NEURAL NETWORK APPROXIMATIONS TO POSTERIOR DENSITIES: AN ANALYTICAL APPROACH

Lennart F. Hoogerheide, Johan F. Kaashoek and Herman K. van Dijk

Econometric and Tinbergen Institutes, Erasmus University Rotterdam

Econometric Institute Report EI 2003-38

KEYWORDS: neural networks, Bayesian inference, importance sampling, Markov chain Monte Carlo.

Introduction

Markov Chain Monte Carlo (MCMC) methods like Metropolis-Hastings (MH) and Gibbs sampling are extensively used in Bayesian analyses of econometric and statistical models. The theory of Markov chain samplers starts with Metropolis et al. (1953) and Hastings (1970). Indirect independence sampling methods such as importance sampling (IS) have also been successfully applied within Bayesian inference. Importance sampling, see Hammersley and Handscomb (1964), has been introduced in Bayesian inference by Kloek and Van Dijk (1978) and is further developed by Van Dijk and Kloek (1980,1984) and Geweke (1989).

However, in practice, the convergence behavior of Monte Carlo methods is still often uncertain. The complex structure of a model or some extraordinary properties of the data may cause this problem. We mention three cases. First, Hobert and Casella (1996) show that the Gibbs sampler does not converge in the case of a hierarchical linear mixed model if the prior is uniform. The reason is that the posterior of (at least) one conditional variance is improper. A second example of a complex model is a set of equations with a near reduced rank structure for the matrix of coefficients. We refer to the studies by Schotman and Van Dijk (1991) and Kleibergen and Van Dijk (1994, 1998). As a third case we mention a multimodal target density, which one may encounter in mixture processes with a small number of observations around one of the different modes. This may cause problems for all methods. If the MH candidate density is unimodal, with a low probability of drawing candidate values in one of the other modes, then this mode may be completely missed, even if the sample size gets very large. In this case importance sampling with a unimodal normal or Student t importance density may yield a sample in which most drawings have a negligible weight and only a few drawings almost completely determine the sampling results.

So, an important problem is the choice of the candidate or importance density, especially when one knows little about the shape of the target density.

In Hoogerheide, Kaashoek and Van Dijk (2002) the class of neural network sampling methods is introduced to sample from a target (posterior) distribution that may be multi-modal or skew, or exhibit strong correlation among the parameters. That is, a class of methods to sample from non-elliptical distributions. In these methods the neural network is used as an importance function in IS or as a candidate density in MH.

In this note we suggest an analytical approach to estimate the moments of a certain (target) distribution, where we mean by 'analytical' that no sampling algorithm like MH or IS is needed. The basic idea of this approach is very simple. First, a neural network is constructed that approximates the target density. An important advantage of neural network functions is their 'universal approximation property'. That is, neural network functions can provide approximations of any square integrable function to any desired accuracy, see Gallant and White (1989). As an application of Kolmogorov's general superposition theorem (see Kolmogorov (1957)), the neural network approximation property is eluded by Hecht-Nielsen (1987). This approximation property implies that the algorithm can handle certain 'strange' target distributions, like multi-modal, extremely skew, strongly correlated or fat-tailed distributions.

Second, the moments of this neural network distribution are computed; these moments are estimates of the moments of the target distribution. The nice property of the standard feed-forward 3-layer network with our choice of activation function is that we have analytical expressions for the moments of the distribution with this neural network function as a density.

The proposed method is applied on a set of illustrative examples. Our results indicate that the neural network approach is feasible, even in a case where a 'standard' Gibbs approach would fail or be extremely slow.

An analytical approach to estimate moments of a distribution

Our approach may be summarized as follows:

Step 1: Construct a neural network approximation to the target density.

Step 2: Compute the moments of the neural network distribution; these are estimates of the moments of the target distribution.

Of course, it is not immediately clear in which cases this approach is feasible, or how to construct a neural network approximation to a target density. First, two assumptions are needed. The target density has to be square integrable, and the domain of the random variable must be bounded. In practice, the second assumption means that there is a certain bounded area, beyond which the probability mass is negligible. Second, in order to approximate a certain target density $f(x)$ of a random vector $X = (X_1, \ldots, X_n)'$, we suggest to use the following type of 3-layer neural network, a feed-forward multilayer perceptron (FMLP):

$$\begin{aligned} & nn(x_1, \ldots, x_n) = \\ = & \sum_{h=1}^{H} c_h \left(\frac{1}{\pi} \arctan(a_h' x + b_h) + \frac{1}{2} \right) + d, \quad (1) \end{aligned}$$

where $a_h \in \mathbb{R}^n$ and $b_h, c_h \in \mathbb{R} (h = 1, \ldots, H), d \in \mathbb{R}$. The reason for choosing the (scaled) arctangent function as an activation function is that it is analytically integrable infinitely many times; we have derived analytical expressions for its integrals, which are given in the sequel of this paper. Therefore we can analytically compute the moments of the distribution of which a density kernel is given by (1) on a bounded region, and zero elsewhere. It should be noted that (1) is not automatically non-negative for all input values x, a requirement for (1) to be a good density kernel. However, requiring the neural network function to be a 'very good' approximation to the target density should result in non-negativity for (almost) all input values.

We suggest to estimate (or 'learn') the parameters (or 'weights') of the neural network by minimizing the sum of squared 'residuals':

$$\sum_{i=1}^{m} \left(nn(x^i) - f(x^i) \right)^2,$$

where $\{x^i | i = 1, \ldots, m\}$ is a set of points in the bounded region to which the random vector X is restricted.

Three questions remain: How to choose the points $x^i (i = 1, \ldots, m)$, how to pick the number of points m, and how to choose the number of hidden cells H? As an answer to these questions, we suggest the following adaptive approach to perform Step 1 above:

Step 1a: Choose initial values of H and m (e.g., $H = 50$ and $m = 2500$).

Step 1b: Estimate the parameters of a neural network with H hidden cells for m points $\{x^i | i = 1, \ldots, m\}$ drawn from a uniform distribution on the domain of X.

Step 1c: Compute the R^2, the squared correlation between the target density f and the neural network nn in the points $\{x^i | i = 1, \ldots, m\}$. If this R^2 is high enough (e.g., $R^2 > 0.995$), then go to Step 1d. If this R^2 is too small (e.g., $R^2 < 0.995$), then add hidden cells to the network (e.g., increase H by 50), and go back to Step 1b.

Step 1d: Compute the R^2 for a set of more than m points (e.g., $2m$ points) in order to check whether the neural network also gives a good approximation to the target density outside the 'estimation set' $\{x^i | i = 1, \ldots, m\}$. If this R^2 is also large enough (e.g. $R^2 > 0.99$), then stop: we will use the neural network at hand to estimate the moments of the target density. If this R^2 is too small, increase the number of points m (e.g. double m) and go back to Step 1b.

The moments of a neural network distribution

We have already mentioned that we have derived analytical expressions for the moments of the distribution of which the density is given by (1) on a bounded region, and zero elsewhere. We shall now formalize this, and give the expressions. Suppose the

vector $X = (X_1, \ldots, X_n)'$ has the following density $p(x_1, \ldots, x_n)$:

$$p(x_1, \ldots, x_n) = \begin{cases} nn(x_1, \ldots, x_n) & \text{if } \underline{x}_i \leq x_i \leq \bar{x}_i \\ & (i = 1, \ldots, n) \\ \\ 0 & \text{else} \end{cases} \tag{2}$$

where $[\underline{x}_i, \bar{x}_i]$ is the interval to which the random variable X_i $(i = 1, 2, \ldots, n)$ is restricted, and where

$$\begin{aligned} & nn(x_1, \ldots, x_n) = \\ = & \sum_{h=1}^{H} \frac{c_h}{\pi} \arctan(a_h' x + b_h) + \frac{1}{2} \sum_{h=1}^{H} c_h + d. \end{aligned} \tag{3}$$

Then the expectation of X_n^k $(k = 1, 2, \ldots)$ is given by:

$$\begin{aligned} & E(X_i^k) = \\ = & \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} \cdots a_{hn}} \sum_{D_1=0}^{1} \cdots \sum_{D_n=0}^{1} (-1)^{D_1 + \cdots + D_n} \times \\ & \times \left[\sum_{m=0}^{k} \left(-\frac{1}{a_{hi}} \right)^m \frac{k!}{(k-m)!} x_{i,D_i}^{k-m} \times \right. \\ & \left. \times J_{n+m} \left(\sum_{l=1}^{n} a_{hl} x_{l,D_l} + b_h \right) \right] \\ & + \left(\frac{1}{2} \sum_{h=1}^{H} c_h + d \right) \frac{1}{k+1} (\bar{x}_i^{k+1} - \underline{x}_i^{k+1}) \times \\ & \times \prod_{l=1; l \neq i}^{n} (\bar{x}_l - \underline{x}_l) \end{aligned} \tag{4}$$

and $E(X_i X_j)$ $(i, j = 1, 2, \ldots, n; i \neq j)$ is equal to:

$$\begin{aligned} & E(X_i X_j) = \\ = & \sum_{h=1}^{H} \frac{c_h}{\pi a_{h1} \cdots a_{hn}} \sum_{D_1=0}^{1} \cdots \sum_{D_n=0}^{1} (-1)^{D_1 + \cdots + D_n} \times \\ & \times \left[x_{i,D_i} x_{j,D_j} J_n \left(\sum_{l=1}^{n} a_{hl} x_{l,D_l} + b_h \right) \right. \\ & - \frac{a_{hi} x_{i,D_i} + a_{hj} x_{j,D_j}}{a_{hi} a_{hj}} J_{n+1} \left(\sum_{l=1}^{n} a_{hl} x_{l,D_l} + b_h \right) \\ & \left. + \frac{1}{a_{hi} a_{hj}} J_{n+2} \left(\sum_{l=1}^{n} a_{hl} x_{l,D_l} + b_h \right) \right] \\ & + \left(\frac{1}{2} \sum_{h=1}^{H} c_h + d \right) \frac{1}{4} (\bar{x}_i^2 - \underline{x}_i^2)(\bar{x}_j^2 - \underline{x}_j^2) \times \\ & \times \prod_{l=1; l \neq i,j}^{n} (\bar{x}_l - \underline{x}_l). \end{aligned} \tag{5}$$

In formulas (4) and (5) we define $x_{i,0} \equiv \bar{x}_i$ and $x_{i,1} \equiv \underline{x}_i$ $(i = 1, 2, \ldots, n)$, the upper and lower bounds of the integration intervals. The function $J_n : \mathbb{R} \to \mathbb{R}$ is the n-th integral of the arctangent function:

$$\begin{aligned} J_n(x) & \equiv \int \cdots \int \arctan(x) dx \cdots dx \\ & = p_n(x) \arctan(x) + q_n(x) \ln(1 + x^2) + r_n(x), \end{aligned} \tag{6}$$

where p_n and q_n are polynomials of degree n and $n-1$, respectively:

$$\begin{aligned} p_n(x) & = p_{n,0} + p_{n,1}\, x + \cdots + p_{n,n-1}\, x^{n-1} + p_{n,n}\, x^n \\ q_n(x) & = q_{n,0} + q_{n,1}\, x + \cdots + q_{n,n-1}\, x^{n-1} \end{aligned}$$

The coefficients $p_{n,k}$ $(k = 0, 1, \ldots, n)$ are:

$$p_{n,k} = \begin{cases} \frac{(-1)^{(n-k)/2}}{(n-k)!k!} & \text{if } n-k \text{ is even,} \\ \\ 0 & \text{if } n-k \text{ is odd,} \end{cases} \tag{7}$$

and the coefficients $q_{n,k}$ $(k = 0, 1, \ldots, n-1)$ are given by:

$$q_{n,k} = \begin{cases} \frac{(-1)^{(n-k+1)/2}}{2(n-k)!k!} & \text{if } n-k \text{ is odd,} \\ \\ 0 & \text{if } n-k \text{ is even.} \end{cases} \tag{8}$$

The polynomial r_n (of degree at most $n-1$) plays the role of the integrating constant. For the proof of this result we refer to Appendix A of Hoogerheide, Kaashoek and Van Dijk (2002).

Example: Bivariate conditionally normal distribution

We shall now apply our approach on an illustrative example of bivariate conditionally normal distributions.

Let x_1 and x_2 be two jointly distributed random variables, for which x_1 is normally distributed given x_2 and vice versa. Then the joint distribution, after location and scale transformations in each variable, can be written as (see Gelman and Meng (1991)):

$$\begin{aligned} f(x_1, x_2) \quad \propto \quad & \exp(-\frac{1}{2}[A x_1^2 x_2^2 + x_1^2 + x_2^2 \\ & -2B x_1 x_2 - 2C_1 x_1 - 2C_2 x_2]), \end{aligned}$$

Just like Gelman and Meng (1991), we consider the symmetric subfamily in which $A = 1, B = 0, C_1 = C_2 = C$, with conditional distributions

$$x_1 | x_2 \quad \sim \quad N\left(\frac{C}{1 + x_2^2}, \frac{1}{1 + x_2^2} \right), \tag{9}$$

$$x_2 | x_1 \quad \sim \quad N\left(\frac{C}{1 + x_1^2}, \frac{1}{1 + x_1^2} \right). \tag{10}$$

We consider three cases: $C = 0, 4, 10$. We apply our analytical neural network approach to each case and compare its results with the real values that are obtained by deterministic integration, which is easy in this illustrative bivariate example. We also apply the Gibbs sampler (see e.g., Geman and Geman (1984)) to each case; we construct two Gibbs sequences and we say that the Gibbs sampler has converged if the two sample means of the Gibbs sequences both differ less than 0.005. The results are in Table 1.

First note that our neural network approach yields quite accurate estimates. Apparently the three densities have been approximated quite well by the corresponding neural networks. This is also indicated by the high R^2 and the contour plots in Figure 1. Second, notice that the larger the value of C, the more hidden cells H and 'estimation points' m are needed to provide a good neural network approximation. However, for larger values of C the Gibbs sampler also needs more drawings to reach convergence. In fact, for $C = 10$ the Gibbs sampler had not converged at all after 100000000 drawings. The reason is that the Gibbs sequences remained in the same mode for 100000000 iterations in a row. Therefore, we may conclude that the Gibbs sampler will take at least many billions of drawings to converge to the real values in this case of bimodality.

Final remarks

We have shown an example in which our analytical neural network approach is feasible, even in a case where a 'standard' Gibbs approach would fail or be extremely slow. In practice, the construction of a neural network with an almost perfect fit often takes much time. In such a case it is computationally more efficient to construct a neural network that provides a 'reasonable' approximation to the posterior density function, and then use this as a candidate density or importance function in MH or IS. For this approach we refer to Hoogerheide, Kaashoek and Van Dijk (2002).

Contact information

Corresponding author: L.F. Hoogerheide, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. E-mail: `lhoogerheide@few.eur.nl`

The paper by Hoogerheide, Kaashoek and Van Dijk (2002) is available as report EI 2002-48 at:
http://www.few.eur.nl/few/research/pubs/ei/2002/reports.htm

References

Gallant, A.R. and H. White (1989): "There exists a neural network that does not make avoidable mistakes", in *Proc. of the International Conference on Neural Networks*, San Diego, 1988 (IEEE Press, New York).

Geman, S. and D. Geman (1984): "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gelman, A. and X. Meng (1991): "A Note on Bivariate Distributions That Are Conditionally Normal", *The American Statistician*, 45, 125-126.

Geweke, J. (1989): "Bayesian inference in econometric models using Monte Carlo integration", *Econometrica*, 57, 1317-1339.

Hammersley, J. and D. Handscomb (1964): "Monte Carlo Methods". Chapman and Hall, London.

Hastings, W.K. (1970): "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika*, 57, 97-109.

Hecht-Nielsen, R. (1987): "Kolmogorov mapping neural network existence theorem", in *Proc. IEEE First International Conference on Neural Networks*, San Diego, 1987, 11-13.

Hobert, J.P. and G. Casella (1996): "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models", *Journal of the American Statistical Association*, 91(436), 1461-1473.

Hoogerheide, L.F., J.F. Kaashoek and H.K. van Dijk (2002): "Functional Approximations to Posterior Densities: A Neural Network Approach to Efficient Sampling", Econometric Institute report 2002-48, Erasmus University Rotterdam.

Kleibergen, F.R., and H.K. Van Dijk (1994): "On the Shape of the Likelihood/Posterior in Cointegration Models", *Econometric Theory*, 10(3-4), 514-551.

Kleibergen, F.R., and H.K. Van Dijk (1998): "Bayesian Simultaneous Equations Analysis using Reduced Rank Structures", *Econometric Theory*, 14(6), 701-743.

Kloek, T., and H.K. Van Dijk (1978): "Bayesian estimates of equation system parameters: an application of integration by Monte Carlo", *Econometrica*, 46, 1-19.

Kolmogorov, A.N. (1957): "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition", *American Mathematical Monthly Translation*, Vol. 28, pp 55-59. (Russian original in Doklady Akademii Nauk SSSR, 144, 953-956)

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953): "Equations of State Calculations by Fast Computing Machines", *Journal of Chemical Physics*, 21, 1087-1091.

Schotman, P.C. and H.K. van Dijk (1991): "A Bayesian Analysis of the Unit Root in Real Exchange Rates", *Journal of Econometrics*, 49, 195-238.

Van Dijk, H.K., and T. Kloek (1980): "Further experience in Bayesian analysis using Monte Carlo integration", *Journal of Econometrics*, 14, 307-328.

Van Dijk, H.K., and T. Kloek (1984): "Experiments with some alternatives for simple importance sampling in Monte Carlo integration", in *Bayesian Statistics 2*, ed. by J. M. Bernardo, M. Degroot, D. Lindley, and A. F. M. Smith, Amsterdam, North-Holland.

Table 1: Results for the distributions with $C = 0$, $C = 4$ and $C = 10$

	C=0			C=4			C=10		
	real values	moments of NN	Gibbs	real values	moments of NN	Gibbs	real values	moments of NN	Gibbs
$E(X_1)$	0.000	0.002	-0.002	1.860	1.865	1.864	4.946	4.936	-
$E(X_2)$	0.000	0.001	-0.003	1.860	1.847	1.855	4.946	4.935	-
$\sigma(X_1)$	0.846	0.839	0.846	1.666	1.670	1.666	4.894	4.894	-
$\sigma(X_2)$	0.846	0.842	0.842	1.666	1.669	1.664	4.894	4.872	-
$\rho(X_1, X_2)$	0.000	-0.004	-0.005	-0.839	-0.835	-0.839	-0.979	-0.982	-
R^2		0.9999			0.9961			0.9956	
H		50			100			150	
m		2500			5000			10000	
drawings			50000			2000000			$> 10^8$

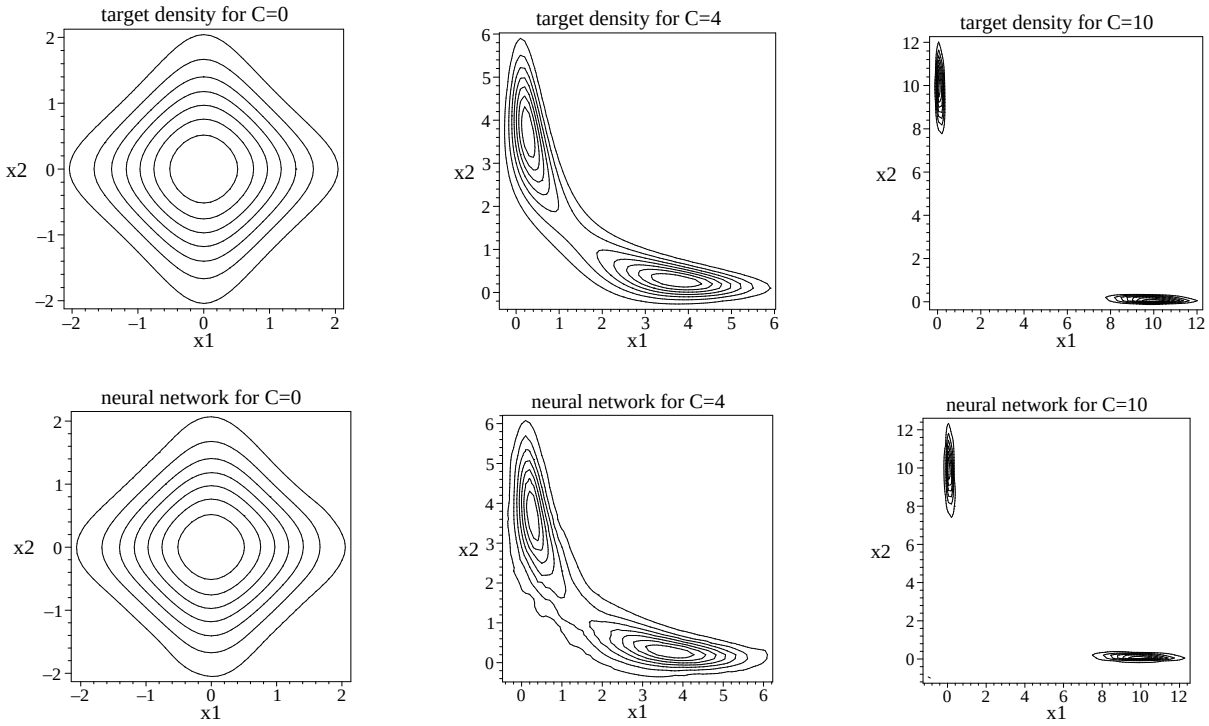

Figure 1: Contour plots of target densities and neural network approximations