Kybernetika

Martin Janžura

Marginal problem, statistical estimation, and Möbius formula

Kybernetika, Vol. 43 (2007), No. 5, 619--631

Persistent URL: http://dml.cz/dmlcz/135802

KYBERNETIKA — VOLUME *43* (2007), NUMBER 5, PAGES 619 – 631

MARGINAL PROBLEM, STATISTICAL ESTIMATION, AND MÖBIUS FORMULA

Martin Janžura

A solution to the marginal problem is obtained in a form of parametric exponential (Gibbs–Markov) distribution, where the unknown parameters are obtained by an optimization procedure that agrees with the maximum likelihood (ML) estimate. With respect to a difficult performance of the method we propose also an alternative approach, providing the original basis of marginals can be appropriately extended. Then the (numerically feasible) solution can be obtained either by the maximum pseudo-likelihood (MPL) estimate, or directly by Möbius formula.

Keywords: Gibbs distributions, maximum entropy, pseudo-likelihood, Möbius formula

AMS Subject Classification: 93E12, 62H12, 60G60

1. INTRODUCTION

In the present paper we address the so-called marginal problem, i. e. the problem of reconstruction of a joint (global) distribution from a collection of marginal (local) ones. To the contrary with some other approaches, where the problem is studied either by graphical or combinatorial reasoning, or by iterative computational algorithms (see, e. g., [10] or [11]), here the problem is studied, more-or-less, from the "statistical" point of view. The "input" information contained in the system of marginal (local) distributions is understood as an evidence, and the problem of finding the unknown joint distribution is re-formulated as a parameter estimation problem.

Namely, in order to find a unique representing joint distribution for the system, we employ the maximum entropy principle (MAXENT). Then, under some technical assumptions, the solution agrees with a parametric exponential (Gibbs) distribution as the most natural and convenient representative. The distribution is also Markovian with the neighborhood system induced by the system of marginals (Section 7). Thus the structure of the distribution is known but the parameters are given only implicitly. In order to fix the parameters, we have to solve the same task as within the problem of statistical estimation. In particular, the parameters are obtained by an optimization procedure that agrees with the maximum likelihood (ML) estimate (as if the marginals were obtained from data). But, as it is well known, under a

certain size of the model, any direct optimization method is unfeasible. Thus, for calculating parameters of the representing distributions in full generality we need to apply some simulation procedures, usually based on the Markov Chain Monte Carlo methods (MCMC, see Section 8).

Nowadays, a standard choice for statisticians is a substitution of the ML estimator for multidimensional models by a more suitable alternative estimator, usually the maximum pseudo-likelihood (MPL) one, which is numerically easily feasible. We show that within the marginal problem the MPL approach would lead to the true parameter as well (Section 9). Unfortunately, the formula for the MPL estimate involves marginals over larger sets of nodes, namely over the neighborhoods of particular nodes. Thus, for an easy calculation of the maximal entropy solution to the marginal problem by the MPL approach, we have, at first, to extend the original marginals to these larger sets, at least approximately.

But, as we show finally in Section 10, once having the needed extended marginals, we can also apply directly the combinatorial Möbius formula for direct evaluating the potentials of the Gibbs distributions, and these potentials are equal exactly to the unknown parameters.

For many topics of the present paper [11] or [14] are the basic references. For exponential distributions see [1] and [7] or, more generally, [3]. For stochastic gradient method see [15] or [14], for general MCMC simulations see [5]. The maximum pseudo-likelihood method was at first mentioned in [2], detailed treatment can be found, e. g., in [6]. For the marginal problem see, e. g., [10] and the references therein.

Personal remark. When I joined the Institute of Information Theory and Automation to pass my postgraduate studies, it was Dr. Albert Perez who was established as my supervisor. Soon, he suggested two research topics to me. The first one was the problem of *simplification of the dependence structure* [12], closely related to the subject of the present paper. Nevertheless, I chose the second one, namely the *Gibbs random fields*, which was still a rather new topic in those days, initiated only about ten years earlier by the pioneering work of Dobrushin [4] and others. But, and it was one of the reasons why we esteemed him, Dr. Perez was able to recognize and anticipate its relevance and importance for the future. Moreover, he was so generous that he decided to reserve the topic for me, which, of course, included continual concern, encouragement and stimulating discussions.

It is great pleasure for me that in this contribution I can demonstrate the tight inter-connection between both the topics, namely the significance of Gibbs distributions as the representatives for collections of prescribed marginals. And I am also really happy to find that Dr. Perez was following in his late work a very similar approach by his concepts of $\mathcal{M}$-construct and explicit expression [13].

2. BASIC DEFINITIONS

Let us consider a finite set S of indices (sites, variables, nodes), and the space of configurations

$$\mathcal{X}_S = \bigotimes_{s \in S} \mathcal{X}_s$$

where $\mathcal{X}_s$ is a finite state space for every $s \in S$. For every $V \subset S$ we denote by $\mathrm{Pr}_V : \mathcal{X}_S \to \mathcal{X}_V$ the projection onto the space $\mathcal{X}_V = \bigotimes_{s\in V} \mathcal{X}_s$, and by $\mathcal{B}_V = \sigma(\mathrm{Pr}_V)$ the σ-algebra of cylinder (local) sets.

Further, by $\mathcal{P}_V$ we denote the class of all probability measures on $\mathcal{B}_V$, and by $\mathcal{F}_V$ the class of all real-valued $\mathcal{B}_V$-measurable functions. (P_V can be alternatively understood as the set of probability measures on $\mathcal{X}_V$, and $\mathcal{F}_V$ as the set of functions on $\mathcal{X}_V$. We shall not distinguish these two modes.) For $P_V \in \mathcal{P}_V$ and $W \subset V$ we shall denote by $P_{V/W} \in \mathcal{P}_W$ its projection into the space $\mathcal{P}_W$, i. e., the corresponding marginal distribution. (Whenever no confusion may occur, we shall write directly P_W.) On the other hand, by $P_{A|B}$ for $A, B \subset S, A \cap B = \emptyset$, we denote the corresponding conditional distribution.

3. PROBLEM

Let us consider a system of (non-void) subsets $\mathcal{V} \subset \exp S$ and a collection of *marginal distributions*

$$\mathcal{Q} = \{Q_V\}_{V\in\mathcal{V}}$$

where

$$Q_V \in \mathcal{P}_V \quad \text{for every } V \in \mathcal{V}.$$

Let us denote

$$\mathcal{P}_\mathcal{Q} = \{P_S \in \mathcal{P}_S; P_{S/V} = Q_V \text{ for every } V \in \mathcal{V}\}.$$

If $\mathcal{P}_\mathcal{Q} \neq \emptyset$ we quote the collection $\mathcal{Q}$ as *strongly consistent*.

The problem to be solved now consists in finding a suitable (in the sense specified below) *representative*

$$\overline{P}_S \in \mathcal{P}_\mathcal{Q},$$

providing $\mathcal{Q}$ is strongly consistent.

4. MAXIMUM ENTROPY PRINCIPLE

Whenever $|\mathcal{P}_\mathcal{Q}| > 1$ we have to employ some additional criterion for selecting $\overline{P}_S$, which, in our case, will be the *maximum entropy principle (MAXENT)*. For a justification of such approach see, e. g., [9] as the standard reference.

Let us recall the formulas for the *entropy* and the *I-divergence*, respectively, namely

$$H(P) = \int -\log P \,\mathrm{d}P = \sum_{x_S\in\mathcal{X}_S} -\log P(x_S)\, P(x_S),$$

and

$$I(P|Q) = \int \log \frac{P}{Q}\,\mathrm{d}P = \sum_{x_S\in\mathcal{X}_S} \log \frac{P(x_S)}{Q(x_S)}\, P(x_S)$$

providing the terms are well defined. Otherwise, we can set $H(P) = 0$ and $I(P|Q) = = \infty$.

Thus, applying the MAXENT, we seek for

$$\overline{P}_S \in \mathrm{argmax}_{P_S\in\mathcal{P}_\mathcal{Q}} H(P_S)$$

or, more generally,

$$\overline{P}_S \in \operatorname{argmin}_{P_S \in \mathcal{P}_{\mathcal{Q}}} I(P_S|R_S)$$

where $R_S \in \mathcal{P}_S$ is some fixed reference probability measure.

For the sake of brevity, we shall deal directly with the first definition, which agrees with the latter one for uniform R_S.

5. MAXIMUM ENTROPY WITH LINEAR CONSTRAINTS

Primarily, let us formulate the solution in a more general framework. Let us consider a collection of statistics

$$\boldsymbol{f} = \{f_j\}_{j\in\mathcal{K}} \quad \text{with } |\mathcal{K}| < \infty,$$

where

$$f_j \in \mathcal{F}_S \quad \text{for every } j \in \mathcal{K}.$$

Moreover, in order to guarantee the basic *regularity (identifiability) condition*, we assume the system

$$1,\ \{f_j\}_{j\in\mathcal{K}}$$

to be linearly independent. (If we assume in addition, e. g., $f_j(\overline{x}_S) = 0$ for every $j \in \mathcal{K}$ and some fixed $\overline{x}_S \in \mathcal{X}_S$, we may omit the constant from the collection.)

Further, let us introduce the *exponential distribution* P_S^α given by

$$P_S^\alpha(x_S) = \exp\left\{\sum_{j\in\mathcal{K}} \alpha_j\, f_j(x_S) - c(\alpha)\right\}$$

where $\alpha = (\alpha_j)_{j\in\mathcal{K}} \in R^{\mathcal{K}}$ is a parameter, and

$$c(\alpha) = \log \sum_{x_S\in\mathcal{X}_S} \exp\left\{\sum_{j\in\mathcal{K}} \alpha_j\, f_j(x_S)\right\}$$

is the appropriate normalizing constant.

Now, thanks to the identifiability condition above, we have a one-to-one relation between the parameter α and the exponential distribution P_S^α. Namely, for $P_S^\alpha = P_S^\beta$ we have $\langle \alpha - \beta, \boldsymbol{f}\rangle = \text{const}$. Further, $c(\alpha)$ is obviously (by the Hölder inequality) convex function of $\alpha \in R^{\mathcal{K}}$, with the gradient $\nabla c(\alpha) = \int \boldsymbol{f}\, \mathrm{d}P_S^\alpha$ and the Hessian matrix $\nabla^2 c(\alpha) = \operatorname{cov}_{P_S^\alpha}(\boldsymbol{f}, \boldsymbol{f})$. Due to the identifiability condition it is also strictly and even strongly (with the positive definite Hessian matrix) convex.

Now, for a collection of constants $\boldsymbol{m} = \{m_j\}_{j\in\mathcal{K}}$ we denote

$$\mathcal{M}(\boldsymbol{m}, \boldsymbol{f}) = \left\{P_S \in \mathcal{P}_S;\, \int f_j\, \mathrm{d}P_S = m_j \text{ for every } j \in \mathcal{K}\right\}.$$

Proposition 1. Let $P_S^\alpha \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$. Then

$$H(P_S^\alpha) \geq H(P_S)$$

for every $P_S \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$ with the equality iff $P_S = P_S^\alpha$.

P r o o f . As it is well-known, we have

$$0 \leq I(P_S|P_S^\alpha) = c(\alpha) - \langle \alpha,\ \boldsymbol{m}\rangle - H(P) = H(P_S^\alpha) - H(P_S)$$

where the inequality turns into equality iff $P_S = P_S^\alpha$. □

Moreover, whenever $P_S^\alpha \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$ exists, it is given uniquely.

Proposition 2. Let $P_S^\alpha,\ P_S^\beta \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$. Then $\alpha = \beta$.

P r o o f . We observe

$$0 \leq I(P_S^\alpha|P_S^\beta) + I(P_S^\beta|P_S^\alpha) = \left\langle \beta - \alpha,\ \int \boldsymbol{f}\,\mathrm{d}P_S^\beta - \int \boldsymbol{f}\,\mathrm{d}P_S^\alpha \right\rangle = 0.$$

Hence $P_S^\alpha = P_S^\beta$, and, due to the identifiability condition, we have $\alpha = \beta$. □

Thus, we may conclude that whenever there exists the *exponential representative* $\overline{P}_S = P_S^{\overline{\alpha}} \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$ then it *satisfies the MAXENT.*

6. EXISTENCE

Let us consider the problem of existence

$$P_S^{\overline{\alpha}} \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$$

for some $\overline{\alpha} \in R^{\mathcal{K}}$. Thus, $\overline{\alpha}$ should be given implicitly as a solution of the system of equations

$$\int \boldsymbol{f}\,\mathrm{d}P^{\overline{\alpha}} = \boldsymbol{m}.$$

Due to the convex property of the normalizing constant $c(\alpha)$ as a function of α, the above condition is equivalent to the variational principle

$$\overline{\alpha} = \operatorname*{argmin}_{\alpha \in R^{\mathcal{K}}} \left\{c(\alpha) - \langle \alpha,\ \boldsymbol{m}\rangle\right\}.$$

We define the closed convex hull

$$C^{\boldsymbol{f}} = \overline{\mathrm{co}}\,\{\boldsymbol{f}(x_S); x_S \in \mathcal{X}_S\} \subset R^{\mathcal{K}},$$

and its (relative) interior $\mathrm{ri}\, C^{\boldsymbol{f}}$.

Then, directly by the definitions

$$\mathcal{M}(\boldsymbol{m}, \boldsymbol{f}) \neq \emptyset \quad \text{iff} \quad \boldsymbol{m} \in C^{\boldsymbol{f}}.$$

Proposition 3. The exponential representation $P^{\overline{\alpha}} \in \mathcal{M}(\boldsymbol{m}, \boldsymbol{f})$ exists for some $\overline{\alpha} \in R^{\mathcal{K}}$ iff $\boldsymbol{m} \in \operatorname{ri} C^{\boldsymbol{f}}$.

Proof. See [1] for the general result, or, e. g. [11], Theorem D.1. □

Remark. The condition of the above proposition is obviously equivalent to the following one:

$$\max_{x_S \in \mathcal{X}_S} \langle \alpha,\ \boldsymbol{f}(x_S) \rangle > \langle \alpha\, \boldsymbol{m} \rangle$$

for every $\alpha \in S_1 = \{\alpha \in R^{\mathcal{K}}; \|\alpha\| = 1\}$, i. e. $\boldsymbol{m}$ can be separated by a hyperplane from any face (external subset) of the convex set $C^{\boldsymbol{f}}$. This equivalence can also serve as a key for the proof of Proposition 3.

Anyhow, the condition $\boldsymbol{m} \in \operatorname{ri} C^{\boldsymbol{f}}$ is the *crucial condition for the existence* of the exponential representation.

7. APPLICATION TO THE MARGINAL PROBLEM

In order to apply the above results to the marginal problem we have to find a suitable collection of statistics $\boldsymbol{f}$ and a collection of constants $\boldsymbol{m}$ so that

$$\mathcal{P}_{\mathcal{Q}} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{f}).$$

Natural candidates for the statistics $\{f_j\}_{j \in \mathcal{K}}$ are the Dirac functions (indicators)

$$\mathcal{D}_{\mathcal{V}} = \{\delta_{x_V}\}_{x_V \in \mathcal{X}_V, V \in \mathcal{V}}$$

but these are apparently linearly dependent. Thus, we have to choose a reasonable basis.

Let us fix a configuration $0_S \in \mathcal{X}_S$. For $V \subset S$ we denote $\mathcal{X}_V^0 = \bigotimes_{v \in V} (\mathcal{X}_v \setminus \{0_v\})$. Further, we denote

$$\overline{\mathcal{V}} = \{W \subset S; \emptyset \neq W \subset V \quad \text{for some } V \in \mathcal{V}\}.$$

Now, we set

$$\mathcal{D}_{\overline{\mathcal{V}}}^0 = \{\delta_{x_W}\}_{x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}}.$$

Proposition 4. We have

i) $\{1,\ \mathcal{D}_{\overline{\mathcal{V}}}^0\}$ linearly independent,

ii) $\mathcal{D}_{\mathcal{V}} \subset \operatorname{Lin}(1,\ \mathcal{D}_{\overline{\mathcal{V}}}^0)$.

Proof. We shall omit the tedious calculations of the general proof. Let us only illustrate the terms for the special case of $S = \{1, 2\},\ \mathcal{V} = \{\{1, 2\}\}$. Then $\overline{\mathcal{V}} = \{\{1\}, \{2\}, \{1, 2\}\}$ and $\mathcal{D}_{\overline{\mathcal{V}}}^0 = \{\delta_{x_1}, \delta_{x_2}, \delta_{x_1 x_2}\}_{x_1 x_2 \in \mathcal{X}_{\{1,2\}}^0}$. Suppose

$$\alpha + \sum_{x_1 \in \mathcal{X}_{\{1\}}^0} \alpha_{x_1} \delta_{x_1} + \sum_{x_2 \in \mathcal{X}_{\{2\}}^0} \alpha_{x_2} \delta_{x_2} + \sum_{x_1 x_2 \in \mathcal{X}_{\{1,2\}}^0} \alpha_{x_1 x_2} \delta_{x_1 x_2} = 0.$$

Then by substituting $(0_1, 0_2)$ we obtain $\alpha = 0$, by substituting $(0_1, y_2)$ we obtain $\alpha_{y_2} = 0$ for every $y_2 \in \mathcal{X}^0_{\{2\}}$, etc. This proves i).

In order to prove ii) we observe

$$\delta_{0_1 x_2} = \delta_{x_2} - \sum_{x_1 \in \mathcal{X}^0_{\{1\}}} \delta_{x_1 x_2} \quad \text{for every} \quad x_2 \in \mathcal{X}^0_{\{2\}},$$

symmetrically for $\delta_{x_1 0_2}$, and

$$\delta_{0_1 0_2} = 1 - \sum_{x_1 \in \mathcal{X}^0_{\{1\}}} \delta_{x_1} - \sum_{x_2 \in \mathcal{X}^0_{\{2\}}} \delta_{x_2} + \sum_{x_1 x_2 \in \mathcal{X}^0_{\{1,2\}}} \delta_{x_1 x_2}.$$

□

From now, we shall understand $\boldsymbol{f} = \{f_j\}_{j \in \mathcal{K}} = \mathcal{D}^0_{\overline{\mathcal{V}}}$ with $\mathcal{K} = \bigcup_{W \in \overline{\mathcal{V}}} \mathcal{X}^0_W$, and consequently, we set

$$\boldsymbol{m} = \{m_{x_W}\}_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}}$$

where $m_{x_W} = Q_{V/W}(x_W)$ for some $V \supset W, V \in \mathcal{V}$. Obviously, due to the (strong) consistency of $\mathcal{Q}$, the above terms are well defined since $Q_{V_1/W} = Q_{V_2/W}$ if $W \subset V_1 \cap V_2$.

Then, provided the crucial condition $\boldsymbol{m} \in \operatorname{ri} C^{\boldsymbol{f}}$ is satisfied, we obtain by the MAXENT the exponential representative $\overline{P}_S = P^{\overline{\alpha}}_S$ in the form

$$P^{\overline{\alpha}}_S(y_S) \propto \exp\left\{ \sum_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}} \overline{\alpha}_{x_W} \delta_{x_W}(y_W) \right\}.$$

If we denote $U^{\overline{\alpha}}_W = \sum_{x_W \in \mathcal{X}^0_W} \overline{\alpha}_{x_W} \delta_{x_W}$, we have $U^{\overline{\alpha}}_W \in \mathcal{F}_W$ and we may write

$$P^{\overline{\alpha}}_S(y_S) \propto \exp\left\{ \sum_{W \in \overline{\mathcal{V}}} U^{\overline{\alpha}}_W(y_W) \right\}.$$

Thus, $P^{\overline{\alpha}}_S$ is the *Gibbs distribution* with the *potential* $U^{\overline{\alpha}} = \{U^{\overline{\alpha}}_W\}_{W \in \overline{\mathcal{V}}}$ (see, e. g., [14] for detailed treatment). Moreover, since

$$P^{\overline{\alpha}}_{\{s\}|S\setminus\{s\}}(y_{\{s\}}|y_{S\setminus\{s\}}) \propto \exp\left\{ \sum_{W \in \overline{\mathcal{V}}, W \ni \{s\}} U^{\overline{\alpha}}_W(y_W) \right\},$$

$P^{\overline{\alpha}}_S$ is also *Markovian* with the *neighborhood system* $\partial = \{\partial(s)\}_{s \in S}$ given by

$$t \in \partial(s) \quad \text{iff} \quad \{t, s\} \subset W \quad \text{for some} \quad W \in \overline{\mathcal{V}}.$$

Hence, the form and the structure of the solution $\overline{P}_S$ is known, and it only "remains" to *identify the unknown parameters* $\overline{\alpha}$.

8. PARAMETER IDENTIFICATION

Let us recall that the collection of parameters $\overline{\alpha}$ is given implicitly as a solution to the system of equations

$$\int \boldsymbol{f}\, \mathrm{d}P^{\overline{\alpha}} = \boldsymbol{m}.$$

or, equivalently, as

$$\overline{\alpha} = \operatorname*{argmin}_{\alpha \in R^{\mathcal{K}}} \left\{ c(\alpha) - \langle \alpha,\ \boldsymbol{m} \rangle \right\}.$$

where $\boldsymbol{m}$ and $\boldsymbol{f}$ are specified in the preceding section. Thus, for real computing we need a convenient numerical method. Any version of the most common Newton's method yields an iterative procedure in the form

$$\alpha^{(n+1)} = \alpha^{(n)} + \rho_n \left(\boldsymbol{m} - \int \boldsymbol{f}\, \mathrm{d}P^{\alpha^{(n)}} \right)$$

where ρ_n should be, in the optimal case, inverse to the Hessian matrix of the function $c(\alpha)$ at $\alpha^{(n)}$. It could be, if needed, substituted by some more simple term but, anyhow, each step of the procedure involves evaluating the expectation

$$\int \boldsymbol{f}\, \mathrm{d}P^{\alpha^{(n)}}$$

which is numerically hardly feasible for large S. Hence, the *stochastic gradient method* (cf. [15] or [14], Section 15.4) was introduced, consisting in substituting the "theoretical" term by its empirical counterpart

$$\int \boldsymbol{f}\, \widehat{\mathrm{d}P^{\alpha^{(n)}}} = \frac{1}{L} \sum_{\ell=1}^{L} \boldsymbol{f}(x_S^{(\ell)})$$

where $x_S^{(1)}, \ldots, x_S^{(L)}$ is a long enough sequence simulated with the distribution $P^{\alpha^{(n)}}$.

The Markov Chain Monte Carlo (MCMC) – or some similar method – can be used for the *simulation* (cf., e. g., [5] for a survey).

With an appropriate choice of ρ_n (cf. [15]) the procedure converges in the a.s. sense but, obviously, it is tedious, time consuming, and it may be unstable. On the other hand, let us emphasize that the exponential form distribution with local statistics is extremely well suited for the MCMC type simulations. Namely, e. g., the most common Metropolis–Hastings algorithm deals at every step with a ratio like

$$\frac{P_S^{\alpha}(x_S)}{P_S^{\alpha}(y_S)} = \exp\left\{ \sum_{j \in \mathcal{K}} \alpha_j\, f_j(x_S) - \sum_{j \in \mathcal{K}} \alpha_j\, f_j(y_S) \right\}$$

that does not involve the normalizing constant and, therefore, can be easily and rapidly evaluated (cf., e. g., [5] or [14]).

9. MAXIMUM PSEUDO–LIKELIHOOD

In principle, the above way of identifying the parameters α agrees with the statistical parameter estimation, namely the *maximum likelihood (ML)*, or, equivalently, the *minimum I-divergence* method. The only difference consists in the fact that within the statistical estimation the collection of constants $\boldsymbol{m}$ is given as the "evidence" obtained from observed data, in particular $m_{x_W} = \hat{P}_{S/W}(x_W)$ for every $x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$ where $\hat{P}_S$ is the empirical distribution. In order to avoid computational problems as indicated in the previous section, the ML approach is sometimes exchanged with the *maximum pseudo-likelihood (MPL)* one. The MPL estimate of the parameter α is given by the following formula:

$$\begin{aligned}\hat{\alpha} \in \operatorname{argmax}_{\alpha \in R^{\mathcal{K}}} \sum_{s \in S} \int \log P^{\alpha}_{\{s\}|S\setminus\{s\}}(y_{\{s\}}|y_{S\setminus\{s\}}) \mathrm{d}\hat{P}_S(y_S) \\ = \operatorname{argmin}_{\alpha \in R^{\mathcal{K}}} \sum_{s \in S} I(\hat{P}_{\{s\}|S\setminus\{s\}}|P^{\alpha}_{\{s\}|S\setminus\{s\}}).\end{aligned}$$

Evidently, the MPL estimate can be also understood as the *minimum conditional I-divergence* estimate.

Since every P^{α}_S, $\alpha \in R^{\mathcal{K}}$, is Markov with the neighborhood system $\partial = \{\partial(s)\}_{s \in S}$, for solving the above optimization problem we actually need to have marginal distributions

$$\left\{\widehat{P}_{S/\overline{\partial}(s)}\right\}_{s \in S}$$

where $\overline{\partial}(s) = \partial(s) \cup \{s\}$.

Let us illustrate how the MPL idea could be applied to our marginal problem. We still assume to have a strongly consistent system $\mathcal{Q} = \{Q_V\}_{V \in \mathcal{V}}$ so that $P^{\overline{\alpha}}_S \in \mathcal{P}_{\mathcal{Q}}$ for some $\overline{\alpha} \in R^{\mathcal{K}}$. Now, let us suppose we are able to extend the system $\mathcal{Q}$ consistently to a system $\mathcal{Q}^{\partial} = \{Q_{\overline{\partial}(s)}\}_{s \in S}$ so that $P^{\overline{\alpha}}_S \in \mathcal{P}_{\mathcal{Q}^{\partial}}$ as well.

Remark. Theoretically, such extension always exists. Namely, we could simply set $Q_{\overline{\partial}(s)} = P^{\overline{\alpha}}_{S/\overline{\partial}(s)}$ for every $s \in S$. On the other hand, the numerical evaluation is hardly feasible without actually having $P^{\overline{\alpha}}_S$ (which represents the final goal). But, fortunately, both the methods described below do not require absolutely precise values of the "input" marginals $\mathcal{Q}^{\partial} = \{Q_{\overline{\partial}(s)}\}_{s \in S}$. With some reasonable approximate values we obtain a reasonable approximation of the true parameter. See also the concluding remark.

Anyhow, under the above assumptions, we can now obtain the unknown parameter $\overline{\alpha} \in R^{\mathcal{K}}$ with the aid of the MPL approach. The statement is worth to be proved.

Proposition 5. Let $P^{\overline{\alpha}}_S \in \mathcal{P}_{\mathcal{Q}} \cap \mathcal{P}_{\mathcal{Q}^{\partial}}$. Then

$$\overline{\alpha} = \operatorname{argmax}_{\alpha \in R^{\mathcal{K}}} \sum_{s \in S} \int \log P^{\alpha}_{\{s\}|\partial(s)}\left(y_{\{s\}}|y_{\partial(s)}\right) \, \mathrm{d}\, Q_{\overline{\partial}(s)}(y_{\overline{\partial}(s)}).$$

P r o o f. Let α^* be the maximizer. Since

$$\sum_{s\in S} I\left(Q_{\{s\}|\partial(s)}\middle|P^{\alpha}_{\{s\}|\partial(s)}\right) \geq 0$$

with the equality iff $Q_{\{s\}|\partial(s)} = P^{\alpha}_{\{s\}|\partial(s)}$ for every $s \in S$, and $P^{\overline{\alpha}}_{S/\overline{\partial}(s)} = Q_{\overline{\partial}(s)}$ by the assumption, we have

$$P^{\alpha^*}_{\{s\}|\partial(s)} = P^{\overline{\alpha}}_{\{s\}|\partial(s)} \quad \text{for every } s \in S.$$

Then $P^{\alpha^*}_S = P^{\overline{\alpha}}_S$ by the Hammersley–Clifford identity:

$$P_S(x_S) = P_S(0_S) \cdot \prod_{s\in S} \frac{P_{\{s\}|\partial(s)}\left(x_{\{s\}}|0_{\partial(s)^+}, x_{\partial(s)^-}\right)}{P_{\{s\}|\partial(s)}\left(0_{\{s\}}|0_{\partial(s)^+}, x_{\partial(s)^-}\right)}$$

where $\partial(s)^- = \{t \in \partial(s); t \prec s\}$ with some fixed linear ordering $\prec$.

Thus $\alpha^* = \overline{\alpha}$ finally by the identifiability condition. □

Since the objective function is concave, the maximum can be obtained also as a solution of the *normal equations*, i. e.

$$m_{x_W} = \frac{1}{|W|} \sum_{s\in W} \left\{P^{\overline{\alpha}}Q\right\}^{(s)}_{\overline{\partial}(s)/W}(x_W)$$

for every $x_W \in \mathcal{X}^0_W$, $W \in \overline{\mathcal{V}}$, where

$$\left\{P^{\overline{\alpha}}Q\right\}^{(s)}_{\overline{\partial}(s)}(x_{\overline{\partial}(s)}) = P^{\overline{\alpha}}_{\{s\}|\partial(s)}\left(x_{\{s\}}|x_{\partial(s)}\right) \cdot Q_{\overline{\partial}(s)/\partial(s)}(x_{\partial(s)}).$$

Let us recall that within the ML approach we have simply $P^{\overline{\alpha}}_{S/W}(x_W)$ on the right hand side.

The main advantage of the MPL approach consists in dealing with the *local characteristics* $P^{\alpha}_{\{s\}|\partial(s)}$ which can be easily evaluated, and, whenever the size of the neighborhoods $\{\partial(s)\}_{s\in S}$ is reasonable, the problem can be *numerically solved rather directly*, without any stochastic algorithm. Moreover, like within the statistical estimation (cf., e. g., [7]), whenever the "input" marginals $\mathcal{Q}^{\partial} = \{Q_{\overline{\partial}(s)}\}_{s\in S}$ are close to the true marginals $\{P^{\overline{\alpha}}_{S/\overline{\partial}(s)}\}_{s\in S}$, then the solution exists and is close to the true parameter $\overline{\alpha}$. (See Remark above.)

10. MÖBIUS FORMULA

Nevertheless, with the information as assumed in the preceding section, there is much more straightforward method, given by *Möbius formula* (see, e. g., [14]), for identifying the parameters. Let us introduce the formula in a general form. We shall denote $\mathcal{S} = \exp S \setminus \{\emptyset\}$ and

$$\mathcal{U}^0 = \{U = (U_A)_{A\in\mathcal{S}}; U_A \in \mathcal{F}_A \quad \text{and} \quad U_A(x_A) = 0 \quad \text{for every} \quad x_A \in \mathcal{X}_A \setminus \mathcal{X}_A^0\}$$

Then $\mathcal{U}^0$ is the space of so-called *vacuum potentials* (see, e. g., [6]). For our purpose it is important that for $U = (U_A)_{A\in\mathcal{S}} \in \mathcal{U}^0$ each $U_A, A \in \mathcal{S}$ can be written as

$$U_A = \sum_{x_A\in\mathcal{X}_A^0} \alpha_{x_A}\delta_{x_A},$$

where $U_A(x_A) = \alpha_{x_A}$ for $x_A \in \mathcal{X}_A^0$ and $U_A(x_A) = 0$ otherwise (see also Section 7).

Proposition 6. (Möbius formula)

i) Let $\Phi \in \mathcal{F}_S$. If we set

$$U_A(x_A) = \sum_{B\subset A} (-1)^{|A\setminus B|} \left[\Phi(x_B, 0_{S\setminus B}) - \Phi(0_S)\right] \quad \text{for every } A \in \mathcal{S} \text{ and } x_A \in \mathcal{X}_A$$

then $U \in \mathcal{U}^0$ and

$$\Phi(x_S) = \Phi(0_S) + \sum_{A\in\mathcal{S}} U_A(x_A) \quad \text{for every } x_S \in \mathcal{X}_S.$$

ii) If

$$\Phi(x_S) = \text{const.} + \sum_{A\in\mathcal{S}} U_A(x_A) \quad \text{for every } x_S \in \mathcal{X}_S$$

where $U \in \mathcal{U}^0$ then

$$U_A(x_A) = \sum_{B\subset A} (-1)^{|A\setminus B|} \left[\Phi(x_B, 0_{S\setminus B}) - \Phi(0_S)\right] \quad \text{for every } A \in \mathcal{S} \text{ and } x_A \in \mathcal{X}_A.$$

P r o o f. The relations can be verified by direct substitutions. □

Now, let us apply the preceding statement ii) to the function

$$\Phi(x_S) = \log P_S^{\overline{\alpha}}(x_S)$$

with $\overline{\alpha} \in R^{\mathcal{K}}$ such that again $P_S^{\overline{\alpha}} \in \mathcal{P}_{\mathcal{Q}} \cap \mathcal{P}_{\mathcal{Q}^\partial}$. We obtain

$$\begin{aligned}
\overline{\alpha}_{x_W} &= \sum_{B\subset W} (-1)^{|A\setminus B|} \left[\log \frac{P_S^{\overline{\alpha}}(x_B, 0_{S\setminus B})}{P_S^{\overline{\alpha}}(0_S)}\right] \\
&= \sum_{B\subset W\setminus\{s\}} (-1)^{|A\setminus B|} \left[\log \frac{P_S^{\overline{\alpha}}(x_B, 0_{S\setminus B})}{P_S^{\overline{\alpha}}(x_{B\cup\{s\}}, 0_{S\setminus\{B\cup\{s\}\}})}\right] \\
&= \sum_{B\subset W\setminus\{s\}} (-1)^{|A\setminus B|} \left[\log \frac{P_{\{s\}|\partial(s)}^{\overline{\alpha}}(0_{\{s\}}|x_{B\cup\{s\}}, 0_{S\setminus\{B\cup\{s\}\}})}{P_{\{s\}|\partial(s)}^{\overline{\alpha}}(x_{\{s\}}|x_{B\cup\{s\}}, 0_{S\setminus\{B\cup\{s\}\}})}\right] \\
&= \sum_{B\subset W\setminus\{s\}} (-1)^{|A\setminus B|} \left[\log \frac{Q_{\{s\}|\partial(s)}(0_{\{s\}}|x_{B\cup\{s\}}, 0_{S\setminus\{B\cup\{s\}\}})}{Q_{\{s\}|\partial(s)}(x_{\{s\}}|x_{B\cup\{s\}}, 0_{S\setminus\{B\cup\{s\}\}})}\right]
\end{aligned}$$

for every $x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$, where $s \in W$ is arbitrary fixed.

Thus, whenever we are able to extend the original system of marginals $\mathcal{Q}$ into the system $\mathcal{Q}^\partial$, we can *calculate the parameters* $\overline{\alpha}$ *directly from the Möbius formula.* Actually, we do not need to know the complete distributions $Q_{\overline{\partial}(s)}, s \in S$, but only $Q_{\overline{\partial}(s)}(x_W, 0_{\overline{\partial}(s)\setminus W})$ for every $W \in \overline{\mathcal{V}}, W \subset \overline{\partial}(s)$, and $x_W \in \mathcal{X}_W$.

Remark (concluding). Obviously, whenever we are not able to calculate the extended marginals, we still can use some approximation (see, e. g., [10] or [8]) in order to obtain at least approximative solution $\hat{\alpha}$. The question of approximation is behind the scope of the present paper. But, anyhow, we may summarize the recommended procedure:

i) Seek for the solution of the marginal problem in the exponential form $P_S^{\overline{\alpha}}$.

ii) Extend the system $\mathcal{Q} = \{Q_V\}_{V \in \mathcal{V}}$ into $\hat{\mathcal{Q}}^\partial = \{\hat{Q}_{\overline{\partial}(s)}\}_{s \in S}$ by some approximative method.

ii) Calculate the (approximate) parameters of the exponential distribution with the aid of Möbius formula:

$$\hat{\alpha}_{x_W} = \sum_{B \subset W \setminus \{s\}} (-1)^{|A \setminus B|} \left[\log \frac{\hat{Q}_{\{s\}|\partial(s)}(0_{\{s\}}|x_{B \cup \{s\}}, 0_{S \setminus \{B \cup \{s\}\}})}{\hat{Q}_{\{s\}|\partial(s)}(x_{\{s\}}|x_{B \cup \{s\}}, 0_{S \setminus \{B \cup \{s\}\}})} \right]$$

for every $x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$, where $s \in W$ is arbitrary fixed.

ACKNOWLEDGEMENT

This research was supported by the Czech Science Foundation under grant No. 201/06/1323 and by the Research Center DAR (Project No. 1M0572 of the Ministry of Education, Youth and Sports of the Czech Republic).

(Received August 3, 2006.)

REFERENCES

[1] O. E. Barndorff-Nielsen: Information and Exponential Families in Statistical Theory. Wiley, New York 1978.
[2] J. Besag: Statistical analysis of non-lattice data. The Statistician *24* (1975), 179–195.
[3] I. Csiszár and F. Matúš: Generalized maximum likelihood estimates for exponential families. Probability Theory and Related Fields (to appear).
[4] R. L. Dobrushin: Prescribing a system of random variables by conditional distributions. Theor. Probab. Appl. *15* (1970), 458–486.
[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (eds.): Markov Chain Monte Carlo in Practice. Chapman and Hall, London 1996.
[6] M. Janžura: Asymptotic results in parameter estimation for Gibbs random fields. Kybernetika *33* (1997), 2, 133–159.

[7] M. Janžura: A parametric model for large discrete stochastic systems. In: Second European Conference on Highly Structured Stochastic Systems, Pavia 1999, pp. 148–150.
[8] M. Janžura and P. Boček: A method for knowledge integration. Kybernetika *34* (1988), 1, 41–55.
[9] E. T. Jaynes: On the rationale of the maximum entropy methods. Proc. IEEE *70* (1982), 939–952.
[10] R. Jiroušek and J. Vejnarová: Construction of multidimensional model by operators of composition: Current state of art. Soft Computing *7* (2003), 328–335.
[11] S. L. Lauritzen: Graphical Models. University Press, Oxford 1006.
[12] A. Perez: ε-admissible simplifications of the dependence structure of random variables. Kybernetika *13* (1979), 439–449.
[13] A. Perez and M. Studený: Comparison of two methods for approximation of probability distributions with prescribed marginals. Kybernetika *43* (2007), 5, 591–618.
[14] G. Winkler: Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer–Verlag, Berlin 1995.
[15] L. Younes: Estimation and annealing for Gibbsian fields. Ann. Inst. H. Poincaré *24* (1988), 2, 269–294.

Martin Janžura, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: janzura@utia.cas.cz