On an approximative solution to the marginal problem

Martin Janžura
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
e-mail:janzura@utia.cas.cz

Abstract

With the aid of the Maximum Entropy principle, a solution to the marginal problem is obtained in a form of parametric exponential (Gibbs-Markov) distribution. The unknown parameters can be calculated by an optimization procedure that agrees with the maximum likelihood estimate but it is numerically hardly feasible for highly dimensional systems. A numerically easily feasible solution can be obtained by the algebraic Möbius formula. The formula, unfortunately, involves terms that are not directly available but can be approximated. And the main aim of the present paper consists in this approximation.

1 Introduction

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., [6] or [7]), here the solution is inspired, more-or-less, by a "statistical" point of view.

In order to find a unique representing joint distribution for the system, we employ the maximum entropy principle. Then, providing some technical assumptions being satisfied, 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 5.). 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). Thus, we may imagine the "input" information contained in the system of marginal (local) distributions as an evidence, and the problem of finding the unknown joint distribution is re-formulated as a parameter estimation problem. But, as it is well known, under a certain size of the model, any direct optimization method is unfeasible. Therefore, for calculating parameters of the representing distributions in full generality we need to apply some simulation procedure, usually based on the Markov Chain Monte Carlo methods (see Section 6.).

But, as we show finally in Section 7., we can also apply the combinatorial Möbius formula for direct evaluating the potentials of the Gibbs distributions, and these potentials are equal exactly to the unknown parameters.

Unfortunately, the formula 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 , we have, at first, to extend the original marginals to these larger sets, at least approximately.

An approximative method of these extension, based partly on the ideas introduced in [6] or [7], is presented in Section 8. as the main goal of this paper.

For many topics of the present paper [7] or [9] are the basic references. For exponential distributions see [1]. For stochastic gradient method see [9] or [10], for general MCMC simulations see [2]. For the marginal problem see, e.g., [6] and the references therein. Some specific approaches can be found, e.g., in [4] or [8].

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. ($\mathcal{P}_V$ 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 $\mathcal{V} \subset \exp \mathrm{S}$, satisfying $V \setminus W \neq \emptyset$ for $V, W \in \mathcal{V}, V \neq W$, 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

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

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 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* . For a justification of such approach see, e.g., [5] 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 set $I(P|Q) = \infty$.

Thus, applying the maximum entropy principle , we seek for

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

or, more generally,

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

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, after all, agrees with the latter one for uniform R_S.

5 Gibbs-Markov distributions

Further, we shall quote $P_S \in \mathcal{P}$ as the *Gibbs distribution* with the *potential* $U = \{U_A\}_{A \in \mathcal{A}}$ where $U_A \in \mathcal{F}_A$ for every $A \in \mathcal{A} \subset \exp \mathrm{S}$ (see, e.g., [9] for detailed treatment) if

$$P_S(y_S) \propto \exp\{\sum_{A \in \mathcal{A}} U_A(y_A)\}.$$

Then we shall write $P_S = P_S^U$. Moreover, since

$$P^U_{\{s\}|S\setminus\{s\}}(y_{\{s\}}|y_{S\setminus\{s\}}) \propto \exp\{\sum_{A\in\mathcal{A}, A\ni s} U_A(y_A)\},$$

P_S^U is also *Markovian* with the *neighborhood system* $\partial = \{\partial(s)\}_{s\in S}$ given by

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

6 Maximum entropy solution

Now, let us fix a configuration $0_S \in \mathcal{X}_S$. For $B \subset S$ we denote $\mathcal{X}_B^0 = \bigotimes_{b\in B}(\mathcal{X}_b \setminus \{0_b\})$. Further, we denote

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

Let as consider the class of potentials

$$\mathcal{U}^0 = \{U = (U_W)_{W\in\overline{\mathcal{V}}};\, U_W \in \mathcal{F}_W \,\text{and}\, U_W(x_W) = 0 \,\text{for every}\, x_W \in \mathcal{X}_W \setminus \mathcal{X}_W^0\}.$$

Then $\mathcal{U}^0$ is the space of so-called *vacuum potentials* (see, e.g., [3]). We may also write $U_W = \sum_{x_W\in\mathcal{X}_W^0} U_W(x_W)\delta_{x_W}$ with some real constants $\{U_W(x_W)\}_{x_W\in\mathcal{X}_W^0}$ for every $W \in \overline{\mathcal{V}}$.

Proposition 1. Let $P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$ for some $\overline{U} \in \mathcal{U}^0$. Then $\overline{U} \in \mathcal{U}^0$ is given uniquely and

$$P_S^{\overline{U}} = \overline{P}_S = \text{argmax}_{P_S\in\mathcal{P}_\mathcal{Q}} H(P_S).$$

Proof. See Proposition 1 and 2 in [3]. □

Remark 1. We shall omit here the question of existence of $P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$ (see also, e.g., [3]). Here it will be simply assumed. Let us emphasize that such assumption involves also the condition

$$Q_V > 0 \quad \text{for every } V \in \mathcal{V}.$$

That will make the further calculations much easier since we do not have to take care about zeros. □

We shall rather discuss the problem of numerical feasibility. Namely, the unknown parameters $\{\overline{U}_W(x_W)\}_{x_W\in\mathcal{X}_W^0, W\in\overline{\mathcal{V}}}$ should be identified by the condition

$$P_S^{\overline{U}} \in \mathcal{P}_\mathcal{Q}$$

which means

$$P_W^{\overline{U}}(x_W) = Q_W(x_W) \quad \text{for every} \quad x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$$

or, equivalently, by

$$\overline{U} = \arg\max_{U \in \mathcal{U}^0} \left[\sum_{W \in \overline{\mathcal{V}}} \sum_{x_W \in \mathcal{X}_W^0} U_W(x_W) Q_W(x_W) - \log \sum_{x_S \in \mathcal{X}_S} \exp\{ \sum_{W \in \overline{\mathcal{V}}} U_W(x_W)\} \right].$$

Both methods contain terms that involve summing over the set $\mathcal{X}_S$ which is numerically hardly feasible for large S.

Hence, the *stochastic gradient method* (cf. [9], Section 15.4, or [10]) was introduced, based on substituting the "theoretical" terms by their simulated counterparts. The Markov Chain Monte Carlo (MCMC) – or some similar method – can be used for the *simulation* (cf., e. g., [2] for a survey).

Let us recall here that, in principle, the above method of identifying the parameters 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 $\{Q_W(x_W)\}_{x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}}$ is given as an "evidence" obtained from observed data, in particular $Q_W(x_W) = \widehat{P}_{S/W}(x_W)$ for every $x_W \in \mathcal{X}_W^0, W \in \overline{\mathcal{V}}$ where $\widehat{P}_S$ is the empirical distribution.

7 Möbius formula

Nevertheless, due to the problems as described above, we prefer much more straightforward method, given by *Möbius formula* (see, e.g., [9]), for identifying the parameters. Let us introduce the formula, which is rather general, in a form suitable for our purposes.

Proposition 2.

Let us denote $\Phi(x_S) = \log P_S^{\overline{U}}(x_S)$ with $\overline{U} \in \mathcal{U}^0$. Then

$$\overline{U}_W(x_W) = \sum_{B \subset W} (-1)^{|W \setminus B|} \left[\Phi(x_B, 0_{S \setminus B}) - \Phi(0_S) \right]$$

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

Proof.

The relation can be verified by direct substitution. See, e.g.,[3] or [9]. □

Now, by elementary rearrangements, we obtain

$$\overline{U}_W(x_W) = \sum_{B \subset W} (-1)^{|W \setminus B|} \left[\log \frac{P_S^{\overline{U}}(x_B, 0_{S \setminus B})}{P_S^{\overline{U}}(0_S)} \right] =$$

$$= \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{P_S^{\overline{U}}(x_B, 0_{S \setminus B})}{P_S^{\overline{U}}(x_{B \cup \{s\}}, 0_{S \setminus \{B \cup \{s\}\}})} \right] =$$

$$= \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{P^{\overline{U}}_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{P^{\overline{U}}_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})} \right]$$

for every $x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}$, where $s \in W$ is arbitrary fixed. For the last expression note that $(W \setminus \{s\}) \subset \partial(s)$ since $s \in W$.

Now, suppose we are able to *extend* the original system of marginals $\mathcal{Q}$ **consistently** into the system

$$\mathcal{Q}^{\partial} = \{Q_{\overline{\partial}(s)}\}_{s \in S},$$

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

$$P^{\overline{U}}_S \in \mathcal{P}_{\mathcal{Q}} \cap \mathcal{P}_{\mathcal{Q}^{\partial}}$$

can be guaranteed. Then we can calculate the parameters $\{\overline{U}_W(x_W)\}_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}}$ directly from the Möbius formula, namely

$$\overline{U}_W(x_W) = \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{Q_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{Q_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})} \right]$$

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

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 $x_W \in \mathcal{X}_W$.

8 Approximation

Unfortunately, the exact extension is usually hardly available, but, from the practical point of view, a reasonable *approximation* can be sufficient. Let us continue with the above reasoning in order to obtain

$$\overline{U}_W(x_W) = \sum_{B \subset W \setminus \{s\}} (-1)^{|W \setminus B|} \left[\log \frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(x_{B\cup\{s\}}, 0_{\overline{\partial}(s)\setminus\{B\cup\{s\}\}})} \right]$$

$$= \sum_{B \subset W} (-1)^{|W \setminus B|} \left[\log \frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(0_{\overline{\partial}(s)})} \right].$$

Now, for approximating $Q_{\overline{\partial}(s)}$ we shall use a rather standard *product form* (see, e.g., [4],[6], or [7]), but, first of all, we need some more notation.

For $s \in S$ we denote $\mathcal{V}_s = \{V \in \mathcal{V}; V \ni s\}$, $v_s = |V_s|$, and I_s the set of all possible enumerations of the elements of $\mathcal{V}_s$. Then, for every $\rho \in I_s$, we may set

$$\widehat{Q}^{\rho}_{\overline{\partial}(s)} = \frac{\prod_{A \in \mathcal{V}_s} Q_A}{\prod_{j=1,\dots,v_s} Q_{B^{\rho}_j}}$$

where $B^{\rho}_j = A_{\rho(j)} \cap \left(\bigcup_{i=1}^{j-1} A_{\rho(i)}\right)$ for every $j = 1, \dots, v_s$, as a natural estimate.

Remark 2. The product form for $\widehat{Q}^{\rho}_{\overline{\partial}(s)}$ can be also justified by the assumption

$$\frac{Q_{\overline{\partial}(s)}(x_B, 0_{\overline{\partial}(s)\setminus B\}})}{Q_{\overline{\partial}(s)}(x_{B\cup\{s\}}, 0_{\overline{\partial}(s)\setminus\{B\cup\{s\}\}})} = \frac{Q_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{Q_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})} =$$

$$= \frac{P^{\overline{U}}_{\{s\}|\partial(s)}(0_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}{P^{\overline{U}}_{\{s\}|\partial(s)}(x_{\{s\}}|x_B, 0_{\partial(s)\setminus\{B\cup\{s\}\}})}$$

where the latter term can be factorized by definition (see Section 5.). □

Further, for every pair $A, W \in \overline{\mathcal{V}}$ let us denote

$$\hat{u}_{W,A}(x_W) = \sum_{B\subset W} (-1)^{|W\setminus B|} \left[\log \frac{Q_A(x_{A\cap B}, 0_{A\setminus B\}})}{Q_A(0_A)}\right].$$

Proposition 3. Let $W \setminus A \neq \emptyset$. Then $\hat{u}_{W,A} \equiv 0$.

Proof. We may write

$$\hat{u}_{W,A}(x_W) = \sum_{B_1\subset W\cap A} \sum_{B_2\subset W\setminus A} (-1)^{|(W\cap A)\setminus B_1|}(-1)^{|(W\setminus A)\setminus B_2|} \left[\log \frac{Q_A(x_{A\cap B_1}, 0_{A\setminus B_1\}})}{Q_A(0_A)}\right].$$

And for $W \setminus A \neq \emptyset$ we have

$$\sum_{B_2\subset W\setminus A} (-1)^{|(W\setminus A)\setminus B_2|} = 0.$$

□

With the above defined terms, by substituting the estimate $\widehat{Q}^{\rho}_{\overline{\partial}(s)}$ into the expression for $\overline{U}_W$, we may introduce the **approximation**

$$\widehat{U}^{s,\rho}_W = \sum_{A\in\mathcal{V}_s, A\supset W} u_{W,A} - \sum_{j=1,\dots,v_s: B^{\rho}_j\supset W} u_{W,B^{\rho}_j}$$

for every $W \in \overline{\mathcal{V}}, s \in W$, and $\rho \in I_s$. Since $s \in W$ and $\rho \in I_s$ are "free parameters", we may, finally **average** over all possible choices, and obtain

$$\widehat{U}_W = |W|^{-1} \sum_{s\in W} |I_s|^{-1} \sum_{\rho\in I_s} \widehat{U}^{s,\rho}_W$$

for every $W \in \overline{\mathcal{V}}$.

Remark 3. It is apparent that for large W many terms disappear. E.g., for $W \in \mathcal{V}$ we have actually $\widehat{U}^{s,\rho}_W = u_{W,W}$ for every $s \in W$ and $\rho \in I_s$, and therefore also $\widehat{U}_W = u_{W,W}$. □

Remark 4. The main advantage of the method is given by its non-sensitivity to the break of assumptions. In practise, we do not have to check the consistency assumption for the original system $\mathcal{Q}$, and the possible zeros can be substituted by some small $\epsilon > 0$. In addition, the model does not require any interconnections between the approximated potential functions $\widehat{U}_W, W \in \overline{\mathcal{V}}$. □

Remark 5. With the model parameters $\{\overline{U}_W(x_W)\}_{x_W \in \mathcal{X}^0_W, W \in \overline{\mathcal{V}}}$ we can easily calculate the relative values of the probability, namely $P^{\overline{U}}_S(x_S)/P^{\overline{U}}_S(y_S)$ for $x_S, y_S \in \mathcal{X}_S$, and the conditional distributions $P^{\overline{U}}_{A|\partial A}$ for "small" $A \subset S$. More complex terms can be again simulated, similarly as in Section 6. □

Acknowledgement.

Supported by grant GA ČR No.201/09/1931 and Research Center DAR (MŠMT ČR Project No. 1M0572).

References

[1] Barndorff–Nielsen, O. E. (1978) *Information and Exponential Families in Statistical Theory.* John Wiley and Sons, New York .

[2] Gilks W. R., Richardson S.,and Spiegelhalter D. J. (eds.) (1996) *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

[3] Janžura M. (2007) *On the connection between marginal problem, statistical estimation, and Möbius formula.* Kybernetika 43(5):619-631.

[4] Janžura M. and Boček P. (1998) *A method for knowledge integration.* Kybernetika 34 , 1, 41–55.

[5] Jaynes E. T. (1982) On the rationale of tmaximum entropy methods. *Proc. IEEE 70*, 939–952.

[6] Jiroušek R. and Vejnarová J. (2003) Construction of multidimensional model by operators of composition: current state of art. *Soft Computing 7*, 328–335.

[7] Lauritzen S. L.(1996) *Graphical Models.* University Press, Oxford .

[8] Perez A. and Studený. M.(2006)Comparison of two methods for approximation of probability distributions with prescribed marginals. *submitted to Kybernetika*

[9] Winkler G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer–Verlag, Berlin .

[10] Younes L. (1988) Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincare 24* , 2, 269–294.