left\{ \phi_h(v - u^\top \beta^{(k)}) + \sum_{j=1}^N \phi_h(\epsilon_j(\beta^{(k)})) \right\}}, \quad i = 1 \dots N. \end{aligned}$$

In the same manner, we can calculate the rescaled version of the influence function on the m -projection by comparing the result of the m -projections with and without

an outlier. The regression coefficients obtained by the m -projection with and without the outlier are denoted as $\beta^{(k+1)}$ and $\beta'^{(k+1)}$, respectively. When the sample is contaminated by an outlier (u, v) , the m -projection projects $h(\epsilon, z; q_1^{(k)} \dots q_{N+1}^{(k)})$ onto $\mathcal{M}_{N+1}$, which is a set of $g_{N+1}(\epsilon, z; \beta^{(k)})$ defined by Eq. (34) parameterized by β . The influence function on the m -projection is then written as

$$\frac{\beta'^{(k+1)} - \beta^{(k+1)}}{\frac{1}{N+1}} = (N+1) \left[\left(q_{N+1}'^{(k+1)} uu^\top + \sum_{i=1}^N q_i'^{(k+1)} x_i x_i^\top \right)^{-1} \left(q_{N+1}'^{(k+1)} vu + \sum_{i=1}^N q_i'^{(k+1)} y_i x_i \right) - \left(\sum_{i=1}^N q_i^{(k+1)} x_i x_i^\top \right)^{-1} \left(\sum_{i=1}^N q_i^{(k+1)} y_i x_i \right) \right]$$

We could derive the influence functions on each of the e - and m -projections, though, there are problems on the above consideration on the effect of an outlier. In the m -projection, we did not take the indirect effect of the outlier account, which is inherited from the e -projection for the outlier-contaminated case. A standard definition of the influence function is associated with the probability measure F for X, Y . The non-parametric nature of the proposed information geometric formulation of the modal linear regression make the conventional robustness analysis difficult.

Towards understanding of the source of robustness of MLR, we consider the influence function of $\hat{\beta}$ defined in Eq. 3, which is a standard ψ -type M-estimator, and the influence function of the estimate by the em algorithm from the viewpoint of the iteratively reweighted least-squares (IRLS) estimator.

It is possible to derive the influence function independently of the solving algorithm. For example, the regression coefficient estimate $\hat{\beta}$ of the MLR model defined by Eq. (3) is a ψ -type M-estimator, and its influence function is expressed as

$$\text{IF}(u, v; F) = \left(\int \frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \hat{\beta}(F)} x x^\top dF(x, y) \right)^{-1} \psi(u, v, \hat{\beta}(F)), \quad (35)$$

$$\text{where } \psi(u, v, \hat{\beta}(F)) = \frac{d}{dz} \phi_h(z)_{z=v-u^\top \hat{\beta}(F)} u. \quad (36)$$

We note that Eq. (35) does not depend on whether the EM algorithm or the steepest descent method is used. Details of the derivation are given in Proposition 1 of Appendix H. To see the effect of an outlier, let us consider a very simple case with $\hat{\beta}(F) = \beta^*$, a predictor variable X and an error variable ϵ are independent. We also assume the Gaussian kernel with bandwidth h is adopted as a kernel function. Then, Eq. (35) is simplified to

$$\text{IF}(u, v; F) = - \frac{h^2(v - u^\top \beta^*) \phi_h(v - u^\top \beta^*)}{\int (\epsilon^2 - h^2) \phi_h(\epsilon) dF_\epsilon(\epsilon)} \left[\int x x^\top dF_X(x) \right]^{-1} u, \quad (37)$$

Fig. 3 The contour plot of the influence function in a simple setting

where F_X , F_ϵ are probability measures of X and ϵ , respectively. Equation (37) shows that given $u_0 \in \mathbb{R}^p$, $\lim_{v \rightarrow \infty} |\text{IF}(u_0, v; F)| = 0$ holds due to $x\phi_h(x) \xrightarrow{x \rightarrow \infty} 0$. We can see the behavior of Eq. (37) in Fig. 3, in which the contour of Eq. (37) is calculated under the model $Y = X \times \beta^* + \epsilon$ with $\beta^* = 1$, $X, Y \in \mathbb{R}$, $\epsilon \sim \mathcal{N}(0, 1^2)$, $X \sim \text{Uniform}(-10, 10)$, and kernel bandwidth is set to $h = 3$. In the figure, the dashed line denotes the ground truth regression line. From this figure, it is seen that outliers around the ground truth regression line have minor impact on the estimate of regression coefficient.

On the other hand, it is also possible to define the estimate $\hat{\beta}$ as the limit of the EM algorithm derived in [23]. In particular, when the Gaussian kernel is adopted for density estimation, the resulting estimator is regarded as an IRLS estimator. Dollinger and Staudte [6] addressed a related problem when they derived the influence function of an IRLS estimator for the linear regression model. They revealed the relationship between the influence functions of the $(k + 1)$ -step and k -step estimates, and derived the influence function for the estimator as its limit.

Following the approach by Dollinger and Staudte [6], we consider the *em* algorithm for MLR as an IRLS procedure, in which an iteration of IRLS corresponds to a pair of *e*- and *m*-projection. Let the initial estimate be $\beta^{(1)}$ and the k -th estimate be $\beta^{(k)}$, respectively. Then, it is shown that the influence functions of estimates $\beta^{(k)}$ and $\beta^{(k+1)}$ satisfy the following recurrence relation:

$$\text{IF}(u, v; \beta^{(k+1)}, F) = z_k + A_k \text{IF}(u, v; \beta^{(k)}, F), \quad (38)$$

$$\text{where } \left\{ \begin{array}{l} z_k = \Sigma_k^{-1} c_k, \\ \Sigma_k = \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) x x^\top dF(x, y), \\ c_k = \phi_h \left(v - u^\top \beta^{(k)}(F) \right) \left(v - u^\top \beta^{(k+1)}(F) \right) u \\ \quad - \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) \left(y - x^\top \beta^{(k+1)}(F) \right) x dF(x, y), \\ C_k = - \int \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta^{(k)}(F)} \left(y - x^\top \beta^{(k+1)}(F) \right) x x^\top dF(x, y), \\ A_k = \Sigma_k^{-1} C_k. \end{array} \right. \quad (39)$$

Details of the derivation are given in Proposition 2 of Appendix H. In the original least squares regression problem dealt in [6], the weighted least-squares (WLS) estimator is proven to be Fisher consistent by using the symmetry of the noise distribution, and Σ_k, c_k, C_k in the recurrence relation are shown to be independent of k . In our MLR problem, though, the noise distribution cannot be assumed to be symmetric. So, we consider the converged value $\beta^{(\infty)}$ to remove the dependency on k . In this limit case, Σ_k, c_k, C_k do not depend on k , and Eq. (38) becomes

$$\text{IF}(u, v; \beta^{(\infty)}, F) = z + A \text{IF}(u, v; \beta^{(\infty)}, F). \quad (40)$$

Now we assume that $|A| < 1$, and the use of Gaussian kernel implies $\frac{d}{dz} \phi_h(z) = -\frac{1}{h^2} z \phi_h(z)$. Then, we obtain

$$\begin{aligned} \text{IF}(u, v; \beta^{(\infty)}, F) &= (\Sigma - C)^{-1} c, \\ &= \left(\int \frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \beta^{(\infty)}} x x^\top dF(x, y) \right)^{-1} \\ &\quad \times \left[\psi(u, v, \beta^{(\infty)}) + \frac{\partial}{\partial \beta} \left\{ \int \phi_h(y - x^\top \beta) dF(x, y) \right\}_{\beta^{(\infty)}} \right]. \end{aligned} \quad (41)$$

If $\beta^{(\infty)}$ is equal to $\hat{\beta}(F)$, then $\frac{\partial}{\partial \beta} \left\{ \int \phi_h(y - x^\top \beta) dF(x, y) \right\}_{\hat{\beta}} = 0$ holds due to the definition of $\hat{\beta}(F)$ and Eq. (42) is consistent with Eq. (35).

In this section, we conjectured that the robustness of MLR is due to the particular structure of the model manifold. To support this conjecture, we attempted to derive the influence function with respect to the e - and the m -steps of the estimation procedure of model coefficient. Currently it is difficult to derive the influence functions, but we showed the difference between the influence function derived from the viewpoint of M-estimator and that derived from the IRLS estimator. The IRLS estimator is composed of a set of e - and m - steps. Our future work is to identify the contribution of each of the e - and the m - steps on the influence function.

7 Conclusions

In this paper, we provide an information geometric perspective on the MLR model, which is a nonparametric method. First, we discuss the MEM algorithm and investigate the relationship between the algorithm and information geometry. We cast the MEM algorithm as a maximum likelihood method by assuming a pseudo-observation. This gives us an empirical density function based on the assumption that one pseudo-observation is obtained. Second, we address the relationship between the MLR model and information geometry. Because the MLR model does not assume a parametric distribution, we cannot construct a corresponding model manifold with conventional approaches. We propose constructing the model manifold based on observations. The empirical density function introduced through a discussion of the MEM algorithm is applied to construct the data manifold. We clarify the relationship between the EM algorithm developed by Yao and Li [23] and the *em* algorithm corresponding to the MLR model.

Elucidating the factors or geometric operation that make the estimator for the MLR model robust remains for future work. We will further investigate the influence function corresponding to the *e*- and *m*- steps for estimating the coefficient of the MLR model.

Acknowledgements The authors are grateful to the anonymous reviewers whose comments led to valuable improvements in the manuscript.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

A $\mathcal{D}$ of MEM is in a mixture family

This section shows that the data manifold $\mathcal{D}$ of MEM is in a mixture family. An element of $\mathcal{D}$ is expressed as:

$$h(x, z; q_1 \dots q_K) = \sum_{i=1}^K q_i \delta(x) \delta_i(z),$$

$$\text{where } \begin{cases} \forall i = 1 \dots K, q_i \geq 0, \\ \sum_{i=1}^K q_i = 1. \end{cases}$$

Because $h(x, z; q_1 \dots q_K)$ is expressed in a convex combination of K probability density functions $\{\delta(x) \delta_i(z)\}_{i=1}^K$, the subset $\mathcal{D}$ is a $(K - 1)$ -dimensional mixture family.

B The ϵ -projection onto $\mathcal{D}$ for the MEM algorithm

This appendix shows that the optimization problem expressed by (19) derives an optimal solution for $\{q_i^{(k)}\}_{i=1}^K$. $D^{(e)}(l(\cdot, \cdot; m^{(k)})||h)$ can be expanded as follows:

$$\begin{aligned} D^{(e)}(l(\cdot, \cdot; m^{(k)})||h) &= D^{(m)}(h||l(\cdot, \cdot; m^{(k)})) \\ &= \sum_{i=1}^K \int h(x, i; q_1 \dots q_K) \log \frac{h(x, i; q_1 \dots q_K)}{l(x, i; m^{(k)})} dx \\ &= \sum_{i=1}^K \int q_i \delta(x) \log [q_i \delta(x)] dx - \sum_{i=1}^K \int q_i \delta(x) \log [\pi_i f_i(x + m^{(k)})] dx \\ &= \sum_{i=1}^K q_i \log \frac{q_i}{\pi_i f_i(m^{(k)})} + \int \delta(x) \log \delta(x) dx. \end{aligned} \quad (43)$$

The divergence $D^{(e)}(l(\cdot, \cdot; m^{(k)})||h)$ can be decomposed into two terms, as expressed in (43). Because the variables $\{q_i\}_{i=1}^K$ are only included in the first term, the optimization problem expressed by (19) is equivalent to

$$\min_{q_1 \dots q_K} \sum_{i=1}^K q_i \log \frac{q_i}{\pi_i f_i(m^{(k)})}, \quad \text{s.t.} \quad \begin{cases} \forall i = 1 \dots K, & q_i \geq 0, \\ \sum_{i=1}^K q_i = 1. \end{cases} \quad (44)$$

In order to obtain an optimal solution, we solve the following problem:

$$\min_{q_1 \dots q_K} L = \sum_{i=1}^K q_i \log \frac{q_i}{\pi_i f_i(m^{(k)})} + \lambda \left[1 - \sum_{i=1}^K q_i \right].$$

We can then obtain $\{q_i^{(k)}\}_{i=1}^K$ that satisfies $\frac{\partial L}{\partial q_i} \Big|_{q_i^{(k)}} = 0$:

$$\begin{aligned} \frac{\partial L}{\partial q_i} &= \log \frac{q_i}{\pi_i f_i(m^{(k)})} + 1 - \lambda, \quad \frac{\partial L}{\partial q_i} \Big|_{q_i^{(k)}} = \log \frac{q_i^{(k)}}{\pi_i f_i(m^{(k)})} + 1 - \lambda = 0, \\ \therefore q_i^{(k)} &= \pi_i f_i(m^{(k)}) e^{\lambda-1}. \end{aligned}$$

The value of λ can be derived from the constraint $\sum_{i=1}^K q_i^{(k)} = 1$:

$$\begin{aligned} e^{\lambda-1} \sum_{i=1}^K \pi_i f_i(m^{(k)}) &= 1 \\ \therefore \lambda &= 1 - \log [f(m^{(k)})]. \end{aligned}$$

Therefore, the optimal solution $\{q_i^{(k)}\}_{i=1}^K$ is expressed as

$$q_i^{(k)} = \pi_i f_i(m^{(k)}) e^{-\log f(m^{(k)})} = \frac{\pi_i f_i(m^{(k)})}{f(m^{(k)})}, \quad i = 1 \dots K.$$

This solution satisfies the constraint $\forall i = 1 \dots K, q_i \geq 0$ and $\sum_{i=1}^K q_i = 1$.

C The m -projection onto $\mathcal{M}$ for the MEM algorithm

This appendix shows that the optimization problem expressed by (20) is equivalent to the one expressed by (22).

We can expand $D^{(m)}(h(\cdot, \cdot; q_1^{(k)} \dots q_K^{(k)}) || l)$ to

$$\begin{aligned} D^{(m)}(h(\cdot, \cdot; q_1^{(k)} \dots q_K^{(k)}) || l) &= \sum_{i=1}^K \int h(x, i; q_1^{(k)} \dots q_K^{(k)}) \log \frac{h(x, i; q_1^{(k)} \dots q_K^{(k)})}{l(x, i; m)} dx \\ &= \sum_{i=1}^K \int q_i^{(k)} \delta(x) \log [q_i^{(k)} \delta(x)] dx - \sum_{i=1}^K \int q_i^{(k)} \delta(x) \log [\pi_i f_i(x + m)] dx \\ &= \sum_{i=1}^K q_i^{(k)} \log q_i^{(k)} + \int \delta(z) \log \delta(z) dz - \sum_{i=1}^K q_i^{(k)} \log \pi_i - \sum_{i=1}^K q_i^{(k)} \log f_i(m). \end{aligned} \quad (45)$$

The divergence $D^{(m)}(h(\cdot, \cdot; q_1^{(k)} \dots q_K^{(k)}) || l)$ can be decomposed into four terms, as expressed in (45). Because the variable m is only included in the fourth term, the optimization problem expressed by (20) is equivalent to

$$\max_{m \in \mathbb{R}^p} \sum_{i=1}^K q_i^{(k)} \log f_i(m).$$

D Proposed model manifold for the MLR model is in a curved exponential family

This appendix shows that the proposed model manifold expressed by (25) is in a curved exponential family. In this paper, we adopt $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$ as a kernel function.

We can expand $\log \phi_h(\epsilon - (y_i - x_i^\top \beta))$:

$$\log \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) = \frac{1}{h^2} \left(-\frac{\epsilon^2}{2} + \epsilon y_i \right) - \sum_{l=1}^p \frac{\beta_l}{h^2} x_{il} \epsilon - \frac{(y_i - x_i^\top \beta)^2}{2h^2} - \log h - \frac{1}{2} \log 2\pi. \quad (46)$$

Then, we show that $\log g(\epsilon, z; \beta)$ can be transformed to the standard form of an exponential family:

$$\begin{aligned} \log g(\epsilon, z; \beta) &= \sum_{i=1}^N \delta_i(z) \log \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) - \log N, \quad \text{from (46)} \\ &= \frac{1}{h^2} \left(-\frac{\epsilon^2}{2} + \sum_{i=1}^N \delta_i(z) y_i \epsilon \right) - \sum_{l=1}^p \frac{\beta_l}{h^2} \left(\sum_{i=1}^N \delta_i(z) x_{il} \epsilon \right) \\ &\quad - \sum_{i=1}^{N-1} \frac{\delta_i(z)}{2} \frac{(y_i - x_i^\top \beta)^2 - (y_N - x_N^\top \beta)^2}{h^2} \\ &\quad - \left(\frac{(y_N - x_N^\top \beta)^2}{2h^2} + \log h + \frac{1}{2} \log 2\pi + \log N \right). \end{aligned} \quad (47)$$

Replacing the following amount as natural parameters results in

$$\begin{cases} \theta^{00} = \frac{1}{h^2}, \\ \theta^{0l} = \frac{\beta_l}{h^2}, \quad l = 1 \dots p, \\ \theta^0 = (\theta^{01} \dots \theta^{0p})^\top, \\ \theta^{li} = \frac{(y_i - x_i^\top \beta)^2 - (y_N - x_N^\top \beta)^2}{h^2}, \quad i = 1 \dots N-1. \end{cases}$$

Now, $y_i - x_i^\top \beta$ is denoted by natural parameters as follows:

$$y_i - x_i^\top \beta = y_i - \sum_{l=1}^p x_{il} \beta_l = y_i - \frac{1}{\theta^{00}} \sum_{l=1}^p x_{il} \theta^{0l} = y_i - \frac{1}{\theta^{00}} x_i^\top \theta^0.$$

Functions $F_{00}(\epsilon, z)$, $F_{0l}(\epsilon, z)$, $F_{li}(\epsilon, z)$, $l = 1 \dots p$, $i = 1 \dots N-1$ denote

$$\begin{cases} F_{00}(\epsilon, z) = -\frac{\epsilon^2}{2} + \sum_{i=1}^N \delta_i(z) y_i \epsilon, \\ F_{0l}(\epsilon, z) = -\sum_{i=1}^N \delta_i(z) x_{il} \epsilon, \\ F_{li}(\epsilon, z) = -\frac{\delta_i(z)}{2}. \end{cases}$$

We replace the last term in (47) with $\psi(\theta)$:

$$\begin{aligned} \frac{(y_N - x_N^\top \beta)^2}{2h^2} + \log h + \frac{1}{2} \log 2\pi + \log N &= \frac{1}{2} \theta^{00} \left(y_N - \frac{1}{\theta^{00}} x_N^\top \theta^0 \right)^2 - \frac{1}{2} \log \theta^{00} \\ &\quad + \frac{1}{2} \log 2\pi + \log N \\ &= \psi(\theta). \end{aligned}$$

Thus, $\log g(\epsilon, z)$ can be expressed by natural parameters and $\psi(\theta)$ as follows:

$$\log g(\epsilon, z) = \sum_{l=0}^p \theta^{0l} F_{0l}(\epsilon, z) + \sum_{i=1}^{N-1} \theta^{1i} F_{1i}(\epsilon, z) - \psi(\theta).$$

This shows that $\mathcal{M}$ expressed by (25) is in an exponential family.

Here, we derive the expectation parameters:

$$\begin{aligned} \eta_{00} &= \mathbb{E}[F_{00}(\epsilon, z)] = \sum_{i=1}^N \int g(\epsilon, i) F_{00}(\epsilon, i) d\epsilon, \quad \because g(\epsilon, i) = \frac{1}{N} \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) \\ &= \frac{1}{N} \sum_{i=1}^N \int \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) \left[-\frac{\epsilon^2}{2} + y_i \epsilon \right] d\epsilon \\ &= -\frac{1}{2N} \sum_{i=1}^N \left(h^2 + (y_i - x_i^\top \beta)^2 \right) + \frac{1}{N} \sum_{i=1}^N y_i (y_i - x_i^\top \beta) \\ \therefore \eta_{00} &= -\frac{h^2}{2} + \frac{1}{2N} \left(y^\top y - (X\beta)^\top (X\beta) \right). \\ \eta_{0l} &= \mathbb{E}[F_{0l}(\epsilon, z)] = \sum_{i=1}^N \int g(\epsilon, i) F_{0l}(\epsilon, i) d\epsilon \\ &= \frac{1}{N} \sum_{i=1}^N \int \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) (-x_{il} \epsilon) d\epsilon \\ \therefore \eta_{0l} &= -\frac{1}{N} v_l^\top (y - X\beta), \quad \text{where } v_l = (x_{1l} \cdots x_{Nl})^\top, \quad l = 1 \dots p. \\ \eta_{1i} &= \mathbb{E}[F_{1i}(\epsilon, z)] = \sum_{j=1}^N \int g(\epsilon, j) F_{1i}(\epsilon, j) d\epsilon = -\frac{1}{2N} \int \phi_h \left(\epsilon - (y_i - x_i^\top \beta) \right) d\epsilon \\ \therefore \eta_{1i} &= -\frac{1}{2N}, \quad i = 1 \dots N. \end{aligned}$$

In order to obtain the dual potential $\varphi(\eta)$, we express $\log h$ as a function of expectation parameters:

$$\begin{aligned}
\eta_0 &= (\eta_{01} \dots \eta_{0p})^\top = -\frac{1}{N} X^\top (y - X\beta) \\
\therefore \beta &= (X^\top X)^{-1} (X^\top y + N\eta_0). \\
\eta_{00} &= -\frac{h^2}{2} + \frac{1}{2N} \left(y^\top y - (X^\top y + N\eta_0)^\top (X^\top X)^{-1} (X^\top y + N\eta_0) \right) \\
h^2 &= -2\eta_{00} + \frac{1}{N} \left(y^\top y - (X^\top y + N\eta_0)^\top (X^\top X)^{-1} (X^\top y + N\eta_0) \right) \\
\therefore \log h &= \frac{1}{2} \log \left[-2\eta_{00} + \frac{1}{N} \left(y^\top y - (X^\top y + N\eta_0)^\top (X^\top X)^{-1} (X^\top y + N\eta_0) \right) \right]. \tag{48}
\end{aligned}$$

Finally, the dual potential $\varphi(\eta)$ is expressed as follows:

$$\begin{aligned}
\varphi(\eta) &= \sum_{l=0}^p \theta^{0l} \eta_{0l} + \sum_{i=1}^{N-1} \theta^{1i} \eta_{1i} - \psi(\theta) \\
&= -\frac{1}{2} - \frac{1}{2N} \sum_{i=1}^N \frac{(y_i - x_i^\top \beta)^2}{h^2} + \frac{1}{N} \sum_{i=1}^N \frac{(y_i - x_i^\top \beta)^2}{h^2} \\
&\quad - \frac{1}{2N} \sum_{i=1}^{N-1} \frac{(y_i - x_i^\top \beta)^2 - (y_N - x_N^\top \beta)^2}{h^2} \\
&\quad - \left(\frac{(y_N - x_N^\top \beta)^2}{2h^2} + \log h + \frac{1}{2} \log 2\pi + \log N \right), \\
&\therefore \sum_{i=1}^{N-1} \frac{(y_i - x_i^\top \beta)^2 - (y_N - x_N^\top \beta)^2}{h^2} = \sum_{i=1}^N \frac{(y_i - x_i^\top \beta)^2 - (y_N - x_N^\top \beta)^2}{h^2}, \\
&= - \left(\log h + \frac{1}{2} + \frac{1}{2} \log 2\pi + \log N \right) \\
\therefore \varphi(\eta) &= -\frac{1}{2} \log \left[-2\eta_{00} + \frac{1}{N} \left(y^\top y - (X^\top y + N\eta_0)^\top (X^\top X)^{-1} (X^\top y + N\eta_0) \right) \right] \\
&\quad - \left(\frac{1}{2} + \frac{1}{2} \log 2\pi + \log N \right).
\end{aligned}$$

The derived natural parameters, expectation parameters, potential function, and dual potential are shown to satisfy the relationship in (12):

$$\begin{aligned}
\frac{\partial \psi}{\partial \theta^{00}} &= \frac{1}{2N} \left(y^\top y - \left(\frac{1}{\theta^{00}} X \theta^0 \right)^\top \left(\frac{1}{\theta^{00}} X \theta^0 \right) \right) - \frac{1}{2\theta^{00}} \\
\therefore \frac{\partial \psi}{\partial \theta^{00}} &= -\frac{h^2}{2} + \frac{1}{2N} \left(y^\top y - (X\beta)^\top (X\beta) \right) = \eta_{00}.
\end{aligned}$$

$$\begin{aligned}
 \frac{\partial \psi}{\partial \theta^0} &= -\frac{1}{N} X^\top y + \frac{1}{N \theta^{00}} X^\top X \theta^0 = -\frac{1}{N} X^\top (y - X \beta) \\
 \therefore \frac{\partial \psi}{\partial \theta^0} &= \eta_0. \\
 \therefore \frac{\partial \psi}{\partial \theta^{1i}} &= -\frac{1}{2N}, \quad i = 1 \dots N. \\
 \frac{\partial \varphi}{\partial \eta_{00}} &= \frac{1}{-2} \frac{-2}{-2\eta_{00} + \frac{1}{N} (y^\top y - (X^\top y + N\eta_0)^\top (X^\top X)^{-1} (X^\top y + N\eta_0))} \\
 \therefore \frac{\partial \varphi}{\partial \eta_{00}} &= \frac{1}{h^2} = \theta^{00}. \\
 \therefore \frac{\partial \varphi}{\partial \eta_0} &= \frac{1}{h^2} (X^\top X)^{-1} (X^\top y + N\eta_0) = \frac{\beta}{h^2}.
 \end{aligned}$$

E Proposed data manifold for the MLR model is in a mixture family

This appendix shows that the data manifold $\mathcal{D}$ for the MLR model introduced in Sect. 5 is in a mixture family. An element of $\mathcal{D}$ is expressed as

$$h(\epsilon, z; q_1 \dots q_N) = \sum_{i=1}^N q_i \delta(\epsilon) \delta_i(z), \quad \text{where} \quad \begin{cases} \forall i = 1 \dots N, & q_i \geq 0, \\ \sum_{i=1}^N q_i = 1. \end{cases}$$

Because $h(\epsilon, z; q_1 \dots q_N)$ is expressed in a convex combination of N probability density functions $\{\delta(\epsilon) \delta_i(z)\}_{i=1}^N$, the subset $\mathcal{D}$ is an $(N - 1)$ -dimensional mixture manifold.

F The e-projection onto $\mathcal{D}$ for the MLR model

This appendix shows that the optimization problem expressed by (29) derives an optimal solution for $\{q_i^{(k)}\}_{i=1}^N$:

$$\min_{q_1 \dots q_N} D^{(e)} \left(g(\cdot, \cdot; \beta^{(k)}) || h(\cdot, \cdot; q_1 \dots q_N) \right), \quad \text{s.t.} \quad \begin{cases} \forall i = 1 \dots N, & q_i \geq 0, \\ \sum_{i=1}^N q_i = 1. \end{cases} \quad (29)$$

We can expand $D^{(e)} (g(\cdot, \cdot; \beta^{(k)}) || h(\cdot, \cdot; q_1 \dots q_N))$ to

$$\begin{aligned}
D^{(e)} \left(g(\cdot, \cdot; \beta^{(k)}) || h(\cdot, \cdot; q_1 \dots q_N) \right) &= D^{(m)} \left(h(\cdot, \cdot; q_1 \dots q_N) || g(\cdot, \cdot; \beta^{(k)}) \right) \\
&= \sum_{i=1}^N \int h(\epsilon, i) \log h(\epsilon, i) d\epsilon - \sum_{i=1}^N \int h(\epsilon, i) \log g(\epsilon, i; \beta^{(k)}) d\epsilon \\
&= \sum_{i=1}^N \int q_i \delta(\epsilon) \log [q_i \delta(\epsilon)] d\epsilon - \sum_{i=1}^N \int q_i \delta(\epsilon) \log g(\epsilon, i; \beta^{(k)}) d\epsilon \\
&= \sum_{i=1}^N q_i \log \frac{q_i}{g(0, i; \beta^{(k)})} + \int \delta(\epsilon) \log \delta(\epsilon) d\epsilon.
\end{aligned}$$

The divergence $D^{(e)}(g||h)$ can be decomposed into two terms. Because $\{q_i\}_{i=1}^N$ is only included in the first term, the optimization problem expressed by (29) is equivalent to

$$\min_{q_1 \dots q_N} \sum_{i=1}^N q_i \log \frac{q_i}{g(0, i; \beta^{(k)})}, \quad \text{s.t.} \quad \begin{cases} \forall i = 1 \dots N, & q_i \geq 0, \\ \sum_{i=1}^N q_i = 1. \end{cases} \quad (49)$$

In order to obtain an optimal solution, we solve the following problem:

$$\min_{q_1 \dots q_N} L = \sum_{i=1}^N q_i \log \frac{q_i}{g(0, i; \beta^{(k)})} + \lambda \left[1 - \sum_{i=1}^N q_i \right].$$

We can obtain $\{q_i^{(k)}\}_{i=1}^N$ that satisfies $\frac{\partial L}{\partial q_i} \Big|_{q_i^{(k)}} = 0$:

$$\begin{aligned}
\frac{\partial L}{\partial q_i} &= \log \frac{q_i}{g(0, i; \beta^{(k)})} + q_i \frac{\frac{1}{g(0, i; \beta^{(k)})}}{\frac{q_i}{g(0, i; \beta^{(k)})}} - \lambda = \log \frac{q_i}{g(0, i; \beta^{(k)})} + 1 - \lambda, \\
\frac{\partial L}{\partial q_i} \Big|_{q_i^{(k)}} &= \log \frac{q_i}{g(0, i; \beta^{(k)})} + 1 - \lambda = 0, \\
\therefore q_i^{(k)} &= g(0, i; \beta^{(k)}) e^{\lambda-1}.
\end{aligned}$$

The value of λ can be derived from the constraint $\sum_{i=1}^N q_i^{(k)} = 1$:

$$\begin{aligned}
e^{\lambda-1} \sum_{i=1}^N g(0, i; \beta^{(k)}) &= 1 \\
\therefore \lambda &= 1 - \log f(0; \beta^{(k)}).
\end{aligned}$$

Therefore, the optimal solution $\{q_i^{(k)}\}_{i=1}^N$ is expressed as

$$q_i^{(k)} = g(0, i; \beta^{(k)}) e^{-\log f(0; \beta^{(k)})} = \frac{g(0, i; \beta^{(k)})}{f(0; \beta^{(k)})} = \frac{\phi_h(y_i - x_i^\top \beta^{(k)})}{\sum_{j=1}^N \phi_h(y_j - x_j^\top \beta^{(k)})}.$$

This solution satisfies the condition $\forall i = 1 \dots N$, $q_i \geq 0$ and $\sum_{i=1}^N q_i = 1$.

G m -projection onto $\mathcal{M}$ for the MLR model

We show that the optimization problem expressed by (31) is equivalent to the one expressed by (32). We can expand $D^{(m)}(h||g)$ to

$$\begin{aligned} D^{(m)}(h||g) &= \sum_{i=1}^N \int h(\epsilon, i; q_1 = q_1^{(k)} \dots q_N = q_N^{(k)}) \log h(\epsilon, i; q_1 = q_1^{(k)} \dots q_N = q_N^{(k)}) d\epsilon \\ &\quad - \sum_{i=1}^N \int h(\epsilon, i; q_1 = q_1^{(k)} \dots q_N = q_N^{(k)}) \log g(\epsilon, i; \beta) d\epsilon \\ &= \sum_{i=1}^N \int q_i^{(k)} \delta(\epsilon) \log [q_i^{(k)} \delta(\epsilon)] d\epsilon - \sum_{i=1}^N \int q_i^{(k)} \delta(\epsilon) \log g(\epsilon, i; \beta) d\epsilon \\ &= \sum_{i=1}^N q_i^{(k)} \log q_i^{(k)} + \int \delta(\epsilon) \log \delta(\epsilon) d\epsilon - \sum_{i=1}^N q_i^{(k)} \log g(0, i; \beta) \\ &= \sum_{i=1}^N q_i^{(k)} \log q_i^{(k)} + \int \delta(\epsilon) \log \delta(\epsilon) d\epsilon + \log N - \sum_{i=1}^N q_i^{(k)} \log \phi_h(y_i - x_i^\top \beta). \end{aligned} \quad (50)$$

The divergence $D^{(m)}(h||g)$ can be decomposed into four terms. Because only the fourth term in (50) includes β , the optimization problem expressed by (31) is equivalent to

$$\max_{\beta} \sum_{i=1}^N q_i^{(k)} \log \phi_h(y_i - x_i^\top \beta).$$

H Influence function of β

We formally derive the influence function of β . We first define the ψ -type M-estimator:

Definition 1 (ψ -type M-estimator) The ψ -type M-estimator $T(F)$ is defined as a solution to the equation with respect to θ :

$$\int_{\mathcal{X}} \psi(x, \theta) dF(x) = 0,$$

where $\psi : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^r$ is a measurable function and F is a probability measure on $\mathcal{X}$.

Proposition 1 (Influence function of MLR coefficient estimation) *To ease the derivation of the influence function, the estimator of the MLR coefficient $\hat{\beta} = \operatorname{argmax}_{\beta \in B} \frac{1}{n} \sum_{i=1}^n \phi_h(y_{(i)} - x_{(i)}^\top \beta)$, $y_{(i)} \in \mathbb{R}$, $x_{(i)} \in \mathbb{R}^p$ is expressed as*

$$\hat{\beta} = \operatorname{argmax}_{\beta \in B} \int \phi_h(y - x^\top \beta) dF_{X,Y}(x, y), \quad (51)$$

where $F_{X,Y}$ is a probability measure for the random variables X, Y . With the Dirac measure $\Delta_{u,v}$, the empirical measure $F_{X,Y} = \frac{1}{n} \sum_{i=1}^n \Delta_{x_{(i)}, y_{(i)}}$ leads to the original form of (3).

To formulate the estimator of (3) as a ψ -type M -estimator, we define the estimator of the MLR coefficient as a solution to

$$\frac{\partial}{\partial \beta} \int \phi_h(y - x^\top \beta) dF_{X,Y}(x, y) = 0. \quad (52)$$

The solution is denoted by $\hat{\beta}(F_{X,Y})$ to make the dependence on the measure $F_{X,Y}$ explicit.

With the measure F , the influence function $IF(u, v; F)$ of the estimator $\hat{\beta}$ with a given $u \in \mathbb{R}^p$, $v \in \mathbb{R}$ is denoted as

$$IF(u, v; F) = \left(\int \frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \hat{\beta}(F)} x x^\top dF(x, y) \right)^{-1} \psi(u, v, \hat{\beta}(F)),$$

$$\text{where } \psi(u, v, \hat{\beta}(F)) = \frac{d}{dz} \phi_h(z)_{z=v-u^\top \hat{\beta}(F)} u.$$

To prove the proposition, we note that

$$\frac{\partial}{\partial \beta} \phi_h(y - x^\top \beta) = \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta} \frac{\partial}{\partial \beta} (y - x^\top \beta) = -\frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta} x.$$

At $\hat{\beta}(F_{X,Y})$, (52) leads to

$$\int \frac{d}{dz} \phi_h(z)_{z=y-x^\top \hat{\beta}(F_{X,Y})} x dF_{X,Y}(x, y) = 0. \quad (53)$$

We write $\psi(x, y, \hat{\beta}(F_{X,Y})) = \frac{d}{dz} \phi_h(z)_{z=y-x^\top \hat{\beta}(F_{X,Y})} x$ henceforth.

The influence function is obtained by letting $F_{X,Y} = (1-t)F + t\Delta_{u,v}$ and evaluating the derivative at $t = 0$. To do so, we expand $\frac{\partial}{\partial \beta} \psi(x, y, \beta)$ as

$$\begin{aligned} \frac{\partial}{\partial \beta} \psi(x, y, \beta) &= \begin{pmatrix} \left(\frac{\partial}{\partial \beta_1} \psi(x, y, \beta) \right)^\top \\ \vdots \\ \left(\frac{\partial}{\partial \beta_p} \psi(x, y, \beta) \right)^\top \end{pmatrix}, \\ \frac{\partial}{\partial \beta_k} \psi(x, y, \beta) &= \frac{\partial}{\partial \beta_k} \begin{pmatrix} \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta x_1} \\ \vdots \\ \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta x_p} \end{pmatrix}, \quad \text{where } x = (x_1 \dots x_p)^\top, \\ &= \begin{pmatrix} x_1 \frac{\partial}{\partial \beta_k} \left[\frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta} \right] \\ \vdots \\ x_p \frac{\partial}{\partial \beta_k} \left[\frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta} \right] \end{pmatrix} \\ &= \begin{pmatrix} x_1 \frac{d}{dz} \left[\frac{d}{dz} \phi_h(z) \right]_{z=y-x^\top \beta} \frac{\partial}{\partial \beta_k} [y - x^\top \beta] \\ \vdots \\ x_p \frac{d}{dz} \left[\frac{d}{dz} \phi_h(z) \right]_{z=y-x^\top \beta} \frac{\partial}{\partial \beta_k} [y - x^\top \beta] \end{pmatrix}, \\ &= -\frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \beta} x_k x, \\ \therefore \frac{\partial}{\partial \beta} \psi(x, y, \beta) &= -\frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \beta} \begin{pmatrix} x_1 x^\top \\ \vdots \\ x_p x^\top \end{pmatrix} = -\frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \beta} x x^\top. \end{aligned} \quad (54)$$

Then, $\frac{d}{dt} \psi(x, y, \hat{\beta}(F_{X,Y}))$ with $F_{X,Y} = (1-t)F + t\Delta_{u,v}$ is given by

$$\frac{d}{dt} \psi(x, y, \hat{\beta}(F_{X,Y}))_{t=0} = \left[\frac{\partial}{\partial \beta} \psi(x, y, \beta)_{\beta=\hat{\beta}(F)} \right]^\top \frac{d}{dt} \hat{\beta}(F_{X,Y})_{t=0} \quad (55)$$

for $t = 0$. Considering that $\frac{d}{dt} \hat{\beta}(F_{X,Y})_{t=0}$ in (55) is the influence function, we substitute $F_{X,Y} = (1-t)F + t\Delta_{u,v}$ into (53) and obtain

$$\begin{aligned} \frac{d}{dt} \left[\int \psi(x, y, \hat{\beta}(F_{X,Y})) dF_{X,Y}(x, y) \right]_{t=0} &= 0, \\ \frac{d}{dt} \left[(1-t) \int \psi(x, y, \hat{\beta}(F_{X,Y})) dF(x, y) + t \int \psi(x, y, \hat{\beta}(F_{X,Y})) d\Delta_{u,v}(x, y) \right]_{t=0} &= 0, \\ - \left[\int \psi(x, y, \hat{\beta}(F_{X,Y})) dF(x, y) \right]_{t=0} &= 0. \end{aligned}$$

$$\begin{aligned}
& + \left[(1-t) \frac{d}{dt} \int \psi(x, y, \hat{\beta}(F_{X,Y})) dF(x, y) \right]_{t=0} \\
& + \left[\int \psi(x, y, \hat{\beta}(F_{X,Y})) d\Delta_{u,v}(x, y) \right]_{t=0} \\
& + \left[t \frac{d}{dt} \int \psi(x, y, \hat{\beta}(F_{X,Y})) d\Delta_{u,v}(x, y) \right]_{t=0} = 0,
\end{aligned}$$

From Eq. (55)

$$\begin{aligned}
& - \int \psi(x, y, \hat{\beta}(F)) dF(x, y) \\
& + \left(\int \frac{\partial}{\partial \beta} \psi(x, y, \beta)_{\beta=\hat{\beta}(F)} dF(x, y) \right)^\top \left(\frac{d}{dt} \hat{\beta}(F_{X,Y})_{t=0} \right) \\
& + \psi(u, v, \hat{\beta}(F)) = 0
\end{aligned} \tag{56}$$

The first term in Eq. (56) is zero due to Eq. (53). Then,

$$\frac{d}{dt} \hat{\beta}(F_{X,Y})_{t=0} = - \left(\left[\int \frac{\partial}{\partial \beta} \psi(x, y, \beta)_{\beta=\hat{\beta}(F)} dF(x, y) \right]^\top \right)^{-1} \psi(u, v, \hat{\beta}(F)).$$

From (54), we obtain

$$\therefore \frac{d}{dt} \hat{\beta}(F_{X,Y})_{t=0} = \left(\int \frac{d^2}{dz^2} \phi_h(z)_{z=y-x^\top \hat{\beta}(F)} x x^\top dF(x, y) \right)^{-1} \psi(u, v, \hat{\beta}(F)). \tag{57}$$

Proposition 2 (Influence function of MLR coefficient estimator defined by IRLS) *Let us consider the estimator $\hat{\beta}$ that is defined by EM algorithm in 2.2:*

$$\begin{aligned}
\beta^{(k+1)} &= (X^\top W_k X)^{-1} X^\top W_k y, \\
\text{where } \begin{cases} W_k = \text{diag}(\pi_1^{(k)} \dots \pi_N^{(k)}), \\ \pi_i^{(k)} = \frac{\phi_h(y_i - x_i^\top \beta^{(k)})}{\sum_{j=1}^N \phi_h(y_j - x_j^\top \beta^{(k)})}, \quad i = 1 \dots N. \end{cases}
\end{aligned} \tag{58}$$

The above estimator can be expressed as

$$\begin{aligned}
& \left[\int \phi_h(y - x^\top \beta^{(k)}(F_{X,Y})) x x^\top dF_{X,Y}(x, y) \right] \beta^{(k+1)}(F_{X,Y}) \\
& = \int \phi_h(y - x^\top \beta^{(k)}(F_{X,Y})) y x dF_{X,Y}(x, y),
\end{aligned} \tag{59}$$

where $F_{X,Y}$ is a probability measure for the random variables X, Y . With the Dirac measure $\Delta_{u,v}$, the empirical measure $F_{X,Y} = \frac{1}{n} \sum_{i=1}^n \Delta_{x(i), y(i)}$ leads to the original form of (58).

Then, the relation between the influence function of $\beta^{(k)}$ and that of $\beta^{(k+1)}$ is denoted as

$$IF(u, v; \beta^{(k+1)}, F) = z_k + A_k IF(u, v; \beta^{(k)}, F),$$

$$\text{where } \begin{cases} z_k = \Sigma_k^{-1} c_k, \\ \Sigma_k = \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) x x^\top dF(x, y), \\ c_k = \phi_h \left(v - u^\top \beta^{(k)}(F) \right) \left(v - u^\top \beta^{(k+1)}(F) \right) u \\ \quad - \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) \left(y - x^\top \beta^{(k+1)}(F) \right) x dF(x, y), \\ C_k = - \int \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta^{(k)}(F)} \left(y - x^\top \beta^{(k+1)}(F) \right) x x^\top dF(x, y), \\ A_k = \Sigma_k^{-1} C_k. \end{cases}$$

The influence function is obtained by letting $F_{X,Y} = (1-t)F + t\Delta_{u,v}$ and evaluating the derivative at $t = 0$. We calculate necessary terms in advance.

$$\begin{aligned} & \frac{d}{dt} \left[\phi_h \left(y - x^\top \beta^{(k)}(F_{X,Y}) \right) \right]_{t=0} \\ &= - \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta^{(k)}(F)} x^\top \frac{d}{dt} \beta^{(k)}(F_{X,Y})_{t=0}, \end{aligned} \quad (60)$$

$$\begin{aligned} & \frac{d}{dt} \left[\left\{ \int \phi_h \left(y - x^\top \beta^{(k)}(F_{X,Y}) \right) x x^\top dF_{X,Y}(x, y) \right\} \beta^{(k+1)}(F_{X,Y}) \right]_{t=0} \\ &= - \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) x^\top \beta^{(k+1)}(F) x dF(x, y) \\ & \quad + \int \left[\frac{d}{dt} \phi_h \left(y - x^\top \beta^{(k)}(F_{X,Y}) \right) \right]_{t=0} x^\top \beta^{(k+1)}(F) x dF(x, y) \\ & \quad + \phi_h \left(v - u^\top \beta^{(k)}(F) \right) u^\top \beta^{(k+1)}(F) u + \Sigma_k \frac{d}{dt} \beta^{(k+1)}(F_{X,Y})_{t=0}, \end{aligned} \quad (61)$$

$$\begin{aligned} & \frac{d}{dt} \left[\int \phi_h \left(y - x^\top \beta^{(k)}(F_{X,Y}) \right) y x dF_{X,Y}(x, y) \right]_{t=0} \\ &= - \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) y x dF(x, y) \\ & \quad + \int \frac{d}{dt} \left[\phi_h \left(y - x^\top \beta^{(k)}(F_{X,Y}) \right) \right]_{t=0} y x dF(x, y) \\ & \quad + \phi_h \left(v - u^\top \beta^{(k)}(F) \right) v u. \end{aligned} \quad (62)$$

Substituting (60), (61), (62) into the derivative of (59) at $t = 0$ leads to

$$\begin{aligned}
& \Sigma_k \frac{d}{dt} \beta^{(k+1)}(F_{X,Y})_{t=0} \\
&= \phi_h \left(v - u^\top \beta^{(k)}(F) \right) \left(v - u^\top \beta^{(k+1)}(F) \right) u \\
&\quad - \int \phi_h \left(y - x^\top \beta^{(k)}(F) \right) \left(y - x^\top \beta^{(k+1)}(F) \right) x dF(x, y) \\
&\quad - \int \frac{d}{dz} \phi_h(z)_{z=y-x^\top \beta^{(k)}(F)} \left(y - x^\top \beta^{(k+1)}(F) \right) x x^\top dF(x, y) \frac{d}{dt} \beta^{(k)}(F_{X,Y})_{t=0} \\
&\therefore \frac{d}{dt} \beta^{(k+1)}(F_{X,Y})_{t=0} = z_k + A_k \frac{d}{dt} \beta^{(k)}(F_{X,Y})_{t=0}.
\end{aligned} \tag{63}$$

References

- Amari, S.: Information geometry of the EM and em algorithms for neural networks. *Neural Netw.* **8**(9), 1379–1408 (1995)
- Amari, S., Nagaoka, H.: *Methods of Information Geometry*. American Mathematical Society, New York (2000)
- Baldauf, M., Silva, J.S.: On the use of robust regression in econometrics. *Econ. Lett.* **114**(1), 124–127 (2012)
- Chen, Y.C., Genovese, C.R., Tibshirani, R.J., Wasserman, L.: Nonparametric modal regression. *Ann. Stat.* **44**(2), 489–514 (2016). <https://doi.org/10.1214/15-AOS1373>
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**(1), 1–22 (1977)
- Dollinger, M.B., Staudte, R.G.: Influence functions of iteratively reweighted least squares estimators. *J. Am. Stat. Assoc.* **86**(415), 709–716 (1991)
- Epanechnikov, V.A.: Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14**(1), 153–158 (1969)
- Fujisawa, H., Eguchi, S.: Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **99**(9), 2053–2081 (2008)
- Gannoun, A., Saracco, J., Yu, K.: On semiparametric mode regression estimation. *Commun. Stat. Theory Methods* **39**(7), 1141–1157 (2010)
- Grasselli, M.R.: Dual connections in nonparametric classical information geometry. *Ann. Inst. Stat. Math.* **62**(5), 873–896 (2010). <https://doi.org/10.1007/s10463-008-0191-3>
- Hampel, F.R.: The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**(346), 383–393 (1974)
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: *Robust Statistics—the Approach Based on Influence Functions*. Wiley, Amsterdam (1986)
- Huber, P.J., Ronchetti, E.M.: *Robust Statistics*. Wiley, Amsterdam (2011)
- Kemp, G.C., Silva, J.S.: Regression towards the mode. *J. Econ.* **170**(1), 92–101 (2012)
- Lee, M.J.: Mode regression. *J. Econ.* **42**(3), 337–349 (1989)
- Lee, M.J.: Quadratic mode regression. *J. Econ.* **57**(1), 1–19 (1993)
- Li, J., Ray, S., Lindsay, B.G.: A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8**, 1687–1723 (2007)
- Matzner-Lofber, E., Gannoun, A., Gooijer, J.G.D.: Nonparametric forecasting: a comparison of three kernel-based methods. *Commun. Stat. Theory Methods* **27**(7), 1593–1617 (1998). <https://doi.org/10.1080/03610929808832180>
- Parzen, E.: On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**(3), 1065–1076 (1962)
- Pistone, G., Sempì, C.: An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **23**(5), 1543–1561 (1995)
- Takano, K., Hino, H., Akaho, S., Murata, N.: Nonparametric e-mixture estimation. *Neural Comput.* **28**(12), 2687–2725 (2016)

22. Wand, M.P., Jones, M.C.: Kernel Smoothing. Taylor & Francis, Milton Park (1994)
23. Yao, W., Li, L.: A new regression model: modal linear regression. *Scand. J. Stat.* **41**(3), 656–671 (2014)
24. Yao, W., Lindsay, B.G., Li, R.: Local modal regression. *J. Nonparametr. Stat.* **24**(3), 647–663 (2012)
25. Zhang, J.: Nonparametric information geometry: from divergence function to referential–representational biduality on statistical manifolds. *Entropy* **15**(12), 5384–5418 (2013). <https://doi.org/10.3390/e15125384>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.