entropy

MDPI

Article

An Information Criterion for Auxiliary Variable Selection in Incomplete Data Analysis

Shinpei Imori [1,3,*] and Hidetoshi Shimodaira [2,3]

1 Graduate School of Science, Hiroshima University, Hiroshima 739-8526, Japan
2 Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan; shimo@i.kyoto-u.ac.jp
3 RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan
* Correspondence: imori@hiroshima-u.ac.jp

Received: 21 February 2019; Accepted: 12 March 2019; Published: 14 March 2019

Abstract: Statistical inference is considered for variables of interest, called primary variables, when auxiliary variables are observed along with the primary variables. We consider the setting of incomplete data analysis, where some primary variables are not observed. Utilizing a parametric model of joint distribution of primary and auxiliary variables, it is possible to improve the estimation of parametric model for the primary variables when the auxiliary variables are closely related to the primary variables. However, the estimation accuracy reduces when the auxiliary variables are irrelevant to the primary variables. For selecting useful auxiliary variables, we formulate the problem as model selection, and propose an information criterion for predicting primary variables by leveraging auxiliary variables. The proposed information criterion is an asymptotically unbiased estimator of the Kullback–Leibler divergence for complete data of primary variables under some reasonable conditions. We also clarify an asymptotic equivalence between the proposed information criterion and a variant of leave-one-out cross validation. Performance of our method is demonstrated via a simulation study and a real data example.

Keywords: Akaike information criterion; auxiliary variables; Fisher information matrix; incomplete data; Kullback–Leibler divergence; misspecification; Takeuchi information criterion

1. Introduction

Auxiliary variables are often observed along with primary variables. Here, the primary variables are random variables of interest, and our purpose is to estimate their predictive distribution, i.e., a probability distribution of the primary variables in future test data, while the auxiliary variables are random variables that are observed in training data but not included in the primary variables. We assume that the auxiliary variables are not observed in the test data, or we do not use them even if they are observed in the test data. When the auxiliary variables have a close relation with the primary variables, we expect to improve the accuracy of predictive distribution of the primary variables by considering a joint modeling of the primary and auxiliary variables.

The notion of auxiliary variables has been considered in statistics and machine learning literature. For example, the "curds and whey" method [1] and the "coaching variables" method [2] are based on a similar idea for improving prediction accuracy of primary variables by using auxiliary variables. In multitask learning, Caruana [3] improved generalization accuracy of a main task by exploiting extra tasks. Auxiliary variables are also considered in incomplete data analysis, i.e., a part of primary variables are not observed; Mercatanti et al. [4] showed some theoretical results to make parameter estimation better by utilizing auxiliary variables in Gaussian mixture model (GMM).

Although auxiliary variables are expected to be useful for modeling primary variables, they can actually be harmful. As mentioned in Mercatanti et al. [4], using auxiliary variables may affect modeling

results adversely because the number of parameters to be estimated increases and a candidate model of the auxiliary variables can be misspecified. Hence, it is important to select useful auxiliary variables. This is formulated as model selection by considering parametric models with auxiliary variables. In this paper, usefulness of auxiliary variables for estimating predictive distribution of primary variables is measured by a risk function based on the Kullback–Leibler (KL) divergence [5] that is often used for model selection. Because the KL risk function includes unknown parameters, we have to estimate it in actual use. Akaike Information Criterion (AIC) proposed by Akaike [6] is one of the most famous criteria, which is known as an asymptotically unbiased estimator of the KL risk function. AIC is a good criterion from the perspective of prediction due to the asymptotic efficiency; see Shibata [7,8]. Takeuchi [9] proposed a modified version of AIC, called Takeuchi Information Criterion (TIC), which relaxes an assumption for deriving AIC, that is, correct specification of candidate model. However, AIC and TIC are derived for primary variables without considering auxiliary variables in the setting of complete data analysis, and therefore, they are not suitable for auxiliary variable selection nor incomplete data analysis.

Incomplete data analysis is widely used in a broad range of statistical problems by regarding a part of primary variables as latent variables that are not observed. This setting also includes complete data analysis as a special case, where all the primary variables are observed. Information criteria for incomplete data analysis have been proposed in previous studies. Shimodaira [10] developed an information criterion based on the KL divergence for complete data when the data are only partially observed. Cavanaugh and Shumway [11] modified the first term of the information criterion of Shimodaira [10] by the objective function of the EM algorithm [12]. Recently, Shimodaira and Maeda [13] proposed an information criterion, which is derived by mitigating a condition assumed in Shimodaira [10] and Cavanaugh and Shumway [11].

However, any of these previously proposed criteria are not derived by taking auxiliary variables into account. Thus, we propose a new information criterion by considering not only primary variable but also auxiliary variables in the setting of incomplete data analysis. The proposed criterion is a generalization of AIC, TIC, and the criterion of Shimodaira and Maeda [13]. To the best of our knowledge, this is the first attempt to derive an information criterion by considering auxiliary variables. Moreover, we show an asymptotic equivalence between the proposed criterion and a variant of leave-one-out cross validation (LOOCV); this result is a generalization of the relationship between TIC and LOOCV [14].

Note that "auxiliary variables" may also be used in other contexts in literature. For example, Ibrahim et al. [15] considered to use auxiliary variables in missing data analysis, which is similar to our usage in the sense that auxiliary variables are highly correlated with missing data. However, they use the auxiliary variables in order to avoid specifying a missing data mechanism; this goal is different from ours, because no missing data mechanism is considered in our study.

The reminder of this paper is organized as follows. Notations as well as the setting of this paper are introduced in Section 2. Illustrative examples of useful and useless auxiliary variables are given in Section 3. The information criterion for selecting useful auxiliary variables in incomplete data analysis is derived in Section 4, and the asymptotic equivalence between the proposed criterion and a variant of LOOCV is shown in Section 5. Performance of our method is examined via a simulation study and a real data analysis in Sections 6 and 7, respectively. Finally, we conclude this paper in Section 8. All proofs are shown in Appendix A.

2. Preliminaries

2.1. Incomplete Data Analysis for Primary Variables

First we explain a setting of incomplete data analysis for primary variables in accordance with Shimodaira and Maeda [13]. Let X denote a vector of primary variables, which consists of two parts as $X = (Y, Z)$, where Y denotes the observed part and Z denotes the unobserved latent part. This setting

reduces to complete data analysis of $X = Y$ when Z is empty. We write the true density function of X as $q_x(x) = q_x(y,z)$ and a candidate parametric model of the true density as $p_x(x;\theta) = p_x(y,z;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is an unknown parameter vector and Θ is its parameter space. We assume that $x = (y,z) \in \mathcal{Y} \times \mathcal{Z}$ for all density functions, where $\mathcal{Y}$ and $\mathcal{Z}$ are domains of Y and Z, respectively. Thus the marginal densities of the observed part Y are obtained by $q_y(y) = \int q_x(y,z)dz$ and $p_y(y;\theta) = \int p_x(y,z;\theta)dz$. For denoting densities, we will omit random variables such as q_y and $p_y(\theta)$. We assume that θ is identifiable with respect to $p_y(\theta)$.

In this paper, we consider only a simple setting of i.i.d. random variables of sample size n. Let $x_i = (y_i, z_i)$, $i = 1, \ldots, n$, be independent realizations of X, where we only observe $y_1, \ldots, y_n$ and we cannot see the values of $z_1, \ldots, z_n$. We estimate θ from the observed training data $y_1, \ldots, y_n$. Then the maximum likelihood estimate (MLE) of θ is given by

$$\hat{\theta}_y = \arg\max_{\theta \in \Theta} \ell_y(\theta) \equiv \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_y(y_i;\theta), \tag{1}$$

where $\ell_y(\theta)$ denotes the log-likelihood function (divided by n) of θ with respect to $y_1, \ldots, y_n$.

If we were only interested in Y, we would consider the plug-in predictive distribution $p_y(\hat{\theta}_y)$ by substituting $\hat{\theta}_y$ into $p_y(\theta)$. However, we are interested in the whole primary variable $X = (Y, Z)$ and its density q_x. We thus consider $p_x(\hat{\theta}_y)$ by substituting $\hat{\theta}_y$ into $p_x(\theta)$, and evaluate the MLE by comparing $p_x(\hat{\theta}_y)$ with q_x. For this purpose, Shimodaira and Maeda [13] derived an information criterion as an asymptotically unbiased estimator of the KL risk function which measures how well $p_x(\hat{\theta}_y)$ approximates q_x.

2.2. Statistical Analysis with Auxiliary Variables

Next, we extend the setting to incomplete data analysis with auxiliary variables. Let A denote a vector of auxiliary variables. In addition to Y, we observe A in the training data, but we are *not* interested in A. For convenience, we introduce a vector of observable variables $B = (Y, A)$ and a vector of all variables $C = (Y, Z, A)$ as summarized in Table 1. Now $c_i = (y_i, z_i, a_i)$, $i = 1, \ldots, n$, are independent realizations of C, and we estimate θ from the observed training data $b_i = (y_i, a_i)$, $i = 1, \ldots, n$. Let $\hat{\theta}_b$ be the MLE of θ by using A in addition to Y. Since we are only interested in the primary variables, we consider the plug-in predictive distribution $p_x(\hat{\theta}_b)$ by substituting $\hat{\theta}_b$ into $p_x(\theta)$, and evaluate the MLE by comparing $p_x(\hat{\theta}_b)$ with q_x.

Table 1. Random variables in incomplete data analysis with auxiliary variables. $B = (Y, A)$ is used for estimation of unknown parameters, and $X = (Y, Z)$ is used for evaluation of candidate models.

	Observed	**Latent**	**Complete**
Primary	Y	Z	X
Auxiliary	A	–	–
All	B	–	C

In order to define the MLE $\hat{\theta}_b$, let us clarify a candidate parametric model with auxiliary variables. We write the true density function of C as $q_c(c) = q_c(y,z,a)$ and a candidate parametric model of the true density as $p_c(c;\beta) = p_c(y,z,a;\beta)$, where $\beta = (\theta^\top, \varphi^\top)^\top \in \mathcal{B} \subset \mathbb{R}^{d+f}$ is an unknown parameter vector with nuisance parameter $\varphi \in \mathbb{R}^f$ and $\mathcal{B}$ is its parameter space. We assume that $c = (y,z,a) \in \mathcal{Y} \times \mathcal{Z} \times \mathcal{A}$ for all density functions, where $\mathcal{A}$ is the domain of A. We also assume that β is identifiable with respect to $p_b(y,a;\beta) = \int p_c(y,z,a;\beta)dz$. Let us redefine $p_x(\theta)$ as $p_x(y,z;\theta) = \int p_c(y,z,a;\beta)da$ and the parameter space of θ as

$$\Theta = \left\{ \theta \;\middle|\; \begin{pmatrix} \theta \\ \varphi \end{pmatrix} \in \mathcal{B} \right\}.$$

Then, $\hat{\theta}_b$ is obtained from the MLE of β given by

$$\hat{\beta}_b = \begin{pmatrix} \hat{\theta}_b \\ \hat{\varphi}_b \end{pmatrix} = \arg\max_{\beta \in \mathcal{B}} \ell_b(\beta) \equiv \arg\max_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} \log p_b(b_i; \beta), \tag{2}$$

where $\ell_b(\beta)$ denotes the log-likelihood function (divided by n) of β with respect to $b_1, \ldots, b_n$.

Finally, we introduce a general notation for density functions. For a random variable, say R, we write the true density function as $q_r(r)$ and a candidate parametric model of q_r as $p_r(r;\theta)$ or $p_r(r;\beta)$. For random variables R and S, we write the true conditional density function of R given $S = s$ as $q_{r|s}(r|s)$ and its corresponding model as $p_{r|s}(r|s;\theta)$ or $p_{r|s}(r|s;\beta)$. For example, a candidate model of C can be decomposed as

$$p_c(y, z, a; \beta) = p_x(y, z; \theta) p_{a|x}(a|y, z; \beta).$$

2.3. Comparing the Two Estimators

We have thus far obtained the two MLEs of θ, namely $\hat{\theta}_y$ and $\hat{\theta}_b$, and their corresponding predictive distributions $p_x(\hat{\theta}_y)$ and $p_x(\hat{\theta}_b)$, respectively. We would like to determine which of the two predictive distributions approximates q_x better than the other. The approximation error of $p_x(\theta)$ is measured by the KL divergence from q_x to $p_x(\theta)$ defined as

$$D_x(q_x; p_x(\theta)) = -\int q_x(x) \log p_x(x;\theta) dx + \int q_x(x) \log q_x(x) dx.$$

Since the last term on the right hand side does not depend on $p_x(\theta)$, we ignore it for computing the loss function of $p_x(\theta)$ defined by

$$\mathcal{L}_x(\theta) = -\int q_x(x) \log p_x(x;\theta) dx.$$

Let $\hat{\theta}$ be an estimator of θ. The risk (or expected loss) function of $p_x(\hat{\theta})$ is defined by

$$\mathcal{R}_x(\hat{\theta}) = E[\mathcal{L}_x(\hat{\theta})], \tag{3}$$

where we take the expectation by considering $\hat{\theta}$ as a random variable. Note that $\hat{\theta}$ in the notation of $\mathcal{R}_x(\hat{\theta})$ indicates the procedure for computing $\hat{\theta}$ instead of a particular value of $\hat{\theta}$. $\mathcal{R}_x(\hat{\theta})$ measures how well $p_x(\hat{\theta})$ approximates q_x on average in the long run.

For comparing the two MLEs, we define $\mathcal{R}_x(\hat{\theta}_y)$ and $\mathcal{R}_x(\hat{\theta}_b)$ by considering that $\hat{\theta}_y$ and $\hat{\theta}_b$ are functions of independent random variables $Y_1, \ldots, Y_n$ and $B_1, \ldots, B_n$, respectively, where $B_i = (Y_i, A_i)$ has the same distribution as B for all $i = 1, \ldots, n$. $\hat{\theta}_b$ is better than $\hat{\theta}_y$ when $\mathcal{R}_x(\hat{\theta}_b) < \mathcal{R}_x(\hat{\theta}_y)$, that is, the auxiliary variable A helps the statistical inference on q_x. On the other hand, A is harmful when $\mathcal{R}_x(\hat{\theta}_b) > \mathcal{R}_x(\hat{\theta}_y)$. Although we focus only on comparison between Y and $B = (Y, A)$ in this paper, if there are more than two auxiliary variables (and their combinations) $A_1, A_2, \ldots$, then we may compare $\mathcal{R}_x(\hat{\theta}_{(y,a_1)}), \mathcal{R}_x(\hat{\theta}_{(y,a_2)}), \ldots$, to determine good auxiliary variables. Of course, the risk functions cannot be calculated in reality because they depend on the unknown true distribution. Thus, we derive a new information criterion as an estimator of the risk function in our setting. Since an asymptotically unbiased estimator of $\mathcal{R}_x(\hat{\theta}_y)$ has been already derived in Shimodaira and Maeda [13], we will only derive an asymptotically unbiased estimator of $\mathcal{R}_x(\hat{\theta}_b)$.

3. An Illustrative Example with Auxiliary Variables

3.1. Model Setting

In this section, we demonstrate parameter estimation by using auxiliary variables in Gaussian mixture model (GMM), which can be formulated in incomplete data analysis. Let us consider a two-component GMM; observed values are generated from one of two Gaussian distributions, where the assigned labels are missing. The observed data and missing labels are realizations of Y and Z, respectively. We estimate a predictive distribution of $X = (Y, Z)$ from the observation of Y, and we attempt improving it by utilizing A in addition to Y. The true density function of primary variables $X = (Y, Z) \in \mathbb{R} \times \{0, 1\}$ is given as

$$\begin{aligned} q_{y|z}(y|z) &= zN(y; -1.2, 0.7) + (1-z)N(y; 1.2, 0.7), \\ q_z(z) &= 0.6z + 0.4(1-z), \end{aligned}$$

where $N(\cdot; \mu, \sigma^2)$ denotes the density function of $N(\mu, \sigma^2)$, i.e., the normal distribution with mean μ and variance σ^2. We consider the following two cases for the true conditional distribution of auxiliary variable A given $X = x$:

Case 1: $q_{a|x}(a|y,z) = q_{a|z}(a|z) = zN(a; 1.8, 0.49) + (1-z)N(a; -1.8, 0.49)$.
Case 2: $q_{a|x}(a|y,z) = q_a(a) = 0.6N(a; 1.8, 0.49) + 0.4N(a; -1.8, 0.49)$.

The random variables X and A are not independent in Case 1 whereas they are independent in Case 2. Hence, A will contribute to estimating θ in Case 1. On the other hand, in Case 2, A must not be useful, and A becomes just noise if we estimate θ from Y and A.

In both cases, we use the following two-component GMM as a candidate model of q_c:

$$\begin{aligned} p_{b|z}(y, a|z; \beta) &= zN_2((y,a)^\top; \mu_1, \Sigma) + (1-z)N_2((y,a)^\top; \mu_2, \Sigma), \\ p_z(z; \theta) &= \pi_1 z + (1-\pi_1)(1-z), \end{aligned} \tag{4}$$

where $N_2(\cdot; \mu_i, \Sigma)$ denotes the density function of bivariate normal distribution $N_2(\mu_i, \Sigma)$, $i = 1, 2$, and the parameters are

$$\mu_1 = \begin{pmatrix} \mu_{1y} \\ \mu_{1a} \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} \mu_{2y} \\ \mu_{2a} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{ya} \\ \sigma_{ya} & \sigma_a^2 \end{pmatrix}.$$

Therefore, $\beta = (\theta^\top, \varphi^\top)^\top$, $\theta = (\pi_1, \mu_{1y}, \mu_{2y}, \sigma_y^2)^\top$ and $\varphi = (\mu_{1a}, \mu_{2a}, \sigma_a^2, \sigma_{ya})^\top$. The true parameters of θ and φ for Case 1 are given by $\theta_0 = (0.6, -1.2, 1.2, 0.7)^\top$ and $\varphi_0 = (1.8, -1.8, 0.49, 0)^\top$, respectively. By considering the joint density function $p_c(y, z, a; \beta) = p_{b|z}(y, a|z; \beta)p_z(z; \theta)$, this candidate model correctly specifies the true density function $q_c(y, z, a) = q_{a|x}(a|y, z)q_{y|z}(y|z)q_z(z)$ in Case 1. On the other hand, the model is misspecified for Case 2, and we cannot think of the true parameters.

3.2. Estimation Results

For illustrating the impact of auxiliary variables on parameter estimation in each case, we generated a typical dataset $c_1, \ldots, c_n$ with sample size $n = 100$ from q_c, which is actually picked from 10,000 datasets generated in the simulation study of Section 6, and details of how to select the typical dataset are also shown there. For each case, we computed the three MLEs $\hat{\theta}_y$, $\hat{\theta}_b$, and $\hat{\theta}_x$, where $\hat{\theta}_x$ is the MLE of θ calculated by using complete data $x_1, \ldots, x_n$ as if labels $z_1, \ldots, z_n$ were available.

The result of Case 1 is shown in Figure 1, where A is beneficial for estimating θ. In the left panel, the two clusters are well separated, which makes parameter estimation stable. The estimated $p_b(\hat{\beta}_b)$ captures the structure of the two clusters corresponding to the label $z_i = 0$ and $z_i = 1$, showing that $p_c(\hat{\beta}_b)$ is estimated reasonably well, and thus $p_x(\hat{\theta}_b)$ is a good approximation of q_x. Looking at the

right panel, we also observe that $p_y(\hat{\theta}_b)$ is better than $p_y(\hat{\theta}_y)$ for approximating $p_y(\hat{\theta}_x)$, suggesting that the auxiliary variable is useful for recovering the lost information of missing data. In fact, the three MLEs are calculated as follows: $\hat{\theta}_y = (0.671, -1.143, 1.324, 0.678)^\top$, $\hat{\theta}_b = (0.613, -1.228, 1.093, 0.744)^\top$, and $\hat{\theta}_x = (0.620, -1.233, 1.141, 0.695)^\top$. By comparing $\|\hat{\theta}_b - \hat{\theta}_x\| = 0.069$ with $\|\hat{\theta}_y - \hat{\theta}_x\| = 0.212$, we can see that $\hat{\theta}_b$ is better than $\hat{\theta}_y$ for predicting $\hat{\theta}_x$ without looking at the latent variable. All these observations indicate that the parameter estimation of θ is improved by using A in Case 1.

The result of Case 2 is shown in Figure 2, where A is harmful for estimating θ. For fair comparison, exactly the same values of $\{(y_i, z_i)\}_{i=1}^{100}$ are used in both cases. Thus, $\hat{\theta}_y$ and $\hat{\theta}_x$ have the same values as in Case 1 whereas $\hat{\theta}_b$ has a different value as $\hat{\theta}_b = (0.581, -0.403, -0.232, 2.015)^\top$. By comparing $\|\hat{\theta}_b - \hat{\theta}_x\| = 2.078$ with $\|\hat{\theta}_y - \hat{\theta}_x\| = 0.212$, we can see that $\hat{\theta}_b$ is worse than $\hat{\theta}_y$ for predicting $\hat{\theta}_x$. This is also seen in Figure 2. In the left panel, the estimated $p_b(\hat{\beta}_b)$ captures some structure of the two clusters, but they do not correspond to the label $z_i = 0$ and $z_i = 1$. As a result, $p_y(\hat{\theta}_b)$ becomes a very poor approximation of $p_y(\hat{\theta}_x)$ in the right panel, indicating that the parameter estimation of θ is actually hindered by using A in Case 2.

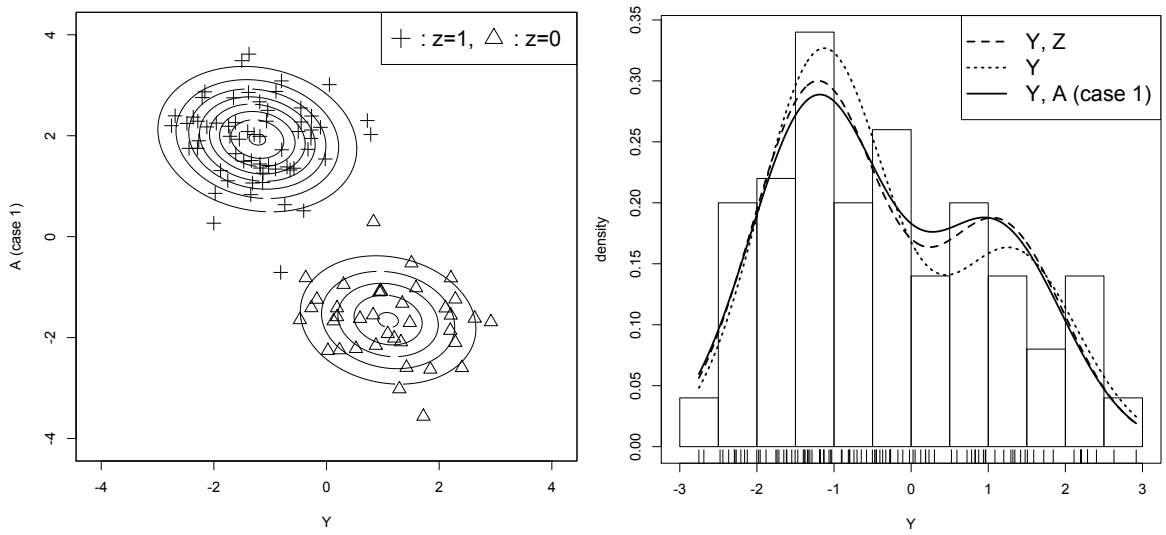

Figure 1. *Useful* auxiliary variable (Case 1). The left panel plots $\{(y_i, a_i)\}_{i=1}^{100}$ with labels indicating z_i. The estimated $p_b(\hat{\beta}_b)$ is shown by the contour lines. The right panel shows the histogram of $\{y_i\}_{i=1}^{100}$, and three density functions $p_y(\hat{\theta}_x)$ (broken line), $p_y(\hat{\theta}_y)$ (dotted line), and $p_y(\hat{\theta}_b)$ (solid line). In Section 4.4, this useful auxiliary variable is selected by our method (Case 1 in Table 2).

Table 2. Comparisons between $\hat{\theta}_b$ and $\hat{\theta}_y$ for predicting X, and that for Y.

	$p_x(\hat{\theta}_b)$ vs. $p_x(\hat{\theta}_y)$	$p_y(\hat{\theta}_b)$ vs. $p_y(\hat{\theta}_y)$
	$\mathbf{AIC}_{x;b} - \mathbf{AIC}_{x;y}$	$\mathbf{AIC}_{y;b} - \mathbf{AIC}_{y;y}$
Case 1	−2.67	−0.96
Case 2	9.86	10.37

These examples suggest that usefulness of auxiliary variables depends strongly on the true distribution and a candidate model. Hence, it is important to select useful auxiliary variables from observed data.

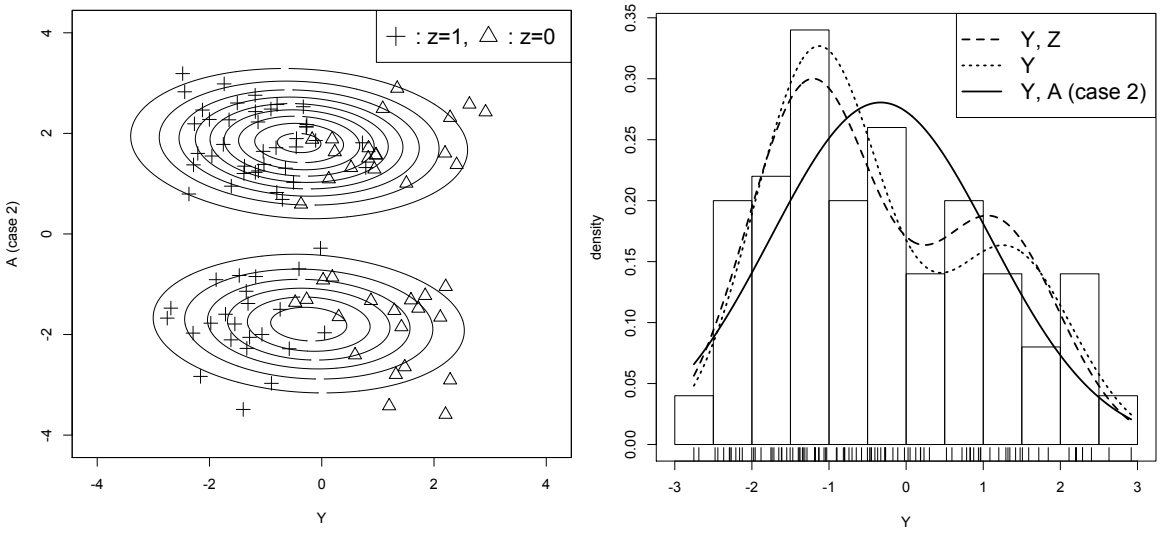

Figure 2. *Useless* auxiliary variable (Case 2). The symbols are the same as Figure 1. In Section 4.4, this useless auxiliary variable is *NOT* selected by our method (Case 2 in Table 2).

4. Information Criterion

4.1. Asymptotic Expansion of the Risk Function

In this section, we derive a new information criterion as an asymptotically unbiased estimator of the risk function $\mathcal{R}_x(\hat{\theta}_b)$ defined in (3). We start from a general framework of misspecification, i.e., without assuming that candidate models are correctly specified, and later we give specific assumptions. Let $\bar{\beta}$ be the optimal parameter value with respect to the KL divergence from q_b to $p_b(\beta)$, that is,

$$\bar{\beta} = \begin{pmatrix} \bar{\theta} \\ \bar{\varphi} \end{pmatrix} = \arg\max_{\beta \in \mathcal{B}} \int q_b(b) \log p_b(b;\beta) db.$$

If the candidate model is correctly specified, i.e., there exists $\beta_0 = (\theta_0^\top, \varphi_0^\top)^\top$ such that $q_b = p_b(\beta_0)$, then $\bar{\beta} = \beta_0$ as well as $\bar{\theta} = \theta_0$.

In this paper, we assume the regularity conditions A1 to A6 of White [16] for q_b and $p_b(\beta)$ so that the MLE $\hat{\beta}_b$ has consistency and asymptotic normality. In particular, $\bar{\beta}$ is determined uniquely (i.e., identifiable) and is interior to $\mathcal{B}$. We assume that I_b and J_b defined below are nonsingular in the neighbourhood of $\bar{\beta}$. Then White [16] showed the asymptotic normality as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_b - \bar{\beta}) \xrightarrow{d} N_{d+f}(0, I_b^{-1} J_b I_b^{-1}), \tag{5}$$

where I_b and J_b are $(d+f) \times (d+f)$ matrices defined by using $\nabla = \partial/\partial\beta$, $\nabla^\top = \partial/\partial\beta^\top$, and $\nabla^2 = \partial^2/\partial\beta\partial\beta^\top$ as

$$I_b = -E[\nabla^2 \log p_b(b;\bar{\beta})], \quad J_b = E[\nabla \log p_b(b;\bar{\beta}) \nabla^\top \log p_b(b;\bar{\beta})].$$

Note that we write derivatives by abbreviated forms, e.g., $\nabla^2 \log p_b(b;\bar{\beta})$ means $\nabla^2 \log p_b(b;\beta)|_{\beta=\bar{\beta}}$ and so on. In addition, we allow interchange of integrals and derivatives rather formally when working with models, although we actually need conditions for the models such as White [16]. Moreover, the condition A7 of White [16] is assumed in order to establish $I_b = J_b$ when considering a situation that the candidate model is correctly specified. We assume the above conditions throughout the paper without explicitly stated.

Let us define three $(d+f)\times(d+f)$ matrices as

$$I_x = -E[\nabla^2 \log p_x(x;\bar{\theta})], \quad I_y = -E[\nabla^2 \log p_y(y;\bar{\theta})], \quad I_{z|y} = -E[\nabla^2 \log p_{z|y}(z|y;\bar{\theta})] = I_x - I_y,$$

which will be used in the lemmas below. Since the derivatives of $\log p_x(x;\theta)$ and $\log p_y(y;\theta)$ with respect to φ is zero, the matrices become singular when $f > 0$, but this is not a problem in our calculation. The following lemma shows that the dominant term of $\mathcal{R}_x(\hat{\theta}_b)$ is $\mathcal{L}_x(\bar{\theta})$ and the remainder terms are of order $O(n^{-1})$, by noting that $\nabla^\top \mathcal{L}_x(\bar{\theta}) = O(1)$ and $E[\hat{\beta}_b - \bar{\beta}] = O(n^{-1})$ in general. The proof is given in Appendix A.1.

Lemma 1. *The risk function $\mathcal{R}_x(\hat{\theta}_b)$ is expanded asymptotically as*

$$\mathcal{R}_x(\hat{\theta}_b) = \mathcal{L}_x(\bar{\theta}) + \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] + \frac{1}{2n}\mathrm{tr}(I_x I_b^{-1} J_b I_b^{-1}) + o(n^{-1}).$$

Just as a remark, the term $\nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] = O(n^{-1})$ above does not appear in the derivation of AIC or TIC, where $B = X$ and thus $\nabla^\top \mathcal{L}_x(\bar{\theta}) = 0$. This term appears when the loss function for evaluation and that for estimation differ, for example, in the derivation of the information criterion under covariate shift; see $K_w^{[1]\top} b_w$ in Equation (4.1) of Shimodaira [17].

4.2. Estimating the Risk Function

For deriving an estimator of $\mathcal{R}_x(\hat{\theta}_b)$, we introduce an additional condition. Let us assume that the candidate model is correctly specified for the latent part as

$$q_{z|y}(z|y) = p_{z|y}(z|y;\bar{\theta}). \tag{6}$$

This is the same condition as Equation (14) of Shimodaira and Maeda [13] except that $\bar{\theta}$ is replaced by

$$\bar{\theta}_y = \arg\max_{\theta\in\Theta} \int q_y(y) \log p_y(y;\theta) dy.$$

Since Z is missing completely in our setting, we need such a condition to proceed further. Although any method cannot detect misspecification of $p_{z|y}$ if p_b is correctly specified, it is often the case that misspecification of $p_{z|y}$ leads to that of p_b, and thus it is detected indirectly as in Case 2 of Section 3.

Note that the symbol of $\bar{\theta}$ in our notation should have been $\bar{\theta}_b$, although we used $\bar{\theta}$ for simplicity, and there is also $\bar{\theta}_x$ defined similarly from $p_x(x;\theta)$. They all differ each other with differences of order $O(1)$ in general, but $\bar{\theta} = \bar{\theta}_y = \bar{\theta}_x = \theta_0$ when $p_c(\beta)$ is correctly specified as $q_c = p_c(\beta_0)$.

Now we give the asymptotic expansion of $E[\ell_y(\hat{\theta}_b)]$, which shows that $-\ell_y(\hat{\theta}_b)$ can be used as an estimator of $\mathcal{L}_x(\bar{\theta})$ but the asymptotic bias is of order $O(n^{-1})$.

Lemma 2. *Assume the condition (6). Then, the expectation of the estimated log-likelihood $\ell_y(\hat{\theta}_b)$ can be expanded as*

$$E[\ell_y(\hat{\theta}_b)] = -\mathcal{L}_x(\bar{\theta}) - C(q_x) - \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] + \frac{1}{n}\mathrm{tr}(I_b^{-1} K_{b,y}) - \frac{1}{2n}\mathrm{tr}(I_y I_b^{-1} J_b I_b^{-1}) + o(n^{-1}),$$

where $K_{b,y} = E[\nabla \log p_b(\bar{\beta}) \nabla^\top \log p_y(\bar{\theta})]$ and $C(q_x) = \int q_x(x) \log q_{z|y}(z|y) dx$.

The proof of Lemma 2 is given in Appendix A.2. By eliminating $\mathcal{L}_x(\bar{\theta})$ from the two expressions in Lemma 1 and Lemma 2, and rearranging the formula, we get the following lemma, which plays a central role in deriving our information criterion.

Lemma 3. *Assume the condition (6). Then, an expansion of the risk function $\mathcal{R}_x(\hat{\theta}_b)$ is given by*

$$\mathcal{R}_x(\hat{\theta}_b) = -E[\ell_y(\hat{\theta}_b)] - C(q_x) + \frac{1}{n}\mathrm{tr}(I_b^{-1}K_{b,y}) + \frac{1}{2n}\mathrm{tr}(I_{z|y}I_b^{-1}J_bI_b^{-1}) + o(n^{-1}). \tag{7}$$

We can ignore $C(q_x)$ for model selection, because it is a constant term which does not depend on the candidate model. Thus, finally, we define an information criterion from the right hand side of (7). The following theorem is an immediate consequence of Lemma 3.

Theorem 1. *Assume the condition (6). Let us define an information criterion as*

$$\widehat{\mathrm{risk}}_{x;b} = -2n\ell_y(\hat{\theta}_b) + 2\mathrm{tr}(I_b^{-1}K_{b,y}) + \mathrm{tr}(I_{z|y}I_b^{-1}J_bI_b^{-1}). \tag{8}$$

Then this criterion is an asymptotically unbiased estimator of $2n\mathcal{R}_x(\hat{\theta}_b)$ by ignoring the constant term $C(q_x)$.

$$E[\widehat{\mathrm{risk}}_{x;b}] = 2n\mathcal{R}_x(\hat{\theta}_b) + 2nC(q_x) + o(1).$$

Note that the subscript of $\widehat{\mathrm{risk}}_{x;b}$, $x;b$ is defined in accordance with Shimodaira and Maeda [13]; thus the former x and the latter b mean random variables used in evaluation and estimation, respectively. This criterion is an extension of TIC because when $X = B = Y$, $\widehat{\mathrm{risk}}_{x;b}$ coincides with TIC of Takeuchi [9] defined as follows:

$$\mathrm{TIC} = -2n\ell_y(\hat{\theta}_y) + 2\mathrm{tr}(I_y^{-1}J_y).$$

4.3. Akaike Information Criteria for Auxiliary Variable Selection

In actual use, $\widehat{\mathrm{risk}}_{x;b}$ may have a too complicated form. Thus, we derive a simpler information criterion by assuming the correctness of the candidate model like as AIC.

Theorem 2. *Suppose $p_c(\beta)$ is correctly specified so that $q_c = p_c(\beta_0)$ for some $\beta_0 \in \mathcal{B}$. Then, we have*

$$J_b = I_b, \quad K_{b,y} = I_y, \tag{9}$$

and thus $\widehat{\mathrm{risk}}_{x;b}$ is rewritten as

$$\mathrm{AIC}_{x;b} = -2n\ell_y(\hat{\theta}_b) + \mathrm{tr}(I_xI_b^{-1}) + \mathrm{tr}(I_yI_b^{-1}). \tag{10}$$

This criterion is an asymptotically unbiased estimator of $2n\mathcal{R}_x(\hat{\theta}_b)$ by ignoring the constant term $C(q_x)$.

$$E[\mathrm{AIC}_{x;b}] = 2n\mathcal{R}_x(\hat{\theta}_b) + 2nC(q_x) + o(1).$$

The proof is given in Appendix A.3. I_x, I_y and I_b are replaced by their consistent estimators in practical situations.

The newly obtained criterion $\mathrm{AIC}_{x;b}$ is a generalization of AIC and some of its variants. If θ is estimated by $\hat{\theta}_y$ instead of $\hat{\theta}_b$, we simply let $B = Y$ in the expression of $\mathrm{AIC}_{x;b}$ so that we get $\mathrm{AIC}_{x;y}$ proposed by Shimodaira and Maeda [13]:

$$\mathrm{AIC}_{x;y} = -2n\ell_y(\hat{\theta}_y) + \mathrm{tr}(I_xI_y^{-1}) + d. \tag{11}$$

Note that if $B = Y$, I_y is not singular because $\beta = \theta$. On the other hand, if there is no latent part, we simply let $X = Y$ in the expression of $\mathrm{AIC}_{x;b}$ so that we get

$$\mathrm{AIC}_{y;b} = -2n\ell_y(\hat{\theta}_b) + 2\mathrm{tr}(I_yI_b^{-1}). \tag{12}$$

This can be used to select useful auxiliary variables in complete data analysis. Moreover, if $X = Y = B$, $\mathrm{AIC}_{x;b}$ reduces to the original AIC proposed by Akaike [6]:

$$\mathrm{AIC}_{y;y} = -2n\ell_y(\hat{\theta}_y) + 2d. \tag{13}$$

It is worth mentioning that $\mathrm{tr}(I_{z|y}I_b^{-1})$ is interpreted as the additional penalty for the latent part:

$$\mathrm{AIC}_{x;b} - \mathrm{AIC}_{y;b} = \mathrm{tr}(I_x I_b^{-1}) - \mathrm{tr}(I_y I_b^{-1}) = \mathrm{tr}(I_{z|y} I_b^{-1}) \geq 0,$$

which is also mentioned in Equation (1) of Shimodaira and Maeda [13] for the case of $B = Y$.

4.4. The Illustrative Example (Cont.)

Let us return to the problem of determining whether to use the auxiliary variables or not, that is, comparison between $p_x(\hat{\theta}_b)$ and $p_x(\hat{\theta}_y)$. By comparing $\mathrm{AIC}_{x;b}$ with $\mathrm{AIC}_{x;y}$, we can determine whether the vector of auxiliary variables A is useful or useless. Thus, only when $\mathrm{AIC}_{x;b} < \mathrm{AIC}_{x;y}$, we conclude that A is useful in order to estimate θ for predicting X.

Let us apply this procedure to the illustrative example in Section 3. The generalized AICs are computed for the two cases of the typical dataset, and the results are shown in Table 2. Looking at the value of $\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}$, it is negative for Case 1, concluding that the auxiliary variable is useful, and it is positive for Case 2, concluding that the auxiliary variable is useless. According to the AIC values, therefore, we use the auxiliary variable of Case 1, but do not use the auxiliary variable of Case 2. This decision agrees with the observations of Figures 1 and 2 in Section 3.2. In fact, the decision is correct, because the value of $\mathcal{R}_x(\hat{\theta}_b) - \mathcal{R}_x(\hat{\theta}_y)$ is negative for Case 1 and positive for Case 2 as will be seen in the simulation study of Section 6.2.

We can also argue the usefulness of the auxiliary variable for predicting Y instead of X, that is, comparison between $p_y(\hat{\theta}_b)$ and $p_y(\hat{\theta}_y)$. By comparing $\mathrm{AIC}_{y;b}$ with $\mathrm{AIC}_{y;y}$, we can determine whether A is useful or useless for predicting Y. Looking at the value of $\mathrm{AIC}_{y;b} - \mathrm{AIC}_{y;y}$ in Table 2, we make the same decision as that for X.

5. Leave-One-Out Cross Validation

Variable selection by cross-validatory (CV) choice [18] is often applied to real data analysis due to its simplicity, although its computational burden is larger than that of information criteria; see Arlot and Celisse [19] for a recent review of cross-validation methods. As shown in Stone [14], leave-one-out cross validation (LOOCV) is asymptotically equivalent to TIC. Because LOOCV does not require calculation of the information matrices of TIC, LOOCV is easier to use than TIC. There are also some literature for improving LOOCV such as Yanagihara et al. [20], which gives a modification of LOOCV to reduce its bias by considering maximum weighted log-likelihood estimation. However, we focus on the result of Stone [14] and extend it to our setting.

In incomplete data analysis, LOOCV cannot be directly used because the loss function with respect to the complete data includes latent variables. Thus, we transform the loss function as follows:

$$\mathcal{L}_x(\theta) = -\int q_y(y) g(y;\theta) dy,$$

where $g(y;\theta) = \log p_y(y;\theta) + f(y;\theta)$ and

$$f(y;\theta) = \int q_{z|y}(z|y) \log p_{z|y}(z|y;\theta) dz.$$

Note that $f(y;\theta) = 0$ when $X = Y$. Using the function $g(y;\theta)$, we then obtain the following LOOCV estimator of the risk function $\mathcal{R}_x(\hat{\theta}_b)$.

$$\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b) = -\frac{1}{n}\sum_{i=1}^{n} g(y_i;\hat{\theta}_b^{(-i)}),$$

where $\hat{\theta}_b^{(-i)}$ is the leave-out-out estimate of θ defined as

$$\hat{\beta}_b^{(-i)} = \begin{pmatrix}\hat{\theta}_b^{(-i)}\\ \hat{\varphi}_b^{(-i)}\end{pmatrix} = \arg\max_{\beta\in\mathcal{B}} \frac{1}{n}\sum_{j\neq i}^{n} \log p_b(b_j;\beta) = \arg\max_{\beta\in\mathcal{B}} \left\{\ell_b(\beta) - \frac{1}{n}\log p_b(b_i;\beta)\right\}.$$

We will show below in this section that $\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b)$ is asymptotically equivalent to $\widehat{\mathrm{risk}}_{x;b}$. For implementing the LOOCV procedure with latent variables, however, we have to estimate $q_{z|y}(z|y)$ by $p_{z|y}(z|y,\hat{\theta}_b)$ in $f(y;\theta)$. This introduces a bias to $\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b)$, and hence, information criteria are preferable to the LOOCV in incomplete data analysis.

Let us show the asymptotic equivalence of $\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b)$ and $\widehat{\mathrm{risk}}_{x;b}$ by assuming that we know the functional form of $f(y;\theta)$. Noting that $\hat{\beta}_b^{(-i)}$ is a critical point of $\ell_b(\beta) - \log p_b(b_i;\beta)/n$, we have

$$\nabla\ell_b(\hat{\beta}_b^{(-i)}) = \frac{1}{n}\nabla \log p_b(b_i;\hat{\beta}_b^{(-i)}) = O_p(n^{-1}).$$

By applying Taylor expansion to $\nabla\ell_b(\beta)$ around $\beta = \hat{\beta}_b$, it follows from $\nabla\ell_b(\hat{\beta}_b) = 0$ that

$$\nabla^2\ell_b(\tilde{\beta}_b^i)(\hat{\beta}_b^{(-i)} - \hat{\beta}_b) = \frac{1}{n}\nabla \log p_b(b_i;\hat{\beta}_b^{(-i)}), \tag{14}$$

where $\tilde{\beta}_b^i$ lies between $\hat{\beta}_b^{(-i)}$ and $\hat{\beta}_b$. We can see from (14) that $\hat{\beta}_b^{(-i)} - \hat{\beta}_b = O_p(n^{-1})$. Next, we regard $g(y_i;\theta)$ as a function of β and apply Taylor expansion to it around $\beta = \hat{\beta}_b$. Therefore, $g(y_i;\hat{\theta}_b^{(-i)})$ can be expressed as follows:

$$g(y_i;\hat{\theta}_b^{(-i)}) = g(y_i;\hat{\theta}_b) + \nabla^\top g(y_i;\tilde{\theta}_b^i)(\hat{\beta}_b^{(-i)} - \hat{\beta}_b), \tag{15}$$

where $\tilde{\theta}_b^i$ lies between $\hat{\theta}_b^{(-i)}$ and $\hat{\theta}_b$ ($\tilde{\theta}_b^i$ does not corresponding to $\tilde{\beta}_b^i$). Then we assume that

$$\frac{1}{n}\sum_{i=1}^{n} \nabla^2\ell_b(\tilde{\beta}_b^i)^{-1}\nabla \log p_b(b_i;\hat{\beta}_b^{(-i)})\nabla^\top g(y_i;\tilde{\theta}_b^i) \xrightarrow{p} -I_b^{-1}E[\nabla \log p_b(b;\bar{\beta})\nabla^\top g(y;\bar{\theta})]. \tag{16}$$

By noting $\hat{\beta}_b^{(-i)} = \hat{\beta}_b + O_p(n^{-1})$, we have $\tilde{\beta}_b^i = \bar{\beta} + O_p(n^{-1/2})$ and $\tilde{\theta}_b^i = \bar{\theta} + O_p(n^{-1/2})$, and thus (16) holds at least formally. With the above setup, we show the following theorem. The proof is given in Appendix A.4.

Theorem 3. *Supposing the same assumptions of Theorem 1 and (16), we have*

$$2n\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b) = \widehat{\mathrm{risk}}_{x;b} - 2\sum_{i=1}^{n} f(y_i;\bar{\theta}) + o_p(1). \tag{17}$$

Because the second term on the right-hand side of (17) does not depend on candidate models under condition (6), this theorem implies that $\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b)$ is asymptotically equivalent to $\widehat{\mathrm{risk}}_{x;b}$ except for the scaling and the constant term. However, someone may wonder why $f(y;\theta)$ is included in $g(y;\theta)$ for comparing models of $p(b;\beta)$. By assuming that $p_{z|y}(\theta)$ is correctly specified for $q_{z|y}$, $f(y;\bar{\theta}) = \int q_{z|y}(z|y)\log q_{z|y}(z|y)dz$ does not depend on the model anymore, so we may simply exclude $f(y;\theta)$ from $g(y;\theta)$, leading to the loss $\mathcal{L}_y(\theta)$ instead. The reason for including $f(y;\theta)$ in $g(y;\theta)$ is

explained as follows. $\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b)$, as well as $\widehat{\mathrm{risk}}_{x;b}$ (and $\mathrm{AIC}_{x;b}$), include the additional penalty for estimating $\hat{\theta}_b$ in $f(y;\hat{\theta}_b)$, which depends on the candidate models even if $p_{z|y}(\theta)$ is correctly specified.

6. Experiments with Simulated Datasets

This section shows the usefulness of auxiliary variables and the proposed information criteria via a simulation study. The models illustrated in Section 3 are used for confirming the asymptotic unbiasedness of the information criterion and the validity of auxiliary variable selection.

6.1. Unbiasedness

At first, we confirm the asymptotic unbiasedness of $\mathrm{AIC}_{x;b}$ for estimating $2n\mathcal{R}_x(\hat{\theta}_b)$ except for the constant term, $C(q_x)$. The simulation setting is the same as Case 1 in Section 3, thus the data generating model is given by

$$q_{b|z}(y,a|z) = zN_2((y,a)^\top;\mu_{10},\Sigma_0) + (1-z)N_2((y,a)^\top;\mu_{20},\Sigma_0),$$
$$q_z(z) = 0.6z + 0.4(1-z),$$

where $\mu_{10} = -\mu_{20} = (-1.2, 1.8)^\top$ and $\Sigma_0 = \mathrm{diag}(0.7, 0.49)$. We generated $T = 10^4$ independent replicates of the dataset $\{(y_i, z_i, a_i)\}_{i=1}^n$ from this model; in fact, we used $\{(y_i, z_i, a_{i,1})\}_{i=1}^n$ generated in Section 6.2. The candidate model is given by (4), which is correctly specified for the above data generating model. Because $\mathrm{AIC}_{x;b}$ is derived by ignoring $C(q_x)$, we compare $E[\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}]$ with $2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{R}_x(\hat{\theta}_y)\}$. The computation of the expectation is approximated by the simulation average as

$$E[\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}] \approx \frac{1}{T}\sum_{t=1}^{T}\{\mathrm{AIC}_{x;b}^{(t)} - \mathrm{AIC}_{x;y}^{(t)}\},$$
$$2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{R}_x(\hat{\theta}_y)\} \approx \frac{2n}{T}\sum_{t=1}^{T}\{\mathcal{L}_x(\hat{\theta}_b^{(t)}) - \mathcal{L}_x(\hat{\theta}_y^{(t)})\},$$

where $\mathrm{AIC}_{x;b}^{(t)}$, $\mathrm{AIC}_{x;y}^{(t)}$, $\hat{\theta}_b^{(t)}$, and $\hat{\theta}_y^{(t)}$ are those computed for the t-th dataset ($t = 1, \ldots, T$).

Here, we remark about calculation of the loss function $\mathcal{L}_x(\hat{\theta})$ in two-component GMM. Let $\hat{\theta} = (\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)^\top$ be an estimator of θ. We expect that the components of GMM corresponding to $Z = 1$ and $Z = 0$ consist of $(\hat{\pi}_1, \hat{\mu}_1, \hat{\sigma}^2)$ and $(1 - \hat{\pi}_1, \hat{\mu}_2, \hat{\sigma}^2)$, respectively. However, we cannot determine the assignment of the estimated parameters in reality, i.e., $(\hat{\pi}_1, \hat{\mu}_1, \hat{\sigma}^2)$ and $(1 - \hat{\pi}_1, \hat{\mu}_2, \hat{\sigma}^2)$ may correspond to $Z = 0$ and $Z = 1$, respectively, because the labels $z_1, \ldots, z_n$ are missing. The assignment is required to calculate $\mathcal{L}_x(\hat{\theta})$ whereas it is not used for $\mathcal{L}_y(\hat{\theta})$ and the proposed information criteria. Hence, in this paper, we define $\mathcal{L}_x(\hat{\theta})$ as the minimum value between $\mathcal{L}(\hat{\theta})$ and $\mathcal{L}(\hat{\theta}')$, where $\hat{\theta}' = (1 - \hat{\pi}_1, \hat{\mu}_2, \hat{\mu}_1, \hat{\sigma}^2)^\top$.

Table 3 shows the result of the simulation for $n = 100, 200, 500, 1000, 2000$, and 5000. For all n, we observe that $E[\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}]$ is very close to $2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{R}_x(\hat{\theta}_y)\}$, indicating the unbiasedness of $\mathrm{AIC}_{x;b}$.

Table 3. Expected Akaike Information Criterion (AIC) difference is compared with the risk difference. The values are computed from $T = 10^4$ runs of simulation with their standard errors in parentheses.

n	100	200	500	1000	2000	5000
$E[\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}]$	−3.559	−3.263	−3.221	−3.197	−3.195	−3.180
	(0.074)	(0.021)	(0.015)	(0.013)	(0.013)	(0.012)
$2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{R}_x(\hat{\theta}_y)\}$	−3.603	−3.333	−3.275	−3.208	−3.182	−3.232
	(0.071)	(0.054)	(0.050)	(0.050)	(0.050)	(0.050)

6.2. Auxiliary Variable Selection

Next, we demonstrate that the proposed AIC selects a useful auxiliary variable (Case 1), while it does not select a useless auxiliary variable (Case 2). In each case, we generated $T = 10^4$ independent replicates of the dataset $\{(y_i, z_i, a_i)\}_{i=1}^n$ from the model. In fact, the values of $\{(y_i, z_i)\}_{i=1}^n$ are shared in both cases, so we generated replicates of $\{(y_i, z_i, a_{i,1}, a_{i,2})\}_{i=1}^n$, where $a_{i,1}$ and $a_{i,2}$ are auxiliary variables for Case 1 and Case 2, respectively. In each case, we compute $\mathrm{AIC}_{x;b}$ and $\mathrm{AIC}_{x;y}$, then we select $\hat{\theta}_b$ (i.e., selecting the auxiliary variable A) if $\mathrm{AIC}_{x;b} < \mathrm{AIC}_{x;y}$ and select $\hat{\theta}_y$ (i.e., not selecting the auxiliary variable A) otherwise. The selected estimator is denoted as $\hat{\theta}_{best}$. This experiment was repeated for $T = 10^4$ times. Note that the typical dataset in Section 3 was picked from the generated datasets so that it has around the median value in each of $\mathcal{L}_x(\hat{\theta}_b) - \mathcal{L}_x(\hat{\theta}_y)$, $\mathcal{L}_y(\hat{\theta}_b) - \mathcal{L}_y(\hat{\theta}_y)$, $\mathrm{AIC}_{x;b} - \mathrm{AIC}_{x;y}$, and $\mathrm{AIC}_{y;b} - \mathrm{AIC}_{y;y}$ in both cases.

The selection frequencies are shown in Tables 4 and 5. We observe that, as expected, the useful auxiliary variable tends to be selected in Case 1, while the useless auxiliary variable tends to be not selected in Case 2.

For verifying the usefulness of the auxiliary variable in both cases, we computed the risk value $\mathcal{R}_x(\hat{\theta})$ for $\hat{\theta} = \hat{\theta}_y$, $\hat{\theta}_b$, and $\hat{\theta}_{best}$. They are approximated by the simulation average as

$$\mathcal{R}_x(\hat{\theta}) \approx \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}_x(\hat{\theta}^{(t)}).$$

The results are shown in Tables 6 and 7. For easier comparisons, the values are the differences from $\mathcal{L}_x(\theta_0)$ with the true value θ_0. For all n, we observe that, as expected, $\mathcal{R}_x(\hat{\theta}_b) < \mathcal{R}_x(\hat{\theta}_y)$ in Case 1, and $\mathcal{R}_x(\hat{\theta}_b) > \mathcal{R}_x(\hat{\theta}_y)$ in Case 2. In both cases, $\mathcal{R}_x(\hat{\theta}_{best})$ is close to $\min\{\mathcal{R}_x(\hat{\theta}_b), \mathcal{R}_x(\hat{\theta}_y)\}$, indicating that the variable selection is working well.

Table 4. *Useful* auxiliary variable (Case 1): selection frequencies of $\hat{\theta}_b$ and $\hat{\theta}_y$.

n	100	200	500	1000	2000	5000
$\hat{\theta}_b$	9230	9475	9649	9687	9711	9727
$\hat{\theta}_y$	770	525	351	313	289	273

Table 5. *Useless* auxiliary variable (Case 2): selection frequencies of $\hat{\theta}_b$ and $\hat{\theta}_y$.

n	100	200	500	1000	2000	5000
$\hat{\theta}_b$	1508	212	1	0	0	0
$\hat{\theta}_y$	8492	9788	9999	10,000	10,000	10,000

Table 6. *Useful* auxiliary variable (Case 1): estimated risk functions of $\hat{\theta}_b$, $\hat{\theta}_y$, and $\hat{\theta}_{best}$, and their standard errors in parenthesis

n	100	200	500	1000	2000	5000
$2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{L}_x(\theta_0)\}$	4.229	4.079	4.051	4.039	4.029	4.033
	(0.032)	(0.030)	(0.029)	(0.028)	(0.029)	(0.028)
$2n\{\mathcal{R}_x(\hat{\theta}_y) - \mathcal{L}_x(\theta_0)\}$	7.831	7.412	7.326	7.247	7.211	7.266
	(0.078)	(0.061)	(0.058)	(0.058)	(0.058)	(0.058)
$2n\{\mathcal{R}_x(\hat{\theta}_{best}) - \mathcal{L}_x(\theta_0)\}$	5.109	4.741	4.501	4.491	4.479	4.454
	(0.052)	(0.045)	(0.041)	(0.042)	(0.042)	(0.041)

Table 7. *Useless* auxiliary variable (Case 2): estimated risk functions of $\hat{\theta}_b$, $\hat{\theta}_y$, and $\hat{\theta}_{best}$, and their standard errors in parenthesis

n	100	200	500	1000	2000	5000
$2n\{\mathcal{R}_x(\hat{\theta}_b) - \mathcal{L}_x(\theta_0)\}$	105.527	214.659	543.685	1091.105	2182.647	5452.623
	(0.111)	(0.167)	(0.301)	(0.474)	(0.723)	(1.151)
$2n\{\mathcal{R}_x(\hat{\theta}_y) - \mathcal{L}_x(\theta_0)\}$	7.831	7.412	7.326	7.247	7.211	7.266
	(0.078)	(0.061)	(0.058)	(0.058)	(0.058)	(0.058)
$2n\{\mathcal{R}_x(\hat{\theta}_{best}) - \mathcal{L}_x(\theta_0)\}$	22.064	11.555	7.375	7.247	7.211	7.266
	(0.358)	(0.304)	(0.079)	(0.058)	(0.058)	(0.058)

7. Experiments with Real Datasets

We show an example of auxiliary variable selection using Wine Data Set available at UCI Machine Learning Repository [21], which consists of 1 categorical variable (3 categories) and 13 continuous variables, denoted as $V_1, \ldots, V_{13}$. For simplicity, we only use the first two categories and regard them as a latent variable $Z \in \{0, 1\}$; the experiment results were similar to the other combinations. The sample size is then $n = 130$ and all variables except for Z are standardized. We set one of the 13 continuous variables as the observed primary variable Y, and set the rest of 12 variables as auxiliary variables $A_1, \ldots, A_{12}$. For example, if Y is V_1, then $A_1, \ldots, A_{12}$ are $V_2, \ldots, V_{13}$. The dataset is now $\{(y_i, z_i, a_{i,1}, \ldots, a_{i,12})\}_{i=1}^n$, which is randomly divided into the training set with sample size $n_{tr} = 86$ (z_i is not used) and the test set with sample size $n_{te} = 44$ ($a_{i,1}, \ldots, a_{i,12}$ are not used).

In the experiment, we compute $\mathrm{AIC}_{x;b_\ell}$ for $B_\ell = (Y, A_\ell)$, $\ell = 1, \ldots, 12$, and $\mathrm{AIC}_{x;y}$ for Y from the training dataset using the model (4). We select $\hat{\theta}_{best}$ from $\hat{\theta}_{b_1}, \ldots, \hat{\theta}_{b_{12}}$ and $\hat{\theta}_y$ by finding the minimum of the 13 AIC values. Thus we are selecting one of the auxiliary variables $A_1, \ldots, A_{12}$ or not selecting any of them. It is possible to select a combination of the auxiliary variables, but we did not attempt such an experiment. For measuring the generalization error, we compute $\mathcal{L}_x(\hat{\theta}_y) - \mathcal{L}_x(\hat{\theta}_{best})$ from the test set as

$$\mathcal{L}_x(\hat{\theta}_y) - \mathcal{L}_x(\hat{\theta}_{best}) \approx -\frac{1}{n_{te}} \sum_{i \in \mathcal{D}^{te}} \{\log p_x(y_i, z_i; \hat{\theta}_y) - \log p_x(y_i, z_i; \hat{\theta}_{best})\},$$

where $\mathcal{D}^{te} \subset \{1, \ldots, n\}$ represents the test set. The assignment problem of $\mathcal{L}_x(\cdot)$ mentioned in Section 6 is avoided by a similar manner.

For each case of $Y = V_\ell$, $\ell = 1, \ldots, 13$, the above experiment was repeated 100 times, and the experiment average of the generalization error was computed. The result is shown in Table 8. A positive value indicates that $\hat{\theta}_{best}$ performed better than $\hat{\theta}_y$. We observe that $\hat{\theta}_{best}$ is better than or almost the same as $\hat{\theta}_y$ for all cases $\ell = 1, \ldots, 13$, suggesting that AIC works well to select a useful auxiliary variable.

Table 8. Experiment average of $n_{te}\{\mathcal{L}(\hat{\theta}_y) - \mathcal{L}_x(\hat{\theta}_{best})\}$ for each case of $Y = V_\ell$, $\ell = 1, \ldots, 13$. Standard errors are in parenthesis.

Y	V_1	V_2	V_3	V_4	V_5	V_6	V_7
$n_{te}\{\mathcal{L}_x(\hat{\theta}_y) - \mathcal{L}_x(\hat{\theta}_{best})\}$	0.13	−0.14	89.71	46.24	−1.76	3.34	76.54
	(0.08)	(0.12)	(3.82)	(4.17)	(2.52)	(1.34)	(6.09)
Y	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	
$n_{te}\{\mathcal{L}_x(\hat{\theta}_y) - \mathcal{L}_x(\hat{\theta}_{best})\}$	13.91	39.45	1.72	111.24	15.48	0.23	
	(2.21)	(3.12)	(0.29)	(8.46)	(2.11)	(0.09)	

8. Conclusions

We often encounter a dataset composed of various variables. If only some of the variables are of interest, then the rest of the variables can be interpreted as auxiliary variables. Auxiliary variables

may be able to improve estimation accuracy of unknown parameters but they could also be harmful. Hence, it is important to select useful auxiliary variables.

In this paper, we focused on exploiting auxiliary variables in incomplete data analysis. The usefulness of auxiliary variables is measured by a risk function based on the KL divergence for complete data. We derived an information criterion which is an asymptotically unbiased estimator of the risk function except for a constant term. Moreover, we extended a result of Stone [14] to our setting and proved asymptotic equivalence between a variant of LOOCV and the proposed criteria. Since LOOCV requires an additional condition for its justification, the proposed criteria are preferable to LOOCV.

This study assumes that variables are different between training set and test set. There are other settings, such as covariate shift [17] and transfer learning [22], where distributions are different between the training set and test set. It will be possible to combine these settings to construct a generalized framework. It is also possible to extend our study for taking account of a missing mechanism. We will leave these extensions as future works.

Author Contributions: Conceptualization, S.I. and H.S.; methodology, S.I. and H.S.; software, S.I.; validation, S.I. and H.S.; formal analysis, S.I. and H.S.; writing-original draft preparation S.I. and H.S.; visualization, S.I. and H.S.

Funding: This research was funded in part by JSPS KAKENHI Grant (17K12650 to S.I., 16H02789 to H.S.) and by "Funds for the Development of Human Resources in Science and Technology" under MEXT, through the "Home for Innovative Researchers and Academic Knowledge Users (HIRAKU)" consortium (to S.I.).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Appendix A.1. Proof of Lemma 1

Proof. Taylor expansion of $\mathcal{L}_x(\theta)$ around $\theta = \bar{\theta}$, by formally taking it as a function of β, gives

$$\mathcal{L}_x(\hat{\theta}_b) = \mathcal{L}_x(\bar{\theta}) + \nabla^\top \mathcal{L}_x(\bar{\theta})(\hat{\beta}_b - \bar{\beta}) + \frac{1}{2}\mathrm{tr}\{I_x(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})^\top\} + o_p(n^{-1}),$$

where $\nabla^2 \mathcal{L}_x(\bar{\theta}) = I_x$ is used above. By taking the expectation of both sides,

$$\begin{aligned} E[\mathcal{L}_x(\hat{\theta}_b)] &= \mathcal{L}_x(\bar{\theta}) + \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] + \frac{1}{2}\mathrm{tr}\{I_x E[(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})^\top]\} + o(n^{-1}) \\ &= \mathcal{L}_x(\bar{\theta}) + \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] + \frac{1}{2n}\mathrm{tr}(I_x I_b^{-1} J_b I_b^{-1}) + o(n^{-1}), \end{aligned}$$

where the asymptotic variance of $\hat{\beta}_b$ in (5) is given as

$$nE[(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})^\top] = I_b^{-1} J_b I_b^{-1} + o(1). \tag{A1}$$

□

Appendix A.2. Proof of Lemma 2

Proof. Taylor expansion of $\ell_y(\theta)$ around $\theta = \bar{\theta}$, by formally taking it as a function of β, gives

$$\ell_y(\hat{\theta}_b) = \ell_y(\bar{\theta}) + \nabla^\top \ell_y(\bar{\theta})(\hat{\beta}_b - \bar{\beta}) - \frac{1}{2}\mathrm{tr}\{I_y(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})^\top\} + o_p(n^{-1}),$$

where $\nabla^2 \ell_y(\bar{\theta}) = -I_y + o_p(1)$ is used above. By taking the expectation of both sides,

$$\begin{aligned} E[\ell_y(\hat{\theta}_b)] &= E[\ell_y(\bar{\theta})] + E[\nabla^\top \ell_y(\bar{\theta})(\hat{\beta}_b - \bar{\beta})] - \frac{1}{2}E[\mathrm{tr}\{I_y(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})^\top\}] + o(n^{-1}) \\ &= E[\ell_y(\bar{\theta})] + E[\nabla^\top \ell_y(\bar{\theta})(\hat{\beta}_b - \bar{\beta})] - \frac{1}{2n}\mathrm{tr}(I_y I_b^{-1} J_b I_b^{-1}) + o(n^{-1}). \end{aligned} \tag{A2}$$

In the last expression, we used (A1) for the asymptotic variance of $\hat{\beta}_b$. For working on the second term in (A2), we first derive an expression of $\hat{\beta}_b - \bar{\beta}$. Taylor expansion of the score function $\nabla \ell_b(\beta)$ around $\beta = \bar{\beta}$ gives

$$\begin{aligned}\nabla \ell_b(\hat{\beta}_b) &= \nabla \ell_b(\bar{\beta}) + \nabla^2 \ell_b(\bar{\beta})(\hat{\beta}_b - \bar{\beta}) + o_p(n^{-1/2}) \\ &= \nabla \ell_b(\bar{\beta}) - I_b(\hat{\beta}_b - \bar{\beta}) + o_p(n^{-1/2}),\end{aligned}$$

where $\nabla^2 \ell_b(\bar{\beta}) = -I_b + o_p(1)$ is used above. By noticing $\nabla \ell_b(\hat{\beta}_b) = 0$, we thus obtain

$$\hat{\beta}_b - \bar{\beta} = I_b^{-1} \nabla \ell_b(\bar{\beta}) + o_p(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} I_b^{-1} \nabla \log p_b(b_i; \bar{\beta}) + o_p(n^{-1/2}), \tag{A3}$$

where $E[\nabla \ell_b(\bar{\beta})] = 0$ and each term in the summation has mean zero, because $E[\nabla \log p_b(b; \bar{\beta})] = \nabla E[\log p_b(b; \bar{\beta})] = 0$. Now we are back to the the second term in (A2). Using (A3), we have

$$\begin{aligned}\nabla^\top \ell_y(\bar{\theta})(\hat{\beta}_b - \bar{\beta}) &= E[\nabla^\top \ell_y(\bar{\theta})](\hat{\beta}_b - \bar{\beta}) + \{\nabla^\top \ell_y(\bar{\theta}) - E[\nabla^\top \ell_y(\bar{\theta})]\}(\hat{\beta}_b - \bar{\beta}) \\ &= E[\nabla^\top \ell_y(\bar{\theta})](\hat{\beta}_b - \bar{\beta}) + \{\nabla^\top \ell_y(\bar{\theta}) - E[\nabla^\top \ell_y(\bar{\theta})]\} I_b^{-1} \nabla \ell_b(\bar{\beta}) + o_p(n^{-1}).\end{aligned} \tag{A4}$$

By noting $E[\nabla \ell_b(\bar{\beta})] = 0$, the expectation of the second term in (A4) is

$$\begin{aligned}E[\{\nabla^\top \ell_y(\bar{\theta}) - E[\nabla^\top \ell_y(\bar{\theta})]\} I_b^{-1} \nabla \ell_b(\bar{\beta})] &= E[\nabla^\top \ell_y(\bar{\theta}) I_b^{-1} \nabla \ell_b(\bar{\beta})] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} E[\nabla^\top \log p_y(y_i; \bar{\theta}) I_b^{-1} \nabla \log p_b(b_j; \bar{\beta})] \\ &= \frac{1}{n} E[\nabla^\top \log p_y(y; \bar{\theta}) I_b^{-1} \nabla \log p_b(b; \bar{\beta})] \\ &= \frac{1}{n} \mathrm{tr}\{I_b^{-1} E[\nabla \log p_b(b; \bar{\beta}) \nabla^\top \log p_y(y; \bar{\theta})]\} \\ &= \frac{1}{n} \mathrm{tr}(I_b^{-1} K_{b,y}).\end{aligned} \tag{A5}$$

Combining (A4) and (A5), we have

$$E[\nabla^\top \ell_y(\bar{\theta})(\hat{\beta}_b - \bar{\beta})] = E[\nabla^\top \ell_y(\bar{\theta})] E[\hat{\beta}_b - \bar{\beta}] + \frac{1}{n} \mathrm{tr}(I_b^{-1} K_{b,y}) + o(n^{-1}). \tag{A6}$$

We next show that $E[\nabla^\top \ell_y(\bar{\theta})] = -\nabla^\top \mathcal{L}_x(\bar{\theta})$. Let us recall that we have assumed $q_{z|y}(z|y) = p_{z|y}(z|y; \bar{\theta})$ in (6), which leads to

$$\begin{aligned}E[\nabla \log p_{z|y}(z|y; \bar{\theta})] &= \int q_y(y) \int p_{z|y}(z|y; \bar{\theta}) \nabla \log p_{z|y}(z|y; \bar{\theta}) \, dz dy \\ &= \int q_y(y) \int \nabla p_{z|y}(z|y; \bar{\theta}) \, dz dy \\ &= \int q_y(y) \nabla \int p_{z|y}(z|y; \bar{\theta}) \, dz dy = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}-\nabla \mathcal{L}_x(\bar{\theta}) &= \nabla E[\log p_x(x; \bar{\theta})] \\ &= E[\nabla \log p_x(x; \bar{\theta})] \\ &= E[\nabla \log p_y(y; \bar{\theta})] + E[\nabla \log p_{z|y}(z|y; \bar{\theta})] \\ &= E[\nabla \ell_y(\bar{\theta})].\end{aligned}$$

Substituting this and (A6) into the second term in (A2), we have

$$\begin{aligned} E[\ell_y(\hat{\theta}_b)] = E[\ell_y(\bar{\theta})] - \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] \\ + \frac{1}{n}\mathrm{tr}(I_b^{-1} K_{b,y}) - \frac{1}{2n}\mathrm{tr}(I_y I_b^{-1} J_b I_b^{-1}) + o(n^{-1}). \end{aligned} \tag{A7}$$

The first term on the right hand side in (A7) is

$$\begin{aligned} E[\ell_y(\bar{\theta})] &= E[\log p_y(y;\bar{\theta})] \\ &= E[\log p_x(x;\bar{\theta})] - E[\log p_{z|y}(z|y;\bar{\theta})] \\ &= -\mathcal{L}_x(\bar{\theta}) - C(q_x), \end{aligned}$$

where (6) is used again in the last term. Finally (A7) is rewritten as

$$\begin{aligned} E[\ell_y(\hat{\theta}_b)] = -\mathcal{L}_x(\bar{\theta}) - C(q_x) - \nabla^\top \mathcal{L}_x(\bar{\theta}) E[\hat{\beta}_b - \bar{\beta}] \\ + \frac{1}{n}\mathrm{tr}(I_b^{-1} K_{b,y}) - \frac{1}{2n}\mathrm{tr}(I_y I_b^{-1} J_b I_b^{-1}) + o(n^{-1}). \end{aligned}$$

□

Appendix A.3. Proof of Theorem 2

Proof. First recall that we have assumed that $q_c(c) = p_c(c;\beta_0)$, which also implies the condition (6) as $q_{z|y}(z|y) = p_{z|y}(z|y;\theta_0)$ with $\bar{\beta} = \beta_0$. Thus Theorem 1 holds. Substituting $J_b = I_b$ and $K_{b,y} = I_y$ in the penalty term of (8), we have

$$2\mathrm{tr}(I_b^{-1} K_{b,y}) + \mathrm{tr}(I_{z|y} I_b^{-1} J_b I_b^{-1}) = 2\mathrm{tr}(I_b^{-1} I_y) + \mathrm{tr}((I_x - I_y) I_b^{-1}) = \mathrm{tr}(I_b^{-1} I_y) + \mathrm{tr}(I_x I_b^{-1}),$$

giving the penalty term of (10). Therefore, we only have to show (9). Noting the identity

$$\nabla^2 \log p_b(b;\beta) = \frac{1}{p_b(b;\beta)} \nabla^2 p_b(b;\beta) - \nabla \log p_b(b;\beta) \nabla^\top \log p_b(b;\beta),$$

it follows from $q_b(b) = p_b(b;\beta_0)$ that

$$\begin{aligned} I_b = -E[\nabla^2 \log p_b(b;\beta_0)] &= -\int \nabla^2 p_b(b;\beta_0) db + E[\nabla \log p_b(b;\beta_0) \nabla^\top \log p_b(b;\beta_0)] \\ &= -\nabla^2 \int p_b(b;\beta_0) db + J_b = J_b. \end{aligned}$$

Note that the same result can be obtained from Theorem 3.3 in White [16]. Next we show $K_{b,y} = I_y$. Since $q_{a|y}(a|y) = p_{a|y}(a|y;\beta_0)$,

$$\int q_{a|y}(a|y) \nabla \log p_{a|y}(a|y;\beta_0) da = \int \nabla p_{a|y}(a|y;\beta_0) da = \nabla \int p_{a|y}(a|y;\beta_0) da = 0.$$

Therefore, we have

$$\begin{aligned} K_{b,y} &= E[\nabla \log p_b(b;\beta_0) \nabla^\top \log p_y(y;\theta_0)] \\ &= E[\nabla \log p_y(y;\theta_0) \nabla^\top \log p_y(y;\theta_0)] + E[\nabla \log p_{a|y}(a|y;\beta_0) \nabla^\top \log p_y(y;\theta_0)] \\ &= I_y + \int q_y(y) \left(\int q_{a|y}(a|y) \nabla \log p_{a|y}(a|y;\beta_0) da \right) \nabla^\top \log p_y(y;\theta_0) dy \\ &= I_y. \end{aligned}$$

□

Appendix A.4. Proof of Theorem 3

Proof. It follows from (14) and (15) that

$$\begin{aligned} g(y_i;\hat{\theta}_b^{(-i)}) &= g(y_i;\hat{\theta}_b) + \frac{1}{n}\nabla^\top g(y_i;\tilde{\theta}_b^i)\nabla^2\ell_b(\tilde{\beta}_b^i)^{-1}\nabla\log p_b(b_i;\hat{\beta}_b^{(-i)}) \\ &= g(y_i;\hat{\theta}_b) + \frac{1}{n}\mathrm{tr}\{\nabla^2\ell_b(\tilde{\beta}_b^i)^{-1}\nabla\log p_b(b_i;\hat{\beta}_b^{(-i)})\nabla^\top g(y_i;\tilde{\theta}_b^i)\}. \end{aligned}$$

This and the assumption (16) imply that

$$\begin{aligned} \mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b) &= -\frac{1}{n}\sum_{i=1}^n g(y_i;\hat{\theta}_b) - \frac{1}{n^2}\sum_{i=1}^n \mathrm{tr}\{\nabla^2\ell_b(\tilde{\beta}_b^i)^{-1}\nabla\log p_b(b_i;\hat{\beta}_b^{(-i)})\nabla^\top g(y_i;\tilde{\theta}_b^i)\} \\ &= -\frac{1}{n}\sum_{i=1}^n g(y_i;\hat{\theta}_b) + \frac{1}{n}\mathrm{tr}\{I_b^{-1}E[\nabla\log p_b(\bar{\beta})\nabla^\top g(y;\bar{\theta})]\} + o_p(n^{-1}). \end{aligned}$$

Under the assumption $q_{z|y}(z|y) = p_{z|y}(z|y;\bar{\theta})$,

$$\nabla f(y;\bar{\theta}) = \int q_{z|y}(z|y)\nabla\log p_{z|y}(z|y;\bar{\theta})dz = \int \nabla p_{z|y}(z|y;\bar{\theta})dz = 0. \tag{A8}$$

This yields that

$$E[\nabla\log p_b(\bar{\beta})\nabla^\top g(y;\bar{\theta})] = E[\nabla\log p_b(\bar{\beta})\nabla^\top\log p_y(\bar{\theta})] = K_{b,y}.$$

Hence, by noting $g(y;\theta) = \log p_y(y;\theta) + f(y;\theta)$, it holds that

$$\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b) = -\ell_y(\hat{\theta}_b) - \frac{1}{n}\sum_{i=1}^n f(y_i;\hat{\theta}_b) + \frac{1}{n}\mathrm{tr}(I_b^{-1}K_{b,y}) + o_p(n^{-1}). \tag{A9}$$

For evaluating the second term on the right hand side, we apply Taylor expansion to $n^{-1}\sum_{i=1}^n f(y_i;\theta)$ around $\theta = \bar{\theta}$ by formally taking it as a function of β. By noting (A8), this gives

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n f(y_i;\hat{\theta}_b) &= \frac{1}{n}\sum_{i=1}^n f(y_i;\bar{\theta}) + \frac{1}{2n}\sum_{i=1}^n(\hat{\beta}_b-\bar{\beta})^\top\nabla^2 f(y_i;\bar{\theta})(\hat{\beta}_b-\bar{\beta}) + o_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^n f(y_i;\bar{\theta}) + \frac{1}{2n}\mathrm{tr}\left\{\sum_{i=1}^n\nabla^2 f(y_i;\bar{\theta})(\hat{\beta}_b-\bar{\beta})(\hat{\beta}_b-\bar{\beta})^\top\right\} + o_p(n^{-1}). \end{aligned}$$

It follows from the law of large numbers that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n\nabla^2 f(y_i;\bar{\theta}) &= \frac{1}{n}\sum_{i=1}^n\int q_{z|y}(z|y_i)\nabla^2\log p_{z|y}(z|y_i;\bar{\theta})dz \\ &\xrightarrow{p} E[\nabla^2\log p_{z|y}(z|y;\bar{\theta})] = -I_{z|y}. \end{aligned}$$

Hence, (A1) indicates that

$$\frac{1}{n}\sum_{i=1}^n f(y_i;\hat{\theta}_b) = \frac{1}{n}\sum_{i=1}^n f(y_i;\bar{\theta}) - \frac{1}{2n}\mathrm{tr}(I_{z|y}I_b^{-1}J_bI_b^{-1}) + o_p(n^{-1}). \tag{A10}$$

By substituting (A10) into (A9), we establish that

$$\mathcal{L}_x^{\mathrm{cv}}(\hat{\theta}_b) = -\ell_y(\hat{\theta}_b) + \frac{1}{n}\mathrm{tr}(I_b^{-1}K_{b,y}) + \frac{1}{2n}\mathrm{tr}(I_{z|y}I_b^{-1}J_bI_b^{-1}) - \frac{1}{n}\sum_{i=1}^n f(y_i;\bar{\theta}) + o_p(n^{-1}).$$

Hence, the proof is complete. □

References

1. Breiman, L.; Friedman, J.H. Predicting multivariate responses in multiple linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1997**, *59*, 3–54. [CrossRef]
2. Tibshirani, R.; Hinton, G. Coaching variables for regression and classification. *Stat. Comput.* **1998**, *8*, 25–33. [CrossRef]
3. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
4. Mercatanti, A.; Li, F.; Mealli, F. Improving inference of Gaussian mixtures using auxiliary variables. *Stat. Anal. Data Min.* **2015**, *8*, 34–48. [CrossRef]
5. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
6. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
7. Shibata, R. An optimal selection of regression variables. *Biometrika* **1981**, *68*, 45–54. [CrossRef]
8. Shibata, R. Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Stat. Math.* **1983**, *35*, 415–423. [CrossRef]
9. Takeuchi, K. Distribution of information statistics and criteria for adequacy of models. *Math. Sci.* **1976**, *153*, 12–18. (In Japanese)
10. Shimodaira, H. A new criterion for selecting models from partially observed data. In *Selecting Models from Data*; Cheeseman, P., Oldford, R.W., Eds.; Springer: New York, NY, USA, 1994; pp. 21–29.
11. Cavanaugh, J.E.; Shumway, R.H. An Akaike information criterion for model selection in the presence of incomplete data. *J. Stat. Plan. Inference* **1998**, *67*, 45–65. [CrossRef]
12. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–38. [CrossRef]
13. Shimodaira, H.; Maeda, H. An information criterion for model selection with missing data via complete-data divergence. *Ann. Inst. Stat. Math.* **2018**, *70*, 421–438. [CrossRef]
14. Stone, M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 44–47. [CrossRef]
15. Ibrahim, J.G.; Lipsitz, S.R.; Horton, N. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2001**, *50*, 361–373. [CrossRef]
16. White, H. Maximum likelihood estimation of misspecified models. *Econometrica* **1982**, *50*, 1–25. [CrossRef]
17. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **2000**, *90*, 227–244. [CrossRef]
18. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* **1974**, *36*, 111–147. [CrossRef]
19. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
20. Yanagihara, H.; Tonda, T.; Matsumoto, C. Bias correction of cross-validation criterion based on Kullback–Leibler information under a general condition. *J. Multivar. Anal.* **2006**, *97*, 1965–1975. [CrossRef]
21. Dua, D.; Karra Taniskidou, E. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 31 July 2017.
22. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]