Consistency of pseudolikelihood estimation of fully visible Boltzmann machines

Aapo Hyvärinen*
HIIT Basic Research Unit
Dept of Computer Science
University of Helsinki, Finland

Neural Computation, in press

13th June 2006

Abstract

Boltzmann machine is a classic model of neural computation, and a number of methods have been proposed for its estimation. Most methods are plagued by either very slow convergence, or asymptotic bias in the resulting estimates. Here we consider estimation in the basic case of fully visible Boltzmann machines. We show that the old principle of pseudo-likelihood estimation provides an estimator that is computationally very simple, yet statistically consistent.

1 Introduction

Assume we observe a binary random vector $\mathbf{x} \in \{-1, +1\}^n$ and we want to model its probability distribution function by

$$P(\mathbf{x}) = \frac{1}{Z(\mathbf{M}, \mathbf{b})} \exp(\frac{1}{2}\mathbf{x}^T\mathbf{M}\mathbf{x} + \mathbf{b}^T\mathbf{x}) \tag{1}$$

The parameter matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$ has to be constrained in some way to make it well-defined, because $\mathbf{M}$ and $\mathbf{M}^T$ give the same probability distribution, and the diagonal elements of $\mathbf{M}$ do not interact with $\mathbf{x}$ at all. We choose the conventional constraint that $\mathbf{M}$ is symmetric and has zero diagonal. The vector $\mathbf{b}$ is an n-dimensional parameter vector. This is a special case ("fully visible", i.e. no latent variables) of the Boltzmann machine framework (Ackley et al., 1985).

*Corresponding author. HIIT Basic Research Unit, Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358-9-191 51120, email: Aapo.Hyvarinen@helsinki.fi.

The central problem in the estimation is that we don't know the constant $Z(\mathbf{M}, \mathbf{b})$. In principle, Z is given by the sum:

$$Z(\mathbf{M}, \mathbf{b}) = \sum_{\boldsymbol{\xi} \in \{-1,+1\}^n} \exp(\frac{1}{2}\boldsymbol{\xi}^T \mathbf{M} \boldsymbol{\xi} + \mathbf{b}^T \boldsymbol{\xi}) \tag{2}$$

whose computation is exponential in the dimension n. Thus, for any larger dimension n, direct numerical computation of Z is out of the question. For continuous-valued variables, we could use score matching (Hyvärinen, 2005), but here we have binary variables.

Maximum likelihood estimation of the model is not possible without some kind of computation of the normalization constant Z, also called the partition function. Typical methods for maximum likelihood estimation are thus computationally very complex, e.g. Markov Chain Monte Carlo (MCMC) methods. Different kinds of approximation methods have therefore been developed, including pseudolikelihood (Besag, 1975), contrastive divergence (Hinton, 2002), and linear response theory (Kappen and Rodriguez, 1998). None of these approximative methods has been shown to be consistent. Our contribution here is to show that pseudolikelihood is consistent, and it is closely connected to contrastive divergence.

2 Pseudolikelihood of the model

In pseudolikelihood estimation (Besag, 1975), we consider the conditional probabilities $P(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n; \boldsymbol{\theta})$, i.e. conditional probabilities of the random variable given all other variables, where $\boldsymbol{\theta}$ denotes the parameter vector. Let us denote by $\mathbf{x}^{\notin i}$ the vector with x_i removed:

$$\mathbf{x}^{\notin i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \tag{3}$$

and the logarithms of the conditional probabilities by

$$C_i(x_i; \mathbf{x}^{\notin i}, \boldsymbol{\theta}) = \log P(x_i|\mathbf{x}^{\notin i}, \boldsymbol{\theta}) \tag{4}$$

We then estimate the model by maximizing these conditional probabilities in the same way as one would maximize ordinary likelihood. Given a sample $\mathbf{x}(1), \ldots, \mathbf{x}(T)$, the pseudolikelihood (normalized as a function of sample size by dividing by T) is thus of the form

$$J_{PL}(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} C_i(x_i(t); \mathbf{x}^{\notin i}(t), \boldsymbol{\theta}) \tag{5}$$

Consistency of the pseudolikelihood has been thoroughly investigated for Markov random fields, see e.g. (Gidas, 1988; Mase, 1995) and the references therein. However, there seem to be few results for the basic case of a random vector.

It is easy to compute the pseudolikelihood for the model in (1). We have

$$P(x_i|\mathbf{x}^{\notin i}, \mathbf{M}, \mathbf{b}) = \frac{\exp(x_i \mathbf{m}_i^T \mathbf{x} + b_i x_i)}{\exp(\mathbf{m}_i^T \mathbf{x} + b_i) + \exp(-\mathbf{m}_i^T \mathbf{x} - b_i)} \tag{6}$$

which gives

$$C_i(x_i|\mathbf{x}^{\notin i}, \mathbf{M}, \mathbf{b}) = x_i \mathbf{m}_i^T \mathbf{x} + b_i x_i - \log\cosh(\mathbf{m}_i^T \mathbf{x} + b_i) - \log 2 \tag{7}$$

and thus, for a given sample $\mathbf{x}(1), \ldots, \mathbf{x}(t)$ of T observations:

$$J_{PL}(\mathbf{M}, \mathbf{b}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} x_i(t) \mathbf{m}_i^T \mathbf{x}(t) + b_i x_i(t) - \log\cosh(\mathbf{m}_i^T \mathbf{x}(t) + b_i) + \text{const.} \tag{8}$$

where the constant does not depend on the parameters.

3 Consistency proof

We now proceed to prove the consistency of the maximum pseudolikelihood estimator obtained by maximization of J_{PL} with respect to the parameters.

The natural starting point is to analyze the point where the gradient of J_{PL} with respect to the parameters is zero. The point of true parameter values is one such point, as shown in the following proposition:

Proposition 1 *Assume data is generated by the distribution in (1) for parameters $\tilde{m}_{ij}$ and $\tilde{b}_i$. Then, the gradient of J_{PL} is zero at $m_{ij} = \tilde{m}_{ij}, b_i = \tilde{b}_i$.*

Proof: We first compute the derivative of the pseudolikelihood with respect to $m_{ij}, i \neq j$:

$$\frac{\partial J_{PL}}{\partial m_{ij}} = \frac{1}{T} \sum_t x_i(t) x_j(t) - x_j(t) \tanh(\mathbf{m}_i^T \mathbf{x}(t) + b_i) \tag{9}$$

A well-known property of Boltmann machines is that

$$E\{x_i|\mathbf{x}^{\notin i}\} = \frac{\exp(\tilde{\mathbf{m}}_i^T \mathbf{x}(t) + \tilde{b}_i) - \exp(-\tilde{\mathbf{m}}_i^T \mathbf{x}(t) - \tilde{b}_i)}{\exp(\tilde{\mathbf{m}}_i^T \mathbf{x}(t) + \tilde{b}_i) + \exp(-\tilde{\mathbf{m}}_i^T \mathbf{x}(t) - \tilde{b}_i)} = \tanh(\tilde{\mathbf{m}}_i^T \mathbf{x} + \tilde{b}_i) \tag{10}$$

At the point where the parameters have the true values, the derivative thus becomes

$$\frac{\partial J_{PL}}{\partial m_{ij}}(\tilde{\mathbf{M}}, \tilde{\mathbf{b}}) = \frac{1}{T} \sum_{t=1}^{T} x_j(t)(x_i(t) - E\{x_i(t)|\mathbf{x}^{\notin i}(t)\}) \tag{11}$$

Now, by the basic properties of conditional expectations, $x_i - E\{x_i|\mathbf{x}^{\notin i}\}$, which is the residual in the best prediction of x_i given $\mathbf{x}^{\notin i}$, is uncorrelated from $\mathbf{x}^{\notin i}$

and thus of x_j. [1] Thus, we have in the limit of $T \to \infty$

$$\frac{\partial J_{PL}}{\partial m_{ij}} = E\{x_j\} E_{x_i}\{x_i - E\{x_i|\mathbf{x}^{\notin i}\}\} = E\{x_j\} \times 0 \tag{12}$$

because the expectation of the residual is zero: $E_{x_i}\{E\{x_i|\mathbf{x}^{\notin i}\}\} = E\{x_i\}$. Thus, the gradient with respect to m_{ij} is zero. As for the b_i, we obtain

$$\frac{\partial J_{PL}}{\partial b_i} = E\{x_i\} - E\{\tanh(\mathbf{m}_i^T \mathbf{x} + b_i)\} \tag{13}$$

which is zero by the same logic. Thus, we have proven the proposition.

We still have to make sure that this critical point is really the global maximum of pseudolikelihood. For this end, we have to make the following assumption. Denote by $\bar{\mathbf{x}}^T = (x_1, \ldots, x_n, 1)^T$ an augmented data vector. We assume

$$E\left\{(\mathbf{q}^T \bar{\mathbf{x}})^2 \cosh^{-2}(\mathbf{m}^T \mathbf{x} + b)\right\} > 0 \tag{14}$$

for any vector $\mathbf{q} \in \mathbb{R}^{n+1}$ of non-zero norm, and for any $\mathbf{m} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. This is not a very strong assumption because obviously, the expectation is always non-negative (cosh is a positive function). Basically, the expectation could be zero only in some pathological cases.

Now we use the concavity of J_{PL}, which is possible due to the following proposition:

Proposition 2 *Assuming (14), and in the limit of an infinite sample, J_{PL} is strictly concave with respect to the vector consisting of the elements of* $\mathbf{M}$ *and* $\mathbf{b}$.

Proof: Since a sum of strictly concave functions is still strictly concave, we can consider each term in the sum with respect to i separately. Each such term is a function of $[\mathbf{m}_i, b_i]$ only. So, we only have to prove that

$$J_i(\mathbf{m}_i, b_i) = E\{x_i \mathbf{m}_i^T \mathbf{x} + b_i x_i - \log\cosh(\mathbf{m}_i^T \mathbf{x} + b_i)\} \tag{15}$$

is strictly concave. The Hessian of J_i with respect to $\mathbf{m}_i$ equals

$$H_{\mathbf{m}_i} J_{PL} = -E\{\mathbf{x}\mathbf{x}^T \cosh^{-2}(\mathbf{m}_i^T \mathbf{x} + b_i)\} \tag{16}$$

The second derivative with respect to b_i equals

$$-E\{\cosh^{-2}(\mathbf{m}_i^T \mathbf{x} + b_i)\} \tag{17}$$

and the cross-derivatives equal

$$\frac{\partial J_i}{\partial \mathbf{m}_i \partial b_i} = -E\{\mathbf{x}^T \cosh^{-2}(\mathbf{m}_i^T \mathbf{x} + b_i)\} \tag{18}$$

[1] In general, we have for any two random variables x, y:
$E_{x,y}\{(E\{y|x\} - y)x\} = \int_x \int_y (\int_{y'} p(y'|x)\, y'\, dy'\, x - xy)\, p(x,y) dx dy$
$= \int_x (\int_{y'} p(y'|x)\, xy'\, dy')(\int_y p(x,y) dy) dx - \int_x \int_y p(x,y) xy\, dy dx.$
Since $\int_y p(x,y) dy\, p(y'|x) = p(x,y')$, the two terms are equal, and the difference is zero.

Collecting these in a single matrix, we see that the total Hessian equals

$$H_{[\mathbf{m}_i,b_i]}J_i = -E\{\bar{\mathbf{x}}\bar{\mathbf{x}}^T \cosh^{-2}(\mathbf{m}_i^T\mathbf{x} + b_i)\} \tag{19}$$

which is, by our assumption in (14), negative-definite for any values of the parameters. A function whose Hessian is always negative-definite is strictly concave. Thus, we have proven the strict concavity of J_{PL}.

This leads us finally to the theorem

Theorem 1 *Assume (14). Then the pseudolikelihood estimator is (globally) consistent for the model in (1).*

Proof: A strictly concave function defined in a real space has a single maximum. If the function is differentiable (as J_{PL} here), the maximum is obtained at the point of zero gradient. This would seem to prove the theorem. However, we have one additional complication because $\mathbf{M}$ is constrained to be symmetric and to have zero diagonal. This is actually not problematic since it only means that the optimization is constrained to a linear subspace. The restriction of a strictly concave function on a linear subspace is still strictly concave. Also, since the gradient is zero for the true parameter values, the projection of the gradient is zero for the true parameter values. Thus, the restrictions of symmetricity and zero diagonal do not change anything. So, we have proven that in the limit of an infinite sample, the pseudolikelihood is maximized by the true parameter values alone. This implies the theorem.

4 Gradient algorithm

Let us briefly consider how pseudolikelihood can be computationally maximized. The simplest way of maximizing the pseudolikelihood is by gradient ascent. The relevant gradients were already given above. However, since $\mathbf{M}$ is constrained to be symmetric and to have zero diagonal, the gradient has to be projected on this linear space. Thus, we compute the symmetrized gradient

$$D(\hat{m}_{ij}) = \frac{1}{2}\frac{J_{PL}}{\partial m_{ij}} + \frac{1}{2}\frac{J_{PL}}{\partial m_{ji}} \tag{20}$$

where the derivatives are given in (9), and evaluated at the current estimates fot the parameters. We then update the current estimates $\hat{m}_{ij}$, for $i \neq j$ only, using this projected gradient in a gradient ascent step:

$$\Delta\hat{m}_{ij} = \mu D(\hat{m}_{ij}) \text{ for all } i \neq j \tag{21}$$

where μ is a step size. As for the b_i, we can use the gradient directly and update

$$\Delta\hat{b}_i = \mu\frac{\partial J_{PL}}{\partial b_i} \tag{22}$$

where the derivative is given in (13).

The algorithm we have given here is a batch algorithm, using the whole sample to calculate the pseudolikelihood. On-line variants are easy to construct as well.

5 Connection to contrastive divergence

Contrastive divergence (Hinton, 2002) is an approximation of MCMC methods. It consists of two related ideas: first, we fix the initial values in the MCMC method to be equal to the sample points themselves, and second, we take a small number of steps in the MCMC method, perhaps just one. This is a general framework that can be applied on non-normalized models with continuous-valued or discrete-valued variables and also in latent variable models.

We shall here prove that for the model in (1), contrastive divergence is equivalent to pseudolikelihood if we use single-step Gibbs sampling, which is the most basic setting.

In the general MCMC setting, the expectation of the gradient of $m_{ij}, i \neq j$ is given by (Ackley et al., 1985)

$$\Delta m_{ij} = \hat{E} x_i x_j - E_M x_i x_j \tag{23}$$

where $\hat{E}$ denotes the expectation over the sample distribution, and E_M denotes the expectation over the distribution given by the model with current parameter values.

In contrastive divergence, the expected gradient update for m_{ij} is given by

$$\Delta m_{ij} = \hat{E} x_i(t) x_j(t) - \hat{E} E_{G(k)} x_i(t) x_j(t) \tag{24}$$

where $E_{G(k)}$ means the expectation under the distribution given by one step of Gibbs sampling on the k-th variable, i.e. replacing $x_k(t)$ by a random variable which follows the conditional distribution of x_k given all other variables. In the simplest random update sheme, the index k is a random variable that has uniform distribution over the indices $1, \ldots, n$. Note that there are two different methods called contrastive divergence defined in (Hinton, 2002): one based on an objective function and the other based on an approximative gradient of that objective function. We consider here the latter because it is the one to be used in practice.

As above, the expectation of the conditional distribution can be computed as

$$E_{G(i)} x_i(t) = \tanh(\mathbf{m}_i^T \mathbf{x}(t) + b_i) \tag{25}$$

while $E_{G(k)} x_i(t) = x_i(t)$ for $k \neq i$. Now, in the second term on the right-hand side of (24) there is a probability of $(n-2)/n$ that the index k is not equal to i or j. Then, the Gibbs sampling has no effect and can be ignored. With probability $1/n$, k equals i and with the same probability, it equals j. Thus,

(24) equals

$$\begin{aligned}\Delta m_{ij} &= \hat{E}x_i(t)x_j(t) - \frac{n-2}{n}\hat{E}x_i(t)x_j(t) \\ &\quad - \frac{1}{n}\hat{E}\tanh(\mathbf{m}_i^T\mathbf{x}(t)+b_i)x_j(t) - \frac{1}{n}\hat{E}x_i(t)\tanh(\mathbf{m}_j^T\mathbf{x}(t)+b_i) \\ &= \frac{2}{n}\left[\hat{E}x_i(t)x_j(t) - \frac{1}{2}\hat{E}\tanh(\mathbf{m}_i^T\mathbf{x}(t)+b_i)x_j(t) - \frac{1}{2}\hat{E}x_i(t)\tanh(\mathbf{m}_j^T\mathbf{x}(t)+b_i)\right]\end{aligned} \tag{26}$$

As for the parameters b_i, we obtain in a similar way

$$\Delta b_i = \hat{E}x_i(t) - \hat{E}E_{G(k)}x_i(t) = \hat{E}x_i(t) - \tanh(\mathbf{m}_i^T\mathbf{x}(t)+b_i) \tag{27}$$

As the gradient step size in contrastive divergence is typically taken from a sequence that converges to zero fast enough, the convergence of contrastive divergence is given by the point where the expected gradient is zero. Now, the expected gradients in (26) and (27) are equal (up the some insignificant multiplicative constants) to the corresponding symmetrized gradients of the pseudolikelihood. So, the two methods converge in the same points.

The convergence of contrastive divergence (the same gradient version as we analyzed here) was analyzed in (Carreira-Perpiñán and Hinton, 2005), with the conclusion that contrastive divergence is asymptotically "biased" for the model in (1). This discrepancy with our results is due to the difference of the definition of biases. In (Carreira-Perpiñán and Hinton, 2005) the bias was computed as the Kullback-Leibler divergence between the distributions given by the model when the estimated parameters for contrastive divergence or likelihood are used. Thus, their conclusion was that contrastive divergence gives, in general, a different estimate than likelihood. However, they also noted that the difference disappears (asymptotically) if the data is really generated by the model, which is the case we consider here. Different variants of contrastive divergence which always give the same estimate as maximum likelihood were further developed in (Carreira-Perpiñán and Hinton, 2005). See also (Welling and Sutton, 2005) for related work.

6 Simulation results

We performed simulation to validate the different estimation methods for the fully visible Boltmann machine. We created random matrices $\mathbf{M}$ so that the elements had independent normal distributions with zero mean and standard deviation of .5. The parameters b_i were randomly generated from the same distribution. The dimension n was set to 5 which is small enough to enable exact sampling from the distribution, which is important in order to be able to reliably validate the estimation results.

We generated data from the distribution in (1) and estimated the parameters using maximum pseudolikelihood. for various sample sizes: 500, 1000, 2000,

4000, 8000, and 16000. We also estimated the parameters using ordinary likelihood for comparison: exact computation of the maximum likelihood estimator was possible due to the small dimension. For each sample size, we created 5 different data sets and ran the estimation once on each data set using a random initial point. For each estimation, the estimation error was computed as the Euclidean distance of the real matrix $[\mathbf{M}, \mathbf{b}]$ and its estimate. Finally, we took the mean of the logarithms of the 5 estimation errors.

The results are shown in Figure 1. The maximum pseudolikelihood estimator seems to be consistent in the sense that the estimation error seems to go to zero when the sample size grows, as implied by our Theorem. Surprisingly, its estimation errors are not really larger than that of ordinary maximum likelihood. Actually the errors are almost identical; they seem to depend more on the random parameters generated than on the method.

7 Conclusion

We have shown that pseudolikelihood, a rather old estimation principle (Besag, 1975), provides a consistent estimator for the fully visible Boltzmann machine. This estimator turns out to be a special case of contrastive divergence. The literature on Boltzmann machines does not seem to have paid much attention to pseudolikelihood estimation so far.

We considered the fully visible case only, because that is where pseudolikelihood estimation can be directly applied. Extensions to hidden variables are an important subject for future work, and have been partly addressed in work on contrastive divergence (Carreira-Perpiñán and Hinton, 2005).

Acknowledgements

This work was supported by the Academy of Finland, Academy Research Fellow position and project #106473. I am grateful to Sam Roweis and Patrik Hoyer for interesting discussions, and to Miguel Carreira-Perpiñán and Geoffrey Hinton for providing access to upublished results.

References

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24:179–195.

Carreira-Perpiñán, M. A. and Hinton, G. E. (2005). On contrastive divergence learning. In *Proc. Workshop on Artificial Intelligence and Statistics (AISTATS2005)*, Barbados.

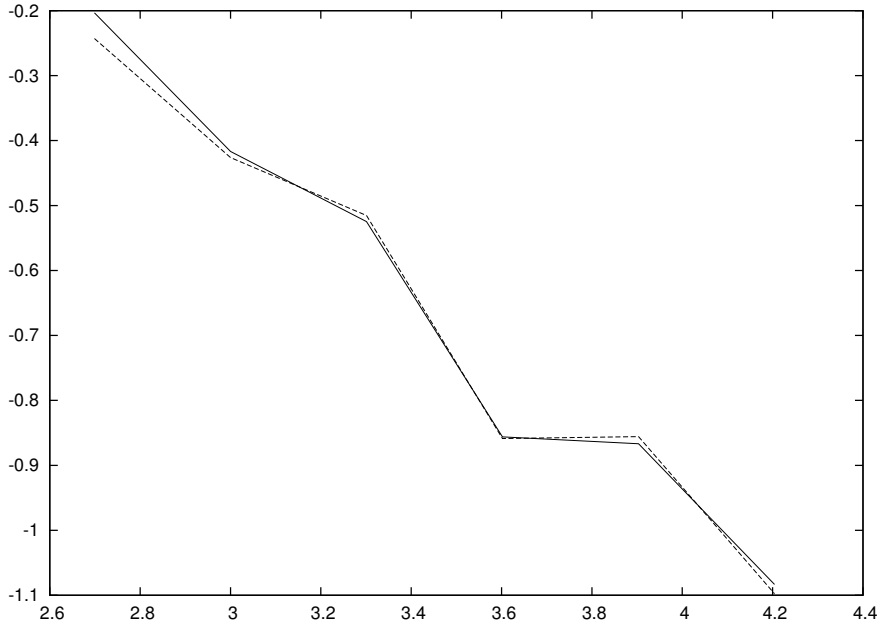

Figure 1: The estimation errors of maximum pseudolikelihood/contrastive divergence (solid line) and maximum likelihood (dashed line). Horizontal axis: $\log_{10}$ of sample size. Vertical axis: $\log_{10}$ of estimation error.

Gidas, B. (1988). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbsian distributions. In Fleming, W. and Lions, P.-L., editors, *Stochastis differential systems, stochastic control theory and applications*. New York: Springer.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *J. of Machine Learning Research*, 6:695–709.

Kappen, H. and Rodriguez, F. (1998). Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156.

Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *The Annals of Applied Probability*, 5(3):603–612.

Welling, M. and Sutton, C. (2005). Learning markov random fields using contrastive free energies. In *Proc. AISTATS*.