Thai Journal of Mathematics : (2017) 1–6
Special Issue on Entropy in Econometrics

http://thaijmath.in.cmu.ac.th
ISSN 1686-0209

Probabilistic Graphical Models Follow Directly from Maximum Entropy

Anh H. Ly[†], **Francisco Zapata**[‡], **Olac Fuentes**[§], **and Vladik Kreinovich**[§,1]

[†]Banking University of Ho Chi Minh City, 56 Hoang Dieu 2
Quan Thu Duc, Thu Duc, Ho Ch Minh City, Vietnam
e-mail : `lynt@buh.edu.vn (A.H. Ly)`

[‡]Department of Industrial, Manufacturing, and Systems Engineering
University of Texas at El Paso 500 W. University, El Paso, TX 79968, USA
e-mail : `fazg74@gmail.com (F. Zapata)`

[§]Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
e-mail: `ofuentes@utep.edu (O. Fuentes)`
`vladik@utep.edu (V. Kreinovich)`

Abstract : Probabilistic graphical models are a very efficient machine learning technique. However, their only known justification is based on heuristic ideas, ideas that do not explain why exactly these models are empirically successful. It is therefore desirable to come up with a theoretical explanation for these models' empirical efficiency. At present, the only such explanation is that these models naturally emerge if we maximize the relative entropy; however, why the relative entropy should be maximized is not clear. In this paper, we show that these models can also be obtained from a more natural – and well-justified – idea of maximizing (absolute) entropy.

Keywords : probabilistic graphical models; maximum entropy.
2010 Mathematics Subject Classification : 94A17; 68T05.

This work was supported in part by NSF grant HRD-1242122 (Cyber-ShARE Center of Excellence).

[1] Corresponding author.

1 Formulation of the Problem

Need to know probabilities. To fully describe the state of an object or a system, we need to know the values of the large number of quantities $x_1, \ldots, x_n$. It is desirable to know which combinations $x = (x_1, \ldots, x_n)$ of values of these values are possible, and what is the frequency with which different combinations appear. In other words, we need to know the probability density $\rho(x)$ of different combinations x.

Bayesian models. In some situations, we know causal relation between the components of the system and thus, between different quantities. For example, we know that:

- x_1 directly influences x_2,
- x_2 directly influences x_3, and
- x_4 and x_5 jointly influence x_6.

This influence can be described by describing the conditional probability densities $\rho_{2|1}(x_2 \,|\, x_1)$, $\rho_{3|2}(x_3 \,|\, x_2$, and $\rho_{6|4,5}(x_6 \,|\, x_2, x_4)$ corresponding to different combinations of the values x_i. Based on these conditional probabilities, we can find the joint distributions of the corresponding sets of quantities:

$$\rho_{1,2}(x_1, x_2) = \rho_{2|1}(x_2 \,|\, x_1) \cdot \rho_1(x_1),$$

$$\rho_{2,3}(x_2, x_3) = \rho_{3|2}(x_3 \,|\, x_2) \cdot \rho_2(x_2),$$

thus

$$\rho_{1,2,3}(x_1, x_2, x_3) = \rho_{3|2}(x_3 \,|\, x_2) \cdot \rho_{2|1}(x_2 \,|\, x_1) \cdot \rho_1(x_1),$$

and

$$\rho_{4,5,6}(x_4, x_5, x_6) = \rho_{6|4,5}(x_6 \,|\, x_4, x_5) \cdot \rho_{4,5}(x_4, x_5).$$

It is reasonable to assume that quantities which are not thus related are independent. For example, since we did not assume any relation between x_4 and x_5, it is reasonable to assume that

$$\rho_{4,5}(x_4, x_5) = \rho_4(x_4) \cdot \rho_5(x_5).$$

Since we did not assume any dependence between the variables from the first group (x_1, x_2, and x_3) and the variables from the second group (x_4, x_5, and x_6), we conclude that

$$\rho(x_1, x_2, x_3, x_4, x_5, x_6) = \rho_{1,2,3}(x_1, x_2, x_3) \cdot \rho_{4,5,6}(x_4, x_5, x_6).$$

Thus, we get

$$\rho(x_1, x_2, x_3, x_4, x_5, x_6) =$$

$$\rho_{3|2}(x_3 \,|\, x_2) \cdot \rho_{2|1}(x_2 \,|\, x_1) \cdot \rho_1(x_1) \cdot \rho_{6|4,5}(x_6 \,|\, x_4, x_5) \cdot \rho_4(x_4) \cdot \rho_5(x_5).$$

In general, the whole function $\rho(x_1, \ldots, x_n)$ is then represented as a product of several functions, each of which depends only on a small number of directly related quantities. Such probabilistic distributions are known as *Bayesian models*.

Probabilistic graphical models. Bayesian models are applicable if the direct influence relation is a strict order, without cycles:

- if x_1 influences x_2, then x_2 cannot influence x_1;
- if x_1 influences x_2 and x_2 influences x_3, then x_3 cannot directly influence x_1, etc.

In many other situations, we know that several quantities are influencing each other. In this case, we cannot use the Bayesian models.

To cover such situations, researchers decided to follow the same pattern: namely, the corresponding probability distribution has the form

$$\rho(x_1, \ldots, x_n) = \prod_C f_C(x_C),$$

where C are small-size subsets of the set $\{1, \ldots, n\}$ and x_C is a combination of variables x_i corresponding to $i \in C$. In this description, each set C represents the set of variables x_i which affect each other.

For example, for $C = \{3, 5, 6\}$, the notation $f_C(x_C)$ means $f_C(x_3, x_5, x_6)$. Such probabilistic models became known as *probabilistic graphical models*; see, e.g., [1].

Probabilistic graphical models: successes and challenges. Probabilistic graphical models turned out to be very efficient: until the recent emergence of deep learning, they were one of the most empirically successful tools in machine learning.

While from the pragmatic viewpoint, probabilistic graphical models have been a great success, from the theoretical viewpoint, they remained a mystery. Yes, we have a heuristic justification – similarity to Bayesian networks. However, usually, each such heuristic justification can be used to justify several slightly different models. So why are necessarily theses models empirically successful?

Natural approach to selecting a single model under uncertainty: maximum entropy approach. In our situation, we only have partial information about the probability distributions – namely, we only have information (in general, partial) about the marginal probability distributions of the combinations of variables x_C corresponding to several small sets C of mutually dependent quantities.

In situations in which we only have partial information about the probability distribution – and thus, several different probability distributions are consistent with this information – a reasonable idea is to select a distribution that retains this uncertainty as much as possible. For example, if all we know about a probability distribution of a single variable is that this variable is always located on the interval $[0, 1]$, and we have no reason to assume that one of the values from this interval is more probable than others, it is reasonable to consider a uniform distribution for which all the values from this interval are equally probable.

In general, the uncertainty of a probability distribution $\rho(x)$ can be described by its *entropy* $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\,dx$; see, e.g., [2, 3]. From this viewpoint, the distribution with the largest uncertainty is the distribution with the largest entropy. Thus, if several probability distributions are consistent with our knowledge, it is reasonable to select a distribution with the largest possible entropy.

What is known. To the best of our knowledge, until now, there has been no justification for these models in terms of the maximum entropy principle. What is known is that these models can be obtained if we maximize *relative* entropy

$$\int \rho(x) \cdot \ln\left(\frac{\rho(x)}{\rho_0(x)}\right) dx$$

for some distribution $\rho_0(x)$; see, e.g., [4, 5, 6, 7].

Remaining problem and what we do in this paper. The main remaining problem is that, in contrast to the (absolute) entropy S whose maximization is well-justified, the reason for maximization of relative entropy is much less clear.

In this paper, we show that the probabilistic graphical models can be justified based on the general maximum entropy principle, without the need to involve relative entropy.

2 Definitions and the Main Result

What partial information we can have: examples. We may have different information about the marginal distribution $\rho_C(x_c) = \int \rho(x_C, x_{-C})\,dx_{-C}$, where $-C$ denotes a complement to the set C. For example:

- We may know moments of this distribution

$$M_{n_i,\ldots,n_j} \stackrel{\text{def}}{=} \int x_i^{n_i} \cdot \ldots \cdot x_j^{n_j} \cdot \rho_C(x_C)\,dx_C.$$

- Alternatively, we may know the conditional probability distribution

$$\rho_{i\,|\,C-i}(x_i\,|\,x_{C-i}) = \frac{\rho_C(x_i, x_{C-i})}{\int \rho_C(x'_i, x_{C-i})\,dx'_i}.$$

What partial information we can have: a general description. In general, for each of the given small sets C of mutually dependent variables, we have one of more constraints of the type

$$F_{C,\alpha} = v_{C,\alpha} \tag{1}$$

corresponding to different indices α, where $v_{C,\alpha}$ is a known value and $F_{C,\alpha}$ is a known functional depending only on the marginal distributions

$$\rho_C(x_C) = \int \rho(x_C, x_{-C})\,dx_{-C}\,. \tag{2}$$

Maximum entropy approach. We want to maximize the entropy

$$S = -\int \rho(x) \cdot \ln(\rho(x))\, dx \tag{3}$$

under:

- the constraints (1) corresponding to different C and α and
- the constraint that the overall probability is 1: $\int \rho(x)\, dx = 1$.

By applying the Lagrange multiplier method to this constraint optimization problem, we can reduce it to the following unconstrained optimization problem of maximizing the expression

$$-\int \rho(x) \cdot \ln(\rho(x))\, dx + \lambda \cdot \int \rho(x)\, dx + \sum_C \sum_\alpha \lambda_{C,\alpha} \cdot (F_{C,\alpha} - v_{C,\alpha}), \tag{4}$$

for some constants λ and $\lambda_{C,\alpha}$ (Lagrange multipliers).

Differentiating the maximized expression with respect to $\rho(x)$, taking into account that the derivative of a constant is 0, and equating the derivative to 0, we conclude that

$$-\ln(\rho(x)) - 1 + \lambda + \sum_C \sum_\alpha \lambda_{C,\alpha} \cdot \frac{\partial F_{C,\alpha}}{\partial \rho(x)} = 0. \tag{5}$$

Since each expression $F_{C,\alpha}$ depends only on the marginal probabilities $\rho_C(x_C)$, we can use the chain rule and conclude that

$$\frac{\partial F_{C,\alpha}}{\partial \rho(x)} = \frac{\partial F_{C,\alpha}}{\partial \rho_C(x_C)} \cdot \frac{\partial \rho_C(x_C)}{\partial \rho(x)}. \tag{6}$$

Due to the formula (2), we have

$$\frac{\partial \rho_C(x_C)}{\partial \rho(x)} = 1,$$

hence

$$\frac{\partial F_{C,\alpha}}{\partial \rho(x)} = \frac{\partial F_{C,\alpha}}{\partial \rho_C(x_C)}. \tag{7}$$

Thus this derivative depends only on the values x_C. Hence, for each set C, the partial sum

$$s_C \overset{\text{def}}{=} \sum_\alpha \lambda_{C,\alpha} \cdot \frac{\partial F_{C,\alpha}}{\partial \rho(x)} \tag{8}$$

also depends only the values x_C: $s_C = s_C(x_C)$. Substituting the expression (8) into the formula (5), we conclude that

$$-\ln(\rho(x)) - 1 + \lambda + \sum_C s_C(x_C) = 0.$$

Thus,

$$\ln(\rho(x)) = -1 + \lambda + \sum_C s_C(x_C).$$

We can move the constant $-1 + \lambda$ into one of the terms $s_{C_0}(x_{C_0})$, so we get

$$\ln(\rho(x)) = \sum_C s'_C(x_C), \tag{9}$$

where:

- $s'_{C_0}(x_{C_0}) = s_{C_0}(x_{C_0}) - 1 + \lambda$ and
- $s'_C(x_C) = s_C(x_C)$ for $C \neq C_0$.

By applying exp to both sides of the formula (9), we get the desired expression

$$\rho(x) = \prod_C f_C(x_C),$$

where $f_C(x_C) \overset{\text{def}}{=} \exp(s'_C(x_C))$. The statement is proven.

References

[1] D. Koller, B. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, Cambridge, Massachusetts, 2009.

[2] E. T. Jaynes, G. L. Bretthorst, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK, 2003.

[3] H. T. Nguyen, V. Kreinovich, B. Wu, G. Xiang, Computing Statistics under Interval and Fuzzy Uncertainty, Springer Verlag, Berlin, Heidelberg, 2012.

[4] A. L. Berger, A. L. Della Pietra, V. J. Della Pietra, A maximum entropy approach to natural language processing, Computational Linguistics 22 (1) (1996) 39–71.

[5] S. Della Pietra, V. Della Pietra, J. Lafferty, Inducing features of random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (4) (1997) 380–393.

[6] A. Ratnapartkhi, A simple introduction to maximum entropy models for natural language processing, University of Pennsylvania, Institute for Research in Cognitive Science, Technical Report 97-08, 1997.

[7] H. M. Wallach, Conditional random fileds: an introduction, University of Pennsylvania, Department of Computer and Information Science, Technical Report (2004), 2-24.

(Received 31 August 2017)
(Accepted 30 October 2017)