Mozeika, Alexander; Dikmen, Onur; Piili, Joonas

Consistent inference of a general model using the pseudolikelihood method

Published in:
Physical Review E

DOI:
10.1103/PhysRevE.90.010101

Published: 01/01/2014

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Mozeika, A., Dikmen, O., & Piili, J. (2014). Consistent inference of a general model using the pseudolikelihood method. Physical Review E, 90(1), 1-3. [010101]. https://doi.org/10.1103/PhysRevE.90.010101

PHYSICAL REVIEW E **90**, 010101(R) (2014)

Consistent inference of a general model using the pseudolikelihood method

Alexander Mozeika,[1] Onur Dikmen,[1] and Joonas Piili[2]
[1]*Department of Information and Computer Science, Aalto University, P.O. Box 15400, FI-00076, Finland*
[2]*Department of Biomedical Engineering and Computational Science, Aalto University, P.O. Box 12200, FI-00076, Finland*
(Received 7 February 2014; published 25 July 2014)

Recently, a maximum pseudolikelihood (MPL) inference method has been successfully applied to statistical physics models with intractable likelihoods. We use information theory to derive a relation between the pseudolikelihood and likelihood functions. Furthermore, we show the consistency of the pseudolikelihood method for a general model.

DOI: 10.1103/PhysRevE.90.010101 PACS number(s): 02.50.Tt, 05.10.−a, 64.60.De

As statistical physics (SP) models started to be widely used not only in their traditional domain but also to describe biological, financial, etc., phenomena, inferring the interactions (model parameters) from the data has become an important research topic in the physics community [1–4]. This also strengthens a connection between SP and the areas of statistics and machine learning where model parameters are inferred from data. In statistics the maximum likelihood (ML) method is a standard approach due to its attractive statistical properties such as consistency, i.e., its ability to recover true parameters of a model, and asymptotic efficiency [5]. Unfortunately, a direct application of this method is usually infeasible as it involves computing the normalization constant of a distribution (partition function in SP) which is nontrivial even for highly stylized models of SP [6].

For equilibrium Gibbs-Boltzmann distribution [7] the ML method is usually implemented by so-called Boltzmann learning [8], which uses samples from the distribution to approximate the gradients of the likelihood. However, it uses the Markov chain Monte Carlo (MCMC) for sampling which can have very long equilibration times even for moderate system sizes. The sampling can be approximated for speedup by stopping the MCMC early, thus leading to the contrastive divergence [9] learning method. This method is very efficient but it is biased and not consistent [9]. Other methods, such as the mean-field approximation [10], allow one to avoid the MCMC sampling, but their statistical properties are generally not known.

In the MPL method [11] one avoids computation of a partition function by replacing the likelihood by a much simpler function of model parameters. Recently, this method was successfully used for protein contact prediction [12], and it seems to outperform other methods for the benchmark Ising-spin models [13]. Furthermore, the MPL method was shown to be consistent for the Ising model used in Boltzmann learning [14] and for the Gibbs-Boltzmann distributions over $\mathbb{Z}^d$ [15].

In this Rapid Communication, we consider ML and MPL methods of inference. We show that both methods are equivalent to the problem of minimization of a relative entropy between the distributions of model and data. This has been known for ML but not for MPL. We use this framework to derive a relation between the likelihood and pseudolikelihood functions. Furthermore, we prove the consistency of the MPL method for a general model.

Let us consider the following inference problem: We are given L samples $\{s^\mu\}_{\mu=1}^L$ drawn independently from the probability distribution $P_{\theta_0}(s)$, where $s = (s_1, s_2, \ldots, s_N)$, and we are required to estimate the *true* parameters θ_0 of this distribution. A classical approach to this problem is to maximize the log likelihood [16] with respect to the parameters for given data,

$$\hat{\theta}_L = \arg\max_\theta \mathcal{L}_L(\theta), \quad \mathcal{L}_L(\theta) = \frac{1}{L}\sum_{\mu=1}^{L} \log P_\theta(s^\mu). \qquad (1)$$

The (ML) *estimator* $\hat{\theta}$ obtained by the above procedure is *weakly consistent* (respectively *strongly consistent*): In the large sample limit $L \to \infty$ we have that $\hat{\theta} \to \theta_0$ in probability (respectively almost surely) for all possible true values of θ_0 [5,17].

With an infinite amount of data ($L = \infty$) the ML procedure (1) allows us to find its true parameters θ_0. To show this we will consider the difference

$$\frac{1}{L}\sum_{\mu=1}^{L}[\log \hat{P}_L(s^\mu) - \log P_\theta(s^\mu)] = \sum_s \hat{P}_L(s) \log \frac{\hat{P}_L(s)}{P_\theta(s)} = D(\hat{P}_L \| P_\theta), \qquad (2)$$

where $\hat{P}_L(s) = \frac{1}{L}\sum_{\mu=1}^{L} \delta_{s,s^\mu}$, with δ_{s,s^μ} denoting the Kronecker delta function, is an *empirical* distribution of data. Thus the maximization of log likelihood in (1) is equivalent to the minimization of the function $D(\hat{P}_L \| P_\theta)$, which is a *relative entropy* (or Kullback-Leibler divergence) of information theory [18]. By the strong law of large numbers we have that $\lim_{L\to\infty} \sum_s \hat{P}_L(s) \log \frac{\hat{P}_L(s)}{P_\theta(s)} = D(P_{\theta_0} \| P_\theta)$, where $D(P_{\theta_0} \| P_\theta) = \sum_s P_{\theta_0}(s) \log \frac{P_{\theta_0}(s)}{P_\theta(s)}$. Note that $D(P_{\theta_0} \| P_\theta) \geqslant 0$ with equality if and only if $P_{\theta_0}(s) = P_\theta(s)$ holds for all s [18]. Furthermore, assuming that the equality of distributions $P_{\theta_0}(s) = P_\theta(s)$ implies the equality of its parameters $\theta_0 = \theta$ (this is the so-called *identifiability* condition) completes the proof. We note that if the limit and maximization operators in $\lim_{L\to\infty} \hat{\theta}_L = \lim_{L\to\infty} \arg\max_\theta \mathcal{L}_L(\theta)$ *commute*, then the above argument also shows a (strong) consistency of the ML estimator $\hat{\theta}_L$. This requirement imposes further conditions on the estimator function $\mathcal{L}_L(\theta)$ [5,17].

Although the ML estimator is consistent, very often the method of inference itself is not practical as it requires the computation of the partition function [19]. One way to circumvent this problem is to use instead of log likelihood a much simpler pseudo-log-likelihood,

$$\hat{\theta}_L = \arg\max_{\theta} \mathcal{PL}_L(\theta),$$
$$\mathcal{PL}_L(\theta) = \frac{1}{L}\sum_{\mu=1}^{L}\sum_{i=1}^{N}\log P_\theta\left(s_i^\mu \middle| s_{-i}^\mu\right), \tag{3}$$

where $P_\theta(s_i|s_{-i}) = P_\theta(s)/\sum_{s_i} P_\theta(s)$, with $s_{-i} = (s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_N)$, is a conditional distribution. The conditional distribution is, by definition, independent of the partition function. As in the case of log likelihood, the pseudo-log-likelihood method (3) is also equivalent to the minimization of a relative entropy. This can be shown as follows.

First, using the relative entropy (2) and equality $P_\theta(s) = P_\theta(s_i|s_{-i})P_\theta(s_{-i})$, which is true for each i, we obtain

$$N\,D(\hat{P}_L\|P_\theta) = N\sum_s \hat{P}_L(s)\log\frac{\hat{P}_L(s)}{P_\theta(s)}$$
$$= \sum_{i=1}^{N}\sum_s \hat{P}_L(s)\log\frac{\hat{P}_L(s)}{P_\theta(s_i|s_{-i})P_\theta(s_{-i})}. \tag{4}$$

Let us now in the above replace the distribution of model $P_\theta(s_{-i})$ by the empirical distribution of data $\hat{P}_L(s_{-i})$. This gives rise to the probability distribution $\hat{P}^i_{\theta|\theta_0}(s) = P_\theta(s_i|s_{-i})\hat{P}_L(s_{-i})$ and immediately leads us to the inequality

$$\sum_{i=1}^{N}\sum_s \hat{P}_L(s)\log\frac{\hat{P}_L(s)}{P_\theta(s_i|s_{-i})\hat{P}_L(s_{-i})}$$
$$= \sum_{i=1}^{N} D\left(\hat{P}_L\,\middle\|\,\hat{P}^i_{\theta|\theta_0}\right) \geqslant 0. \tag{5}$$

Clearly the minimum of the above sum of relative entropies, with respect to the model parameters θ, corresponds to the maximum of the pseudo-log-likelihood function used in (3). This sum is also a lower bound for the (rescaled by N) relative entropy (4). In order to show this we consider the difference

$$N\,D(\hat{P}_L\|P_\theta) - \sum_{i=1}^{N} D\left(\hat{P}_L\,\middle\|\,\hat{P}^i_{\theta|\theta_0}\right)$$
$$= \sum_{i=1}^{N}\sum_s \hat{P}_L(s)\ln\frac{\hat{P}_L(s_{-i})}{P_\theta(s_{-i})}$$
$$= \sum_{i=1}^{N}\sum_{s_{-i}} \hat{P}_L(s_{-i})\ln\frac{\hat{P}_L(s_{-i})}{P_\theta(s_{-i})} \geqslant 0. \tag{6}$$

The inequality in the above is due to the last line being a sum of relative entropies.

A consequence of the inequality (6) is the relation

$$\mathcal{PL}_L(\theta) - \sum_{i=1}^{N} H_i(\hat{P}_L) \geqslant N\mathcal{L}_L(\theta), \tag{7}$$

where $H_i(\hat{P}_L) = -\sum_{s_{-i}}\hat{P}_L(s_{-i})\ln\hat{P}_L(s_{-i})$ is a Shannon entropy of the empirical distribution $\hat{P}_L(s_{-i}) = \sum_{s_i}\hat{P}_L(s)$, between the objective functions of the ML (1) and MPL (3) methods. Furthermore, using the inequality (5), we can show that the MPL procedure (3) recovers the true parameters θ_0 with an infinite amount of data. To show this we consider the sum of relative entropies

$$\sum_{i=1}^{N} D\left(P_{\theta_0}\,\middle\|\,P^i_{\theta|\theta_0}\right) = \sum_{i=1}^{N}\sum_s P_{\theta_0}(s)\log\frac{P_{\theta_0}(s)}{P_\theta(s_i|s_{-i})P_{\theta_0}(s_{-i})}$$
$$= \sum_{i=1}^{N}\sum_s P_{\theta_0}(s)\log\frac{P_{\theta_0}(s_i|s_{-i})P_{\theta_0}(s_{-i})}{P_\theta(s_i|s_{-i})P_{\theta_0}(s_{-i})}$$
$$= \sum_{i=1}^{N}\sum_s P_{\theta_0}(s)\log P_{\theta_0}(s_i|s_{-i})$$
$$- \sum_{i=1}^{N}\sum_s P_{\theta_0}(s)\log P_\theta(s_i|s_{-i})$$
$$= Q_0(\theta_0) - Q_0(\theta) \geqslant 0, \tag{8}$$

where $Q_0(\theta) = \lim_{L\to\infty}\mathcal{PL}_L(\theta)$. Thus, $Q_0(\theta) \leqslant Q_0(\theta_0)$ and if $P_{\theta_0}(s_i|s_{-i}) \neq P_\theta(s_i|s_{-i})$ implies that $\theta_0 \neq \theta$ then $\theta = \theta_0$ is the unique maximum of $Q_0(\theta)$. We note that this proves condition (i) of Theorem 1 in the Appendix. We will use this theorem to show the (weak) consistency of the MPL estimator (3).

Let us assume that $\theta \in \Theta$, where Θ is a compact set [this is condition (ii) of Theorem 1] and define $\hat{Q}_L(\theta) = \mathcal{PL}_L(\theta)$. If $Q_0(\theta)$ is a continuous function of θ and $\hat{Q}_L(\theta)$ converges uniformly in probability to $Q_0(\theta)$, i.e., $\sup_{\theta\in\Theta}|\hat{Q}_L(\theta) - Q_0(\theta)| \xrightarrow{\text{Prob.}} 0$ as $L\to\infty$, then conditions (iii) and (iv) of Theorem 1 are satisfied. In order to prove these conditions we will use Lemma 1 in the Appendix. To this end we define $q(s,\theta) = \log\prod_{i=1}^{N} P_\theta(s_i|s_{-i})$ and hence $\hat{Q}_L(\theta) = (1/L)\sum_{\mu=1}^{L} q(s^\mu,\theta)$. Now let us assume that the function $q(s,\theta)$ is continuous at each $\theta\in\Theta$ and consider

$$|q(s,\theta)| = \left|\log\prod_{i=1}^{N} P_\theta(s_i|s_{-i})\right| \leqslant \sup_{\theta\in\Theta}\left|\log\prod_{i=1}^{N} P_\theta(s_i|s_{-i})\right|$$
$$= d(s), \tag{9}$$

then if $\sum_s P_{\theta_0}(s)d(s) < \infty$, we have that $Q_0(\theta)$ is continuous and $\hat{Q}_L(\theta)$ converges uniformly in probability to $Q_0(\theta)$ by Lemma 1. Thus if (a) $P_{\theta_0}(s_i|s_{-i}) \neq P_\theta(s_i|s_{-i})$ implies that $\theta_0 \neq \theta$, (b) $\theta\in\Theta$, where Θ is a compact set, (c) $\log\prod_{i=1}^{N} P_\theta(s_i|s_{-i})$ is continuous, and (d) $\sum_s P_{\theta_0}(s)\sup_{\theta\in\Theta}|\log\prod_{i=1}^{N} P_\theta(s_i|s_{-i})| < \infty$, then the MPL estimator (3) is weakly consistent, i.e., $\hat{\theta} \xrightarrow{\text{Prob.}} \theta_0$ as $L\to\infty$.

To summarize, we mapped the maximum likelihood (ML) and maximum pseudolikelihood (MPL) methods of inference onto the information theory framework which allows us to investigate the relation between these two methods. In this framework, for both methods, the relative entropy is an objective function, the minimization of which is equivalent to ML and MPL. Furthermore, we derive an inequality which establishes a relation between the likelihood and pseudolikelihood functions. Finally, we prove the (weak) consistency of the pseudolikelihood method for a general probability distribution. We envisage that the strong consistency of MPL can also be proven by, for example, adopting the consistency proof of ML in Ref. [5]. Also, all derivations in this Rapid Communication are for the distributions of discrete variables, but we expect that extending these results to the case of continuous variables is a straightforward matter.

The work of A.M. and O.D. is supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170). A.M. was also supported in part by the Academy of Finland, Grant No. 256287. The work of J.P. is supported by The Emil Aaltonen Foundation. A.M. is thankful to Manfred Opper for interesting and helpful discussions. A.M. and O.D. would like to thank Rémi Lemoy for discussions.

APPENDIX: THEOREM AND LEMMA

Here we state the theorem and lemma (on pp. 2121 and 2129 of Ref. [17], respectively) which are used in the main text.

Theorem 1. If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0, (ii) $\theta \in \Theta$, where Θ is a compact set, (iii) $Q_0(\theta)$ is continuous, and (iv) $\hat{Q}_L(\theta)$ converges uniformly in probability to $Q_0(\theta)$, then $\hat{\theta} \xrightarrow{\text{Prob.}} \theta_0$ as $L \to \infty$.

Lemma 1. If s^μ, where $\mu = 1, \ldots, L$, are drawn independently from the probability distribution $P(s)$, Θ is a compact set, $q(s^\mu,\theta)$ is continuous at each $\theta \in \Theta$ with probability one, there is $d(s)$ with $|q(s,\theta)| \leqslant d(s)$ for all $\theta \in \Theta$ and $\sum_s P(s)d(s) < \infty$, then $\sum_s P(s)q(s,\theta)$ is continuous and

$$\sup_{\theta\in\Theta} \left| \frac{1}{L} \sum_{\mu=1}^{L} q(s^\mu,\theta) - \sum_s P(s)q(s,\theta) \right| \xrightarrow{\text{Prob.}} 0$$

as $L \to \infty$.

[1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).
[2] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. USA **108**, E1293 (2011).
[3] I. Mastromatteo and M. Marsili, J. Stat. Mech.: Theory Exp. (2011) P10012.
[4] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, Proc. Natl. Acad. Sci. USA **109**, 4786 (2012).
[5] T. S. Ferguson, *A Course in Large Sample Theory* (CRC Press, Boca Raton, FL, 1996), Vol. 38.
[6] F. Barahona, J. Phys. A: Math. Gen. **15**, 3241 (1982).
[7] The Gibbs-Boltzmann distribution is given by $P(s) = e^{-\beta E(s)}/Z$, where β is inverse temperature, $E(s)$ is an energy function, and $Z = \sum_s e^{-\beta E(s)}$ is a partition function.
[8] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, Cognit. Sci. **9**, 147 (1985).
[9] M. A. Carreira-Perpinan and G. E. Hinton, in *AISTATS 2005: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, edited by R. Cowell and Z. Ghahramani (Society for Artificial Intelligence and Statistics, Savannah Hotel, Barbados, 2005), pp. 33–40.
[10] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice* (MIT Press, Cambridge, MA, 2001).
[11] J. Besag, Statistician **24**, 179 (1975).
[12] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, Phys. Rev. E **87**, 012707 (2013).
[13] E. Aurell and M. Ekeberg, Phys. Rev. Lett. **108**, 090201 (2012).
[14] A. Hyvärinen, Neural Comput. **18**, 2283 (2006).
[15] B. Gidas, in *Stochastic Differential Systems, Stochastic Control Theory and Applications* (Springer, Berlin, 1988), pp. 129–145.
[16] This is equivalent to maximizing the likelihood function $\prod_{\mu=1}^{L} P_\theta(s^\mu)$.
[17] W. K. Newey and D. McFadden, in *Handbook of Econometrics* (Elsevier, Amsterdam, 1994), Vol. 4, pp. 2111–2245.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, Hoboken, NJ, 2012).
[19] For, e.g., the partition function $Z = \sum_s e^{-\beta E(s)}$, where $s \in \{-1,1\}^N$, of the Ising model is a sum over 2^N states.