A latent-observed dissimilarity measure

Yasushi Terazono*

Abstract

Quantitatively assessing relationships between latent variables and observed variables is important for understanding and developing generative models and representation learning. In this paper, we propose latent-observed dissimilarity (LOD) to evaluate the dissimilarity between the probabilistic characteristics of latent and observed variables. We also define four essential types of generative models with different independence/conditional independence configurations. Experiments using tractable real-world data show that LOD can effectively capture the differences between models and reflect the capability for higher layer learning. They also show that the conditional independence of latent variables given observed variables contributes to improving the transmission of information and characteristics from lower layers to higher layers.

1 Introduction

Models with latent variables have been proposed and investigated for explaining, understanding, or classifying observed data. If a model is a generative model, observed data are modeled to be as if they were generated by latent variables through parameterized probability distributions. Popular criteria for learning generative models include likelihood or posterior probability, which both evaluate the probability of the given observed data or parameters. Another kind of criteria is mutual information. Mutual information has been used to learn non-linear generative models [14] in which relationships between observed and latent variables are directly evaluated. It has also been used to learn linear encoding (recognition) models [2, 12].

The relationships between observed and latent variables have greater importance in more complex generative models, e.g., deep learning models [6, 9]. In the pre-training of deep belief networks (DBNs), one of the models or techniques of deep learning, posterior samples of latent variables in the lower layer are used as samples of observed variables in the next, higher layer. For successive layer learning to be possible, latent variables should possess properties that enable such learning. It is crucial and fundamental for multiple layer learning theory to assess which observed variable properties are preserved, discarded, or modified

*Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo

in latent variables. For this purpose, it is necessary to have good measures that capture the capability of higher layer learning and to know the configurations of models suitable for higher layer learning. Unfortunately, mutual information is not an adequate measure for this purpose. The maximization of mutual information is known to yield independent latent variables under certain conditions [12], however, if latent variables are independent of each other, successive learning exploiting their correlations becomes impossible.

In this paper, we propose a novel measure to capture the dissimilarity between latent and observed variables in two-layer models. We refer to the proposed measure as latent-observed dissimilarity (LOD). The key idea is to define a "virtual-latent" probability mass function (pmf) over observed variables, using the conditionally expected information of latent variables. This definition provides us with a new pmf for which we can measure the dissimilarity from the original pmf. The dissimilarity between these two pmfs can be regarded as the dissimilarity between the latent and observed variables, since the defined pmf reflects the conditionally expected information of latent samples, while the original pmf reflects the self-information of observed samples. We applied LOD to four essential types of two-layer models: 1) a single-latent-variable model (SL), 2) a multi-latent-variable model whose latent variables are independent of each other (IL), 3) a multi-latent-variable model whose latent variables are conditionally independent given observed variables (CI), and 4) a multi-latent-variable model whose latent variables are independent of each other and conditionally independent given observed variables (ICI). These four types cover the major possible combinations of independence or conditional independence in two-layer models. In our experiments, LOD clearly reflected the difference between these four model types. LOD was also shown to reflect the latent layer's capability for higher layer learning. Our experiments also revealed that the conditional independence of latent variables given observed variables, particularly for CI models, contributes to the improvement of higher layer learning, improving LOD and the mutual information between lower and higher layers.

2 Latent-observed dissimilarity

2.1 Definition of LOD

Let $p_{\mathrm{G}}(X, Y)$ denote the probability mass function (pmf) of a generative model where X denotes observed variables and Y denotes latent variables. When an observation X is received, its self information under a model p_{G} is given as $-\log p_{\mathrm{G}}(X)$. We first define the corresponding expected information for latent variables. Let $f(X)$ denote the expected information of Y given X,

$$f(X) = E_{Y|X}\left[-\log p_{\mathrm{G}}(Y)\right] \tag{1}$$

$$= -\sum_{Y} p_{\mathrm{G}}(Y|X) \log p_{\mathrm{G}}(Y) \tag{2}$$

where $f(X)$ may be said to be the expected surprise of the latent layer given X, while $-\log p(X)$ is the surprise of the observed layer given X.

We then define a pmf $q(X)$ based on $f(X)$. To measure the distance between some pmf and $f(X)$, preprocessing is necessary because the function $f(X)$ is not guaranteed to be a pmf. Based on the fact that $f(X)$ represents the expected information, we define the following pmf,

$$q(X) = \frac{\exp(-f(X))}{C}, \tag{3}$$

where $C = \sum_X \exp(-f(X))$. Let $\tilde{p}(X)$ denote a data distribution. That is, we assume $\frac{1}{T}\sum_{t=1}^{T} g(X(t)) = \sum_X \tilde{p}(X)\, g(X)$ for any function g. Using $q(X)$, we define the dissimilarity between the observed and latent variables for a dataset using KL-divergence,

$$\mathrm{LOD}(X, Y) = D(\tilde{p}(X)\, || q(X)). \tag{4}$$

2.2 Characteristics of LOD

Single variable example. We now study the differences between LOD and mutual information using single variable examples.

The proposed measure, LOD, behaves differently from the mutual information of X and Y. When the joint probability of X and Y is defined by $p_{\mathrm{G}}(X, Y)$, the mutual information $I(X;Y)$ between X and Y is

$$I(X;Y) = \sum_X p_{\mathrm{G}}(X)\, D_Y(p_{\mathrm{G}}(Y|X)\, || p_{\mathrm{G}}(Y)), \tag{5}$$

where D denotes the Kullback-Leibler divergence. A more data-based evaluation is possible if the data distribution $\tilde{p}(X)$ is employed

$$\mathrm{MI}(X, Y) = \textstyle\sum_X \tilde{p}(X) D_Y(p_{\mathrm{G}}(Y|X)\, || p_{\mathrm{G}}(Y)). \tag{6}$$

We also refer to MI as (data-based) mutual information.

Consider the difference between LOD and MI in the simplest case. Consider a model consisting of a single observed variable and a single latent variable. Let $X \in \{x_1, x_2, \ldots, x_6\}$ and $Y \in \{y_1, y_2, \ldots, y_3\}$. Define the probabilities $\tilde{p}(X = x_1), \tilde{p}(X = x_2), \ldots, \tilde{p}(X = x_6)$ as $1/21, 2/21, \ldots, 6/21$, respectively. For simplicity, we assume the mapping from X to Y to be deterministic, so each $p_{\mathrm{G}}(Y|X)$ is either 0 or 1. From among all possible $p_{\mathrm{G}}(Y|X)$ under this assumption, let $p_1(Y, X)$ denote the one that realizes the best LOD, and let $p_2(Y, X)$ denote the one that realizes the best MI. The joint and marginal probabilities of p_1 and p_2 as well as the transformed probabilities $q(X)$ are shown in Table 1. Note that since $p_1(X) = p_2(X) = \tilde{p}(X)$ by assumption, the log likelihood is maximized for both p_1 and p_2, as $\sum_X \tilde{p}(X) \log \tilde{p}(X) = \sum_X \tilde{p}(X) \log p_1(X) = \sum_X \tilde{p}(X) \log p_2(X)$. The scores of LOD and MI are shown in Table 2.

From these results, we can confirm the differences between the minimum LOD model and the minimum MI model. The model $p_1(X, Y)$ that minimizes LOD provides a $q_1(X)$ that has a distribution similar to $\tilde{p}(X)$. The model

Table 1: Joint and marginal probabilities of p_1 and p_2, and transformed probabilities $q(X)$. Top: the best similarity assignment. Bottom: the best mutual information assignment. Note that $a = 1/21$, $b = 1/42$.

$p_1(X,Y)$	x_1	x_2	x_3	x_4	x_5	x_6	$p_1(Y)$
y_1	a	$2a$	0	0	0	0	$3a$
y_2	0	0	$3a$	$4a$	0	0	$7a$
y_3	0	0	0	0	$5a$	$6a$	$11a$
$p_1(X)$	a	$2a$	$3a$	$4a$	$5a$	$6a$	
$q_1(X)$	$3b$	$3b$	$7b$	$7b$	$11b$	$11b$	

$p_2(X,Y)$	x_1	x_2	x_3	x_4	x_5	x_6	$p_2(Y)$
y_1	a	0	0	0	0	$6a$	$7a$
y_2	0	$2a$	0	0	$5a$	0	$7a$
y_3	0	0	$3a$	$4a$	0	0	$7a$
$p_2(X)$	a	$2a$	$3a$	$4a$	$5a$	$6a$	
$q_2(X)$	$7b$	$7b$	$7b$	$7b$	$7b$	$7b$	

Table 2: Scores for p_1 and p_2. LOD: smaller is better. MI: larger is better.

	p_1	p_2
LOD	0.0137	0.129
MI	0.983	1.10

$q_2(X)$ from p_2 that minimizes MI is far from similar, though the fact that MI is minimized in p_2 means that knowing Y in the p_2 model reduces the uncertainty of X more than in the p_1 model.

Sizes of latent/observed space. The proposed dissimilarity measure LOD achieves zero when $-\log \tilde{p}(X) = f(X)$. However, there are other cases where LOD also achieves zero. An illustrative case is the *expanding* case where the size of the latent space in the model is an integer multiplication of the size of the observed space. Let K_A denote the total number of states of observed variables, $K_A = \prod_i K_i$, and let L_A denote the total number of states of latent variables, $L_A = \prod_j L_j$. Suppose $L_A = \alpha K_A$ by an integer $\alpha \geq 1$ and $p_{\mathrm{G}}(X, Y)$ is defined as

$$p_{\mathrm{G}}(Y = l | X = k) = \begin{cases} 1/\alpha, & \text{if } \alpha(k-1) + 1 \leq l \leq \alpha k. \\ 0, & \text{otherwise.} \end{cases}$$

This leads to $p_{\mathrm{G}}(Y = \alpha(k-1) + l) = p(X = k)/\alpha$ for $l = 1, \ldots, \alpha$. In this case, $f(X = k) = \log p_{\mathrm{G}}(X = k) - \log \alpha$, and hence $q(X = k) = p_{\mathrm{G}}(X = k)$, yielding LOD $= 0$. The *shrinking* case, where $K_A = \beta L_A$ by an integer $\beta \geq 1$, is also possible, which we shall omit the explanation. The *expanding/shrinking* cases show an invarance aspect of LOD, which imply the potential advantage of LOD as an optimization criterion for the expansion and reduction of latent representation spaces.

3 Models

In this section, the model types used in our experiments (Section 4) are defined. These model types differ in the independence or conditional independence of their latent variables. By comparing these models in our experiments, we hope to determine which configurations affect the relationships between observed, latent, and higher latent variables.

We consider the unsupervised learning of two-layer generative models with four different configurations of latent and observed variables. One of the layers is of observed, or manifest variables, X, and the other is of latent, or hidden variables, Y. The stochastic variables X and Y are assumed to be finite and discrete, and X and Y may consist of multiple variables. Let N_x be the number of observation variables and N_y be that of latent variables. In addition, let K_i, $i = 1, \ldots, N_x$ be the number of states X_i can take and L_j, $j = 1, \ldots, N_y$ be the number that Y_j can take. We denote a model probability by $p_{\mathrm{G}}(X, Y)$.

The models and satisfied constraints are summarized in Table 3.

Single-label models (SL). The most simple of these models has a single latent variable where each observed variable is conditioned only by the latent variable. A Bayesian network representation of this model is shown in Figure 1a. This model is a type of mixture model and is called a latent class or naive Bayes

Table 3: Models and satisfied constraints.

Constraint	SL	IL	CI	ICI
$p\left(X\vert Y\right) = \prod_i p(X_i\vert Y)$	✓	✓	✓	✓
$p\left(Y\vert X\right) = \prod_j p(Y_j\vert X)$	(✓)	-	✓	✓
$p\left(Y\right) = \prod_j p(Y_j)$	(✓)	✓	-	✓

model in different contexts. The model assumes the conditional independence of X given Y,

$$p_{\mathrm{G}}\left(X|Y\right) = \textstyle\prod_{i=1}^{N_x} p(X_i|Y). \tag{7}$$

The joint probability of the model is

$$p\left(X, Y\right) = \left\{\textstyle\prod_i p_{\mathrm{G}}\left(X_i|Y\right)\right\} p_{\mathrm{G}}\left(Y\right). \tag{8}$$

We define each conditional probability by a conditional probability table,

$$p_{\mathrm{G}}(X_i = x|Y = y) = \Theta_{i,x}^{y}. \tag{9}$$

We call this model the single label model (SL). If L, the number of values Y can take, is sufficiently large, say $L \geq \prod_i K_i$, then the model can realize any $p\left(X\right)$.

Independent label models (IL). There are several ways to add more latent variables to single-label models. One is to add latent variables as indicated in Figure 1b. Though the extension seems simple and straightforward in the graphical representation, the graph indicates the additional assumption that Y is independent, that is, $p_{\mathrm{G}}\left(Y\right) = \prod_j p_{\mathrm{G}}\left(Y_j\right)$. The joint probability is thus

$$p_{\mathrm{G}}(X, Y) = \left\{\textstyle\prod_{i=1}^{N_x} p_{\mathrm{G}}(X_i|Y)\right\}\left\{\textstyle\prod_{j=1}^{N_y} p_{\mathrm{G}}(Y_j)\right\}. \tag{10}$$

Models in this form have been proposed in different contexts, including the probabilistic formulation of the quick medical reference network (QMR-DT) [18, 10], and the partially observed bipartite network (POBN) used for the analysis of transcriptional regulatory networks [1]. These models usually further restrict the form of probability. In this paper, however, we do not restrict $p_{\mathrm{G}}(X_i|Y)$ and $p_{\mathrm{G}}(Y_j)$ to some specific form. We define each conditional probability by a conditional probability table, $p_{\mathrm{G}}(X_i = x|Y_j = y) = \Theta_{i,x}^{j,y}$, and $p(Y_j)$ is defined as $p(Y_j = y) = \Phi_{j,y}$. We call this model the independent label model (IL).

Conditionally independent label models (CI). If independence is not assumed on multiple latent variables, a model takes the form shown in Figure 2a. However, since Z is latent and unsupervised learning is assumed, models of this form are just equivalent to "large" single-label models. A possible constraint other than independence is conditional independence of Z given X.

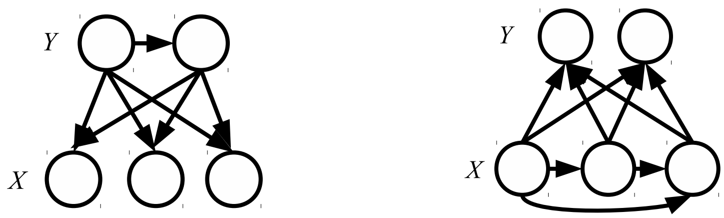
$$p_{\mathrm{G}}(X, Y) = \left\{\textstyle\prod_i p_{\mathrm{G}}(X_i|Y)\right\}p(Y) \tag{11}$$

$$= \left\{\textstyle\prod_j p_{\mathrm{G}}(Y_j|X)\right\}p(X). \tag{12}$$

(a) a single-label model (SL). (b) an independent-label model (IL).

Figure 1: Bayesian network representations of single label and independent label models.

That is, latent variables are conditionally independent given observed variables, while observed variables are conditionally independent given latent variables. These two kinds of conditional independence are impossible to capture in a single Bayesian network representation; two Bayesian networks are necessary to illustrate two-way conditional independence. Figure 2a illustrates conditional independence in a generative model and Figure 2b illustrates it in a recognition model. We call this model a conditionally independent label model (CI).



(a) Generative model, in which observed variables are conditionally independent given latent variables.

(b) Recognition model, in which latent variables are conditionally independent given observed variables.

Figure 2: Two Bayesian network representations of a single probability model.

Joint probabilities satisfying these two-way constraints do exist. An example class is that of the restricted Boltzmann machines (RBMs) [19, 7]. In an RBM, a joint probability of X and Y is defined as $p_{\mathrm{G}}\left(X, Y\right) = \frac{1}{D}\exp\bigl(-\boldsymbol{a}^{\mathrm{T}}X - \boldsymbol{b}^{\mathrm{T}}Y - Y^{\mathrm{T}}\boldsymbol{W}X\bigr)$, where D is the normalizing constant. This is often called a partition function. Constraints (11) and (12) are consistently satisfied by RBMs.

If the generative part of a model is defined in the most general form, that is, if it is parameterized as $p_{\mathrm{G}}\left(X_i = x_i | Y = y\right) = \Theta^{y}_{i,x_i}$ and $p_{\mathrm{G}}\left(Y = y\right) = \Phi_y$, the parameters Θ^{y}_{i,x_i} and Φ_y should be constrained to satisfy the recognition conditional independence (12). It is almost impossible to solve such constraints analytically; however, a numerical, and perhaps approximate, satisfaction of the constraints is possible through the framework of (stochastic) Helmholtz machines (HMs) and the wake-sleep algorithm [8, 4, 3].

Independent and conditionally independent label models If the independence and conditional independence constraints are assumed simultaneously, the model satisfies

$$p_{\mathrm{G}}(X,Y) = \left\{\prod_i p(X_i|Y)\right\}\left\{\prod_j p(Y_j)\right\} \tag{13}$$
$$= \left\{\prod_j p(Y_j|X)\right\}p(X). \tag{14}$$

We call this model the independent and conditionally independent label model (ICI). Learning and (approximate) realization of this class of models are also possible using the wake-sleep algorithm.

If X and Y are continuous and linearly mapped each other, i.e., $X = \boldsymbol{A}Y$ and $Y = \boldsymbol{W}X$ where $\boldsymbol{A}$ and $\boldsymbol{W}$ are matrices, the model represents independent component analysis (ICA) [12]. In ICA, only $\boldsymbol{W}$ is learned using some independence criterion. The relationship between ICA and Helmholtz machines has been investigated in, for example, [21] and [13].

4 Experiments

4.1 Two-layer models

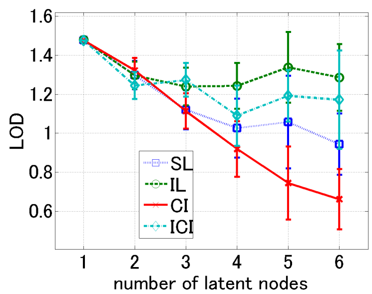
We first considered the two-layer models described in Section 3. The models were trained on patches from images in the MNIST handwritten digits database[11].

Experimental settings. We preprocessed the images by quantizing them to three levels per pixel. From each 28×28 pixel image, a 2×2 pixel image patch was taken from a fixed location. Thus, $N_x = 4$, $K_1 = \cdots = K_4 = 3$ for observed variable X. We used all the training samples in the database, so the number of samples T was 60000. Thus, a patch set consisted of 60000 samples of four observed variables, where each variable is a "trit" (i.e., takes one of three values). Eight non-overlapping locations were employed to yield eight such patch sets. To avoid the local minimum problem, twenty trials were made for each patch set, changing the initial parameters for the EM and the wake-sleep algorithm, and the trial with the best log likelihood $(1/T)\sum_{t=1}^{T}\log p\left(X\left(t\right)\right)$ was chosen for each patch set.

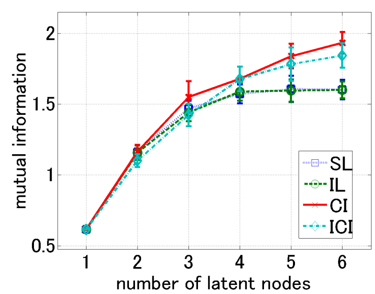
The four kinds of models described in Section 3 were tested. For the IL, CI, and ICI models, L_j $(j = 1, \ldots, N_y)$ were fixed to two, and N_y was varied from one to six. For the SL model, L, the number of values Y could take were $2^1, 2^2, \ldots, 2^6$. The SL and IL models were trained using the EM algorithm, while the CI and ICI models were trained using the wake-sleep algorithm.

After learning, we evaluated the learned models using the following quantities: a) log likelihood $(1/T)\sum_{t=1}^{T}\log p_{\mathrm{G}}\left(X\left(t\right)\right)$, b) data-based mutual information MI (6), c) the proposed dissimilarity measure LOD (4).

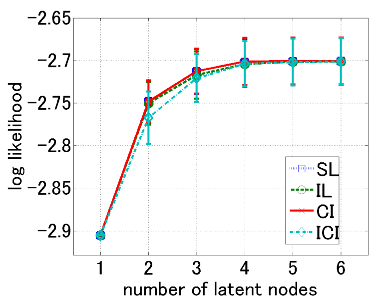
To remove any large deviation caused by different patch sets, an offset removal procedure was performed as follows. Let $V\left(m,s,n\right)$ denote the raw evaluation values, where m denotes the model, s denotes model size, and n denotes patch set number. 1) The average of values of the smallest model in the series was measured over the patch sets, $V_a\left(m\right) = \frac{1}{N}\sum_n V\left(m,1,n\right)$. 2) From the evaluated values of a patch set model, the value of the smallest model was subtracted, $V_r(m,s,n) = V(m,s,n) - V(m,1,n)$. 3) The average V_a was then added back to V_r, $V_q(m,s,n) = V_a(m) + V_r(m,s,n)$. Means and standard deviations were calculated for V_q using $\frac{1}{N}\sum_n V_q(m,s,n)$ and plotted.



(a) LOD for different model configurations.



(b) Mutual information between latent and observed variables.



(c) Log likelihood for different model configurations.

Results: LOD. Figure 3a shows LOD scores for the tested models. CI has a lower LOD than the other models for $N_y \geq 3$. The graphs are, as a whole, decreasing for N_y, but monotonic decrease holds only for CI. For $N_y \geq 4$, four types kept the order of CI < SL < ICI < IL. This suggests that the conditional independence of latent variables given observed variables improves LOD because the essential difference between CI and SL as well as between ICI and IL is the conditional independence. Compared to MI and log likelihood, LOD clearly captured the difference between model types. The difference between LOD and log likelihood (Figure 3c) indicates that the minimization of LOD may lead to a model different from the maximum log likelihood model. The incorporation of LOD into log likelihood as a regularization may also be a future topic of discussion.

Results: Mutual Information. Figure 3b shows the mutual information between the latent and observed variables for the tested models. All models show a monotonic increase of mutual information for N_y. For $N_y \geq 4$, models appear to form two groups: CI and ICI, and SL and IL. The CI-ICI group took larger values than the SL-IL group, and in the CI-ICI group, CI was larger. These phenomena can be explained as follows. First, the conditional independence of the latent variables contributed to a larger MI. Secondly, the independence assumption on latent variables did not affect MI as much as it affected LOD. In fact, recalling the equivalence of ICA and mutual information maximization [12], the independence assumption probably does not disturb the increase of mutual information.

Results: log likelihood. Figure 3c shows the log likelihood of the tested models. All models showed almost equally high likelihood for the same model size; of these, ICI had a slightly lower value. This is because ICI is the most restricted model among these four types and the log likelihood was the objective of the optimization.

For LOD, MI, and log likelihood, CI almost always yielded the best results. This supports the incorporation of conditional independence into models to improve the information transmission from the observed to latent variables without penalizing the log likelihood too much.

4.2 Learning of the higher (third) Layer

Next, we performed learning of SL models on top of the two-layer models learned in 4.1, and evaluated how the characteristics of the lower layers are preserved or reflected in the higher layers.

Learning and evaluation procedures. Let us refer to the two-layer models learned in 4.1 as the "lower" models, and denote their probability as $p_{\mathrm{L}}(X, Y)$. After learning these lower models, a learning process similar to greedy layer-wise learning in deep belief networks [9] was carried out. We applied each model's posterior distribution $p_{\mathrm{L}}(Y|X)$ to the dataset used in 4.1 to derive $\tilde{p}(Y) := \sum_X \tilde{p}(X) p_{\mathrm{L}}(Y|X) = (1/T) \sum_t p_{\mathrm{L}}(Y|X = \boldsymbol{x}(t))$. For the derived $\tilde{p}(Y)$ of each model, we learned a "higher" SL model, $p_{\mathrm{H}}(Y, Z) = \left(\prod_j p_{\mathrm{H}}(Y_j|Z)\right) p_{\mathrm{H}}(Z)$, to maximize $\sum_Y \tilde{p}(Y) \log p_{\mathrm{H}}(Y)$, where Z denotes a set of the third layer latent variables. The learning of $p_{\mathrm{H}}(Y, Z)$ based on $\tilde{p}(Y)$ is essentially equivalent to the learning based on the samples Y from $p_{\mathrm{L}}(Y|X)$ for the dataset; however, as model sizes are assumed to be small and tractability is ensured, we can directly store and calculate $\tilde{p}(Y)$ and do not need the actual samples from $p_{\mathrm{L}}(Y|X)$.

The learning procedure yields the higher two-layer SL models $p_{\mathrm{H}}(Y, Z)$ on top of the lower two-layer model $p_{\mathrm{L}}(X, Y)$. We evaluated the correlations between the lower model score $S(X, Y)$ for $p_{\mathrm{L}}(X, Y)$ and the connected model score $S(X, Z)$ for $p_{\mathrm{C}}(X, Y, Z)$, where the score was either LOD or MI. The probability of a connected model p_{C} is defined by

$$p_{\mathrm{C}}(X, Y, Z) = p_{\mathrm{H}}(Z|Y)\, p_{\mathrm{L}}(Y|X)\, \tilde{p}(X). \tag{15}$$

In (15), the lower and higher models are used as encoders, because here we are focusing on how the higher layers preserve the characteristics of the lower layers and not on the generative properties of the models.

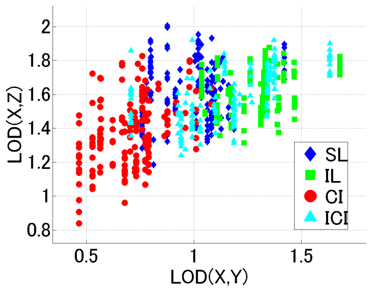
In the four lower model types, higher model learning is impossible for the lower SL models as they are, since SL models only have a single latent variable. To make the learning of higher models possible, the lower SL models were converted into multiple latent variable models as follows. For the models whose number of states of Y was 2^m, a corresponding model with m binary latent variables as in Figure 2a was defined. Let $Y' = \left(Y'_1, \ldots, Y'_m\right)$ denote its latent variables. The states of Y can be mapped to the states of Y' in a bijective (one-to-one and onto) manner. Once such a bijection is determined, the m-latent variables model and the SL model are equivalent as generative models for X. To determine a bijection for each lower model, we first prepared twenty random bijections as the candidates. For each bijection, learning a higher SL model with a single binary latent variable was performed, and the bijection yielding the largest mean log likelihood $\sum_Y \tilde{p}(Y) \log p_{\mathrm{H}}(Y)$ was selected from the twenty candidates.

Experimental settings. The experiment was configured as follows. The number of datasets was eight, as in 4.1, and the lower models with $N_y = 3, 4, 5, 6$ were used. For each lower model, SL models with $K_z = 2, 3, \ldots, 2^{N_y - 2}$ were learned. The number of the models used was thus $8 \times (1 + 3 + 7 + 15) = 208$ for each lower model type (SL, IL, CI, and ICI). Higher SL models were learned using the EM algorithm, which we ran twenty times with different initial values, picking the run that gave the best log likelihood. For the lower and higher models, LOD and mutual information were evaluated using (15) for between X-Y and X-Z.

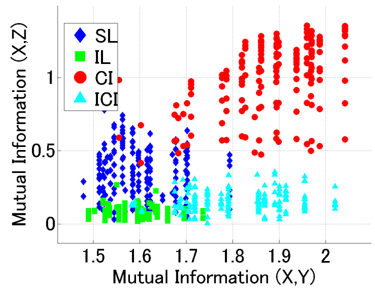
Results. Figure 3d shows the relations between $\mathrm{LOD}(X, Y)$ and $\mathrm{LOD}(X, Z)$. Figure 3e shows the relations between $\mathrm{MI}(X, Y)$ and $\mathrm{MI}(X, Z)$. Their correlations are shown in Table 4.

In Figures 3d and 3e, CI models achieved the lowest X-Z dissimilarity and the highest X-Z mutual information among the four model types. This indicates that latent variables encoded by CI models keep more aspects of the information of the observed variables than the other model types do. From Table 4, the CI models had larger correlation coefficients than those of SL models for both LOD and MI. This relationship was also true for the ICI models and IL models. The capability of Y to provide information to Z was improved by the incorporation of the conditional independence of the latent variables given observed variables.

LOD for (X, Y) and (X, Z) showed significant ($p < 0.05$) correlations for all of the four model types, whereas MI showed significant correlations only for CI and ICI models. These results indicates that, along with dissimilarity itself, LOD also represents how well similarity can be transmitted to the higher layer, whereas MI does not necessarily represent such a capability of transmission.



(d) LOD (X, Y) and LOD (X, Z).



(e) MI (X, Y) and MI (X, Z).

Figure 3: LOD and MI for different model configurations.

Table 4: Correlations between Score (X, Y) and Score (X, Z). Score is either LOD or MI. r means the correlation coefficient, and p means the p-value. See Figures 3d and 3e for the source data.

	LOD		MI	
MODEL	r	p	r	p
(SL)	0.157	0.037	-0.0664	0.381
(IL)	0.373	< 0.001	-0.0586	0.440
(CI)	0.410	< 0.001	0.510	< 0.001
(ICI)	0.602	< 0.001	0.156	0.039

5 Conclusions

We proposed latent-observed dissimilarity (LOD), a dissimilarity measure between latent and observed variables in generative models, to evaluate the relationships between latent and observed variables. LOD compares the self-information of an observation with the expected information of a latent layer given that observation. We numerically evaluated four types of two-layer models (SL, IL, CI, and ICI) using log likelihood, mutual information, and LOD. The results suggested an advantage of using LOD as a measure for multi-layer learning; the LOD between observed and latent variables had significant correlation with the LOD between observed and higher layer latent variables for all four types of models, while mutual information had significant correlation only for CI models. The results also suggested the conditional independence of latent variables given observed variables facilitates the transmission of a layer's characteristics to the higher layers. This fact sheds new light on the advantages of conditional independence, of which usually only its computational advantage is emphasized.

Acknowledgement

This work was supported by MEXT KAKENHI Grant Number 23240019.

References

[1] Angel Alvarez and Peter J. Woolf. Partially observed bipartite network analysis to identify predictive connections in transcriptional regulatory networks. *BMC Systems Biology*, 2011.

[2] Pierre Baldi and Kurt Hornik. Learning in linear neural networks: a survey. *IEEE Transactions on neural networks*, 6(4):837–858, 7 1995.

[3] Peter Dayan and Geoffrey E. Hinton. Varieties of helmholtz machine. *Neural Networks*, 9(8):1385–1403, 1996.

[4] Peter Dayan, Geoffrey E. Hinton, Radford M. Neal, and Richard S. Zemel. The helmholtz machine. *Neural Computation*, 7:889–904, 1995.

[5] Finale Doshi-Velez and Zoubin Ghahramani. Correlated non-parametric latent feature models. *UAI 2009*, 2009.

[6] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimension of data with neural networks. *Science*, 313:504–507, 2006.

[7] Geoffery E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[8] Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Neal Radford M. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

[9] Geoffrey E. Hinton, S. Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[10] Tommi S. Jaakkola and Michael I. Jordan. Variational probabilistic inference and the qmr-dt network. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

[11] Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The mnist database. *http://yann.lecun.com/exdb/mnist/*.

[12] D. Obradovic and G. Deco. Information maximization and independent component analysis: Is there a difference? *Neural Computation*, 10:2085–2101, 1998.

[13] Masashi Ohata, Toshiharu Mukai, and Kiyotoshi Matsuoka. Independent component analysis on the basis of helmholtz machine. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003.

[14] Nicolas Pinchaud. Information theoretic learning of robust deep representations. *NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] Novi Quadrianto, Viktoriaa Sharmanska, David. KnKnowles, and Zoubin Ghahramani. The supervised ibp: Neighbourhood preserving infinite latent feature models. *Uncertainty in Artificial Intelligence UAI 2013*, 2013.

[16] Ruslan Salakhutdinov and Geoffrey Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. *AI and Statistics (AISTATS) 2007*, 2007.

[17] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 7 2007.

[18] Michael Shwe and Gregory F. Cooper. An empirical analysis of likelihood-weighting simulation on a large, multiply-connected belief network. *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, 1990.

[19] Paul Smolensky. *Parallel Distributed ProProcess: Explorations in the Microstructure of Cognition Volume 1*, volume 1. MIT Press, 1986.

[20] Lucas Theis, Sebastian Gerwinn, Fabian Sinz, and Matthias Bethge. In all likelihood, deep belief is not enough. *Journal of Machine Learning Research*, 12:3071–3096, 2011.

[21] Lei Xu. Bayesian kullback ying-yang dependence reduction theory. *Neurocomputing*, 22(1-3):81–111, 11 1998.