A Geometric Approach to Statistical Estimation

Rudolf Kulhavý[1]

Institute of Information Theory and Automation, Academy of Sciences
of the Czech Republic, P. O. Box 18, 182 08 Prague, Czech Republic
kulhavy@utia.cas.cz

Abstract

The role of Kerridge inaccuracy, Shannon entropy and Kullback-Leibler distance in statistical estimation is shown for both dicrete and continuous observations. The cases of data independence and regression-type dependence are considered in parallel. Pythagorean-like relations valid for probability distributions are presented and their importance for estimation under compressed data is indicated.

1. Introduction

Rules of probability theory provide a fundamental tool for statistical estimation. It is the computational complexity of these rules, however, that makes estimation algorithms often infeasible. Modified rules are perhaps needed that would be easier to implement, yet close enough to what probability does. An appealing way of approximate inference would be to merge statistical and computational uncertainty. The question is how the two kinds of uncertainty should be translated into one language.

One possible approach is to use concepts of information theory, namely, to view estimation as calculation of a certain distance between the empirical and model distributions of data. The approach is far from being new. In statistics, minimum distance estimation and its consistency for a large class of distances was studied very early [1]. In robust statistics, D-estimators were studied including the question how the choice of a particular distance affects robustness [2], [3]. In system identification, information-theoretic distances were used in structure determination [4] and approximation [5], [6].

In this paper, we make use of three information measures — inaccuracy, entropy and Kullback-Leibler (K-L) distance and show how they are related to likelihood. Then we consider the case when a sample average of some prespecified functions is known rather than a complete empirical distribution. A distance between the empirical and model distributions is decomposed into a sum of two distances in a Pythagorean-like way. In [7], [8], [9], a Pythagorean relation was shown to hold for K-L distances. Here we make a slight extension presenting a Pythagorean-like theorem that links inaccuracy and K-L distance. Apart from giving another unified view of parameter estimation, the result provides a tool for possible approximation of likelihood.

[1] Supported in part by Grant 102/94/0314 of the Czech Grant Agency and Grant 275109 of the Academy of Sciences of the Czech Republic.

2. Parameter estimation revisited

Independent observations

Consider a sequence of random variables $Y^N = (Y_1, \ldots, Y_N)$ with values in a set $\mathscr{Y}$. Suppose that Y_k are independent and identically distributed according to a common probability distribution S_θ parametrized by an unknown parameter $\theta \in \Theta$. To cover the cases of discrete and continuous Y with one notation, we introduce densities $s_\theta = \frac{\mathrm{d}S_\theta}{\mathrm{d}\upsilon}$, $\theta \in \Theta$ as Radon-Nikodym densities of S_θ with respect to a common dominating measure υ on $\mathscr{Y}$. In particular, when Y is discrete and υ is a counting measure, $s_\theta(y)$ are probability mass functions, when Y is continuous and μ is a Lebesgue measure, $s_\theta(y)$ are probability density functions.

Owing to the independence assumption, the joint density $p_\theta(y^N)$ with respect to υ^N is simply

$$p_\theta(y^N) = \prod_{k=1}^{N} s_\theta(y_k). \tag{1}$$

The product can be expressed in a form more convenient for later approximation. We introduce an *empirical density* of observed data as

$$r_N(y) = \frac{1}{N} \sum_{k=1}^{N} \delta_{y_k}(y)$$

where δ_{y_k} is a Radon-Nikodym density with respect to the measure υ of a point-mass distribution concentrated at the point $\{y_k\}$. When Y is discrete, $\delta_{y_k}(y)$ is 1 for $y = y_k$ and 0 elsewhere. When Y is continuous, $\delta_{y_k}(y)$ is a Dirac function, i.e., $\delta_{y_k}(y) = 0$ for $y \neq y_k$ and $\int_{\mathscr{Y}} \delta_{y_k}(y)\, \upsilon(\mathrm{d}y) = 1$. Next we define *Kerridge inaccuracy* of r relative to s [10]

$$K(r{:}s) = -\int_{\mathscr{Y}} r(y) \log s(y)\, \upsilon(\mathrm{d}y).$$

With the two notions, the joint density can be rewritten as

$$p_\theta(y^N) = \exp\Big(-N\, K(r_N{:}s_\theta)\Big). \tag{2}$$

Remark 2.1 Inaccuracy $K(r\!:\!s)$ is closely linked with *Shannon entropy* of r [11]

$$H(r) = -\int_{\mathscr{Y}} r(y)\log r(y)\,\upsilon(\mathrm{d}y)$$

and *Kullback-Leibler (K-L) distance* of r and s [12]

$$D(r\|s) = \int_{\mathscr{Y}} r(y)\log\frac{r(y)}{s(y)}\,\upsilon(\mathrm{d}y).$$

Indeed, when Y is discrete, we have

$$K(r_N\!:\!s_\theta) = H(r_N) + D(r_N\|s_\theta). \tag{3}$$

An analogous formula does not hold for continuous Y. A formal evaluation gives $H(r_N) = -\infty$ and $D(r_N\|s_\theta) = \infty$. Yet, $K(r_N\!:\!s_\theta)$ is finite under a weak assumption that $s_\theta(y_k) > 0$ for $k = 1,\ldots,N$.

Remark 2.2 Given a particular sequence y^N, the joint density p_θ can be regarded as a function of the unknown parameters θ known as *likelihood*, $l_N(\theta) = p_\theta(y^N)$. Substituting (2) for p_θ, we have

$$K(r_N\!:\!s_\theta) = -\frac{1}{N}\log l_N(\theta). \tag{4}$$

Thus, maximizing likelihood is equivalent to minimizing inaccuracy

$$\arg\max_{\theta\in\Theta} l_N(\theta) = \arg\min_{\theta\in\Theta} K(r_N\!:\!s_\theta).$$

provided the extremum points exist. For discrete Y, it follows from (3) that maximizing likelihood is equivalent to minimizing K-L distance

$$\arg\max_{\theta\in\Theta} l_N(\theta) = \arg\min_{\theta\in\Theta} D(r_N\|s_\theta)$$

since $H(r_N)$ is independent of θ.

Remark 2.3 Owing to (3), inaccuracy can be regarded as a combined measure of uncertainty of Y. While $H(r_N)$ measures the intrinsic uncertainty of Y caused by its stochastic behaviour, $D(r_N\|s_\theta)$ quantifies the increase of uncertainty due to the use of a wrong distribution to predict Y. From the statistical point of view, inaccuracy (4) is a negative normalized log-likelihood. The minimum inaccuracy achievable within a class of densities s_θ is just a transformed value of the maximum likelihood over the class. In coding theory, inaccuracy gives an average length of code designed for s_θ rather than r_N. While $H(r_N)$ gives the minimum average code length achievable with r_N, $D(r_N\|s_\theta)$ measures the increase of an average length of code designed for s_θ (see Theorem 5.4.3 in [13]). The last interpretation links inaccuracy with the minimum description length principle [14].

Example 2.1 (Bernoulli distribution) Consider a simple model of coin tossing where $\mathscr{Y} = \{\mathrm{Head},\mathrm{Tail}\}$ and $s_\theta(y)$ is θ if $y = \mathrm{Head}$ and $1-\theta$ if $y = \mathrm{Tail}$. Let $\hat\theta_N$ be the relative frequency of heads observed in the sequence of trials y^N. The inaccuracy of the corresponding probability vectors is then

$$\begin{aligned}K(r_N\!:\!s_\theta) &= K([\hat\theta_N, 1-\hat\theta_N]\|[\theta,1-\theta])\\ &= H\big([\hat\theta_N,1-\hat\theta_N]\big) + D\big([\hat\theta_N,1-\hat\theta_N]\|[\theta,1-\theta]\big).\end{aligned}$$

It is not difficult to see that

$$\begin{aligned}&K(r_N\!:\!s_\theta) - K(r_N\!:\!s_{\hat\theta_N})\\ &\quad= D\big([\hat\theta_N,1-\hat\theta_N]\|[\theta,1-\theta]\big) \ge 0.\end{aligned}$$

Thus, $\theta = \hat\theta_N$ minimizes inaccuracy over Θ.

Example 2.2 (Normal distribution) Let Y be normally distributed with an unknown mean θ, $Y \sim \mathrm{N}(\theta,\sigma^2)$. Straightforward calculations yield

$$\begin{aligned}K(r_N\!:\!s_\theta) \;=\; &\frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2\sigma^2}V_N\\ &+\frac{1}{2\sigma^2}\big(\theta-\hat\theta_N\big)^2\end{aligned}$$

with $\hat\theta_N = E_N(Y)$, $V_N = E_N(Y^2) - E_N(Y)^2$ where $E_N(X) = \frac{1}{N}\sum_{k=1}^N X_k$ stands for the empirical mean of a random variable X. Because of the inequality

$$K(r_N\!:\!s_\theta) - K(r_N\!:\!s_{\hat\theta_N}) = \frac{1}{2\sigma^2}\big(\theta-\hat\theta_N\big)^2 \ge 0,$$

$\hat\theta^N$ is the minimum inaccuracy estimate of θ.

Dependent observations

Consider sequences of random variables

$$Y^N = (Y_1,\ldots,Y_N),\quad U^N = (U_1,\ldots,U_N)$$

with values in sets $\mathscr{Y}$ and $\mathscr{U}$, respectively. Suppose that the output values Y_k depend on past data U^k, Y^{k-1} only through a known vector function $Z_k = z(U^k, Y^{k-1}) \in \mathscr{Z}$. Let the distribution S_θ of Y_k given Z_k be parametrized by $\theta\in\Theta$. Let the distribution G of U_k given Y^{k-1}, U^{k-1} be independent of θ. We introduce densities $s_\theta = \frac{\mathrm{d}S_\theta}{\mathrm{d}\upsilon}$, $\theta\in\Theta$ and $g = \frac{\mathrm{d}G}{\mathrm{d}\mu}$ as Radon-Nikodym densities of S_θ and G with respect to corresponding dominating measures υ on $\mathscr{Y}$ and μ on $\mathscr{U}$, respectively. Let ζ be a common dominating measure for distributions considered on $\mathscr{Z}$.

By elementary rules of probability theory, the density $p_\theta(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)$ with respect to $\upsilon^N\times\mu^N$ is

$$\begin{aligned}&p_\theta(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)\\ &\quad= \prod_{k=m+1}^{N+m} s_\theta(y_k|z_k)\prod_{k=m+1}^{N+m} g(u_k|y^{k-1},u^{k-1}).\end{aligned} \tag{5}$$

Here m denotes the minimum number of samples for which z_{m+1} is defined. Thanks to the product form of (5), the θ-dependent part of it can be rewritten as follows. We introduce an *empirical density* of observed data

$$r_N(y,z) = \frac{1}{N}\sum_{k=m+1}^{N+m}\delta_{y_k,z_k}(y,z)$$

where δ_{y_k,z_k} is Radon-Nikodym density with respect to $\upsilon\times\zeta$ of a point mass concentrated at $\{(y_k\,,z_k)\}$. We define *conditional Kerridge inaccuracy* as

$$\bar{K}(r\!:\!s) = -\int_{\mathscr{Y}\times\mathscr{Z}} r(y,z)\log s(y|z)\,\upsilon(\mathrm{d}y)\,\zeta(\mathrm{d}z).$$

With these notions, the product of conditional sampling densities $s_\theta(y_k\,|z_k)$ can be put in the form

$$\prod_{k=m+1}^{N+m} s_\theta(y_k\,|z_k) = \exp\Big(-N\,\bar{K}(r_N\!:\!s_\theta)\Big). \qquad (6)$$

Remark 2.4 Again, there is a close connection between the conditional inaccuracy $\bar{K}(r\!:\!s)$, *conditional Shannon entropy* of r

$$\bar{H}(r) = -\int_{\mathscr{Y}\times\mathscr{Z}} r(y,z)\log\frac{r(y,z)}{r(z)}\,\upsilon(\mathrm{d}y)\,\zeta(\mathrm{d}z)$$

and *conditional K-L distance* of r and s

$$\bar{D}(r\|s) = \int_{\mathscr{Y}\times\mathscr{Z}} r(y,z)\log\frac{r(y,z)}{s(y\,|z)\,r(z)}\,\upsilon(\mathrm{d}y)\,\zeta(\mathrm{d}z)$$

where $r(z) = \int_{\mathscr{Y}} r(y,z)\,\upsilon(\mathrm{d}y)$ is a marginal density. When Y is discrete, the three quantities are related by

$$\bar{K}(r_N\!:\!s_\theta) = \bar{H}(r_N) + \bar{D}(r_N\|s_\theta). \qquad (7)$$

An analogous formula does not hold for continuous Y since $\bar{H}(r) = -\infty$ and $\bar{D}(r\|s) = \infty$ then. Yet, $\bar{K}(r\!:\!s)$ is finite provided that $s_\theta(y_k\,|z_k) > 0$ for $k = m{+}1,\ldots,N{+}m$.

Remark 2.5 Given particular sequences y^{N+m} and u^{N+m}, the density p_θ can be regarded as a function of the unknown parameters θ, i.e., *likelihood* $l_N(\theta) = p_\theta(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m, u^m)$. Substituting (6) for p_θ gives

$$\bar{K}(r_N\!:\!s_\theta) = -\frac{1}{N}\log l_N(\theta) + c \qquad (8)$$

where c is a constant independent of θ. Thus, maximizing likelihood is equivalent to minimizing conditional inaccuracy

$$\arg\max_{\theta\in\Theta} l_N(\theta) = \arg\min_{\theta\in\Theta} \bar{K}(r_N\!:\!s_\theta)$$

provided the extremum points exist. For discrete Y, it follows from (7) that maximizing likelihood is equivalent to minimizing conditional K-L distance

$$\arg\max_{\theta\in\Theta} l_N(\theta) = \arg\min_{\theta\in\Theta} \bar{D}(r_N\|s_\theta)$$

since $\bar{H}(r_N)$ is independent of θ.

Remark 2.6 Most of what was said in Remark 2.3 applies to the case of dependent data straightforwardly. Note, however, that the conditional entropy $\bar{H}(r_N)$ depends now on the *structure* of a model defined by the choice of Z_k. When estimating simultaneously the structure and parameters of a model, we have thus to pay attention to both the entropy and K-L distance components of inaccuracy. One notable example is Akaike's information criterion [4] which follows essentially by taking expectation of (approximate) conditional inaccuracy.

Example 2.3 (Markov chain) Consider a simple model of "weather forecast" where Y_k denotes weather on k-th day, $Z_k = Y_{k-1}$, $\mathscr{Y} = \mathscr{Z} = \{\mathrm{Sunny}, \mathrm{Rainy}\}$, $s_\theta(y|z) = \theta$ if $y = z$ and $1-\theta$ if $y \neq z$. Thus θ stands for the probability of steady weather. Let σ_N be the relative frequency of sunny days followed by another sunny day, ρ_N be the relative frequency of rainy days followed by another rainy day, Σ_N be the relative frequency of sunny days followed by any day. Then

$$\begin{aligned}\bar{K}(r_N\!:\!s_\theta) &= -\Sigma_N\bar{K}([\sigma_N, 1-\sigma_N]\|[\theta,1-\theta]) \\ &\quad -(1-\Sigma_N)\bar{K}([\rho_N\,,1-\rho_N]\|[\theta,1-\theta]) \\ &= \bar{K}([\hat{\theta}_N, 1-\hat{\theta}_N]\!:\![\theta,1-\theta]) \\ &= H([\hat{\theta}_N, 1-\hat{\theta}_N]) + D([\hat{\theta}_N\,,1-\hat{\theta}_N]\|[\theta,1-\theta])\end{aligned}$$

where

$$\hat{\theta}_N = \Sigma_N\,\sigma_N + (1-\Sigma_N)\,\rho_N.$$

The obvious inequality

$$\begin{aligned}&\bar{K}(r_N\!:\!s_\theta) - \bar{K}(r_N\!:\!s_{\hat{\theta}_N}) \\ &\quad = D\big([\hat{\theta}_N\,,1-\hat{\theta}_N]\|[\theta,1-\theta]\big) \geq 0.\end{aligned}$$

implies that $\theta = \hat{\theta}_N$ minimizes inaccuracy over Θ.

Example 2.4 (ARX model) Let $Y_k|Z_k$ be normally distributed with the conditional mean $\theta^T Z_k$, $Y_k \sim \mathrm{N}(\theta^T Z_k\,,\sigma^2)$. Provided $E_N\big(ZZ^T\big) > 0$, we have

$$\begin{aligned}\bar{K}(r_N\!:\!s_\theta) \;=\; &\frac{1}{2}\log 2\pi + \frac{1}{2}\log\sigma^2 + \frac{1}{2\sigma^2}V_N \\ &+\frac{1}{2\sigma^2}\big(\theta-\hat{\theta}_N\big)^T C_N\big(\theta-\hat{\theta}_N\big)\end{aligned}$$

with

$$\begin{aligned}\hat{\theta}_N \;&=\; C_N^{-1}\,E_N(ZY) \\ V_N \;&=\; E_N(Y^2) - E_N(YZ^T)C_N^{-1}E_N(ZY) \\ C_N \;&=\; E_N(ZZ^T)\end{aligned}$$

where $E_N(X) = \frac{1}{N}\sum_{k=m+1}^{N+m} X_k$ stands for the empirical mean of a random variable X. Clearly, for every $\theta\in\Theta$

$$\begin{aligned}&\bar{K}(r_N\!:\!s_\theta) - \bar{K}(r_N\!:\!s_{\hat{\theta}_N}) \\ &\quad = \frac{1}{2\sigma^2}\big(\theta-\hat{\theta}_N\big)^T C_N\big(\theta-\hat{\theta}_N\big) \geq 0.\end{aligned}$$

For $\theta = \hat{\theta}^N$ inaccuracy achieves its minimum over Θ.

3. Pythagorean-like relations

Independent observations

Let the sample average of a given vector function $h\colon \mathscr{Y} \to \mathbb{R}^n$

$$\frac{1}{N}\sum_{k=1}^{N} h(y_k) = \bar{h}$$

be the only information available from observed data. We denote by $\mathscr{R}_{\bar{h}}$ the set of all densities $r(y)$ such that

$$\int_{\mathscr{Y}} r(y)\,h(y)\,\upsilon(\mathrm{d}y) = \bar{h}. \qquad (9)$$

Obviously, $r_N \in \mathscr{R}_{\bar{h}}$.

Consider an exponential family $\mathscr{S}_h$ composed of densities

$$s_\lambda(y) = s_0(y)\, \exp\Big(\lambda^T h(y) - \psi(\lambda)\Big) \qquad (10)$$

where $s_0(y)$ is a fixed density and $\psi(\lambda)$ is logarithm of the normalizing constant

$$\psi(\lambda) = \log \int_{\mathscr{Y}} s_0(y)\, \exp\Big(\lambda^T h(y)\Big)\, \upsilon(\mathrm{d}y).$$

Let Λ be the set of $\lambda \in \mathbb{R}^n$ such that $\psi(\lambda) < \infty$.

We say that $s_{\hat{\lambda}}(y)$ is a *h-projection* of $r_N(y)$ onto $\mathscr{S}_h$ if both densities give the same expectation of $h(y)$

$$\int_{\mathscr{Y}} \Big(r_N(y) - s_{\hat{\lambda}}(y)\Big) h(y)\, \upsilon(\mathrm{d}y) = 0. \qquad (11)$$

Clearly, $s_{\hat{\lambda}}$ lies in the intersection $\mathscr{R}_{\bar{h}} \cap \mathscr{S}_h$.

Theorem 3.1 *Let $s_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathscr{S}_h$. Then for every $r \in \mathscr{R}_{\bar{h}}$ and every $s_\lambda \in \mathscr{S}_h$*

$$K(r{:}\, s_\lambda) = K(r{:}\, s_{\hat{\lambda}}) + D(s_{\hat{\lambda}} \| s_\lambda). \qquad (12)$$

Proof. Combining (11), (9) and (10) we have

$$\int_{\mathscr{Y}} \Big(r(y) - s_{\hat{\lambda}}(y)\Big) \log \frac{s_{\hat{\lambda}}(y)}{s_\lambda(y)}\, \upsilon(\mathrm{d}y) = 0.$$

The proposition follows by definitions of inaccuracy and K-L distance. □

A h-projection is a solution to two closely related optimization problems. To show it, we need only the following fundamental property of K-L distance (see, e.g., Theorem 2.6.3 in [13]).

Lemma 3.1 *For any two densities $r(y)$ and $s(y)$, it holds $D(r\|s) \geq 0$. The equality occurs if and only if $r(y) = s(y)$ almost everywhere with respect to υ.*

Corollary 3.1 *Let $s_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathscr{S}_h$. Then for every $r \in \mathscr{R}_{\bar{h}}$*

$$K(r{:}\, s_{\hat{\lambda}}) = \min_{\lambda \in \Lambda} K(r{:}\, s_\lambda). \qquad (13)$$

Proof. Theorem 3.1 and Lemma 3.1 together imply

$$K(r{:}\, s_{\hat{\lambda}}) = K(r{:}\, s_\lambda) - D(s_{\hat{\lambda}} \| s_\lambda) \leq K(r{:}\, s_\lambda)$$

for every $\lambda \in \Lambda$ with equality if and only if $s_\lambda(y) = s_{\hat{\lambda}}(y)$ almost everywhere with respect to υ. □

With regard to Remark 2.2, the minimum inaccuracy estimate $\hat{\lambda}$ is the maximum likelihood estimate for the family $\mathscr{S}_h$. Note that $K(r{:}\, s_{\hat{\lambda}})$ is independent of θ.

Corollary 3.2 *Let $s_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathscr{S}_h$. Then for every $s_\lambda \in \mathscr{S}_h$*

$$D(s_{\hat{\lambda}} \| s_\lambda) = \min_{r \in \mathscr{R}_{\bar{h}}} D(r\|s_\lambda) \qquad (14)$$

Proof. Theorem 3.1 implies

$$D(s_{\hat{\lambda}} \| s_\lambda) = K(r{:}\, s_\lambda) - K(r{:}\, s_{\hat{\lambda}}).$$

Restricting to $r \in \mathscr{R}_{\bar{h}}$ such that $D(r\|s_\lambda) < 0$, we have by (3)

$$D(s_{\hat{\lambda}} \| s_\lambda) = D(r\|s_\lambda) - D(r\|s_{\hat{\lambda}}).$$

Therefore, by Lemma 3.1

$$D(s_{\hat{\lambda}} \| s_\lambda) \leq D(r\|s_\lambda)$$

with equality if and only if $r(y) = s_{\hat{\lambda}}(y)$ almost everywhere with respect to υ. □

A h-projection $s_{\hat{\lambda}}$ is thus also a minimum K-L distance (or generalized maximum entropy) estimate of $r \in \mathscr{R}_{\bar{h}}$ relative to s_λ.

Corollary 3.3 *Let $s_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathscr{S}_h$. Then for every $\lambda \in \Lambda$, the likelihood value satisfies*

$$l_N(\lambda) = l_N(\hat{\lambda})\, \exp\Big(-N\, D(s_{\hat{\lambda}} \| s_\lambda)\Big) \qquad (15)$$

Proof. The proposition follows by taking together the definition $l_N(\lambda) = p_\lambda(y^N)$, (2) and Theorem 3.2. □

Dependent observations

Let the sample average of a given vector function $h{:}\, \mathscr{Y} \times \mathscr{Z} \to \mathbb{R}^n$

$$\frac{1}{N} \sum_{k=m+1}^{N+m} h(y_k, z_k) = \bar{h}$$

be the only information available from observed data y^{N+m}, u^{N+m}. Let $\mathscr{R}_{\bar{h}}$ be the set of $r(y,z)$ such that

$$\int_{\mathscr{Y} \times \mathscr{Z}} r(y,z)\, h(y,z)\, \upsilon(\mathrm{d}y)\, \zeta(\mathrm{d}z) = \bar{h}. \qquad (16)$$

Obviously, $r_N \in \mathscr{R}_{\bar{h}}$.

Consider an exponential family $\mathscr{Q}_h$ composed of densities

$$q_\lambda(y,z) = s_0(y|z)\, w(z)\, \exp\Big(\lambda^T h(y,z) - \psi(\lambda)\Big) \qquad (17)$$

where $s_0(y|z)$ and $w(z)$ are fixed densities and $\psi(\lambda)$ is logarithm of the normalizing constant

$$\psi(\lambda) = \log \int_{\mathscr{Y} \times \mathscr{Z}} s_0(y|z)\, w(z)\, \exp\Big(\lambda^T h(y,z)\Big) \cdot \upsilon(\mathrm{d}y)\, \zeta(\mathrm{d}z).$$

Let Λ be the set of $\lambda \in \mathbb{R}^n$ such that $\psi(\lambda) < \infty$.

One can easily compute the conditional density

$$s_\lambda(y|z) = s_0(y|z)\, \exp\Big(\lambda^T h(y,z) - \psi(\lambda,z)\Big) \qquad (18)$$

and the marginal density

$$w_\lambda(z) = w(z)\, \exp\Big(\psi(\lambda,z) - \psi(\lambda)\Big) \qquad (19)$$

where

$$\psi(\lambda, z) = \log \int_{\mathscr{Y}} s_0(y|z)\, \exp\Big(\lambda^T h(y,z)\Big)\, \upsilon(\mathrm{d}y).$$

Assume that even the marginal density (19) is exponential, i.e., $\psi(\lambda, z)$ can be factorized as follows

$$\psi(\lambda, z) = \eta(\lambda)^T h^*(z) + \eta_0(\lambda). \tag{20}$$

Assume, moreover, that the functions $h_i^*(z)$ are linear combinations of functions $h_j(y,z)$, i.e.,

$$\mathrm{span}\{h_i^*(z)\} \subset \mathrm{span}\{h_j(y,z)\}. \tag{21}$$

It can be ensured by adding $h_i^*(z)$ to $h_j(y,z)$ if necessary.

We say that $q_{\hat\lambda}(y,z)$ is a *h-projection* of $r_N(y,z)$ onto $\mathscr{Q}_h$ if r_N and $q_{\hat\lambda}$ give the same expectation of $h(Y,Z)$

$$\int_{\mathscr{Y}\times\mathscr{Z}} \Big(r_N(y,z) - q_{\hat\lambda}(y,z)\Big)\, h(y,z)\, \upsilon(\mathrm{d}y)\, \zeta(\mathrm{d}z) = 0. \tag{22}$$

Clearly, $q_{\hat\lambda}$ lies in the intersection $\mathscr{R}_{\bar h} \cap \mathscr{Q}_h$. Owing to the assumption (21), relation (22) implies also

$$\int_{\mathscr{Z}} \Big(r_N(z) - w_{\hat\lambda}(z)\Big)\, h^*(z)\, \zeta(\mathrm{d}z) = 0. \tag{23}$$

In other words, $w_{\hat\lambda}(z)$ is a h^**-projection* of $r_N(z)$ onto $\mathscr{W}_{h^*} = \{w_\lambda(z)\}$.

We define conditional K-L distance of s and s' given r as

$$\bar D(s\|s'|r) = \int_{\mathscr{Y}\times\mathscr{Z}} s(y|z)\, r(z) \log \frac{s(y|z)}{s'(y|z)}\, \upsilon(\mathrm{d}y)\, \zeta(\mathrm{d}z).$$

Lemma 3.2 *Let $s_\lambda(y|z)$, $s_{\lambda'}(y|z)$ be any two densities from $\mathscr{S}_h$. Under the assumptions (20)–(21),*

$$D(s_\lambda\|s_{\lambda'}|r) = D(s_\lambda\|s_{\lambda'}|r_N)$$

for every $r \in \mathscr{R}_{\bar h}$.

Proof. Substituting (18) for s_λ gives

$$D(s_\lambda\|s_{\lambda'}|r) = \int_{\mathscr{Z}} r(z)\, \Big((\lambda-\lambda')^T \hat h(\lambda, z) - \psi(\lambda, z) + \psi(\lambda, z)\Big)\, \zeta(\mathrm{d}z)$$

where

$$\hat h(\lambda, z) = \int_{\mathscr{Y}} s_\lambda(y|z)\, h(y,z)\, \upsilon(\mathrm{d}y).$$

One easily verifies that $\hat h(\lambda, z) = \nabla_\lambda \psi(\lambda, z)$. The proposition follows then by substituting (20) for $\psi(\lambda, z)$ and taking (21) into account. □

Theorem 3.2 *Let $s_{\hat\lambda}(y|z)\, w_{\hat\lambda}(z)$ be a h-projection of $r_N(y,z)$ onto $\mathscr{Q}_h$. Then, for every $r \in \mathscr{R}_{\bar h}$ and every $s_\lambda \in \mathscr{S}_h$, the following holds independently of w*

$$\bar K(r{:}s_\lambda) = \bar K(r{:}s_{\hat\lambda}) + \bar D(s_{\hat\lambda}\|s_\lambda|r). \tag{24}$$

Proof. By Theorem 3.1, the h-projection (22) satisfies the joint Pythagorean relation

$$K(r{:}s_\lambda w_\lambda) = K(r{:}s_{\hat\lambda} w_{\hat\lambda}) + D(s_{\hat\lambda} w_{\hat\lambda}\|s_\lambda w_\lambda).$$

Also by Theorem 3.1, the h^*-projection (23) satisfies the marginal Pythagorean relation

$$K(r{:}w_\lambda) = K(r{:}w_{\hat\lambda}) + D(w_{\hat\lambda}\|w_\lambda)$$

where r stands for the marginal density $r(z)$. Subtracting both equations, we get a conditional Pythagorean relation

$$\bar K(r{:}s_\lambda) = \bar K(r{:}s_{\hat\lambda}) + \bar D(s_{\hat\lambda}\|s_\lambda|w_{\hat\lambda}).$$

Finally, by Lemma 3.2,

$$\bar D(s_{\hat\lambda}\|s_\lambda|w_{\hat\lambda}) = \bar D(s_{\hat\lambda}\|s_\lambda|r)$$

for every $r \in \mathscr{R}_{\bar h}$. □

A h-projection solves two dual optimization problems again. To solve them, we need the following modification of Lemma 3.1.

Lemma 3.3 *For any $r(y,z)$ and $s(y|z)$, $\bar D(r\|s) \geq 0$ with equality if and only if $r(y,z) = s(y|z)\, r(z)$ almost everywhere with respect to $\upsilon \times \zeta$. Similarly, for any $s(y|z)$, $s'(y|z)$ and $r(z)$, $\bar D(s\|s'|r) \geq 0$ with equality if and only if $s(y|z)\, r(z) = s'(y|z)\, r(z)$ almost everywhere with respect to $\upsilon \times \zeta$.*

Corollary 3.4 *Let $s_{\hat\lambda} w_{\hat\lambda}$ be a h-projection of r_N onto $\mathscr{Q}_h$. Then for every $r \in \mathscr{R}_{\bar h}$*

$$\bar K(r{:}s_{\hat\lambda}) = \min_{\lambda\in\Lambda} \bar K(r{:}s_\lambda). \tag{25}$$

Proof. Theorem 3.2 and Lemma 3.3 together imply

$$\bar K(r{:}s_{\hat\lambda}) = \bar K(r{:}s_\lambda) - \bar D(s_{\hat\lambda}\|s_\lambda|r) \leq \bar K(r{:}s_\lambda)$$

for every $\lambda \in \Lambda$ where equality holds if and only if $s_\lambda(y|z)\, r(z) = s_{\hat\lambda}(y|z)\, r(z)$ almost everywhere with respect to $\upsilon \times \zeta$. □

With regard to Remark 2.5, the minimum conditional inaccuracy estimate $\hat\lambda$ is the maximum likelihood estimate for the family $\mathscr{S}_h = \{s_\lambda(y|z)\}$. Note that $\bar K(r{:}s_{\hat\lambda})$ is independent of θ.

Corollary 3.5 *Let $s_{\hat\lambda} w_{\hat\lambda}$ be a h-projection of r_N onto $\mathscr{Q}_h$. Then for every $s_\lambda \in \mathscr{S}_h$ and every $r \in \mathscr{R}_{\bar h}$*

$$\bar D(s_{\hat\lambda}\|s_\lambda|r) = \min_{\tilde r\in\mathscr{R}_{\bar h}} \bar D(\tilde r\|s_\lambda). \tag{26}$$

Proof. Theorem 3.2 implies

$$\bar D(s_{\hat\lambda}\|s_\lambda|r) = \bar K(r{:}s_\lambda) - \bar K(r{:}s_{\hat\lambda}).$$

Restricting to $r \in \mathscr{R}_{\bar h}$ such that $\bar D(r\|s_\lambda) < 0$, we have by (7)

$$\bar D(s_{\hat\lambda}\|s_\lambda|r) = \bar D(r\|s_\lambda) - \bar D(r\|s_{\hat\lambda}).$$

Owing to Lemma 3.3

$$\bar{D}(s_{\hat{\lambda}}\|s_\lambda|r) \le \bar{D}(r\|s_\lambda)$$

with equality if and only if $r(y,z) = s_{\hat{\lambda}}(y|z)\,r(z)$ almost everywhere with respect to $\upsilon\times\zeta$. The proposition follows by Lemma 3.2. □

A h-projection $s_{\hat{\lambda}}$ is thus also a minimum conditional K-L distance (generalized maximum entropy) estimate of $r \in \mathscr{R}_{\bar{h}}$ relative to s_λ.

Corollary 3.6 *Let $s_{\hat{\lambda}}\,w_{\hat{\lambda}}$ be a h-projection of r_N onto $\mathscr{Q}_h$. Then for every $\lambda\in\hat{\Lambda}$, the likelihood value satisfies*

$$l_N(\lambda) = l_N(\hat{\lambda})\,\exp\Big(-N\,\bar{D}(s_{\hat{\lambda}}\|s_\lambda|r)\Big) \qquad (27)$$

where r is an arbitrary density from $\mathscr{R}_{\bar{h}}$.

Proof. It follows easily by combining the definition $l_N(\lambda) = p_\lambda(y_{m+1}^{N+m}, u_{m+1}^{N+m}|y^m,u^m)$, relation (6) and Theorem 3.2. □

4. Estimation with compressed data

Exponential family $\mathscr{S}_h = \{s_\lambda\}$
Suppose that the family of sampling distributions is exponential and we are to estimate the parameters λ. What Corollaries 3.3 and 3.6 say is that as far as we know the minimum inaccuracy (maximum likelihood) estimate $\hat{\lambda}$ of λ, the whole likelihood can be restored by evaluating (possibly conditional) K-L distance between $s_{\hat{\lambda}}$ and s_λ. Combined with Corollaries 3.2 and 3.5, we see that the target object is K-L distance

$$D(s_{\hat{\lambda}}\|s_\lambda) = \min_{r\in\mathscr{R}_{\bar{h}}} D(r\|s_\lambda)$$

in the case of independent data and conditional K-L distance

$$\bar{D}(s_{\hat{\lambda}}\|s_\lambda|r) = \min_{\tilde{r}\in\mathscr{R}_{\bar{h}}} \bar{D}(\tilde{r}\|s_\lambda),\ r\in\mathscr{R}_{\bar{h}}.$$

in the case of regression-type dependence. In both cases, the sample average (empirical mean) $\bar{h}$ carries sufficient information for exact restoration of the above functions.

General family $\mathscr{S} = \{s_\theta\}$
Even if $\mathscr{S}$ is not an exponential family or cannot be imbedded in an exponential family of sufficiently low dimension, Theorems 3.1 and 3.2 can be applied — separately for each particular density s_θ.

For independent data, choosing $s_0(y) = s_\theta(y)$ in (10) defines an exponential family $\mathscr{S}_{\theta,h}$ going through the point s_θ. By Corollary 3.3 and 3.2, we have

$$l_N(\theta) = l_N(\theta,\hat{\lambda})\,\exp\Big(-N\min_{r\in\mathscr{R}_{\bar{h}}} D(r\|s_\theta)\Big). \qquad (28)$$

Analogously, for regression-type dependence, choosing $s_0(y|z) = s_\theta(y|z)$ in (17) defines an exponential family $\mathscr{S}_{\theta,h}$ going through s_θ. By Corollary 3.6 and 3.5, we have

$$l_N(\theta) = l_N(\theta,\hat{\lambda})\,\exp\Big(-N\min_{r\in\mathscr{R}_{\bar{h}}} \bar{D}(r\|s_\theta)\Big). \qquad (29)$$

Compared with (28), K-L distance is replaced by a conditional one.

Note that $l_N(\theta,\hat{\lambda})$ is the maximum value of likelihood for the family $\mathscr{S}_{\theta,h}$. Its value depends on θ, but with carefully chosen functions h_i, its variation can be neglected, $l_N(\theta,\hat{\lambda}) \approx \text{const},\ \theta\in\Theta$. This suggests to approximate likelihood by the minimum K-L distance between $\mathscr{R}_{\bar{h}}$ and s_θ. Note that another argument for this conclusion can be found in large deviation theory [15].

5. References

[1] J. Wolfowitz, "The minimum distance method", *Ann. Math. Statist.*, vol. 28, pp. 75–88, 1957.

[2] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.

[3] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer, Dordrecht, 1989.

[4] H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automat. Control*, vol. 19, pp. 716–723, 1974.

[5] B. Hanzon, "A differential-geometric approach to approximate nonlinear filtering", in *Geometrization of Statistical Theory*, C. T. J. Dodson, Ed., pp. 219–224. ULDM Publications, Lancaster, England, 1987.

[6] A. A. Stoorvogel and J. H. van Schuppen, "System identification with information theoretic criteria", Report BS-R9513, CWI, Amsterdam, 1995.

[7] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*, vol. 53 of *Transl. of Math. Monographs*, Amer. Math. Soc., Providence, RI, 1982.

[8] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems", *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[9] S. Amari, *Differential-Geometrical Methods in Statistics*, vol. 28 of *Lecture Notes in Statistics*, Springer-Verlag, Berlin, second edition, 1990.

[10] D. F. Kerridge, "Inaccuracy and inference", *J. Roy. Statist. Soc. Ser. B*, vol. 23, pp. 284–294, 1961.

[11] C. E. Shannon, "A mathematical theory of communication", *Bell System Tech. J.*, vol. 26, pp. 379–423, 623–656, 1948.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency", *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[14] J. Rissannen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[15] R. Kulhavý, "A Kullback-Leibler distance approach to system identification", in *Preprints of the IFAC Symposium on Adaptive Systems in Control and Signal Processing*, Budapest, Hungary, 1995, pp. 55–66.