KYBERNETIKA — VOLUME 22 (1986), NUMBER 5

ESTIMATING INTERACTIONS IN BINARY DATA SEQUENCES

MARTIN JANŽURA

Gibbs random sequences with pair interactions, as defined in frame of statistical mechanics (cf. e.g. Föllmer [5] and Preston [9]), are used to form probability models for dependent binary data. The appropriate probability measure is uniquely determined by the vector of interactions which describe its dependence structure. An applicable method for estimating the interactions is developed, and properties of the obtained estimate are derived. Direct instruction for implementation is given and demonstrated by a numerical example.

1. INTRODUCTION

A sequence of binary data is supposed to be generated by a discrete time stochastic process assuming only values zero and one. On base of the given finite sequence we try to find distribution of the stochastic process.

This could be hardly done without any additional assumptions on the class of distributions under consideration within the problem. Generally, the assumptions should involve some kind of homogeneity and weak dependence. Gibbs random processes, studied in frame of statistical mechanics, seem to satisfy the conditions mentioned above. The homogeneity is ensured by time stationarity, and the dependence structure being described by pair interactions of some fixed finite range, the requirement of rather weak dependence is satisfied as well. Besides, the class of Gibbs random processes is wide enough, including e.g. both the i.i.d. and the Markovian cases.

The problem of finding the unknown distribution is transformed to problem of estimating the interactions, the role of which is in some sense similar to that of covariances within the time series theory.

Since no "good" direct method seems to be available, the interactions will be estimated via estimating probabilities of some suitably chosen sets. Then the estimate of interactions is obtained by numerical minimization of a given convex function.

After the basic definitions and results in Section 2 the estimate is constructed

and its properties are proved in Section 3. The following Section 4 contains several remarks to the implementation of the method. The numerical example in Section 5 shows, in addition, a possible application of the method to testing "quality" of some generators of pseudo-random numbers.

Most of the known theoretic results are adopted from Föllmer [5] and Preston [9], while all applications of these models in statistics up to now seem to be either hard if not impossible to be implemented (cf. Mase [7]) or considerable approximative (cf. Besag [1]).

2 PRELIMINARIES

Binary random sequence (*b.r.s.*) will be a probability measure μ defined on the product measurable space $(X^{\mathscr{Z}}, \mathscr{F}^{\mathscr{Z}})$ where $X = \{0, 1\}$ is the state space, $\mathscr{F} = \exp X$ is the σ-algebra of all its subsets, and $\mathscr{Z}$ denotes the set of all integers.

For $\mathscr{V} \subset \mathscr{Z}$ we denote by $Pr_{\mathscr{V}}: X^{\mathscr{Z}} \to X^{\mathscr{V}}$ the corresponding projection function. For the sake of brevity we shall write $x_{\mathscr{V}}$ instead of $Pr_{\mathscr{V}}(x_{\mathscr{Z}})$ for $x_{\mathscr{Z}} \in X^{\mathscr{Z}}$. Further, we shall write x instead of $x_{\mathscr{Z}}$, and simply x_n instead of $x_{\{n\}}$ for one-point subsets of $\mathscr{Z}$. For $x_{\mathscr{V}} \in X^{\mathscr{V}}$ we denote by $\bar{x}_{\mathscr{V}} = Pr_{\mathscr{V}}^{-1}(x_{\mathscr{V}}) \in \mathscr{F}^{\mathscr{Z}}$ the corresponding measurable cylinder.

Furthermore, we shall use short notation for the conditional distributions, i.e. $\mu(x_{\mathscr{V}} \mid y_{\mathscr{Z}\setminus\mathscr{V}})$ means $\mathsf{E}_{\mu}[I_{(x \in \bar{x}_{\mathscr{V}})} \mid Pr_{\mathscr{Z}\setminus\mathscr{V}}^{-1}(\mathscr{F}^{\mathscr{Z}\setminus\mathscr{V}})](y)$ for every $\mathscr{V} \subset \mathscr{Z}$, $x_{\mathscr{V}} \in X^{\mathscr{V}}$, $y \in X^{\mathscr{Z}}$, where I is used for the indicator function.

A b.r.s. μ is called *R-Markovian* $(R \geqq 1)$ if

$$\mu(x_n \mid x_{\mathscr{Z}\setminus\{n\}}) = \mu(x_n \mid x_{[n-R,n+R]\setminus\{n\}})$$

for every $x \in X^{\mathscr{Z}}$, $n \in \mathscr{Z}$.

An R-Markovian b.r.s. μ is *Gibbs b.r.s. with pair interactions* if

$$\mu(x_n \mid x_{[n-R,n+R]\setminus\{n\}}) = \frac{\exp\{x_n(U_0 + \sum_{i=1}^{R} U_i(x_{n+i} + x_{n-i}))\}}{1 + \exp\{U_0 + \sum_{i=1}^{R} U_i(x_{n+i} + x_{n-i})\}}$$

for every $x \in X^{\mathscr{Z}}$, $n \in \mathscr{Z}$, where $U = (U_0, \ldots, U_R) \in \mathscr{R}^{R+1}$ ($\mathscr{R}$ denotes the set of all reals) is $(R + 1)$-tuple of parametets called (*pair*) *interactions*. The number R is called *range* of the interactions. In the sequel, R being fixed, we shall denote $\mathscr{R}^{R+1}$ simply by $\mathscr{E}$.

For every $U \in \mathscr{E}$ there is *exactly one* Gibbs b.r.s. with pair interactions given by U (Theorem 3 in Dobrushin [2]). This uniquelly defined b.r.s. which will be denoted by μ_U is *stationary* (Proposition 5.4 in Preston [9]), i.e.

$$\mu_U T^{-1} = \mu_U ,$$

where T: $X^{\mathscr{Z}} \to X^{\mathscr{Z}}$ is the shift on $X^{\mathscr{Z}}$ defined through $T(x)_n = x_{n+1}$ for every $x \in X^{\mathscr{Z}}$, $n \in \mathscr{Z}$.

Moreover, according to Proposition 4.1 in Preston [9], the b.r.s. μ_U is *ergodic*, i.e. its restriction to σ-algebra of invariant sets assumes only values zero or one, i.e. if $\mu_U(F) > 0$ then $\mu_U(F) = 1$ for every $F \in \mathscr{S} = \{E \in \mathscr{F}^{\mathscr{Z}};\ T^{-1}E = E\}$.

For every $U \in \mathscr{E}$ the limit

$$\lim_{N\to\infty} N^{-1} \log \sum_{x_{[1,N]}} \exp\{\sum_{i=0}^{R} U_i(\sum_{j=1}^{N-i} x_j x_{j+i} + \sum_{j=N-i+1}^{N} x_j z_{j+1})\} = p(U)$$

exists (due to Proposition 4.24 in Föllmer [5]), where the $z_{[N+1,\infty)} \in X^{[N+1,\infty)}$ may be taken arbitrarily.

Following e.g. Mayer [8], Section I.2.1., we obtain

$$p(U) = R^{-1} \log \lambda_{\max}(M_U)\,,$$

where M_U is strictly positive-valued $(2^R \times 2^R)$-matrix with elements defined through the formula

$$M_U(x_{[1,R]}, z_{[1,R]}) = \exp\{\sum_{i=0}^{R} U_i(\sum_{j=1}^{R-i} x_j x_{j+i} + \sum_{j=R-i+1}^{R} x_j z_{j+i-R})\}$$

$$\text{for every} \quad x_{[1,R]}, z_{[1,R]} \in X^{[1,R]}\,,$$

and $\lambda_{\max}(M_U)$ is its uniquely defined strictly positive eigenvalue larger in absolute value than all other eigenvalues of the matrix M_U ($\lambda_{\max}(M_U)$ exists due to the well-known Perron-Frobenius theorem).

The latter formula for $p(U)$ shows, according to known theorems on analytic behaviour of matrices depending on parameters, that the function p: $\mathscr{E} \to \mathscr{R}$ is real analytic in all variables $U_0, \ldots, U_R$. Explicitly, denoting

$$C_i = \{x \in X^{\mathscr{Z}};\ x_0 \,.\, x_i = 1\} \in \mathscr{F}^{\mathscr{Z}} \quad \text{for} \quad i = 0, \ldots, R\,,$$

it holds

$$\frac{\partial p}{\partial U_i}(U) = \mu_U(C_i)\,,$$

and

$$\frac{\partial^2 p}{\partial U_i\, \partial U_j}(U) = \sum_{k=-\infty}^{\infty} [\mu_U(C_i \cap T^{-k}C_j) - \mu_U(C_i)\,\mu_U(C_j)]$$

(cf. Theorem 5.1 in Künsch [6]).

Furthermore, the function p is strictly convex (cf. Lemma 8.6 and Lemma 8.7 in Preston [9]) wherefrom the one-to-one correspondence between U and μ_U follows, and even strongly convex (cf. Dobrushin and Nahapetian [3]) wherefrom it especially follows that $\partial^2 p/\partial U_i\, \partial U_j(U^0) > 0$ for every $U^0 \in \mathscr{E}$.

3. ESTIMATION OF INTERACTIONS

A sequence $x_1, \ldots, x_n \in \{0, 1\}$ of binary data is now supposed to be generated by a stochastic process with distribution given by a Gibbs b.r.s. μ_{U^0} with pair interactions U^0 of some fixed finite range $R \geqq 1$. Considering the interactions as vector parameter, we have obtained a parameter estimation problem.

Let us define the transform $\Phi: \mathscr{E} \to \mathscr{E}$ given by

$$\Phi(U) = (\mu_U(C_0), \ldots, \mu_U(C_K))$$

for every $U \in \mathscr{E}$.

Proposition 3.1. The transform Φ is one-to-one.

Proof. Let $\Phi(U) = \Phi(\bar{U})$ hold. Then, according to the formula 4.25 in Föllmer [5], it follows

$$0 \leqq H(\mu_U \mid \mu_{\bar{U}}) + H(\mu_{\bar{U}} \mid \mu_U) = \sum_{i=0}^{R} (\bar{U}_i - U_i)(\mu_U(C_i) - \mu_{\bar{U}}(C_i)) = 0\,,$$

where $H(\cdot \mid \cdot)$ is the relative entropy rate (information gain) given by

$$H(\mu \mid \nu) = \lim_{n \to \infty} n^{-1} \mathsf{E}_\mu \left[\log \frac{\mu(\bar{x}_{[1,n]})}{\nu(\bar{x}_{[1,n]})} \right]$$

providing the expressions make sense and the limit exists. Therefore $U = \bar{U}$ due to Theorem 4.27 in Föllmer [5]. □

If we denote by $\mathsf{D}(U) = ((\partial \Phi_i / \partial U_j)(U))_{i,j=0}^{R}$ the matrix of the first partial derivatives of the partial functions constituting the transform Φ, it holds that

$$\mathsf{D}(U)_{i,j} = \frac{\partial^2 p}{\partial U_i \, \partial U_j}(U) \quad \text{for every} \quad i, j = 0, \ldots, R$$

is a continuous function due to the properties of the function p mentioned in Section 2. Moreover $\mathsf{D}(U)$ is positive definite matrix which follows also from the strong convexity of the function p.

Thus, we have proved that the transform Φ is so called *regular mapping* on $\mathscr{E}$. This especially yields that the image $\Phi(\delta)$ of every open $\delta \subset \mathscr{E}$ is again an open subset of $\mathscr{E}$.

The aim of introducing the transform Φ consists in the fact that the unknown parameter U^0 will be estimated via estimating the transformed parameter $\beta^0 = \Phi(U^0)$.

Let us define $\hat{\beta}_i^n = (n - i)^{-1} \sum_{j=1}^{n-i} x_j x_{j+i}$ for every $i = 0, \ldots, R$ and every $x_{[1,n]} \in \in \{0, 1\}^n$, $n \in \mathscr{X}$, $n > R$.

The estimate $\hat{\beta}^n = (\hat{\beta}_0^n, \ldots, \hat{\beta}_R^n)$ is *consistent* if $\hat{\beta}^n \to \beta^0$ a.s. $[\mu_{U^0}]$, and *asymptotically normal* if

$$\mathscr{L}(n^{1/2}(\hat{\beta}^n - \beta^0)) \Rightarrow \mathsf{N}_{R+1}(\mathbb{O}, \mathbb{V})\,, \quad \text{i.e.}$$

$n^{1/2}(\hat{\beta}^n - \beta^0)$ converges in distribution to $(R+1)$-dimensional normal distribution with zero vector of mean values and covariance matrix $\mathbb{V}$.

Theorem 3.1. For every $U^0 \in \mathscr{E}$ the estimate $\hat{\beta}^n$ is consistent and asymptotically normal with asymptotic covariance matrix given by $\mathsf{D}(U^0)$.

Proof. The consistency follows immediately from the well-known ergodic theorem due to which it holds

$$n^{-1} \sum_{i=1}^{n} f \circ T^i \to \mathsf{E}_\mu[f] \quad \text{a.s.} \quad [\mu]$$

for every ergodic μ and bounded measurable f. Thus, we may substitute for f subsequently all I_{C_i}, $i = 0, \ldots, R$, and since $\hat{\beta}_i^n = (n-i)^{-1} \sum_{j=1}^{n-i} I_{C_i} \cdot T^j$ holds, the consistency is proved.

The asymptotic normality follows after some easy rearrangements from the central limit theorem for one-dimensional Gibbs random fields (cf. e.g. Theorem 3 in Dobrushin and Tirozzi [4]). □

Now, the estimate $\hat{U}^n$ of the parameter $U^0 \in \mathscr{E}$ will be obtained by means of the inverse transform of the estimate $\hat{\beta}^n$. For $\hat{\beta}^n \notin \Phi(\mathscr{E})$, the probability of which tends to zero, we may define the estimate $\hat{U}^n$ arbitrarily.

Theorem 3.2. For every $U^0 \in \mathscr{E}$ the estimate $\hat{U}^n$ is consistent and asymptotically normal with asymptotic covariance matrix given by $\mathsf{D}(U^0)^{-1}$.

Proof. The statement of the theorem follows immediately from the properties of the transform Φ and known theorems. □

Corollary. For every $U^0 \in \mathscr{E}$ it holds

$$\mathscr{L}(n(\hat{U}^n - U^0)^{\mathrm{T}} \mathsf{D}(U^0)(\hat{U}^n - U^0)) \Rightarrow \chi^2_{R+1},$$

i.e. the asymptotic distribution on $n(\hat{U}^n - U^0)^{\mathrm{T}} \mathsf{D}(U^0)(\hat{U}^n - U^0)$ is the *chi-square* with $R+1$ degrees of freedom.

The proof is again an easy consequence of known limit theorems. □

4. IMPLEMENTATION

The method of estimation proposed in the previous section seems to be suitable enough from the point of view of the standard properties of the estimate. The crucial role within the method is played by the transform Φ. Its properties are known and serve to transfer properties of the estimate of the transformed parameter to the estimate of the original parameter. However, no explicit formula has been introduced which could enable us to calculate the inverse transformation. Thus, while implementing the method, we have to follow a slightly different way.

First, let us define for every fixed $\beta^0 \in \mathscr{E}$ the function

$$F_{\beta^0}\colon \mathscr{E} \to \mathscr{R} \quad \text{through the following formula}$$

$$F_{\beta^0}(U) = p(U) - \sum_{i=0}^{R} U_i \beta_i^0 \quad \text{for every } U \in \mathscr{E}\,.$$

Proposition 4.1. For every $U^0 \in \mathscr{E}$ it holds

$$F_{\Phi(U^0)}(U^0) = \min_{U \in \mathscr{E}} F_{\Phi(U^0)}(U)\,.$$

Proof. Cf. Theorem 4.27 in Föllmer [5]. □

Corollary. For every $\beta^0 \in \mathscr{E}$ it holds

$$\beta^0 = \Phi(U^0) \quad \text{iff} \quad F_{\beta^0}(U^0) = \min_{U \in \mathscr{E}} F_{\beta^0}(U)\,.$$

Proof. Let $F_{\beta^0}(U^0) = \min_{U \in \mathscr{E}} F_{\beta^0}(U)$. Then $0 = \partial F_{\beta^0}/\partial U_i = \Phi_i(U^0) - \beta_i^0$ for every $i = 0, \ldots, R$. Thus the sufficiency is proved, while the necessity follows immediately from the preceding proposition. □

Thus, following the corollary above, we shall minimize the function $F_{\hat{\beta}^n}$ to find $\hat{U}^n = \Phi^{-1}(\hat{\beta}^n)$. (Providing the minimum does not exist, some stopping rule in the minimization algorithm will give a result "better" than any arbitrary definition of $\hat{U}^n$ in this case.) The only problem might be with calculating the function p. But, the definition of $p(U)$ through the matrix M_U for every $U \in \mathscr{E}$ (cf. Sec. 2) is rather easy to be dealt with providing the range R is not too large. Thus, realizing this, we conclude that the method is, in fact, not difficult to be implemented.

Remark. The approach involving minimization of the function $F_{\hat{\beta}^n}$ may be viewed on as the estimation based on "minimum distance method" (cf. e.g. Vajda [10]). Let the given collection of observations generate some stationary "empirical b.r.s." $\hat{\mu}$. We shall look for the b.r.s. from the class of Gibbs b.r.s's with interactions of range R which is the closest one to $\hat{\mu}$ in sense of distance measured by the relative entropy rate (information gain) $H(\cdot \mid \cdot)$ as defined e.g. in the proof of Proposition 3.1. But, to minimize $H(\hat{\mu} \mid \mu_U)$ over $U \in \mathscr{E}$ means to minimize function $p(U) - \sum_{i=0}^{R} U_i \hat{\mu}(C_i)$ (cf. formula 4.25 in Föllmer [5]). Hence, it is not necessary to construct the empirical b.r.s. $\hat{\mu}$. The values $\hat{\mu}(C_0), \ldots, \hat{\mu}(C_R)$, for which we may substitute our estimate $\hat{\beta}^n$ (cf. Sec. 3), are sufficient for our purposes.

5. EXAMPLE

The introduced method will be now demonstrated with the aid of a rather simple example. We shall estimate interactions in sequences of binary data obtained from some generators of pseudo-random numbers. The role of interactions will be easily visible from the results which can be more or less expected.

The considered generator is given by the recurrent formula

$$y_k = (c \,.\, y_{k-1}) \text{ MOD } 2^N ,$$

and produces numbers from 0 to $2^N - 1$.

For every $k = 1, \ldots, n$ the binary datum x_k is given by checking which half of the interval the value y_k belongs to, i.e. $x_k = y_k \text{ DIV } 2^{N-1}$ (by DIV we mean the integer division).

Four generators with various $c = 3, 11, 67, 259$, respectively, and constant $N = 16$ were investigated. For every considered generator three simulations with various initial values y_0 were performed. Further, we fixed the number of observations $n = 1000$ and the range of intearctions $R = 3$.

Assuming the i.i.d. sample case with $\mu(0) = \mu(1) = 0.5$, all the interactions should be zero, i.e. $U^0 = (0, 0, 0, 0)$ and easy calculation shows that

$$\mathrm{D}(U^0)^{-1} = \begin{pmatrix} 52 & -16 & -16 & -16 \\ -16 & 16 & 0 & 0 \\ -16 & 0 & 16 & 0 \\ -16 & 0 & 0 & 16 \end{pmatrix}$$

is the asymptotic covariance matrix of the vector $n^{1/2}(\hat{U}^n - U^0)$.

This enables us to express the statistic

$$\chi^2 = n(\hat{U}^n - U^0)^{\mathrm{T}} \mathrm{D}(U^0)(\hat{U}^n - U^0) =$$

$$= 1000 \,.\, \tfrac{1}{4}[(\hat{U}_0 + \hat{U}_1 + \hat{U}_2 + \hat{U}_3)^2 + \tfrac{1}{4}((\hat{U}_1)^2 + (\hat{U}_2)^2 + (\hat{U}_3)^2)]$$

(writing simply $\hat{U}_i$ instead of $\hat{U}_i^{1000}$ for every $i = 0, \ldots, 3$), the distribution of which should be approximately χ_4^2.

Thus, we may test the U^0-hypothesis, comparing value of the statistic χ^2 with the corresponding quantile. For the standard 0·05 level the quantile is $\chi_4^2(0{\cdot}05) = 9{,}49$. In the following table the significant values of the statistics χ^2 are underlined.

	$c = 3$			$c = 11$		
$\hat{U}_0$	−1·29	−1·51	−1·18	−0·58	−0·14	−0·39
$\hat{U}_1$	1·30	1·56	1·37	0·30	0·22	0·63
$\hat{U}_2$	0·16	−0·11	−0·17	0·07	−0·01	−0·05
$\hat{U}_3$	−0·13	0·09	0·00	0·17	−0·11	−0·06
χ^2	109·35	153·84	119·35	8·18	4·58	29·49

	$c = 67$			$c = 259$		
$\hat{U}_0$	−0·21	0·04	−0·22	0·03	−0·24	−0·12
$\hat{U}_1$	−0·23	−0·10	0·19	0·06	0·00	−0·01
$\hat{U}_2$	0·07	0·00	0·03	0·03	0·02	0·08
$\hat{U}_3$	−0·03	−0·11	−0·04	−0·13	0·17	0·08
χ^2	3·72	1·75	2·96	1·46	2·46	1·02

One can deduce from the results that for small $c = 3$ the sequence is significantly non − i.i.d., but approximately Markovian. This tendency, however weakened,

remains more or less true for $c = 11$ as well. For $c = 67$ and $c = 259$, respectively, the i.i.d. hypothesis cannot be rejected, but for the latter case the longer distance interactions start to play more important role.

Now, let us write in one sequence all one-letter binary words followed by all two-letter ones, three-letter ones ... etc., for every length the words being written in the lexicographical ordering, i.e.

$$0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ \ldots$$

We have done so up to the length six and considered the obtained sequence to be a sequence of binary data. The method described above showed very good "random-like" properties of this deterministic sequence.

$\hat{U}_0$	0·05
$\hat{U}_1$	0·00
$\hat{U}_2$	−0·02
$\hat{U}_3$	−0·02
χ^2	0·04

6. CONCLUDING REMARK

The idea of the introduced method depends neither on the assumption of binary data nor on the assumption of pair interaction. The generalization is straightforward, only the notation and the expressions are much more complicated.

(Received June 28, 1985.)

REFERENCES

[1] J. E. Besag: Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Statist. Soc. Ser. *B 36* (1974), 192—236.

[2] R. L. Dobrushin: The description of the random field by its conditional distributions and its regularity conditions (in Russian). Teor. Veroyat. i Primenen. *12* (1968), 2, 201—229.

[3] R. L. Dobrushin and B. S. Nahapetian: Strong convexity of the pressure for lattice systems of classical statistical physics (in Russian). Teoret. Mat. Fiz. *20* (1974), 223—234.

[4] R. L. Dobrushin and B. Tirozzi: The central limit theorem and the problem of equivalence of ensembles. Comm. Math. Phys. *54* (1977), 174—192.

[5] H. Föllmer: On entropy and information gain in random fields. Z. Wahrsch. verw. Gebiete *26* (1973), 207—217.

[6] H. Künsch: Decay of correlations under Dobrushin's uniqueness condition and its applications. Comm. Math. Phys. *84* (1982), 207—222.

[7] S. Mase: Locally asymptotic normality of Gibbs models on a lattice. Adv. in Appl. Probab. *16* (1984), 3, 585—602.

[8] D. H. Mayer: The Ruelle-Araki Transfer Operator in Classical Mechanics. (Lecture Notes in Physics 123.) Springer-Verlag, Berlin 1980.

[9] C. Preston: Random Fields. (Lecture Notes in Math. 534.) Springer-Verlag, Berlin 1976.

[10] I. Vajda: A new general approach to minimum distance estimation. In: Trans. of the Ninth Prague Conference, ..., 1982, Academia, Prague 1983, pp. 103—112.

RNDr Martin Janžura, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.