Research Memorandum No.674
May 1998

Mean Field Approximation in Bayesian Variable Selection

Yukito Iba

The Institute of Statistical Mathematics
106-8569, Minami-Azabu, Minatoku, Tokyo,
Japan
Email: iba@ism.ac.jp
URL: http://www.ism.ac.jp/~iba/

KEYWORDS
Probabilistic and Statistical Methods, Learning and Generalization, Mean Field Approximation, Model Selection, Multiple Regression, Bayes

(*to be submitted to ICONIP'98*)

Mean Field Approximation in Bayesian Variable Selection

Yukito Iba

Email:iba@ism.ac.jp

The Institute of Statistical Mathematics
106-8569, Minami-Azabu, Minatoku, Tokyo, Japan

ABSTRACT

Variable selection for a multiple regression model (Noisy Linear Perceptron) is studied with a mean field approximation. In our Bayesian framework, variable selection is formulated as estimation of discrete parameters that indicate a subset of the explanatory variables. Then, a mean field approximation is introduced for the calculation of the posterior averages over the discrete parameters. An application to a real world example, Boston housing data, is shown.

KEYWORDS: Probabilistic and Statistical Methods, Learning and Generalization, Mean Field Approximation, Model Selection, Multiple Regression, Bayes

1. Mean Field Approximation

Mean Field Approximation (MFA) is a well-known technique in statistical physics. In general, MFA is defined as a collection of techniques to calculate averages over multivariate distributions, which uses the idea of ignoring higher order correlations and solving self-consistent equations. In this paper, we restrict ourselves to a type of MFA, which treat distributions with discrete variables, specifically binary variables.

MFA has a long history in physics. It is, however, a recent topic in the area of statistical information processing. It is in the late 1980s that the mean field approximation is used in the study of Boltzmann machine learning [14]. An application in image processing [2] also draws much attention of researchers in the area and works with higher order versions of MFA, say, Bethe approximation (TAP equation) [20] appeared. Recently, TAP equation is also applied to the decoding of error correcting codes [8]. A relation between TAP equation and belief propagation in Bayesian network is also discussed in [8].

The aim of this paper is to apply the idea of MFA to a *variable selection* problem. In our Bayesian treatment, variable selection for multiple regression is formulated as estimation of discrete parameters that indicate a subset of the explanatory variables. Then, a mean field approximation is introduced for the calculation of the posterior marginals on the space of submodels, each of which is specified by a subset of the explanatory variables.

Although MFA is closely related to the Hopfield-Tank approach [4] to combinatorial optimization, there is an important conceptual difference [1]. While the Hopfield-Tank method is a tool for optimization, MFA is a method of approximation of the averages over a distribution. MFA is a tool that is useful in the problems where averages over a multivariate distribution play important roles. Two important examples of such problems are the calculation of the averages over a Gibbs distribution in statistical physics and the calculation of the posterior averages in Bayesian statistics. In the present context, MFA gives posterior averages of discrete parameters, each of which gives the weight or "the probability of the existence" of the corresponding explanatory variable. These averages take the values between 0 and 1. At this point, our work is distinguished from other works on variable selection with a Hopfield-Tank method, say, Sakai [16].

[1] In a sense, the relation of MFA to the Hopfield-Tank method is similar to that of MCMC to simulated annealing.

2. Generalization in Neural Networks

Here we overview the issues on the generalization in neural network models and present the motivation of our work in the context of the interpolation between ridge regression and variable selection.

Generalization from a finite set of examples is one of the most important topics in the study of artificial neural network in these ten years. Actually, it is a repeat of a major theme in statistics — How can we beat overfitting and get a better performance of prediction ? For multiple regression models (noisy linear perceptrons) and other types of feed-foward network models, two extreme approaches are known.

An extreme is *ridge regression* or *linear weight decay*, where a penalty function of a quadratic form of synaptic weights (regression coefficients) is added to the target function that is optimized in the learning process [10]. Information from given data is shared among a number of weights with this way of learning. The other extreme is *variable selection* or *pruning* , where we select a subnetwork (a submodel) with as small number of adjustable connections as that is enough to represent the essence of the data. Information is concentrated on small number of weights with the way of learning.

It is natural to interpolate these two extremes and develop methods, such as a "fuzzy variable selection". A number of works in this direction are classified into two categories. A way to realize to such a interpolation is the use of *generalized ridge regression*, or, *nonlinear weight decay*, where a nonquadratic penalty function is used as the penalty to the magnitude of synaptic weights (regression coefficients) [6, 18, 13, 19, 7, 12]. Another approach to this problem is the introduction of weighted averages over a set of models with different architectures (with different sets of explanatory variables), instead of using a single "best" model. Methods of this category are most easily formulated in a Bayesian framework, where the posterior distribution on a space of submodels are treated as well as the posterior distribution on a parameter space of each submodel [11, 5, 3, 9, 15].

Our work belongs to the second category. We are mostly interested in the algorithm for the computation of posterior marginals in a space of submodels. The introduction of the posterior distribution on a set of submodels results in combinatorial explosion of the requirement of computer resources, when the submodels are not linearly ordered. The calculation with exact enumeration is possible only when the number of the variables is small [11]. A potentially powerful tool for this problem is Markov Chain Monte Carlo algorithm (MCMC) [3, 9, 15, 5]. The Monte Carlo approach is, however, still computationally expensive and often suffers from slow convergence.

At this point we introduce a Mean Field Approximation. As already suggested in the previous section, the essence of our idea is the application MFA for calculation of the posterior average on the space of submodels. In the next section, we will give a formulation and basic notations for the implementation of the idea to a linear network, i.e., a multiple regression model.

3. Variable Selection for a Multiple Regression Model

The likelihood function of a multiple regression model is

$$-E_L = -\frac{1}{2\sigma^2}\sum_{\alpha=1}^{N}(y^\alpha - \sum_{i=1}^{M} a_i x_i^\alpha)^2 \ , \tag{1}$$

$$L(\{y^\alpha\}|\{a_i\}) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp(-E_L) \ , \tag{2}$$

where y^α and $\{x_i\}^\alpha$ are the αth observation of the response variable (the output of the network) and the explanatory variables (the inputs of the network) respectably. N is the number of the observations and M is the maximum number of explanatory variables available in data. We introduce a set of indicators $\{\widetilde{S}_i\}$, $\widetilde{S}_i \in \{0,1\}$ for the description of submodels. If and only if $\widetilde{S}_i = 1$, the variable x_i appears in the submodel specified by $\{\widetilde{S}_i\}$. Using the indicators $\{\widetilde{S}_i\}$, a submodel is written as

$$-\widetilde{E}_L = -\frac{1}{2\sigma^2}\sum_{\alpha=1}^{N}(y^\alpha - \sum_{i=1}^{M} a_i x_i^\alpha \widetilde{S}_i)^2 \ , \tag{3}$$

$$\widetilde{L}(\{y_\alpha\}|\{a_i\},\{\widetilde{S}_i\}) = \frac{1}{(2\pi\sigma^2)^{N/2}}\exp(-\widetilde{E}_L) \ . \tag{4}$$

With this submodel, the maximum likelihood estimator of the regression coefficient a_i is obtained as the solution of the corresponding normal equation

$$V_{x_i x_i} a_i + \sum_j V_{x_i x_j} a_j \widetilde{S}_j = V_{x_i y} \ , \tag{5}$$

when $\widetilde{S}_i = 1$. Here, sufficient statistics $V_{y^2}, V_{x_i y}, V_{x_i x_j}$ are defined as $V_{y^2} = \sum_\alpha (y^\alpha)^2$, $V_{x_i y} = \sum_\alpha (x_i^\alpha y^\alpha)^2$, $V_{x_i x_j} = \sum_\alpha (x_i^\alpha x_j^\alpha)^2$.

When we define $S_i = 2\widetilde{S}_i - 1$ and rearrange the terms, we have

$$-\widetilde{E}_L = \frac{1}{2\sigma^2}(C + \sum_i H_i S_i + \sum_i \sum_{j>i} J_{ij} S_i S_j) \ , \tag{6}$$

$$C = -V_{y^2} + \sum_i a_i V_{x_i y} - \frac{1}{2}\sum_i \sum_{j>i} a_i a_j V_{x_i x_j} - \frac{1}{2}\sum_i a_i^2 V_{x_i x_i} \ , \tag{7}$$

$$H_i = a_i V_{x_i y} - \frac{1}{2}\sum_{j\neq i} a_i a_j V_{x_i x_j} - \frac{1}{2} a_i^2 V_{x_i x_i} \ , \tag{8}$$

$$J_{ij} = -\frac{1}{2} a_i a_j V_{x_i x_j} \tag{9}$$

using $\widetilde{S}_i^2 = \widetilde{S}_i$. For later convenience, the indicator S_i that takes the values of ± 1 is introduced.

To define a Bayesian model, we should specify prior distributions. The prior distribution of the indicators $\{S_i\}$ is assumed as

$$\pi(\{S_i\}) = \frac{\exp(-h\sum_i S_i)}{(\exp(h) + \exp(-h))^M} \ , \tag{10}$$

where the hyperparameter h determines the degree of the penalty to submodels with larger number of explanatory variables. In this paper, we also assume that the prior of the coefficients $\{a_i\}$ is the uniform density in $\mathbf{R}^M$. Then the posterior distribution $\widetilde{P}_{pos}(\{a_i\},\{S_i\})$ of $\{S_i\}, \{a_i\}$ is formally defined as a Gibbs distribution

$$\widetilde{P}_{pos}(\{a_i\},\{S_i\}) = \frac{\exp(-\widetilde{E}_{pos})}{\widetilde{Z}_{pos}} \tag{11}$$

with the energy $-\widetilde{E}_{pos} = -\widetilde{E}_L - h\sum_i S_i$ and the normalization $\widetilde{Z}_{pos} = \sum_{\{S_i\}} \int \Pi_i da_i \exp(-\widetilde{E}_{pos})$ [2]. Here and hereafter, $\sum_{\{S_i\}}$ denote the summation over 2^M configurations of $\{S_i\}$.

4. MFA for Bayesian Variable Selection

Now, we discuss an application of Mean Field Approximation to variable selection for multiple regression. First, we consider the posterior distribution of $\{S_i\}$ conditioned with a given set $\{a_i\}$ of regression coefficients,

$$P_S(\{S_i\}|\{a_i\}) = \frac{\exp(-\widetilde{E}_{pos})}{Z_S} \ . \tag{12}$$

The normalization constant Z_S is defined by

$$Z_S = \sum_{\{S_i\}} \exp(-\widetilde{E}_{pos}) \ . \tag{13}$$

Consider a distribution

$$Q(\{S_i\}) = \Pi_i \frac{1 + m_i \cdot S_i}{2} \tag{14}$$

and the KL-divergence

$$D(Q;P_S) = \sum_{\{S_i\}} Q(\{S_i\}) \log \frac{Q(\{S_i\})}{P_S(\{S_i\})|\{a_i\})} \tag{15}$$

between Q and P_S. An easy calculation shows that

$$\begin{aligned} D(Q;P_S) = \ & -\tfrac{1}{2\sigma^2}(C + \textstyle\sum_i \sum_{j>i} J_{ij} m_i m_j + \sum_i H_i m_i) \\ & + h\textstyle\sum_i m_i + \log Z_S \\ & + \textstyle\sum_i \left\{ \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2} \right\} \end{aligned} \tag{16}$$

[2] The integral in the definition of $\widetilde{Z}_{pos}$ diverges with the uniform prior for $\{a_i\}$. Apparently, this is not significant in our approximation, where a point estimate $\{a_i^*\}$ of the regression coefficients is used. It might cause, however, some difficulties, e.g. in a Bayesian interpretation of the strength h of the penalty.

The values $\{m_i^*\}$ of $\{m_i\}$ that minimize $D(Q;P_S)$ is a solution of the "self-consistent equation"

$$m_i^* = \tanh\left(\frac{\sum_{j\neq i} J_{ij}m_j^* + H_i}{2\sigma^2} - h\right) . \quad (17)$$

From the inequality $D(Q;P_S) \geq 0$, the relation $\log Z_S \geq \log Z_{MF}^*$ is derived, where

$$\log Z_{MF}^* = \frac{1}{2\sigma^2}(C + \sum_i H_i m_i^* + \sum_i \sum_{j>i} J_{ij} m_i^* m_j^*) - h\sum_i m_i^* - \sum_i \left\{ \frac{1+m_i^*}{2}\log\frac{1+m_i^*}{2} + \frac{1-m_i^*}{2}\log\frac{1-m_i^*}{2} \right\} . \quad (18)$$

The equality is hold only when the distributions Q and P_S are identical. The form of the distribution (14) suggests that the ith component m_i^* of a solution of the equation (17) is regarded as an approximation of the averages of the indicator S_i over the distribution (12). Thus the posterior probability p_i that the variable x_i is contained in a submodel is approximated by

$$p_i^* = \frac{1+m_i^*}{2} . \quad (19)$$

We can also regard $\log Z_{MF}^*$ as an approximation of $\log Z_S$. The self-consistent equation (17) and the expressions (19) and (18) are the essence of the MFA in our problem.

So far, we are working with a given set of regression coefficients $\{a_i\}$. In our problem, they are unknown parameters. A way to deal with them is to replace them with point estimates that maximize the likelihood

$$l(\{a_i\},\sigma^2) = \log \sum_{\{S_i\}} \widetilde{L}(\{y^\alpha\}|\{a_i\},\{S_i\})\pi(\{S_i\}) \quad (20)$$

marginalized over $\{S_i\}$ [3]. It is represented by the normalization constant Z_S of (13) as

$$l(\{a_i\},\sigma^2) = \log Z_S - M\log(e^h + e^{-h}) - \frac{N}{2}\log 2\pi\sigma^2 . \quad (21)$$

This marginal likelihood (21) is approximated by the MFA as

$$l_{MF}(\{a_i\},\sigma^2) = \log Z_{MF}^* - M\log(e^h + e^{-h}) - \frac{N}{2}\log 2\pi\sigma^2 . \quad (22)$$

The values $\{a_i^*\}$ that maximize (22) is a solution of a "soft version" of the normal equation [4]

$$V_{x_i x_i} a_i^* + \sum_{j\neq i} V_{x_i x_j} a_j^* p_j^* = V_{x_i y} . \quad (23)$$

The equation (23) is, in fact, a equation nonlinear in $\{a_i^*\}$, because the probability $\{p_j^*\}$ is a function of $\{a_i^*\}$ implicitly determined with (19) and (17,8,9). In actual implementation, the following algorithm is used to solve the set of equations.

(1) Set the initial values $\{m_i^0\}$ of $\{m_i\}$ and $t := 0$.

(2) Solve the equation of $\{a_i^t\}$

$$V_{x_i x_i} a_i^t + \sum_{j\neq i} V_{x_i x_j} a_j^t p_j^t = V_{x_i y} \quad (24)$$

with $p^t = (1+m_i^t)/2$.

(3) Iterate the equation

$$m_i^{t+1} := (1-\delta)\cdot m_i^t + \delta \cdot \tanh\left(\frac{\sum_{j\neq i} J_{ij}m_j^t + H_i}{2\sigma^2} - h\right) . \quad (25)$$

Here, the variables $\{H_i\}$,$\{J_{ij}\}$ are defined by (8,9) with $a_i = a_i^t$. δ is a constant that satisfies $0 < \delta \leq 1$.

(4) Set $t := t+1$. Check the convergence and return to the step (2), if necessary. If convergence is ensured, set $p_i^* := (1+m_i^t)/2$ and $a_i^* := a_i^t$ and exit.

We can also estimate the variance σ^2 of the residual with the maximization of (22). For this purpose, we add a step for the estimation of σ^2 to the iteration. In our implementation, the value of σ^2 is updated once per the iteration.

5. An Example

In this section, we discuss an application of the proposed method to a real world example. We use the dataset taken from p.244 of Belsley et al. [1]. This dataset, known as Boston housing data, is also analyzed by other authors with different methods [1, 12, 17]. There are thirteen explanatory variables ($M = 13$) and one response variable, the logarithm of the median value of owner-occupied homes for each area of Boston. The sample size is $N = 506$.

Our results on this data are shown in Fig.1 and Fig.2. In the experiment, the variance σ^2 is estimated from the data. Although we observe local optima with some parameters of h, they are not serious and the solution with the highest frequency at each value of h is plotted in the figures.

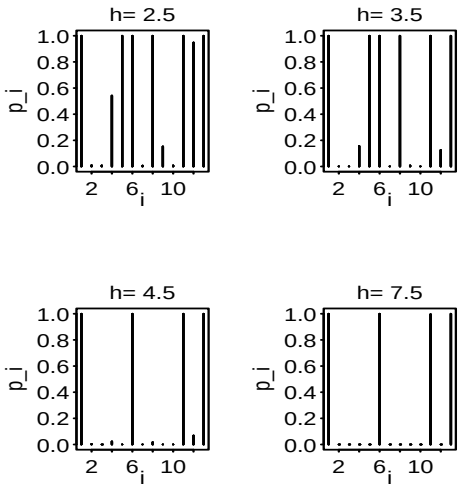

Fig.1: The results with some typical values of h. The index i of a variable is shown in the horizontal axis and the probability $p_i^* = (1+m_i^*)/2$ is shown in the vertical axis.

[3] Although the use of a point estimation of $\{a_i\}$ is not fully justified in a Bayesian paradigm and leads to an approximation beyond the conventional use of MFA, it greatly simplified the situation.

[4] Note that the values $\{m_i^*\}$ should be considered as a function of $\{a_i^*\}$ in the derivation. However, with the use of $\frac{\partial D(Q;P_S)}{\partial m_i}|_{\{m_i\}=\{m_i^*\}} = 0$, the result coincides with that of a naive calculation.

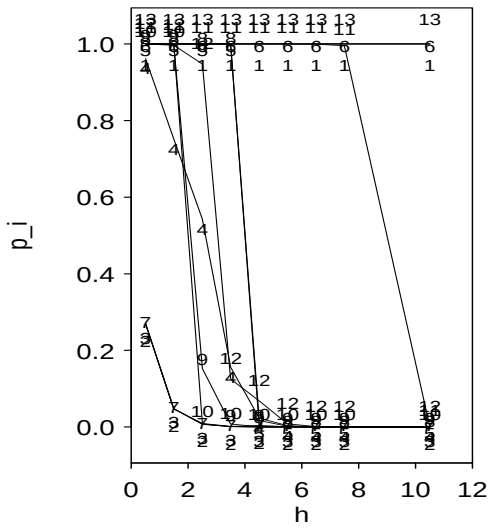

Fig.2: The change of p_i^* is shown with h. The number used as a symbol indicates the index i of the variable. The positions of the symbols are slightly perturbed by noise to prevent complete overlap between them. Lines are accurate.

In Fig.2, the probability p_i^* takes the value $p_i^* \sim 1$, for $i = 1, 6, 13$, even with a larger value of the penalty h. It seems consistent with the result in [17]. For most of the other variables, the probability p_i^* shows a sudden jump between 0 and 1, or, a simple decay with $p_i^* \sim \frac{\exp(-h)}{\exp(h)+\exp(-h)}$, as the value of h is increased. However, for some i, say $i = 4$, it shows a gradual decay and has nontrivial values between 0 and 1 in a range of h.

6. Future Problems

In the present treatment, the strength h of the penalty to the number of variables is a free hyperparameter. It is not a serious fault when we use the proposed method as a tool for *regression diagnostics* , i.e., we observe the behavior of solutions in different values of h and get information on data. A criterion to determine the optimal value of h is, however, required, when the method is used as a fully automatic tool for the prediction. Although there are a few possible approaches, their implementations and tests are left for future studies.

Another important problem is the study of the multiple solutions in our iterative procedure. For example, when near loss of the linear independence among explanatory variables (*colinearity*) exists, the algorithm is likely to have multiple fixed points. Note that the existence of local optima is usually regarded as a mere difficulty of algorithms in information processing. There seems, however, often a possibility of exploration of the nature of the data from the emergence of multiple extremum. It is interesting if the proposed algorithm provides an example for such a study. For this purpose, it requires more investigation into the behavior of the algorithm in complex situations.

We acknowledge Dr.Kabashima for fruitful discussions on the work.

References

[1] Belsley,D.A., Kuh,E. and Welsch,R.E., *Regression Diagnostics* , Wiley, New York (1980).

[2] Geiger,D. and Girosi,F., "Parallel and deterministic algorithms for MRF's : surface reconstruction and integration", *IEEE Trans. Pattern Anal. Machine Intell.*, **Vol. PAMI-13** , 401–412 (1991).

[3] George,E.I. and McCulloch,R.E. "Variable selection via Gibbs sampling", *Journal of the American Statistical Association*, **Vol.88**, 881–889 (1993).

[4] Hopfield,J.J. and Tank,D.W.(1985) " "Neural" computation of decisions in optimization problems", *Biological Cybernetics*, **Vol.52**, 141–152.

[5] Iba, Y., "Metropolis-type Monte Carlo algorithm and quasi-Bayesian estimation procedure : An application to a change-point problem", *Proceedings of the Institute of Statistical Mathematics*, **Vol.39**, No.2, 225–244 (1991) (in Japanese with English summary).

[6] Ishiguro, M., "A Bayesian multiple regression model", *Proceedings of the Institute of Statistical Mathematics*, **Vol.33** No.1, 8–10 (1985) (in Japanese).

[7] Ishikawa,M. , "Structural learning with forgetting", *Neural Networks*, **Vol.9**, 509–521 (1996).

[8] Kabashima, Y. and Saad,D. , "Belief propagation .vs. TAP for decoding corrupted messages", submitted to *Europhysics Letters* (1998)

[9] Kuo,L and Mallick,B. , "Variable selection for regression models", preprint (1996).

[10] MacKay, D.J.C. , "A practical Bayesian framework for backprop networks", *Neural Computation* , **Vol.4**, 448–472 (1992).

[11] Mitchell,J. and Beauchamp,J.J. , "Bayesian variable selection in linear regression" *Journal of the American Statistical Association*, **Vol.83**, 1023–1036 (1988) (with discussion).

[12] Neal,R.M. , *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics 118, Springer (1996).

[13] Nowlan,S.J. and Hinton,G.E. , "Simplifying neural networks by soft weight-sharing", *Neural Computation*, **Vol.4**, No.4, 473–493 (1992).

[14] Peterson,C. and Hartman,E. , "Explorations of the mean field theory learning algorithm", *Neural Networks*, **Vol.2**, 475–494 (1989).

[15] Phillips,D.B. and Smith,A.F.M. , "Bayesian model comparison via jump diffusions", *Markov Chain Monte Carlo in Practice*, Chap.13, Eds. Gilks,W.R., Richardson,S. and Spiegelhalter,D.J., Chapman and Hall (1996).

[16] Sakai,H., "An application of BIC type method to harmonic analysis and a new criterion for order determination of an AR process", *ISM Cooperative Research Report 20*, The Institute of Statistical Mathematics(Tokyo) (1993).

[17] Shimodaira, H. "A model search technique based on confidence set and map of models", *Proceedings of the Institute of Statistical Mathematics*, **Vol.41**, 131–147 (1993) (in Japanese).

[18] Weigend,A.S. and Rumelhart,D.E. , "Generalization through minimal networks with application to forecasting", *Computing Science and Statistics*, Proceedings of the 23rd symposium on the interface, 156–163, Interface Foundation, Fairfax Station, Va. (1991).

[19] Williams,P.M., "Bayesian regularization and pruning using a Laplace prior", *Neural Computation*, **Vol.7**, No.1, 117-143 (1995).

[20] Zhang,J., "The mean field theory in EM procedures for Markov random fields" *IEEE Transactions on Signal Processing*, **Vol.40**, 2570–2583 (1992).