MAXIMUM LIKELIHOOD ESTIMATION OF THE CERTAIN MODEL OF CONDITIONAL INDEPENDENCE

M.I. Schlesinger
http://www.irtc.org.ua/image/

Abstract. The certain model of a conditional independence is investigated as well as the problem of its identifiability, when the certain subset of its parameter is hidden and not observable. The identification problem for such type of model is formulated as a special case of the well - known problem of non-supervised learning or self-learning. Commonly used procedures of nonsupervised learning are based on the maximum likelihood estimation of the statistical parameters of the model. These procedures perform the certain type of hill-climbing and converge to the local but not global maximum of the likelihood function. The main new result of the investigation consists in, that for the model under consideration there exists the algorithm of self-learning, that finds a global maximum of likelihood function though the likelihood function is not unimodal.

1 Description of the model

Let x, y and k be three parameters, which describe some object under investigation. The variables x, y and k take their values from some finite sets X, Y and K, the set K consisting of only two values: 1 и 2. The parameters x and y are observable parameters. Below they are referred to as features of the object. The parameter k is unobservable and is referred to as a state of the object. The three-tuple x, y and k is random and takes values from the set $X \times Y \times K$ in accordance with the probability distribution p^*: $X \times Y \times K \to R$. For any $x \in X$, $y \in Y$, $k \in K$ the number $p^*(x,y,k)$ means a probability of the situation, when the object's state is k and its features are equal x and y. These probabilities are assumed to exist though to be unknown.

The probabilities $p^*(x,y,k)$, $x \in X$, $y \in Y$, $k \in K$ define uniquely the marginal and conditional probabilities of the form

$$p^*_{XY}(x,y) = \sum_{k \in K} p^*(x,y,k) \;, \quad p^*_X(x) = \sum_{y \in Y} p^*_{XY}(x,y) \;, \quad p^*_Y(y) = \sum_{x \in X} p^*_{XY}(x,y) \;;$$

$$p^*_K(k) = \sum_{x \in X} \sum_{y \in Y} p^*(x,y,k) \;, \quad p^*_{XY/K}(x,y/k) = \frac{p^*(x,y,k)}{p^*_K(k)} \;,$$

$$p^*_{X/Y}(x/y) = \frac{p^*_{XY}(x,y)}{p^*_Y(y)} \;, \quad p^*_{Y/X}(y/x) = \frac{p^*_{XY}(x,y)}{p^*_X(x)}$$

and similarly other marginal and conditional probabilities.

One can see, that the lower indices in these functions' identifiers and list of arguments are the same. Because of that these lower indices will be omitted and it will not cause the ambiguous

understanding. For example, the designation $p^*(x,y\ /\ k)$ will be used instead of $p^*_{XY/K}(x,y\ /\ k)$ and so on.

A community of probabilities $p^*(x,y,k)$, $x \in X$, $y \in Y$, $k \in K$, will be called statistical model of the object under investigation. In this article the model is investigated for the case, when the model describes a conditional independence of the features, namely, when in the every state k the observable features do not depend one on another. Formally, we will consider the models, for which the equality

$$p^*(x,y\ /\ k) = p^*(x\ /\ k) \cdot p^*(y\ /\ k) \tag{1}$$

is valid. The conditional independence of features x and y under every fixed state k does not mean their unconditional independence. In no way the supposition (1) implies

$$p^*(x,y) = p^*(x) \cdot p^*(y)\ . \tag{2}$$

The conditional independence of observable features x and y of the object does not deny their dependence in general, but states only, that this dependence is carried out through the dependence on the unobservable state k of the object.

The aim of the article is to answer the question, whether the statistical model of the object of such type can be identified in the situation, when one can observe only the features of the object and never can observe its state. We would like know, in what degree it is possible to estimate consistently not only the probabilities $p^*(x,y)$ but the probabilities $p^*(x,y,k)$ too in the situation, when only the sequence of observations (x_1,y_1), (x_2,y_2), ..., (x_i,y_i), ... is available, which does not contain in itself an information about an occurrence of unobservable state. In other words, for the situation, when the state k of the object is never observed, it is necessary to restore its probability distribution $p^*(k)$ and to detect, how the observable features x and y depend on this state.

The second and not less important aim is to construct an algorithm, that fulfills such identification. We formulate this algorithm, as a special case of the more general algorithm for mixture estimation [1]. This more general algorithm was constructed, as a solver of some well-defined optimization problem. In general case the algorithm converges to some local maximum of the likelihood function. An important result of the present investigation consists in that for the model under investigation all local maxima are equivalent in some sense and because of that any local maximum is a global maximum too.

The model under investigation is probably the simplest model of conditional independence. The obtained results give a hope, that it will be possible to generalize them for the more complex models (at least for the case when $|K| > 2$). However, such generalization is hardly possible without essential additional efforts. As we will see, the analysis of unimodality even for the simplest case is not too easy.

2. Identifiability of the model

We need the following additional assumption, that will be called an assumption about existence of ideal representatives. Namely, it will be assumed, that one of the features, we will say a feature x, has such two values x_1 and x_2, that

$$p^*(x_1 / k=1) \neq 0,\ p^*(x_1 / k=2)=0,\ p^*(x_2 / k=1)=0,\ p^*(x_2 / k=2)\neq 0 \ . \tag{3}$$

The value x_1 will be called an ideal representative of the first state and x_2 is an ideal representative of the second state. It is supposed only, that such two values exist, but it is not supposed, that these two values are known. As to feature y the more weak assumption is made: it will be supposed only, that this feature depends on the state k, it means, that equality

$$p^*(y / k=1) \neq p^*(y / k=2) \tag{4}$$

holds at least for some value $y \in Y$.

And at last, quite natural is to suppose, that

$$p^*(k=1) \neq 0 \quad \text{и} \quad p^*(k=2)\neq 0 \ . \tag{5}$$

Under these assumptions the statistical model of object is identifiable completely. On the base of random but long enough sequence (x_1,y_1), (x_2,y_2), ..., (x_i,y_i), ... the probabilities $p^*(x,y)$ may be restored, and then such probabilities $p^*(x,y,k)$ may be calculated, that for every pair $(x,y) \in X \times Y$ the equality

$$p^*(x,y)=\sum_k p^*(x,y,k) \tag{6}$$

holds. We will show, that only two functions $p^*(x,y,k)$ exist, which satisfy the conditions (1), (3)-(6), and one of these functions can be obtained from another only by permutation of state values k. The following theorem proof formulates this idea more strictly.

Theorem 1. *For any community of numbers $p^*(x,y)$, $x \in X$, $y \in Y$, no more then two different communities $p^*(x,y,k)$, $x \in X$, $y \in Y$, $k \in K$, exist, which satisfy the condition*

$$p^*(x,y)=\sum_k p^*(x,y,k)\ , \tag{7}$$

$$\exists\, x \in X \left(p^*(x / k=1) \neq 0 \ \&\ p^*(x / k=2)=0 \right)\ , \tag{8}$$

$$\exists\, x \in X \left(p^*(x \;/\; k=1) = 0 \;\&\; p^*(x \;/\; k=2) \neq 0 \right) , \tag{9}$$

$$\exists\, y \in Y \left(p^*(y \;/\; k=1) \neq p^*(y \;/\; k=2) \right) , \tag{10}$$

$$p^*(k=1) \neq 0,\; p^*(k=2) \neq 0 , \tag{11}$$

$$\forall (x,y,k) \in X \times Y \times K \left(p^*(x,y,k) = p^*(k) \cdot p^*(x \;/\; k) \cdot p^*(y \;/\; k) \right) . \tag{12}$$

Moreover, if $p_1^*(x,y,k)$ *and* $p_2^*(x,y,k)$, $x \in X$, $y \in Y$, $k \in K$ *, are such two communities, then it is valid, that*

$$\forall x,y \left(p_1^*(x,y,k=1) = p_2^*(x,y,k=2) \right) , \tag{13}$$

$$\forall x,y \left(p_2^*(x,y,k=2) = p_1^*(x,y,k=1) \right) . \tag{14}$$

The considerations, which prove an identifiability of the model of a conditional independence when an unobservable parameter k takes only two values, can be certainly generalized for the case, when this parameter takes an arbitrary but beforehand defined number of values.

In the next section an algorithm of a model identification will be described and analyzed. Unfortunately, the problem will have been analyzed only for the case when $K = \{1,2\}$ and now no evident ways for a generalization of the obtained results to the more complex cases are known.

3. The problem formulation and ways for its solution

It will be natural to consider the problem as a special case of the mixture estimation problem. Really, there are two populations. The first of them is a population of pairs x, y observed under the condition, that object stays in the first state. A probability distribution in this population has a form $p(x \;/\; k=1) \cdot p(y \;/\; k=1)$. The second population is a population of these pairs, observed under the condition, that the object stays in the second state. In the second population a probability distribution is of the form $p(x \;/\; k=2) \cdot p(y \;/\; k=2)$. From these two population a mixture with weights $p(k=1)$ and $p(k=2)$ has been formed and so the new population with the probability distribution $p(k=1) \cdot p(x \;/\; k=1) \cdot p(y \;/\; k=1) + p(k=2) \cdot p(x \;/\; k=2) \cdot p(y \;/\; k=2)$ has been obtained. A sequence of independent observations (x_1, y_1), (x_2, y_2), ..., (x_m, y_m) from this mixed population is obtained. Using this sequence as a base, the probabilities $p(x \;/\; k=1)$, $p(y / k=1)$, $p(x \;/\; k=2)$, $p(y \;/\; k=2)$ are to be estimated, which describe the unmixed populations, i.e. the mixture components, as well as the probabilities $p(k=1)$ and $p(k=2)$, which define the weights of these components in the mixture. It is natural to formulate this task as a looking for such probabilities $p(x \;/\; k)$, $p(y \;/\; k)$, $p(k)$, $k = 1,2$, which maximize the

given sequence (x_1,y_1), (x_2,y_2), …, (x_m,y_m) of observations. It means, that the problem is reduced to the maximization of the function

$$F = \sum_i \log \sum_k p(k)\cdot p(x_i \ / k)\cdot p(y_i \ / k) \ . \tag{15}$$

This problem had been formulated in [1] in the more general form and the general algorithm for its solution had been described. The algorithm has the following form for the model under the present analysis.

Let us denote as $p^*(x,y)$ a number $\frac{m(x,y)}{m}$, where $m(x,y)$ shows, how many times the pair (x,y) occurs in the sequence (x_1,y_1), (x_2,y_2), ..., (x_m,y_m) of the observations. So the function (15) under maximization may be evidently written in the form

$$F = \sum_{x,y} p^*(x,y)\log \sum_k p_k \cdot p(x \ / k)\cdot p(y \ / k) \ . \tag{16}$$

The community of variables $(p(k),\ p(x \ / k),\ p(y \ / k),\ x \in X,\ y \in Y,\ k \in K)$ will be denoted as A . The algorithm begins to work with an arbitrary value of A^0 , and then constructs step by step the sequence A^0, A^1 , A^2 , ..., A^t ,... and so on. The following calculations are to be done to obtain A^{t+1} when the previous value A^t is already obtained:

$$\alpha_k(x,y) = \frac{p^t(k)\cdot p^t(x \ / k)\cdot p^t(y \ / k)}{\sum_{k\in K} p^t(k)\cdot p^t(x \ / k)\cdot p^t(y \ / k)},\ k \in K,\ x,y \in X \times Y \ ; \tag{17}$$

$$p^{t+1}(k) \ = \ \sum_{x,y} \alpha_k(x,y)\cdot p^*(x,y), \quad k \in K \ ; \tag{18}$$

$$p^{t+1}(x/k) \ = \ \frac{\sum_y a_k(x,y)\cdot p^*(x,y)}{p^{t+1}(k)}, \quad k \in K,\ x \in X \ ; \tag{19}$$

$$p^{t+1}(y/k) \ = \ \frac{\sum_x a_k(x,y)\cdot p^*(x,y)}{p^{t+1}(k)}, \quad k \in K,\ y \in Y \ . \tag{20}$$

These calculations, which transform A^t into A^{t+1} , will be denoted as T , so that $T(A^t) = A^{t+1}$, and a value A , that satisfies the equation $A = T(A)$, will be called a stable point of algorithm. In [2] it had been proved, that for any initial value A the sequence A, $T(A)$, $T^2(A)$,..., $T^t(A)$,... converges to some stable point of algorithm.

The following part of the report is devoted, firstly, to the proof of the fact, that the stability of the point coincides with the well-known necessary conditions for maximum of functions (15) and (16). This statement had been already proved in [1], but for the special case under the given investigation this statement can be proved much more simpler and, consequently, this statement becomes more convincing. Then it will be proved a new and much more important statement, that under certain additional conditions this necessary condition is sufficient too.

The expression (16) is a variable, that depends on $\left(|X|+|Y|+1\right)\cdot|K|$ variables $\left(p(k),\ k\in K\right),\ \left(p(x\ /\ k),\ x\in X, k\in K\right),\ \left(p(y\ /\ k),\ y\in Y, k\in K\right)$. This expression is to be maximized under the following $2\cdot|K|+1$ conditions

$$\sum_{k\in K} p(k)=1\ ,$$
$$\sum_{x\in X} p(x\ /\ k)=1,\ k\in K\ ,$$
$$\sum_{y\in Y} p(y\ /\ k)=1,\ k\in K\ .$$

The function by Lagrange for such conditioned maximization has a form

$$\Phi(A)=\sum_{x,y} p^*(x,y)\ \log\sum_k p(k)\cdot p(x\ /\ k)\cdot p(y\ /\ k)+$$
$$+\lambda\cdot\sum_{k\in K} p(k)+\sum_k \gamma_k\cdot\sum_{x\in X} p(x\ /\ k)+\sum_k \eta_k\sum_{y\in Y} p(y\ /\ k)\ ,$$

and necessary conditions of maximum are expressed by the following system of equations:

$$\left.\begin{array}{ll}
\dfrac{\partial\ \Phi(A)}{\partial\ p(k)}=\sum_{x,y}\dfrac{p^*(x,y)}{p(x,y)}\cdot p(x\ /\ k)\cdot p(y\ /\ k)+\lambda\ =0,\ \ k\in K\ , & \text{(a)}\\[2ex]
\dfrac{\partial\ \Phi(A)}{\partial\ p(x\ /\ k)}=\sum_{y}\dfrac{p^*(x,y)}{p(x,y)}\cdot p(k)\cdot p(y\ /\ k)+\gamma_k\ =0,\ \ x\in X,\ k\in K\ , & \text{(b)}\\[2ex]
\dfrac{\partial\ \Phi(A)}{\partial\ p(y\ /\ k)}=\sum_{x}\dfrac{p^*(x,y)}{p(x,y)}\cdot p(k)\cdot p(x\ /\ k)+\eta_k\ =0,\ \ y\in Y,\ k\in K\ , & \text{(c)}\\[2ex]
\sum_k p(k)=1\ , & \text{(d)}\\[2ex]
\sum_x p(x/k)=1,\ k\in K\ , & \text{(e)}\\[2ex]
\sum_y p(y\ /\ k)=1,\ k\in K\ , & \text{(f)}
\end{array}\right\}(21)$$

An interdependence between the condition (21) and the stability of the value A is expressed by the following theorem.

Theorem 2. *If for every* $k\in K\ \ p(k)\neq 0,$ *then the condition* $A=T(A)$ *and the system of equations (21) are equivalent.*

We have proved, that the stability of points is not only a necessary condition of maximum of a likelihood function (16), but in the certain sense a sufficient condition of its global maximum too. Namely, the condition of the stability of a point can be complemented with the certain easy recognizable conditions so, that after such complement they become sufficient for the global maximum. An exact formulation of this statement is given by the following theorem.

Theorem 3. *Let a function* $p^*(x,y)$ *is such, that such values* $p^*(k)$ *and functions* $p^*(x/k),\ p^*(y/k),\ k=1,2$ *, exist, that* $p^*(x,y)=\sum_k p^*(k)\cdot p^*(x/k)\cdot p^*(y/k)$ *; let numbers* $p(k)$ *and functions* $p(x/k),\ p(y/k),\ k=1,2$ *, satisfy the conditions*

$$\sum_x \frac{p^*(x,y)}{\sum_k p(k)\cdot p(x/k)\cdot p(y/k)}\cdot p(x/k)=1 \quad , \tag{22}$$

for every y *and* k *and*

$$\sum_y \frac{p^*(x,y)}{\sum_k p(k)\cdot p(x/k)\cdot p(y/k)}\cdot p(y/k)=1 \tag{23}$$

for every x *and* k *;*

at this case either for every x *,* y *and* k *the equalities* $p(x/k)=p(x)$ *and* $p(y/k)=p(y)$ *are fulfilled, or for any numbers* $p'(k)$ *and functions* $p'(x/k),\ p'(y/k)$ *the following inequality is valid:*

$$\sum_{x,y} p^*(x,y)\log\sum_k p(k)\cdot p(x/k)\cdot p(y/k)\ge\sum_{x,y} p^*(x,y)\log\sum_k p'(k)\cdot p'(x/k)\cdot p'(y/k)\ . \tag{24}$$

The theorem 3 makes our knowledge about algorithms for mixture identification essentially more deep and exact. Up today it was known only, that the global maximum is one of the possible stable points of the mixture identification algorithm. Virtually, such knowledge is of the same level, that was achieved in the first publication about the described here algorithm [1]. However such knowledge is sometimes insufficient for its convenient application in the firm belief, that the algorithm really gives the required results, because it was known from the very beginning, that generally the false stable points can occur, which do not coincide with the global maxima. As a result of just now reported investigation, the exhaustive description of all stable points is received. Namely, it has been clarified, that only those solutions $p(k),\ p(x/k),\ p(y/k)$ can be a stable points of the algorithm for which

a). either $p(k)=0$ for some k ;

b). or $p(x / k) = p(x)$ and $p(y / k) = p(y)$ for all x, y, k ;

c). or the solution $p(k), p(x / k), p(y / k)$ provides the global maximum of the likelihood function $\sum_{x,y} p^*(x,y) \cdot \log \sum_k p(k) \cdot p(x / k) \cdot p(y / k)$.

Consequently, if the decision of the algorithm does not satisfy the easily recognizable conditions a. or b. the obtained solution is a point of the global maximum of the likelihood function. This conclusion is a main positive result of the present investigation.

References

1. Schlesinger M.I. Interdependence of supervised and non-supervised learning in pattern recognition, Cybernetic.-1968, N 2, pp. 81-88, Kiev.
2. Schlesinger M.I. Mathematical tools for picture processing Kiev, Naukova Dumka, 1989. - 198 pages.
3. A.P.Dempster, N.M.Laird, D.B.Rubin, Maximum likelihood from in complete data via the EM algorithm (with Discussion). J.R.Statist. Soc.B. 39 (1977) pp. 1-38.