Consistency of pseudolikelihood estimation of fully visible **BM**s

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

Neural Computation, in press

13th June 2006

Abstract

BM 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 BMs. We show that the old principle of PLE provides an estimator that is computationally very simple , yet statistically consistent.

1 Int roduct ion

Assume a binary r.v. $\mathbf{x} \in \{-1, +1\}^n$ and pdf:

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

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 BM 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 $Z(\mathbf{M}, \mathbf{b})$:

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

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.

MLE of the model is not possible without some kind of computation of the normalization constant Z, the partition function. Typical methods for MLE are thus computationally very complex , e.g. MCMC methods . Different kinds of approximation methods have therefore been developed : PL (Besag , 1975), CD (Hinton , 2002), and linear response theory (Kappen and Rodriguez , 1998). None of these approximative methods has been shown to be consistent .

***contribution**: PL is consistent, and it is closely connected to CD.*

2 Pseudolikelihood

PLE (Besag1975):the conditional probabilities $P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n;\boldsymbol{\theta})$. Let:

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

and the logarithms of the conditional probabilities by

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

Given a sample $\mathbf{x}(1),\ldots,\mathbf{x}(T)$, the PL:

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

Consistency of the PL has been thoroughly investigated for MRF see e.g. (Gidas, 1988; Mase, 1995) and the references therein.

PL for the model in (1):

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

==>

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

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

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

3 Consistency proof

We now proceed to prove the consistency of the MPLE obtained by maximization of J_{PL} wrt the parameters.

The natural starting point is to analyze the point where the gradient of J_{PL} wrt the parameters $=0$. 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: the derivative, $i \neq j$:

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

A well-known property of BMs:

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

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. ==> $T \to \infty$

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

because the expectation $E_{x_i}\{E\{x_i|\mathbf{x}^{\in/i}\}\} = E\{x_i\}$. As for the b_i,

$$\frac{\partial J_{PL}}{\partial b_i} = E\{x_i\} - E\{\tanh(\mathbf{m}_i^T\mathbf{x} + b_i)\} = 0 \tag{13}$$

by the same logic. □

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.

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

Proof: Since a sum of strictly concave functions is still strictly concave. 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:

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

$$\text{The second derivative wrt } b_i = -E\{\cosh^{-2}(\mathbf{m}^T{}_i\,\mathbf{x} + b_i)\} \tag{17}$$

and the cross-derivatives:

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

==> the total Hessian:

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

which is, by our assumption in (14), n.d. for any values of the parameters.

A function whose Hessian is n.d. is strictly concave.

==>

Theorem 1 *Assume (14). Then the PLE 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 $=0$ for the true parameter values, the projection of the gradient $=0$ 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 PL is maximized by the true parameter values alone.

4 Gradient algorithm

The simplest way of maximizing the PL 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 of the parameters $\hat{}m_{ij}$, for $i \neq j$ only, using this projected gradient:

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

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 PL. Online variants are easy to construct as well.

5 Connection t oCD

CD (Hinton, 2002) is an approximation of MCMC methods. It consists of two related ideas : 1, we fix the initial values in the MCMC method to = the sample points themselves , 2, 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.

Fact. for the model in (1), CD == PL 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(Ackley et al., 1985)

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

where $E\hat{}$: the expectation over the sample distribution,
E_M: the expectation over the distribution of the model with current paramete.

In CD, the expected gradient update for m_{ij} is

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

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 rv which follows the conditional distribution of x_k given all other variables . In the simplest random update sheme , the index k : rv ~unif $\{1, \ldots, n\}$. Note that there are two different methods called CD (Hinton , 2002): 1 based on an objective function , 2 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:

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

while $E_{G(k)} x_i(t) = x_i(t)$ for $k \neq i$. Now, in the second term on the rhs of (24) there is a probability of $(n-2)/n$ that the index $k \neq i$ or j. Then, the Gibbs sampling has no eff ect and can be ignored. With probability $1/n, k = i$, with the same probability , $k = j$.
==>

$$\begin{aligned}(24): \ \Delta m_{ij} &= \hat{E} x_i(t) x_j(t) - \frac{n-2}{n} \hat{E} x_i(t) x_j(t) \\ &\quad - \frac{1}{n} \hat{E} \tanh(\mathbf{m}_i^T \mathbf{x}(t) + b_i) x_j(t) - \frac{1}{n} \hat{E} x_i(t) \tanh(\mathbf{m}_j^T \mathbf{x}(t) + b_i) \\ &= \frac{2}{n} \left[\hat{E} x_i(t) x_j(t) - \frac{1}{2} \hat{E} \tanh(\mathbf{m}_i^T \mathbf{x}(t) + b_i) x_j(t) - \frac{1}{2} \hat{E} x_i(t) \tanh(\mathbf{m}_j^T \mathbf{x}(t) + b_i) \right]\end{aligned} \tag{26}$$

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 CD is typically taken from a sequence $\to 0$ fast enough, the convergence of CD is given by the point where the expected gradient $=0$. Now, the expected gradients in (26) and (27) are equal (up the some insignificant multiplicative constants) to the corresponding symmetrized gradients of the PL. So, the two methods converge in the same points.

The convergence of CD (the same gradient version as we analyzed here) was analyzed in (Carreira-Perpiñán and Hinton, 2005), with the conclusion that CD 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 KL divergence between the distributions given by the model when the estimated parameters for CD or likelihood are used. Thus, their conclusion was that CD 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 CD which always give the same estimate as ML were further developed in (Carreira-Perpiñán and Hinton, 2005). See also (Welling and Sutton, 2005).

6 Simulation results

We performed simulation to validate the different estimation methods for the fully visible BM. We created random matrices $\mathbf{M} \sim$ N(0,.5). The parameters $b_i \sim$N(0,.5). The dimension $n = 5$ 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 MPL. 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 MLE 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 MPLE seems to be consistent in the sense that the estimation error seems to go to 0 when the sample size grows, as implied by our Theorem. Surprisingly, its estimation errors are not really larger than that of ordinary ML. 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 PL (Besag, 1975), provides a consistent estimator for the FVBM. This estimator turns out to be a special case of CD. The literature on BMs does not seem to have paid much attention to PLE so far.

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