A generalisation of independence in statistical models for categorical distribution

Yu Fujimoto*
Aoyama Gakuin University,
5-10-1 Fuchinobe, Chuo, Sagamihara, Kanagawa 252-5258, Japan
E-mail: yu.fujimoto@it.aoyama.ac.jp
*Corresponding author

Noboru Murata
Waseda University,
3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, Japan
E-mail: noboru.murata@eb.waseda.ac.jp

Abstract: In this paper, generalised statistical independence in statistical models for categorical distributions is proposed from the viewpoint of generalised multiplication characterised by a monotonically increasing function and its inverse function, and it is implemented in naive Bayes models. This paper also proposes an idea of their estimation method which directly uses empirical marginal distributions to retain simplicity of calculation. This method is interpreted as an optimisation of a rough approximation of the Bregman divergence so that it is expected to have a kind of robust property. Effectiveness of proposed models is shown by numerical experiments on some benchmark datasets.

Keywords: independent model; naive Bayes model; generalised independence; copula; Bregman divergence.


Biographical notes: Yu Fujimoto received his PhD in Engineering in 2007 at Waseda University, Tokyo, Japan. Currently, he is an Assistant Professor at Aoyama Gakuin University, Kanagawa, Japan. His research interests are focused on statistical model estimation and evaluation.

Noboru Murata received his DrEng degree from the University of Tokyo, Japan in 1992. He is a Professor at Waseda University, Japan. His research interest has been on learning systems, particularly focusing on the dynamics and statistical properties. He is currently working on machine learning and statistical data analysis.

This paper is a revised and expanded version of a paper presented at IDEAL2010, the 11th International Conference on Intelligent Data Engineering and Automated Learning which took place at the University of the West of Scotland, September 1st to 3rd 2010.
1 Introduction

Statistical models based on some kind of independence, such as naive Bayes (NB) models, Bayesian networks (Pearl, 1988; Jensen, 2001) or aspect models (Hofmann, 2001), are broadly used in various situations; the assumption of independence is attractive in modelling relation between categorical variables with a lot of categories because the composed independence model may have a significantly smaller number of parameters than the model denoting dependence. Technically, the assumption of independence in these models should be introduced by analysing a dataset. However, in practical scenes, the models are casually used without rigorous analysis; e.g., in classification problems, it is known that the NB model shows good performance even if the assumption is violated (Domingos and Pazzani, 1997). In this paper, we introduce a generalisation of independence and propose an extension of statistical models based on statistical independence to express weak special dependence with the small number of parameters.

In the statistical inference, we naturally use arithmetic operators, such as multiplication or division, for probability values. For instance, statistical independence can be defined with multiplication of marginal probabilities. We can generalise these operators with an appropriate monotonically increasing function \( u(\cdot) \) and its inverse function \( \xi(\cdot) \). For example, multiplication between two positive values \( a \) and \( b \) are generalised as follows.

\[
\begin{align*}
\text{generalisation} & \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \ quad
KL divergence that is a special case of the Bregman divergence, but sometimes results in poor estimation, especially when the sample set is very small or includes outliers. Some divergences belonging to the Bregman class show robustness against small sample sets and outliers, and with such divergences, statistical algorithms like the boosting or the EM algorithm are generalised and improved (Murata et al., 2004; Takenouchi et al., 2008; Fujimoto and Murata, 2007).

Let $X = \{X^1, \ldots, X^M\}$ be a set of $M$ categorical random variables where $X^m$ has a domain $X^m = \{x^m_i\}_{i=1}^I$, and $p_X(x)$ be a joint probability of $x \in X$ where $X$ indicates a set of joint events

$$X = \{(x^1, \ldots, x^M) \mid x^m \in X^m \ (m = 1, \ldots, M)\}.$$ 

Then, a space of discrete joint probability distributions is defined as follows,

$$P_X = \left\{p_X \mid X \to \mathbb{R}_+, \sum_{x \in X} p_X(x) = 1 \right\}.$$ 

We also consider a space of positive finite measures over $X$,

$$F_X = \left\{f_X \mid X \to \mathbb{R}_+, \sum_{x \in X} f_X(x) < \infty \right\}.$$ 

When the context is clear, $p_X \in P_X$ and $f_X \in F_X$ are denoted as $p$ and $f$ for simplicity. The Bregman divergence between two functions in $F_X$ is defined as follows.

**Definition 1 [Bregman divergence (Murata et al., 2004)]:** Let $U(\cdot)$ be a strictly convex and differentiable function on $\mathbb{R}$, and let us denote its derivative by $u(\cdot) = U'(\cdot)$ and the inverse of $u(\cdot)$ by $\xi(\cdot) = u^{-1}(\cdot)$, respectively. The Bregman divergence between two functions $f, g \in F_X$ is defined by

$$D_U(f, g) = \sum_{x \in X} \left\{U(\xi(g(x))) - U(\xi(f(x))) - f(x)(\xi(g(x)) - \xi(f(x))) \right\}.$$ 

The definition of the Bregman divergence can be naturally extended to the case where $x$ takes continuous values, however, for the sake of simplicity, this paper only deals with the case where $x$ is discrete.

The Bregman divergences show various behaviours depending on functions $U$ (or $u$, or $\xi$). Particularly, functions with one parameter $\pi$ are frequently used. Typical examples of functions $u$ and $\xi$, appeared in Fujisawa and Eguchi (2005) and Takenouchi et al. (2008), are listed in Table 1. Note that, the KL divergence between two distributions $p, q \in P_X$, given as

$$D_{KL}(p, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)},$$

is included in $D_U$ as a special case and is broadly used for the estimation and the evaluation of statistical models.
Table 1  Examples of functions $u$ and $\xi$

<table>
<thead>
<tr>
<th></th>
<th>$u(z)$</th>
<th>$\text{dom}(u)$</th>
<th>$\text{range}(u)$</th>
<th>$\xi(z) = u^{-1}(z)$</th>
<th>$\text{dom}(\xi)$</th>
<th>$\text{range}(\xi)$</th>
<th>$\text{dom}(\pi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL</td>
<td>$\exp(z)$</td>
<td>$(-\infty, \infty)$</td>
<td>$(0, \infty)$</td>
<td>$\log(z)$</td>
<td>$(0, \infty)$</td>
<td>$(-\infty, \infty)$</td>
<td>$-\infty$</td>
</tr>
<tr>
<td>Ex.1</td>
<td>$(\pi z + 1)^{\frac{1}{\pi}}$</td>
<td>$(-\frac{1}{\pi}, \pi)$</td>
<td>$[0, \infty)$</td>
<td>$\frac{z - 1}{\pi}$</td>
<td>$(0, \infty)$</td>
<td>$(-\frac{1}{\pi}, \pi)$</td>
<td>$(0, \infty)$</td>
</tr>
<tr>
<td>Ex.2</td>
<td>$\exp(z) + \pi$</td>
<td>$(-\infty, \alpha)$</td>
<td>$(\pi, \infty)$</td>
<td>$\log(z - \pi)$</td>
<td>$(\pi, \infty)$</td>
<td>$(-\infty, \infty)$</td>
<td>$(-\infty, \infty)$</td>
</tr>
</tbody>
</table>

Here, the following two representations for $f \in \mathcal{F}_{\mathcal{X}}$ play important roles to understand properties of the Bregman divergence (Murata et al., 2004), given as

$m$-representation: $f$,

$u$-representation: $\xi(f)$.

With these representations, a function $f \in \mathcal{F}_{\mathcal{X}}$ corresponds to a point $p \in \mathcal{P}_{\mathcal{X}}$ by an appropriate projection from $\mathcal{F}_{\mathcal{X}}$ to $\mathcal{P}_{\mathcal{X}}$. Following two examples are natural projections from $\mathcal{F}_{\mathcal{X}}$ to $\mathcal{P}_{\mathcal{X}}$ associated with $m$- and $u$-representations, which are called $m$- and $u$-normalisations respectively in this paper,

$m$-normalisation: $T_m(f(x)) = \frac{f(x)}{c_m},$ (1)

$u$-normalisation: $T_u(f(x)) = u(\xi(f(x)) - c_u),$ (2)

where $T_m(\cdot)$ and $T_u(\cdot)$ are normalisation operators and $c_m$ and $c_u$ are normalising constants to hold $\sum_{x \in \mathcal{X}} T_m(f(x)) = 1$ and $\sum_{x \in \mathcal{X}} T_u(f(x)) = 1$ respectively.

3 Generalised arithmetic operators for probabilities

In this section, we introduce generalised arithmetic operators derived from $u$- and $\xi$-functions.

3.1 Generalised arithmetic operators

With the definition of $m$-representation, addition of two functions $f, g \in \mathcal{F}_{\mathcal{X}}$ is calculated by conventional addition in $m$-representation of functions. In the same way, multiplication of two functions is regarded as addition in $u$-representation based on the KL divergence, that is given by

$$ f \cdot g = \exp(\log(f) + \log(g)). $$

From this fact, the basic arithmetic rules, i.e., multiplication and division can be generalised based on the Bregman divergence.

**Definition 2 (U-multiplication and U-division):** Generalised multiplication and division of $f, g \in \mathcal{F}_{\mathcal{X}}$ are given as follows,

$$ f \otimes g = u(\xi(f) + \xi(g)),$$ (3)

$$ f \oslash g = u(\xi(f) - \xi(g)),$$ (4)

where $\otimes$ and $\oslash$ are multiplication and division operators based on the function $U$.

In this paper, the generalised operators $\otimes$ and $\oslash$ are called $U$-multiplication and $U$-division.
3.2 Generalised product rule for probabilities

Given a joint probability distribution \( p_X \in \mathcal{P}_X \), its marginal probability distribution \( p_{X^m} \in \mathcal{P}_{X^m} \) is given by

\[
p_{X^m}(x^m) = \sum_{i \neq m} \sum_{x \in X^i} p_X(x^1, \ldots, x^M).
\] (5)

Let \( \hat{X}^m \) be a set of joint events of \( M - 1 \) variables other than \( X^m \). With equation (4), a conditional function space of \( \hat{X}^m \) given \( x^m \in X^m \), written as \( \mathcal{F}_{\hat{X}^m|X^m} \), is defined by \( p_X \in \mathcal{P}_X \) and \( p_{X^m} \in \mathcal{P}_{X^m} \) as follows,

\[
\mathcal{F}_{\hat{X}^m|X^m} = \{ f_{\hat{X}^m|X^m} | f_{\hat{X}^m|X^m}(\hat{x}^m|x^m) = p_X(x) \otimes p_{X^m}(x^m) \},
\]

where \( f_{\hat{X}^m|X^m} \) is derived based on the \( U \)-division;

\[
p_{X}(x) \otimes p_{X^m}(x^m) = u(\xi(p_X(x)) - \xi(p_{X^m}(x^m))).
\] (6)

In the case of the KL divergence, equation (6) is given by

\[
\exp(\log(p_X(x)) - \log(p_{X^m}(x^m))) = \frac{p_X(x)}{p_{X^m}(x^m)} = p_{X^m|X^m}(\hat{x}^m|x^m),
\]

and \( \mathcal{F}_{\hat{X}^m|X^m} \) is the conventional conditional probability distribution space.

With definitions given by equations (3), (5) and (6), the joint probability \( p_X \) has the following relation,

\[
p_X(x) = f_{\hat{X}^m|X^m}(\hat{x}^m|x^m) \otimes p_{X^m}(x^m) = u(\xi(p_X(x)) - \xi(p_{X^m}(x^m))) + \xi(p_{X^m}(x^m)))
\]

Note that, \( \mathcal{F}_{\hat{X}^m|X^m} \) is generally in the function space and should be projected to the distribution space with \( m \)- or \( u \)-normalisations to obtain a probability distribution for statistical inference. For example, we can use \( u \)-normalisation given by equation (2) to obtain the conditional distribution, that is

\[
p_{X^m|X^m}(\hat{x}^m|x^m) = T_u \left( f_{\hat{X}^m|X^m}(\hat{x}^m|x^m) \right) = u(\xi(p_X(x)) - \xi(p_{X^m}(x^m))) - c_u.
\]

In this case, the following natural relation holds,

\[
T_u(p_{X^m|X^m}(\hat{x}^m|x^m) \otimes p_{X^m}(x^m))
\]

\[
= T_u(T_u(f_{\hat{X}^m|X^m}(\hat{x}^m|x^m) \otimes p_{X^m}(x^m))
\]

\[
= u(\xi(p_X(x)) - \xi(p_{X^m}(x^m))) - c_u + \xi(p_{X^m}(x^m)) - c_u'
\]

\[
= u(\xi(p_X(x))) = p_X(x),
\]

where \( c_u' = -c_u \). Note that this natural relation does not hold when we use \( m \)-normalisation given by equation (1), that is,

\[
T_m(T_m(f_{\hat{X}^m|X^m}(\hat{x}^m|x^m) \otimes p_{X^m}(x^m)) \neq p_X(x).
\]
A generalisation of independence in statistical models 177

Therefore, we use \( u \)-normalisation when we need to normalise functions in this paper. With equations (3) to (6), the Bayes’ theorem,

\[
p_{X^m|X^m}(\bar{x}^m|x^m) = \frac{p_{X^m|X^m}(x^m|\bar{x}^m)p_{X^m}(\bar{x}^m)}{\sum_{x^m' \in X^m} p_{X^m|X^m}(x^m'|\bar{x}^m)p_{X^m}(\bar{x}^m)},
\]
is naturally generalised as follows.

Proposition 1 (generalised Bayes’ theorem): The conditional function \( f_{X^m|X^m} \) based on the \( U \)-operators is given by

\[
f_{X^m|X^m}(\bar{x}^m|x^m) = f_{X^m|X^m}(x^m|\bar{x}^m) \otimes \left( \sum_{x^m' \in X^m} f_{X^m|X^m}(x^m'|\bar{x}^m) \otimes p_{X^m}(\bar{x}^m) \right).
\]

With marginal probability distributions, statistical independence is defined as follows.

Definition 3 (independence): Let \( p_X \) be the joint probability distribution defined with a product of marginal probability distributions \( p_{X^1}, \ldots, p_{X^M} \) as

\[
p_X(x; p_{X^1}, \ldots, p_{X^M}) = \prod_{m=1}^{M} p_{X^m}(x^m)
\]

\[
= \exp\left( \sum_{m=1}^{M} \log(p_{X^m}(x^m)) \right) \quad (\forall x \in X).
\]

Variables \( X^1, \ldots, X^M \) are mutually independent if their joint probability distributions \( p_X \) has the following property,

\[
p_X(x) = p_X(x) \quad (\forall x \in X).
\]

Equations (7) and (8) indicate that the sum of logarithmic marginal probabilities in the function \( \exp(\cdot) \) defines statistical independence. Lastly, by using \( U \)-multiplication, we can generalise Definition 3, as follows.

Definition 4 (U-independence): Let \( p_X^\otimes \) be the joint probability distribution defined by using functions \( u \) and \( \xi \) as
\[ p_\otimes(x; p_{X^1}, \ldots, p_{X^M}, u) = T_u \left( p_{X^1}(x^1) \otimes p_{X^2}(x^2) \otimes \cdots \otimes p_{X^M}(x^M) \right) = T_u \left( \bigotimes_{m=1}^{M} p_{X^m}(x^m) \right) = u \left( \sum_{m=1}^{M} \xi(p_{X^m}(x^m)) - c_u \right) \in \mathcal{P}_X \quad (\forall x \in X), \]

where \( \bigotimes_{m=1}^{M} \) is the operator of \( U \)-multiplication for \( m = 1, \ldots, M \). Variables \( X^1, \ldots, X^M \) are called mutually \( U \)-independent if their joint probability distribution \( p_X \) has the following property,

\[ p_X(x) = p_\otimes(x) \quad (\forall x \in X). \]

Note that we assume that range \( \sum_{m=1}^{M} (p_{X^m} - c_u) \subseteq \text{dom}(u) \) and range\( (u) \supseteq \text{range}(p_\otimes) \) hold.

4 Properties of \( U \)-independence

In this section, we focus on \( U \)-independence for expression of probability distributions and discuss some of its properties.

4.1 Expression of weak dependence

Expression of \( U \)-independence by using \( \xi(\cdot) \neq \log(\cdot) \) is usually different from conventional statistical independence. In other words, we can say that \( U \)-independence indicates a kind of weak dependence between random variables.

Figure 1 shows intuitive differences between conventional independence and \( U \)-independence based on Ex.1 in Table 1 (see Figure 2 for shapes of functions \( \xi \) and \( u \)). The graphs in the figure show the \( U \)-independent distributions \( p_{\otimes} = T_u(p_{X^1} \otimes p_{X^2}) \) for \( \pi = \{-1, -0.5, 0\} \) constructed with exactly the same marginals \( p_{X^1} = (0.167, 0.333, 0.5) \) and \( p_{X^2} = (0.1, 0.2, 0.3, 0.4) \). As shown in the figure, changing functions \( u \) and \( \xi \) in \( U \)-multiplication overstates, or understates, probabilities in marginal distributions. Therefore \( U \)-independence is interpreted as an expression of weak special dependence between variables. From another perspective, we can say that \( U \)-independence constructs a kind of copula (Nelsen, 2006) based on probability distributions instead of cumulative distribution functions.

The intuitive interpretation of \( U \)-independence is given as follows. Let \( X \) be a set of two discrete random variables, \( X = \{X^1, X^2\} \). And let \( \mathcal{P}_u^\otimes \) be a set of \( U \)-independent distributions with a fixed \( u \)-function, given as

\[ \mathcal{P}_u^\otimes = \{ p_\otimes(x; p_{X^1}, p_{X^2}, u) \}. \]

The definition indicates that \( p_\otimes \in \mathcal{P}_u^\otimes \) is denoted only with two marginal distributions \( p_{X^1} \in \mathcal{P}_{X^1} \) and \( p_{X^2} \in \mathcal{P}_{X^2} \) as shown in Figure 3(a). As shown in Figure 1, there are various joint distributions given by \( U \)-independence with specific marginals according to the difference of \( u \)-functions. Therefore, a subspace \( \mathcal{P}_u^\otimes \) is intuitively interpreted as a curved surface in \( \mathcal{P}_X \) as shown in Figure 3(b).
A generalisation of independence in statistical models

Figure 1 Examples of \( p_\otimes = T_u(p_{X_1} \otimes p_{X_2}) \) constructed with Ex.1 in Table 1, where \( p_{X_1} = (0.167, 0.333, 0.5) \) and \( p_{X_2} = (0.1, 0.2, 0.3, 0.4) \) (a) \( \pi = -1.0 \) (b) \( \pi = -0.5 \) (c) \( \pi = 0.0 \) (conventional independence) (d) plot of \( p_\times (x) - p_\otimes (x) \) of \( \pi = -1.0 \), which shows difference between (c) and (a); red parts indicate positive values and blue parts indicate negative values (see online version for colours).

Note: The difference indicates that \( p_\otimes (x) \) tends to be larger than \( p_\times (x) \) when \( p_{X_1}(x^1) \) or \( p_{X_2}(x^2) \) is small, and vice versa.

Figure 2 Shapes of \( u \)- and \( \xi \)-functions of Ex.1 in Table 1 (a) \( \xi \)-function (b) \( u \)-function

Figure 3 Intuitive interpretation of \( U \)-independence (a) \( p_\otimes \) is given as a point on a plane \( \mathcal{P}_\otimes \) (b) \( \mathcal{P}_\otimes^U \) is given as a curved surface in \( \mathcal{P}_X \)

4.2 Empirical marginals

Consider a set of two discrete variables \( X = \{X^1, X^2\} \). Let \( \mathcal{P}_\times \) and \( \mathcal{P}_\otimes (\mathcal{U}) \) be sets of independent and \( U \)-independent distributions, given as

\[
\mathcal{P}_\times = \{ p_\times (x; p_{X_1}, p_{X_2}) \} \\
\mathcal{P}_\otimes (\mathcal{U}) = \{ p_\otimes (x; p_{X_1}, p_{X_2}, u) \mid u \in \mathcal{U} \},
\]
where $\mathcal{U}$ is a set of $u$-functions. Let $\tilde{p}_X$ and $\tilde{p}_{X^m}$ be empirical joint and marginal distributions, given as

$$
\tilde{p}_X(x) = \frac{n_x}{\sum_{x' \in X} n_{x'}} \quad \text{and} \quad \tilde{p}_{X^m}(x^m) = \frac{n_{x^m}}{\sum_{x'^m \in X^m} n_{x'^m}},
$$

where $n_x$ is the frequency of the observed event $x \in X$ and $n_{x^m}$ is that of $x^m \in X^m$. The maximum likelihood (ML) estimate of the conventional independent model $q_X \in \mathcal{P}_X$ is given with empirical marginals as follows,

$$
q_X = \underset{q \in \mathcal{P}_X}{\operatorname{argmin}} D_{KL}(\tilde{p}_X, q) = \underset{q \in \mathcal{P}_X}{\operatorname{argmin}} \sum_{x \in X} \tilde{p}_X(x) \log \frac{\tilde{p}_X(x)}{q(x)},
$$

which is the KL divergence between $\tilde{p}_X$ and $q$. Figure 4(a) shows how to interpret the ML estimation of $q_X$. On the other hand, to obtain the ML estimate of the $U$-independent model $q_\otimes \in \mathcal{P}_\otimes(\mathcal{U})$, we need to solve a non-linear optimisation problem, that is

$$
q_\otimes = \underset{q \in \mathcal{P}_\otimes(\mathcal{U})}{\operatorname{argmin}} D_{KL}(\tilde{p}_X, q) = \tilde{u}(\tilde{\xi}(\tilde{p}_{X^1}) + \tilde{\xi}(\tilde{p}_{X^2}) - c_u), \quad (11)
$$

with respect to marginals $\tilde{p}_{X^1}, \tilde{p}_{X^2}$ and the function $\tilde{u}$ for multiplication operator. Figure 4(b) shows an intuitive interpretation of the ML estimation of $q_\otimes$. To avoid this non-linear optimisation problem, we use empirical marginals and only focus on searching a function $u$ from a one parameter family $\mathcal{U} = \{u(z; \pi)\}$ which is given as an example in Table 1; i.e., we find the estimate

$$
\tilde{q}_\otimes = \underset{q \in \mathcal{P}_\otimes(\mathcal{U})}{\operatorname{argmin}} D_{KL}(\tilde{p}_X, q) = \tilde{u}(\tilde{\xi}(\tilde{p}_{X^1}) + \tilde{\xi}(\tilde{p}_{X^2}) - c_u), \quad (13)
$$

where $\tilde{p}_\otimes$ is constructed from empirical marginals $\tilde{p}_{X^1}$ and $\tilde{p}_{X^2}$ as

$$
\tilde{p}_\otimes(\mathcal{U}) = \{\tilde{p}_\otimes(x; u) = T_u(\tilde{p}_{X^1}(x^1) \otimes \tilde{p}_{X^2}(x^2)) \mid u \in \mathcal{U}\}.
$$

The solution of equation (13) with respect to $\tilde{u}$ can be obtained by cross-validation (CV), that is much easier than solving equation (12) with respect to $\tilde{u}$, $\tilde{p}_{X^1}$ and $\tilde{p}_{X^2}$. We call the solution of equation (13) an empirical $U$-independent model. We show an interpretation of the ML estimation of $q_\otimes$ in Figure 4(c).

This approximated inference with $\mathcal{P}_\otimes(\mathcal{U})$ is supported by the following fact. Let us consider a minimisation problem of the Bregman divergence $D_U$ with the same function $U$ as applied to construct $\mathcal{P}_\otimes$ in $U$-multiplication. The minimiser of the Bregman divergence in $\mathcal{P}_\otimes$ is given as follows,

$$
\operatorname{argmin}_{q \in \mathcal{P}_\otimes} D_U(\tilde{p}_X, q) = \operatorname{argmin}_{q \in \mathcal{P}_\otimes} \sum_{x \in X} \{U(\xi(q(x))) - \tilde{p}(x)\xi(q(x))\}. \quad (14)
$$

The right-hand side of equation (14) is roughly approximated as follows,
\[
\sum_{x^1 \in \mathcal{X}^1} \sum_{x^2 \in \mathcal{X}^2} U \left( \xi \left( p \left( x^1 \right) \right) + \xi \left( p \left( x^2 \right) \right) - c_u \right) \\
- \sum_{x^1 \in \mathcal{X}^1} \sum_{x^2 \in \mathcal{X}^2} \tilde{p} \left( x^1, x^2 \right) \left( \xi \left( p \left( x^1 \right) \right) + \xi \left( p \left( x^2 \right) \right) - c_u \right) \\
\equiv \sum_{x^1 \in \mathcal{X}^1} U \left( \xi \left( p \left( x^1 \right) \right) \right) + \sum_{x^2 \in \mathcal{X}^2} U \left( \xi \left( p \left( x^2 \right) \right) \right) \\
- \sum_{x^1 \in \mathcal{X}^1} \sum_{x^2 \in \mathcal{X}^2} \tilde{p} \left( x^1, x^2 \right) \left( \xi \left( p \left( x^1 \right) \right) + \xi \left( p \left( x^2 \right) \right) \right) + C \\
= H_U \left( \tilde{p}_\mathcal{X}, q \right),
\]

where \( C \) is a constant. Therefore, the minimiser of equation (14) is approximated as

\[
\arg\min_{q \in \mathcal{P}_U} D_U \left( \tilde{p}_\mathcal{X}, q \right) \cong \arg\min_{q \in \mathcal{P}_U^*} H_U^* \left( \tilde{p}_\mathcal{X}, q \right) = \tilde{q}_\emptyset,
\]

and it is given by the empirical marginals. This interpretation indicates that \( U \)-independent models inherit some nice properties like robustness from the Bregman divergence.

**Figure 4** Intuitive interpretation of the ML estimation; arrows show the KL divergence from \( \tilde{p}_\mathcal{X} \) to the nearest point in \( \mathcal{P}_\mathcal{X}, \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) and \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) respectively (a) estimation of \( q_\emptyset \); \( q_\emptyset \) is given as a point in \( \mathcal{P}_\emptyset \) denoted with \( \tilde{p}_\mathcal{X}_1 \) and \( \tilde{p}_\mathcal{X}_2 \) (b) estimation of \( q_\emptyset \); \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) is a set of \( \mathcal{P}_\emptyset \) with various \( u(\cdot) \); \( q_\emptyset \) is the nearest point in \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) from \( \tilde{p}_\mathcal{X} \), which is given as a solution of a non-linear minimisation problem (c) estimation of \( q_\emptyset \); \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) is a set of points in \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \) denoted with \( \tilde{p}_\mathcal{X}_1 \) and \( \tilde{p}_\mathcal{X}_2 \); \( q_\emptyset \) is the nearest point in \( \mathcal{P}_\emptyset \left( \mathcal{U} \right) \)

Note that the set \( \mathcal{P}_\emptyset \) defined with any \( u \)-function has the following property. Let us define uniform distributions on respective domains as

\[
\nu_{\mathcal{X}^m} \left( x^m \right) = \frac{1}{I^m} \quad (\forall x^m \in \mathcal{X}^m, m = 1, 2), \quad \nu_{\mathcal{X}} \left( x \right) = \frac{1}{I^1 \times I^2} \quad (\forall x \in \mathcal{X}),
\]

where \( I^1 \) and \( I^2 \) are the number of elements in \( \mathcal{X}^1 \) and \( \mathcal{X}^2 \). Then, the following property holds with any types of \( u \)-function,

\[
\nu_{\mathcal{X}} \left( x \right) = T_u \left( \nu_{\mathcal{X}^1} \left( x^1 \right) \otimes \nu_{\mathcal{X}^2} \left( x^2 \right) \right).
\]

As denoted in the previous subsection, the joint expression of the \( U \)-independent distribution is affected by the form of the function \( u \), however it is reduced to \( \nu_{\mathcal{X}} \) for any function \( u \) in the case that all the marginals are uniform distributions. This fact
indicates that the space of empirical $U$-independence $\tilde{P}_U(U)$ is not a rich subspace in $P_X$ if the empirical marginals are close to uniform. On the other hand, when the marginals are far from uniform and have extremely high (or low) probabilities because of small sample sets or outliers, empirical $U$-independent models can be flexible and convenient candidates.

4.3 Robust estimation of joint distribution with small samples

Now, we show an experimental result of the empirical $U$-independent model with various $u$-functions based on a small sample set. Let $X^1$ and $X^2$ be two discrete random variables, both variables with 20 categories. Figure 5(a) shows the true distribution $p_X$ which is strictly independent given by equation (7). At first, 400 samples were generated from the true distribution $p_X$, and the empirical distribution $\tilde{p}_X$ was obtained with those samples. Figure 5(b) shows the empirical distribution $\tilde{p}_X$ with 400 samples, which is very sparse and has a lot of sampling zeroes. We estimate empirical $U$-independent models $\tilde{q}_\otimes$ by equation (13) where its $U$-multiplication is defined by Ex.1 in Table 1 with various $\pi$ values. Note that $\pi = 0$ corresponds to the conventional independent model $q_X$.

Figure 5(c) shows $D_{KL}(p_X; \tilde{q}_\otimes)$ which measures the discrepancy between the true distribution and empirical $U$-independent models with various $\pi$ values. The result shows that the empirical $U$-independent model could be better than the conventional independent model even though the true distribution $p_X$ is strictly independent. Figure 5(d) shows $D_{KL}(\nu_X; \tilde{q}_\otimes)$ where $\nu_X$ is the uniform joint distribution of variables $X^1$ and $X^2$, and it depicts that the estimated model becomes closer to the uniform distribution as the $\pi$ value becomes smaller. Therefore, improvement of $D_{KL}(p_X; \tilde{q}_\otimes)$ in this experiment could be attributed to appropriate uniformisation of $\tilde{q}_\otimes$ with $U$-multiplication when the empirical distribution has many sampling zeroes. To obtain the optimum $\pi$ value with a given dataset, the CV or the bootstrap evaluation are available; especially the Bayesian bootstrap (Rubin, 1981) is a useful tool for an extremely small dataset like this experiment.

Figure 5 Experimental setup and result (a) $p_X$; strictly independent distribution $p_X$ (b) $\tilde{p}_X$; empirical distribution $\tilde{p}_X$ based on small samples (c) $D_{KL}(p_X; \tilde{q}_\otimes)$; KL divergence between $p_X$ and $\tilde{q}_\otimes$ (d) $D_{KL}(\nu_X; \tilde{q}_\otimes)$; KL divergence between $\nu_X$ and $\tilde{q}_\otimes$
5 \textit{U}-independence in NB model

As an application of \textit{U}-independence in statistical models, we introduce the NB model.

5.1 Extension of NB model

Let $Y$ be a categorical class variable, and $X = \{X^1, \ldots, X^M\}$ be a set of categorical variables. Then, a set of NB models is defined as follows,

$$P_{\text{NB}} = \left\{ p_{\text{NB}} \mid p_{\text{NB}}(x, y) = p_Y(y) \prod_{m=1}^{M} p_{X^m \mid Y}(x^m \mid y) \right\}. \tag{15}$$

The NB has some convenient properties, such as simple structure, easy estimation and scalability. And it is also known as a simple but robust classification tool (Domingos and Pazzani, 1997). With the empirical joint distribution $\tilde{p}_{X \mid Y}$, the ML estimate of the NB model is given by

$$q_{\text{NB}} = \arg\min_{q \in P_{\text{NB}}} D_{\text{KL}}(\tilde{p}_{X \mid Y}, q).$$

The concrete form of $q_{\text{NB}}(x, y)$ is composed of empirical marginals

$$\tilde{p}_Y(y) = \frac{n(y)}{\sum_{y' \in Y} n(y')} \quad \text{and} \quad \tilde{p}_{X^m \mid Y}(x^m \mid y) = \frac{n(x^m, y) + \alpha}{\sum_{x^m \in X^m} (n(x^m, y) + \alpha)}, \tag{16}$$

where $n(y)$ and $n(x^m, y)$ are the numbers of observations of events $Y = y$ and $(X^m = x^m, Y = y)$ respectively, and $\alpha \in [0, 1]$ is a Laplace smoother for estimation of $\tilde{p}_{X^m \mid Y}(x^m \mid y)$.

Now, we consider an extension of the NB model by using \textit{U}-independence. For example, assume that all the elements in the variable set $X$ are mutually conditional \textit{U}-independent given $Y$, we can derive the expression,

$$\tilde{p}_U = \left\{ \tilde{p}_U \mid \tilde{p}_U(x, y) = \tilde{p}_Y(y) T_u \left( \bigotimes_{m=1}^{M} \tilde{p}_{X^m \mid Y}(x^m \mid y) \right) \right\}. \tag{17}$$

For another example, assume that only some of the elements in $X$ are conditionally \textit{U}-independent, given as

$$\tilde{p}_S_l = \left\{ \tilde{p}_S_l \mid \tilde{p}_S_l(x, y) = \tilde{p}_Y(y) \prod_{l \in S_l} \tilde{p}_{X^l \mid Y}(x^l \mid y) T_u \left( \bigotimes_{m \in S_l} \tilde{p}_{X^m \mid Y}(x^m \mid y) \right) \right\}, \tag{18}$$

where $S_l$ is an index set of weakly dependent variables in $X^1, \ldots, X^M$ and $S_l$ is an index set of the strictly independent variables, i.e. $S_l = \{1, \ldots, M\} \setminus S_l$. Note that equations (15) and (17) are the special cases of equation (18) with $S_l = \emptyset$ and $S_l = \emptyset$ respectively. In this paper, we directly use empirical distributions given by equation (16) in these models to avoid a non-linear optimisation problem for derivation of the ML
estimate in an analogous way as the discussion of the empirical $U$-independent model. Hence, given $p_{XY}$, the ML estimates of equations (17) and (18) are derived by
\[
\tilde{q}_U = \arg\min_{q \in P_U} D_{\text{KL}}(\tilde{p}_{XY}, q) \quad \text{and} \quad \tilde{q}_{S_I} = \arg\min_{q \in P_{S_I}} D_{\text{KL}}(\tilde{p}_{XY}, q).
\]
The extended NB models with empirical marginals are called empirical $U$-NB models in this paper. In the next subsection, we experimentally evaluate empirical $U$-NB models by using benchmark datasets.

5.2 Numerical experiments

We compare empirical $U$-NB models given by equation (18) with various $S_I$ from the viewpoint of classification error by using four benchmark datasets ‘monks-1’ (MO1), ‘monks-2’ (MO2), ‘car evaluation’ (CAR) and ‘nursery’ (NUR) distributed in UCI ML repository (Asuncion and Newman, 2007). We try to tune the function $w$ in the $U$-NB by using a small dataset as shown in Table 2. In this experiment, we find an optimal $w$ with respect to $\pi$ in a one-parameter family Ex.1 in Table 1.

Table 2 Datasets used in experiment

<table>
<thead>
<tr>
<th>Dataset</th>
<th>M</th>
<th># Train data</th>
<th># Test data</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>MO1</td>
<td>6</td>
<td>124</td>
<td>432</td>
<td>$Y = 1$ when $(X^1 = X^2) \text{ or } X^5 = 1$</td>
</tr>
<tr>
<td>MO2</td>
<td>6</td>
<td>169</td>
<td>432</td>
<td>$Y = 1$ when exactly two of ${X^m}$ are 1</td>
</tr>
<tr>
<td>CAR</td>
<td>6</td>
<td>300</td>
<td>1,728</td>
<td></td>
</tr>
<tr>
<td>NUR</td>
<td>8</td>
<td>300</td>
<td>12,960</td>
<td></td>
</tr>
</tbody>
</table>

For each dataset, we prepare empirical $U$-NBs given by equation (18) for all the possible $S_I$. At first, we find the optimal $\hat{\alpha}$ for $q_{NB}$ by using 10-fold CV of classification error in training datasets. Secondly, we also find the optimal $\tilde{\pi}$ for each model $\tilde{q}_{S_I}$ under the fixed Laplace smoother ($\alpha = 10^{-8}$) by using 10-fold CV of classification error in training datasets and we choose the best $U$-NB. Then, we evaluate error rates of MAP estimates of the classes from the viewpoint of classification task in test datasets. Table 3 shows the classification error rates and the KL divergences $D_{\text{KL}}(p^\ast, q_{NB})$ and $D_{\text{KL}}(p^\ast, \tilde{q}_{S_I})$ where $p^\ast$ is an empirical joint distribution of the test dataset. As shown in Table 3, the selected $U$-NBs show improvement in error rates of classification for all the datasets and they also show improvement in KL divergences for MO1, MO2 and CAR. In Table 3, we also show the result of the $U$-NB with $S_I = \{1, 2\}$ for MO1 though it is not selected by CV. As shown in Table 2, MO1 do have dependence between $X^1$ and $X^2$ with respect to $Y$, however, the selected model with $S_I = \{1, 4, 5, 6\}$ is better than the model with $S_I = \{1, 2\}$. This result seems to be affected by the bias of the training dataset; the pseudo weak conditional dependence between $X^1$, $X^4$, $X^5$ and $X^6$ is captured based on empirical marginals constructed from the training dataset, and the derived model works better than the model with $S_I = \{1, 2\}$. The results indicate that the model $\tilde{q}_{S_I}$ can be a useful tool to improve the conventional NB by modelling weak dependence in variables by only using empirical marginals.
Table 3 Results of experiment

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NB ((q = q_{NB}))</th>
<th>(U)-NB ((\tilde{q} = \tilde{q}_{NB}))</th>
<th>NB ((q = q_{NB}))</th>
<th>(U)-NB ((\tilde{q} = \tilde{q}_{NB}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>(D_{KL}(p^*, q))</td>
<td>(\tilde{S}_I)</td>
<td>(\tilde{S}_I)</td>
<td>(\tilde{S}_I)</td>
</tr>
<tr>
<td>MO1</td>
<td>0.2870</td>
<td>0.5796</td>
<td>1.00</td>
<td>0.2431</td>
</tr>
<tr>
<td></td>
<td>0.2431</td>
<td>0.5407</td>
<td>1.00</td>
<td>0.2431</td>
</tr>
<tr>
<td>MO2</td>
<td>0.3866</td>
<td>0.6529</td>
<td>0.60</td>
<td>0.3287</td>
</tr>
<tr>
<td></td>
<td>0.3287</td>
<td>0.6529</td>
<td>0.60</td>
<td>0.3287</td>
</tr>
<tr>
<td>CAR</td>
<td>0.1870</td>
<td>0.6635</td>
<td>0.69</td>
<td>0.1766</td>
</tr>
<tr>
<td></td>
<td>0.1766</td>
<td>0.6479</td>
<td>0.69</td>
<td>0.1766</td>
</tr>
<tr>
<td>NUR</td>
<td>0.1191</td>
<td>0.4337</td>
<td>0.87</td>
<td>0.1017</td>
</tr>
<tr>
<td></td>
<td>0.1017</td>
<td>0.4367</td>
<td>0.87</td>
<td>0.1017</td>
</tr>
</tbody>
</table>

Notes: For NB, \(\tilde{\alpha}\) is selected by 10-fold CV. For \(U\)-NB, \(\tilde{S}_I\) and \(\tilde{\pi}\) are also selected by 10-fold CV; \(\alpha\) is fixed as \(\alpha = 10^{-8}\) (\(\tilde{\pi}\) we also show the result of \(U\)-NB with \(\tilde{S}_I = \{X_1, X_2\}\) for MO1 though it is not selected by CV).

6 Discussion and concluding remarks

We introduced generalised multiplication based on the monotonically increasing functions which is deeply related to the Bregman divergence and proposed some extensions of independence in statistical models. To reduce computational cost, we also proposed the empirical \(U\)-independent model which has robust property attributable to the approximated Bregman divergence. In addition, we show effectiveness of \(U\)-independence in the NB model for simple classification tasks by using some benchmark datasets.

When we use the \(U\)-NB model as a classifier in a practical scene, we have to handle very small probability values, particularly, when the dimension of \(X\), given as \(M\), is very large. In the conventional NB model, the log-transformation is a convenient tool to handle these small values because the NB model is decomposable by using the logarithm function as follows,

\[
\log(q_{NB}(x, y)) = \log(p_Y(y)) + \sum_{m=1}^{M} \log(p_{X^m | Y}(x^m | y)).
\]

However, the log-transformation is ineffective for \(U\)-NBs like equations (17) and (18) since we still need to handle \(U\)-multiplied small probabilities in these models. One way to avoid this problem is given as follows; we can extend the \(U\)-NB in the following way,

\[
\mathcal{P}_{U} = \left\{ \tilde{p}_{U} \mid \tilde{p}_{U}(x, y) = T_u \left( p_Y(y) \otimes \left( \bigotimes_{m=1}^{M} p_{X^m | Y}(x^m | y) \right) \right) \right\}.
\] (19)

In a classification task based on a dataset with large \(M\), the model given by equation (19) has some computational advantages. Since \(u(\cdot)\) is a monotonically increasing function, the MAP estimate of the class \(y\) for given \(x\) is derived as follows,
\[
\hat{y} = \arg \max_{y \in \mathcal{Y}} p_{Y|X}(y|x) = \arg \max_{y \in \mathcal{Y}} \frac{T_u \left( p_Y(y) \otimes \left( \bigotimes_{m=1}^{M} p_{X_m|Y}(x^m|y) \right) \right)}{p_X(x)}
\]

\[
= \arg \max_{y \in \mathcal{Y}} u \left( \xi(p_Y(y)) + \sum_{m=1}^{M} \xi(p_{X_m|Y}(x^m|y)) - c_u \right)
\]

\[
= \arg \max_{y \in \mathcal{Y}} \xi(p_Y(y)) + \sum_{m=1}^{M} \xi(p_{X_m|Y}(x^m|y))
\]

Note that equation (20) is derived because \( p_X(x) \) and \( c_u \) are constants in this situation. With this expression, we can handle small probabilities by the \( \xi \)-transformation like the log-transformation in the conventional NB. However, we cannot describe appropriate strictly independent variable subset \( S_I \) in this model as shown in our experiments. In this case, it may be important to choose appropriate marginal distributions though it remains as the future work.

In the same manner with the proposed models in this paper, we can also extend loglinear models and Bayesian networks. Compared with cross terms in loglinear models or link expression in Bayesian network, \( U \)-independence has some advantages in the number of parameters and in robustness; e.g., we expect simple description of Bayesian networks by omitting some links with weak dependence by using \( U \)-independence. From the viewpoint of copulas (Nelsen, 2006), \( U \)-independence is also deeply related to Archimedean copula. Further research is needed to discuss statistical properties of \( U \)-independence as a kind of copula. In our implementation, a function \( u \) is selected from a one-parameter family, particularly, from the family of Ex.1 in Table 1. The selection of an optimal function \( \hat{u} \) is another interesting topic.

Acknowledgements

This work was partially supported by JSPS, Grant-in-Aid for Young Scientists (B), 22700292.

References

A generalisation of independence in statistical models


