Open Journal of Statistics, 2015, 5, 464-473
Published Online August 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.55048

Fluctuation-Model-Based Discrete Probability Estimation for Small Samples

Takashi Isozaki

Sony Computer Science Laboratories, Inc., Tokyo, Japan
Email: isozaki@csl.sony.co.jp

Received 10 July 2015; accepted 17 August 2015; published 20 August 2015

Abstract

A robust method is proposed for estimating discrete probability functions for small samples. The proposed approach introduces and minimizes a parameterized objective function that is analogous to free energy functions in statistical physics. A key feature of the method is a model of the parameter that controls the trade-off between likelihood and robustness in response to the degree of fluctuation. The method thus does not require the value of the parameter to be manually selected. It is proved that the estimator approaches the maximum likelihood estimator at the asymptotic limit. The effectiveness of the method in terms of robustness is demonstrated by experimental studies on point estimation for probability distributions with various entropies.

Keywords

Probability Estimation, Fluctuation Models, Helmholtz Free Energy, Canonical Distributions

1. Introduction

For categorical observational data analysis, it is often necessary to deal with multivariate systems since variables of such data generally depend on each other. Highly predictive statistical inference requires parameters that achieve low-entropy, so it is preferable to use data of many variables because the following relationship in Shannon entropies H of random variables X and Y holds if X and Y depend on each other: $H(X|Y) \le H(X)$ [1], where $H(X)$ denotes Shannon entropy of X and $H(X|Y)$ denotes the conditional entropy of X given Y. These are respectively defined with marginal, joint, and conditional probability mass functions P as follows [1]:

$$H(X) = -\sum_i P(X=i)\log P(X=i),$$

How to cite this paper: Isozaki, T. (2015) Fluctuation-Model-Based Discrete Probability Estimation for Small Samples. *Open Journal of Statistics*, **5**, 464-473. http://dx.doi.org/10.4236/ojs.2015.55048

$$H\left(X\middle|Y\right)=-\sum_{i,j}P\left(X=i,Y=j\right)\log P\left(X=i\middle|Y=j\right),$$

where *i* and *j* are indices of discrete states of *X* and *Y*. When we estimate probabilities in discrete probabilistic models with many-variables (e.g., Markov network and Bayesian network models [2]), statistical estimation of conditional and joint probabilities often needs exponentially large data because of the combinatorial explosion of binding events in variables. Therefore, the models are often inferred from insufficient data. The maximum likelihood (ML) method provides an estimated probability function, which of *X* with *k* discrete states is expressed by

$$P\left(X=i\right)=\frac{n_i}{n},\quad i=1,\cdots,k$$

$$\sum_{i=1}^{k}P\left(X=i\right)=1,$$

where *n* and n_i respectively denote a sample size and the frequency of occurrences in *i* state. The estimated probability is correct in the large sample limit. However, the ML methods suffer from short size data, and few robust methods have been investigated for such data in estimation of discrete probability functions as far as we know, although many robust methods for outliers such as *M*-estimators have been developed [3] [4] in parametric continuous distributions. The maximum entropy method [5], which may be applied to small datasets, was originally appropriate for data with missing information.

In the present study, a new robust method is proposed for estimating discrete probability functions for small samples. The method uses a parameterized objective function based on Kullback-Leibler divergence [6] and Shannon entropy. The function has a similar form to the Helmholtz free energy function that appears in statistical physics [7] [8]. A key feature of the method is a model of the parameter that controls the trade-off between likelihood and robustness in response to the degree of data fluctuation. The method thus does not require the value of the parameter to be manually selected. This model is a modification of a preceding work [9], in which the parameter is represented by an artificial model containing a free hyperparameter.

In the domain of machine learning, although several methods slightly similar to ours have been proposed [10]-[14], there is a critical distinction between these methods and ours. Many studies that have applied free energy to statistical inference have not included the similar trade-off parameter or have treated it as a fixed value, a manually controlled parameter, or a free parameter. Regarding the existing methods, we thus consider that the potentials of free-energy-like functions have not been well extracted. Other similar methods have been developed in the context of robust estimation for outliers, in which a free parameter is introduced in an analogous fashion [3] [15] [16]. However, the problem of how to determine the value of the free parameters remains.

This paper is organized as follows. In the next section, an objective function with parameter β is introduced for robust estimation, and then probability functions obtained by the proposed method are shown to be formally equivalent to the canonical distributions that appear in statistical physics. In Section 3, a new representation of β is presented as a data-fluctuation-model, and the preferable asymptotic property of β is proved. In Section 4, some characteristic properties among quantities used in the proposed method are provided. In Section 5, we perform experiments using the proposed probability estimation method. In Section 6, conclusions regarding the estimation method are given.

2. Probability Estimation with Parameter *β*

Note that in this paper a capital letter (such as *X*) denotes a random discrete variable, a non-capital letter (such as *x*) denotes the special state of that variable, a bold capital letter (such as $\boldsymbol{Y}$) denotes a set of variables, and a bold non-capital letter (such as $\boldsymbol{y}$) denotes configurations of that set.

To construct a method for estimating finite-discrete-probability distributions of random variable *X* from sample set of finite size *n*, the following quantities are defined. $P\left(X\right)$ denotes a discrete-probability function estimated by a proposed method that is described below. A function of *X* is defined as $U_0\left(X\right)$ on the basis of the Kullback--Leibler (KL) divergence [6] between empirical functions and $P\left(X\right)$ as follows:

$$U_0\left(X\right)=D\left(P\left(X\right)\middle\|\tilde{P}\left(X\right)\right)=\sum_{x}P\left(x\right)\log\frac{P\left(x\right)}{\tilde{P}\left(x\right)},\tag{1}$$

where $\tilde{P}(X)$ is a empirical distribution function. For non-parametric discrete distributions, the empirical distributions are equivalent to relative frequencies; *i.e.*, they are equivalent to the maximum likelihood (ML) distributions. $\tilde{P}(X)$ can thus be replaced by ML distributions denoted by $\hat{P}(X)$. An objective function $F(X)$ is defined as follows:

$$F(X) = U_0(X) - \frac{1}{\beta_0(X)} H(X), \tag{2}$$

where $U_0(X)$ is defined by Equation (1), $H(X)$ is the Shannon entropy [1] of the estimated functions given as $H(X) = -\sum_x P(x) \log P(x)$, and $\beta_0(X)$ is a parameter that is defined so that $0 < \beta_0(X) < \infty$. $U(X)$ and $\beta(X)$ are introduced for later convenience. $U(X)$ is represented by a cross entropy $H\left(P(X), \hat{P}(X)\right)$ and $\beta(X)$ is a normalization parameter of $\beta_0(X)$ as follow:

$$U(X) = H\left(P(X), \hat{P}(X)\right) = -\sum_x P(x) \log \hat{P}(x) \tag{3}$$

and

$$\beta(X) = \beta_0(X) \big/ \left(1 + \beta_0(X)\right), \tag{4}$$

where $0 < \beta(X) < 1$. $F(X)$ can be rewritten by using $U(X)$ and $\beta(X)$ as:

$$F(X) = U(X) - \frac{1}{\beta(X)} H(X). \tag{5}$$

In addition, the following quantity $\epsilon(x)$ is defined as

$$\epsilon(x) = -\log \hat{P}(x). \tag{6}$$

$U(X)$ is also written with ϵ as $U(X) = \langle \epsilon \rangle (X)$; that is, $\langle \rangle$ denotes an expectation value in respect to $P(X)$.

The estimator of probability functions, $P(X)$, is defined so as to minimize Lagrangian L consisting of $F(X)$ as $\partial L / \partial P(X) = 0$. L is expressed as

$$L = F + \lambda\left(\sum_x P(x) - 1\right) = \frac{1}{\beta} \sum_x P(x) \log P(x) - \sum_x P(x) \log \hat{P}(x) + \lambda\left(\sum_x P(x) - 1\right), \tag{7}$$

where λ is the Lagrange multiplier. $P(X)$ is thereby obtained as

$$P(x) = \frac{\exp\left(-\beta\left(-\log \hat{P}(x)\right)\right)}{\sum_{x'} \exp\left(-\beta\left(-\log \hat{P}(x')\right)\right)} = \frac{\exp\left(-\beta \epsilon(x)\right)}{\sum_{x'} \exp\left(-\beta \epsilon(x')\right)}, \tag{8}$$

where $\epsilon(x)$ as expressed in Equation (6) is used. Equation (8) is equivalent to a form known as the canonical distribution, which is also called Gibbs distribution, in statistical physics. The following equivalent form is more convenient for practical use:

$$P(x) = \frac{\left[\hat{P}(x)\right]^{\beta}}{\sum_{x'} \left[\hat{P}(x')\right]^{\beta}}. \tag{9}$$

For estimating conditional and joint-probability functions, conditional entropy $H(X|\boldsymbol{Y})$, which is defined as: $H(X|\boldsymbol{Y}) = -\sum_{x,\boldsymbol{y}} P(x,\boldsymbol{y}) \log P(x|\boldsymbol{y})$, and conditional KL divergence:

$$D\left(P(X|\boldsymbol{Y}) \| Q(X|\boldsymbol{Y})\right) = \sum_{x,\boldsymbol{y}} P(x,\boldsymbol{y}) \log\left(P(x|\boldsymbol{y}) / Q(x|\boldsymbol{y})\right)$$

are used. β for $P(x|\boldsymbol{y})$ is defined as

$$\beta(X|\boldsymbol{Y}) = \beta_0(X|\boldsymbol{Y}) \big/ \left(\beta_0(X|\boldsymbol{Y}) + 1\right).$$

The formula for estimating conditional probabilities is therefore obtained by using the conditional entropy and KL divergence and $\beta(X|\boldsymbol{Y})$ in the following form:

$$P(x|\boldsymbol{y})=\frac{\exp\left(-\beta\left(-\log\hat{P}(x|\boldsymbol{y})\right)\right)}{\sum_{x'}\exp\left(-\beta\left(-\log\hat{P}(x'|\boldsymbol{y})\right)\right)}. \tag{10}$$

Joint probability can be calculated by using Equations (8) and (10) and the definite relation $P(X,Y)=P(X|Y)P(Y)$. In general, it is calculated using decomposition rules such that

$$P(X_1,X_2,\cdots,X_n)=P(X_n|X_{n-1},\cdots,X_2,X_1)\cdots P(X_2|X_1)P(X_1).$$

3. Model of β

$P(X)$ approaches $\hat{P}(X)$ when β in Equation (9) approaches 1. On the other hand, $P(X)$ approaches the uniform distribution if $\hat{P}(X)>0$ and β approaches 0. If β close to 1 represents that the data size is sufficiently large and if β close to 0 represents that the data size is very small, β has favorable properties for accurate and robust estimation. This is because the ML estimators generally have preferable consistency and asymptotic efficiency and the distributions close to the uniform can be regarded as the ones that have robustness for small size data. Before β is defined, the following quantity $P_n^G(X)$ with *n* data size is defined by a geometric mean as

$$P_n^G(X)=Z_n^G\left(\prod_{i=0}^{n}P_i(X)\right)^{\frac{1}{n+1}}, \tag{11}$$

where Z_n^G denotes a normalization constant, and $P_i(X)$ denotes the estimated function obtained from Equation (9) with initial *i* data. It is defined that $P_0^G(X)=P_0(X)=1/|X|$, where $|X|$ denotes the number of states of variable *X*. Fluctuation δ of *X* with *n* data is given as

$$\delta_n(X)=\log P_{n-1}^G(X)-\log\hat{P}(X).$$

Then, $\beta_0(X)$ for *n* data is defined by using an expectation value of $\delta_n(X)$ with respect to $P_{n-1}^G(X)$, *i.e.*, the following KL divergence:

$$\beta_0(X)=1/D\left(P_{n-1}^G(X)\|\hat{P}(X)\right)=1/\sum_x P_{n-1}^G(x)\log\frac{P_{n-1}^G(x)}{\hat{P}(x)}, \tag{12}$$

where $n\geq 1$, and $\hat{P}(X)$ is the ML estimator function obtained from *n* data. It is assumed that $D\left(P_{n-1}^G(X)\|\hat{P}(X)\right)>0$. The normalized β_0, that is, β, is defined by Equation (4). Note that the canonical distribution $P(x)$ expressed by Equation (9) can be determined, without any free parameters, by using Equations (9), (11), (12), (4), and $P(X)=1/|X|$ for data size $n=0$. It is also defined that in Equation (10) for conditional data size $n=0$ given $\boldsymbol{Y}=\boldsymbol{y}$, $P(x|\boldsymbol{y})=1/|X|$ for any $\boldsymbol{y}$ in the same manner as $P(x)$.

Objective function *F* is rewritten in the same form as that in statistical mechanics as follows:

$$F=U-\frac{1}{\beta}H=-\frac{1}{\beta}\log Z, \tag{13}$$

where *Z* is the partition function, which is a similar function well known in statistical mechanics, defined for single or multivariate probabilities as

$$Z(\boldsymbol{X})=\sum_{\boldsymbol{x}}\left[\hat{P}(\boldsymbol{x})\right]^{\beta}, \tag{14}$$

and for conditional probabilities as

$$Z(X|\boldsymbol{Y})=\sum_{x}\left[\hat{P}(x|\boldsymbol{Y})\right]^{\beta}. \tag{15}$$

A significant feature of β is confirmed as follows. The lemma needed for this proof is stated as

Lemma 1.

$P_i(X)$ is denoted as the canonical distribution estimated from Equation (8) with *i* data. For data size $n \to \infty$, $P_n^G(x)$, defined by Equation (11), converges to a definite value $P^G(x)$ when $P_i(x) > 0$ for integers *i* such that $i \geq 0$ and any state *x* of *X*.

Proof.

$$\begin{aligned} &\log P_n^G(x) - \log P_{n-1}^G(x) \\ &= \frac{1}{n+1}\sum_{i=0}^{n}\log P_i(x) - \frac{1}{n}\sum_{i=0}^{n-1}\log P_i(x) + \log Z_n^G - \log Z_{n-1}^G \\ &= \left(\frac{n}{n+1} - 1\right)\log P_{n-1}^G + \frac{1}{n+1}\log P_n(x) + \log Z_n^G - \log Z_{n-1}^G. \end{aligned} \tag{16}$$

Because $P_{n-1}^G(x) > 0$ and $P_n(x) > 0$, $\log P_{n-1}^G(x)$ and $\log P_n(x)$ are definite values. Thus both terms on the right-hand side of Equation (16) converge to $\log Z_n^G - \log Z_{n-1}^G$, which is a constant if $n \to \infty$. Hence, $P_n^G(x) = bP_{n-1}^G$, with constant *b*, can be written. Therefore, $b = 1$ and $P_n^G(x) \to P^G(x)$ in order that $P_n^G(x)$ converges not to 0 or 1 but to definite values. □

Theorem 1. *At the asymptotic limit (i.e., large sample limit),* β *converges to* 1 *when* $P_i(x) > 0$ *for integers i such that* $i \geq 0$ *and any state x, where* $P_i(x)$ *is denoted as the canonical distribution represented by* Equation (8) *from i data.*

Proof. According to Lemma 1, $P_n^G(x) \to P^G(x)$ at the limit $n \to \infty$, where $P^G(x)$ is a definite value for any *x*. The following Equation is thus obtained from Equation (11) with $P_i(x)$ as follows:

$$\log P^G(x) = \lim_{n\to\infty}\frac{1}{n+1}\sum_{i=0}^{n}\log P_i(x) + \text{constant}, \tag{17}$$

$P_n(x)$ thus converges to a definite value, and $P^G(x)$ converges to $P_n(x)$ (and the constant in Equation (17) goes to 0) at $n \to \infty$. Meanwhile, ML estimator $\hat{P}(x)$ converges to true distribution $P_t(x)$ due to the consistency of the ML estimators, and β thus converges to a definite value according to Equation (9). $P_n(x)$ at $n \to \infty$ is denoted as $P_\infty(x)$. Therefore, the following Equation is derived from Equations (4), (9) and (12) at $n \to \infty$,

$$\frac{1}{\beta_0} = \frac{1-\beta}{\beta} = D\left(P_\infty(X) \| P_t(X)\right) = \sum_x \frac{\left[P_t(x)\right]^\beta}{Z}\log\frac{\left[P_t(x)\right]^{\beta-1}}{Z} = \text{a definite value} \tag{18}$$

for $P(x) > 0$ and any state X. Equation (18) requires $\beta \to 0$ or $\beta \to 1$ in order that $\left[P_t(x)\right]^\beta$ or $\left[P_t(x)\right]^{\beta-1}$ is a constant for any probability distribution $P_t(x)$. However, $\beta \to 0$ does not satisfy Equation (18), while $\beta \to 1$ satisfies it. Accordingly, $\beta \to 1$ at the asymptotic limit. □

According to Theorem 1, the more data are obtained, the more β approaches 1, and the more the estimator approaches the ML estimator. The estimator that is obtained by the proposed method therefore has the same preferable asymptotic properties, namely, consistency and efficiency, as the ML estimators have. For insufficient data size, β_0 is probably small due to the influence of the uniform distributions given by Equation (12), so β is also small. The estimated probability functions by the proposed method are thus interpreted as adaptively tempered ML estimator functions in response to the degree of data fluctuation. The proposed estimation method does thereby not require manually selecting the value of parameter β, and is called "*ATML*," which is abbreviated as the "adaptively tempered ML" method. ATML has an advantage of simpleness over methods that need complicated algorithms (e.g., [17]).

The role of β can be seen as a trade-off parameter between likelihood and robustness by referring to another expression of Equation (2) as follows:

$$F = \beta D\left(P(X) \| \hat{P}(X)\right) + (1-\beta) D\left(P(X) \| P_u(X)\right) + \text{constant}, \tag{19}$$

where $P_u(X)$ denotes the uniform distribution function, which contributes to the robustness, while $\hat{P}(X)$ contributes to the likelihood. Additionally, objective function *F* can also be interpreted as a KL-based divergence measure since β is also represented by a KL divergence.

ATML has an analogy with statistical physics, since the canonical distribution for the estimator is obtained from Equation (8). Actually, *U*, *H*, β, and *F* respectively play similar roles to (internal) energy, entropy, (inverted) temperature, and Helmholtz free energy in statistical physics. Solving Equation $\partial L / \partial P(X) = 0$ for obtaining $P(X)$ mathematically corresponds to employing the minimum-free-energy (MFE) principle [7] in thermal physics.

ATML may seem analogous to Jaynes' maximum entropy (ME) methods [5], which are well known as least-biased inference methods. However, the constraints on which ME methods are based may not be reliable for small samples and thus may be biased, although this kind of bias is not usually considered. On the other hand, ATML is thus designed so that even the bias can be corrected by using parameter β.

4. Characteristic Properties of ATML

The canonical distribution expressed as Equation (8) can provide some characteristic properties of ATML, which are similar to those in statistical physics. The following notations are defined for later convenience. Probability mass functions that are estimated by ATML, denoted by P_k, have discrete states denoted as index *k*. The corresponding ML estimator is denoted by $\hat{P}_k$, and $\{\beta, U\}$ is used instead of $\{\beta_0, U_0\}$.

In statistical physics, (inverted) temperature β is usually defined as [7] [8]

$$\beta := \frac{\partial H}{\partial U}. \tag{20}$$

If the canonical distribution of P_k, which takes the form of Equation (8), is used, Equation (20) is automatically satisfied as follows:

Lemma 2. $H = \beta U + \log Z$ *under the MFE condition, where H, U,* β*, and Z are defined in the previous sections.*

Proof. Since probability mass function P_k has a canonical form under the MFE condition, it follows that

$$\begin{aligned} H &= -\sum_k P_k \log P_k = -\sum_k \frac{\hat{P}_k^\beta}{Z} \log \frac{\hat{P}_k^\beta}{Z} \\ &= \beta \cdot \left(-\sum_k \frac{\hat{P}_k^\beta}{Z} \log \hat{P}_k \right) + \left(\sum_k \frac{\hat{P}_k^\beta}{Z} \right) \cdot \log Z = \beta U + \log Z. \end{aligned} \tag{21}$$

□

Theorem 2. Equation (20) *is automatically satisfied under the MFE condition.*

Proof. Partially differentiating both sides of Equation (21) with respect to *U* gives

$$\frac{\partial H}{\partial U} = \beta + U \frac{\partial \beta}{\partial U} + \frac{\partial}{\partial U} \log Z. \tag{22}$$

$$\frac{\partial}{\partial U} \log Z = \frac{\partial \beta}{\partial U} \frac{\partial}{\partial \beta} \log Z. \tag{23}$$

$$\frac{\partial}{\partial \beta} \log Z = \frac{1}{Z} \frac{\partial}{\partial \beta} \left(\sum_k \hat{P}_k^\beta \right) = \frac{1}{Z} \sum_k \hat{P}_k^\beta \log \hat{P}_k = \sum_k \frac{\hat{P}_k^\beta}{Z} \log \hat{P}_k = -U.$$

It follows that

$$\frac{\partial H}{\partial U} = \beta.$$

□

In the same way, it can be proved that $\beta_0 = \partial H / \partial U_0$.

$\langle \epsilon^2 \rangle - \langle \epsilon \rangle^2$, which is called energy fluctuations in statistical mechanics, is shown to have the following relation, where $\langle \, \rangle$ denotes an expectation value with respect to the canonical distributions.

$$\langle \epsilon^2 \rangle - \langle \epsilon \rangle^2 = -\frac{\partial U}{\partial \beta}. \tag{24}$$

In regard to ϵ defined in the proposed estimation method, namely, Equation (6), the same relation as that shown here is satisfied as follows:

$$
\begin{aligned}
-\frac{\partial U}{\partial \beta} &= \sum_k \left(\log \hat{P}_k\right)\left\{\frac{1}{Z}\hat{P}_k^{\beta}\log \hat{P}_k - \left(\frac{1}{Z^2}\right)\sum_m \frac{\partial \hat{P}_m^{\beta}}{\partial \beta}\right\} \\
&= \sum_k \left(\log \hat{P}_k\right)\left\{\frac{\hat{P}_k^{\beta}}{Z}\log \hat{P}_k - \frac{\hat{P}_k^{\beta}}{Z}\sum_m \frac{\hat{P}_m^{\beta}}{Z}\log \hat{P}_m\right\} \\
&= \sum_k \frac{\hat{P}_k^{\beta}}{Z}\log \hat{P}_k\left(\log \hat{P}_k - \sum_m \frac{\hat{P}_m^{\beta}}{Z}\log \hat{P}_m\right) \\
&= \left\langle \epsilon^2 \right\rangle - \left\langle \epsilon \right\rangle^2,
\end{aligned}
$$

Equation (24) is therefore proved.

Fisher information $\tilde{I}(\beta)$ with a parameter β is defined in the usual way as

$$
\tilde{I}(\beta) := \sum_k f_k(\beta)\left(\frac{\partial}{\partial \beta}\log f_k(\beta)\right)^2, \tag{25}
$$

where *f* is the likelihood function. We define tempered Fisher information $I(\beta)$ as Fisher information where the likelihood function is replaced with the canonical distributions with parameter β. It is shown that $I(\beta) = \left\langle \epsilon^2 \right\rangle - \left\langle \epsilon \right\rangle^2$ as follows:

$$
\begin{aligned}
I(\beta) &= \sum_k \frac{\hat{P}_k^{\beta}}{Z}\left(\frac{\partial}{\partial \beta}\log \frac{\hat{P}_k^{\beta}}{Z}\right)^2 = \sum_k \frac{\hat{P}_k^{\beta}}{Z}\left(\log \hat{P}_k - \frac{1}{Z}\sum_m \hat{P}_m^{\beta}\log \hat{P}_m\right)^2 \\
&= \sum_k \frac{\hat{P}_k^{\beta}}{Z}\left(-\log \hat{P}_k\right)^2 - \left(\sum_k \frac{\hat{P}_k^{\beta}}{Z}\left(-\log \hat{P}_k\right)\right)^2 = \left\langle \epsilon^2 \right\rangle - \left\langle \epsilon \right\rangle^2.
\end{aligned} \tag{26}
$$

The tempered Fisher information is therefore identical to Equation (24).

It is noteworthy that ATML has other mathematical similarities with statistical physics. That is, the same relationships that appear in statistical physics listed as follows hold.

- The following relation is easily derived from the definition of partition function *Z*:

$$
U = -\frac{\partial}{\partial \beta}\log Z. \tag{27}
$$

- The following relation, known as the Gibbs-Helmholtz relation, is derived from Equations (13) and (27) as

$$
U = \frac{\partial}{\partial \beta}(\beta F). \tag{28}
$$

- The following relation is simply obtained from Equations (5) and (28) as

$$
H = \beta^2 \frac{\partial F}{\partial \beta}. \tag{29}
$$

- The tempered Fisher information is represented by the second-order differential of the partition function for β as

$$
I(\beta) = \left\langle \epsilon^2 \right\rangle - \left\langle \epsilon \right\rangle^2 = \frac{\partial^2}{\partial \beta^2}\log Z. \tag{30}
$$

5. Examples

Numerical experiments are performed to demonstrate the robustness of ATML for small samples, in comparison with the ML and ME methods [5].

X is assumed to have three internal states and four probability mass functions with a variety of entropies denoted as $H(X)$ in natural logarithms as

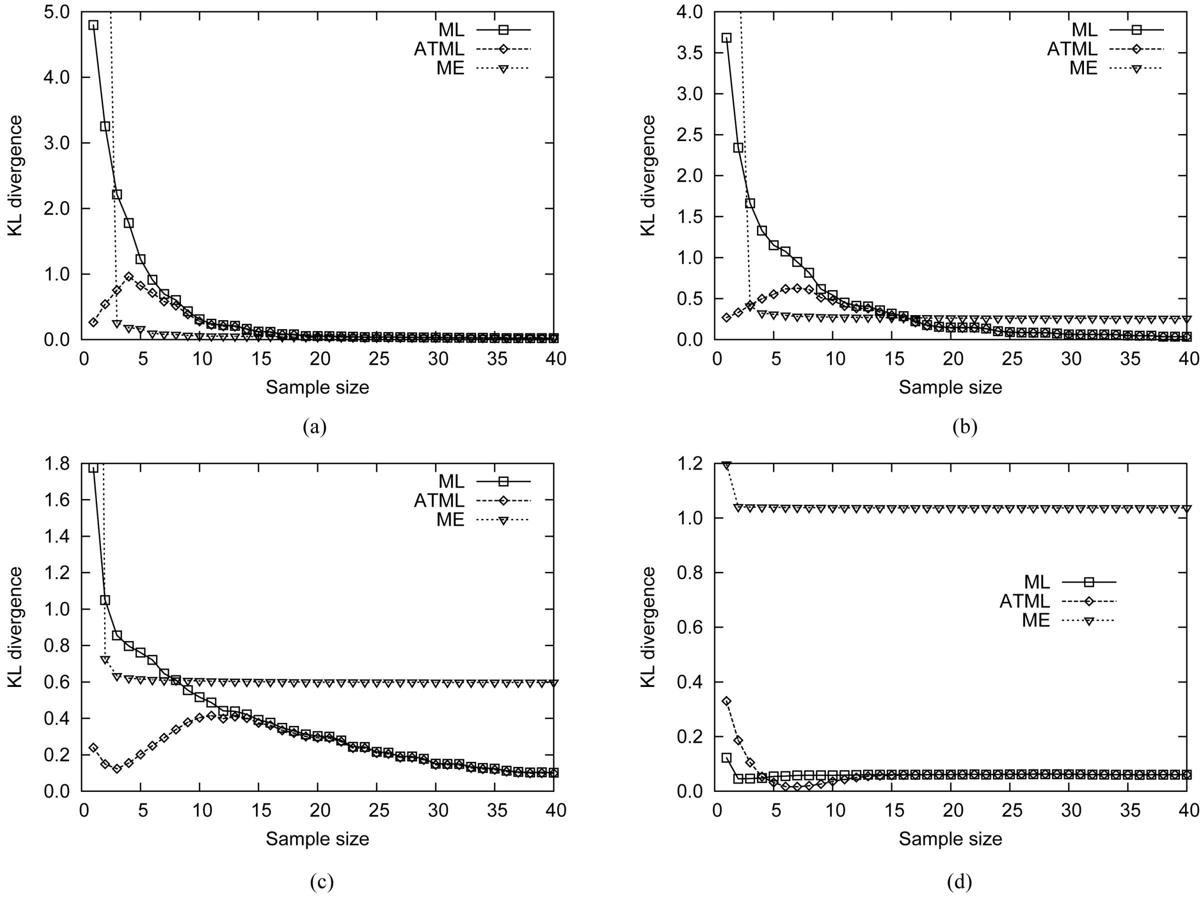
1. $P(x=0)=0.431, P(x=1)=0.337, P(x=2)=0.232, H(X)=1.07$,
2. $P(x=0)=0.677, P(x=1)=0.206, P(x=2)=0.117, H(X)=0.841$,
3. $P(x=0)=0.851, P(x=1)=0.117, P(x=2)=0.0320, H(X)=0.498$,
4. $P(x=0)=0.9898, P(x=1)=0.00810, P(x=2)=0.00210, H(X)=0.0621$.

Data from each function was sampled, and probabilities were estimated from given data sets with various data sizes. According to convention, averaged outputs $\langle X \rangle$ were set as the constraint in the ME method as follows: $\langle X \rangle := (1/N)\sum_{d=1}^{N} X_d$, where X_d denotes *d*-th sample's output, and *N* denotes sample size. After that, true and estimated probabilities were compared by using KL divergence as a metric with the following form:

$$D\left(P(X) \| P_{\mathrm{e}}(X)\right) = \sum_{x} P(x) \log \frac{P(x)}{P_{\mathrm{e}}(x)}, \tag{31}$$

where $P(X)$ is the true distribution, and $P_{\mathrm{e}}(X)$ is the distribution estimated by ML, ME, or ATML. For avoiding zero probabilities, probabilities were smoothed by adding 0.0001 to the counts.

The KL divergences are shown in **Figure 1**, where they are averaged values from 100 samples at each sample size from identical distributions. It can be seen that the ML estimators are inferior to ATML due to overfitting, except for the distribution having very small entropy. Even the degree of superiority of the ML estimation in (d) is relatively smaller than that of inferiority in other distributions. The ME estimators showed the opposite behaviors to those by the ML estimators, and showed some relatively poor results in large-sample regions. ML methods tend to fit data and can thus more accurately estimate distributions with very low entropies than others in small-sample cases. For example, if a true entropy equals zero, the ML method can estimate the exact true



(a) (b)

(c) (d)

Figure 1. KL divergences between true probability mass functions and probability mass functions estimated by using ML, ATML, and ME. The horizontal axes denote sample sizes. *H* denotes Shannon entropy in natural logarithms. (a) *H* = 1:07; (b) *H* = 0:841; (c) *H* = 0:498; (d) *H* = 0:0621.

distribution from only one sample. On the other hand, ME methods tend to increase entropies and can thereby accurately estimate distributions with high entropies close to the uniform distributions. Hence, the ML method tends to overfit data, and the ME method tends to underfit data in the view of misestimation. Even so, ATML showed relative stability in terms of both sample sizes and distributions. This result indicates the effectiveness of ATML as a probability estimation method.

6. Conclusion

A robust method for estimating discrete probability functions, called "adaptively tempered maximum likelihood" method (ATML for short), is proposed. The estimators obtained in this method minimize a parameterized objective function similar to Helmholtz free energies that appear in statistical physics. The key feature of the proposed method is a model of the parameter as a fluctuation of finite size data. The parameter that is modeled plays an important role in determining the appropriate trade-off between likelihood and robustness in response to the degree of the fluctuations. ATML does thereby not require manually selecting the value of the parameter. It is also proved that the obtained estimator approaches the maximum likelihood estimator at the asymptotic limit. The effectiveness of ATML in terms of robustness was demonstrated by experimental studies on point estimation for probability distributions with various entropies.

Acknowledgements

The author thanks Mario Tokoro and Hiroaki Kitano of Sony Computer Science Laboratories, Inc. for their support as well as Te Sun Han, Hiroshi Nagaoka, Tomohiro Ogawa and Jun Suzuki of the University of Electro-Communications for their valuable comments and discussions.

References

[1] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell Systems Technical Journal*, **27**, 379-423, 623-656.

[2] Pearl, J. (1988) Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

[3] Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998) Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika*, **85**, 549-559. http://dx.doi.org/10.1093/biomet/85.3.549

[4] Beran, R. (1977) Minimum Hellinger Distance Estimates for Parametric Models. *Annals of Statistics*, **5**, 445-463. http://dx.doi.org/10.1214/aos/1176343842

[5] Jaynes, E.T. (1957) Information Theory and Statistical Mechanics. *Physical Review*, **106**, 620-630. http://dx.doi.org/10.1103/PhysRev.106.620

[6] Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86. http://dx.doi.org/10.1214/aoms/1177729694

[7] Callen, H.B. (1985) Thermodynamics and an Introduction to Thermostatistics. 2nd Edition, John Wiley & Sons, Hoboken, NJ.

[8] Kittel, C. and Kroemer, H. (1980) Thermal Physics. W. H. Freeman, San Francisco, CA.

[9] Isozaki, T., Kato, N. and Ueno, M. (2009) "Data Temperature" in Minimum Free Energies for Parameter Learning of Bayesian Networks. *International Journal on Artificial Intelligence Tools*, **18**, 653-671. http://dx.doi.org/10.1142/S0218213009000342

[10] Hofmann, T. (1999) Probabilistic Latent Semantic Analysis. *Proceedings of Conference on Uncertainty in Artificial Intelligence* (*UAI*-99), Stockholm, 30 July-1 August 1999, 289-296.

[11] LeCun, Y. and Huang, F.J. (2005) Loss Functions for Discriminative Training of Energy-Based Models. *Proceedings of International Workshop on Artificial Intelligence and Statistics* (*AISTATS*-05), Barbados, 6-8 January 2005, 206-213.

[12] Pereira, F., Tishby, N. and Lee, L. (1993) Distributional Clustering of English Words. In: *Proceedings of Annual Meeting on Association for Computational Linguistics* (*ACL*-93), Association for Computational Linguistics, Stroudsburg, 183-190. http://dx.doi.org/10.3115/981574.981598

[13] Ueda, N. and Nakano, R. (1995) Deterministic Annealing Variant of the EM Algorithm. *Proceedings of Advances in Neural Information Processing Systems* 7 (*NIPS* 7), Denver, 29 November-1 December 1994, 545-552.

[14] Watanabe, K., Shiga, M. and Watanabe, S. (2009) Upper Bound for Variational Free Energy of Bayesian Networks. *Machine Learning*, **75**, 199-215. http://dx.doi.org/10.1007/s10994-008-5099-x

[15] Jones, M.C., Hjort, N.L., Harris, I.R. and Basu, A. (2001) A Comparison of Related Density-Based Minimum Divergence Estimators. *Biometrika*, **88**, 865-873. http://dx.doi.org/10.1093/biomet/88.3.865

[16] Windham, M.P. (1995) Robustifying Model Fitting. *Journal of the Royal Statistical Society B*, **57**, 599-609.

[17] Pöschel, T., Ebeling, W., Frömmel, C. and Ramírez, R. (2003) Correction Algorithm for Finite Sample Statistics. *The European Physical Journal E*, **12**, 531-541. http://dx.doi.org/10.1140/epje/e2004-00025-4