

# A Finite Population Estimation Study with Bayesian Neural Networks

FAMING LIANG and ANTHONY YUNG CHEUNG KUK<sup>1</sup>

## ABSTRACT

In this article, we study the use of Bayesian neural networks in finite population estimation. We propose estimators for finite population mean and the associated mean squared error. We also propose to use the student  $t$ -distribution to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys. Numerical results show that Bayesian neural networks have made a significant improvement in finite population estimation over linear regression based methods.

**KEY WORDS:** Bayesian model averaging; Bayesian neural network; Evolutionary Monte Carlo; Finite population; Markov Chain Monte Carlo; Prediction.

## 1. INTRODUCTION

Regression estimation is widely used in sample surveys for incorporating auxiliary population information (Cochran 1977) with the underlying model

$$y_t = \beta_0 + x_{t1} \beta_1 + \dots + x_{tp} \beta_p + \epsilon_t, \quad t = 1, 2, \dots, n, \quad (1)$$

where  $y_t$  is the survey variable for the  $t^{\text{th}}$  element of a population,  $\mathbf{x}_t = (x_{t1}, \dots, x_{tp})$  is the vector of auxiliary variables associated with  $y_t$ ,  $\beta_0, \beta_1, \dots, \beta_p$  are the regression coefficients, and  $\epsilon_t$  is the independent disturbance with zero mean and common variance. Although this model generally performs well, it has several inherent limitations. First, the model is specified linearly and thus can not capture some types of nonlinear relationship, which may be essential in some applications. Second, the least squares estimate, which is widely used for the model (1), may not be reliable in the presence of collinearity among the auxiliary variables. In this case, techniques, such as condition number reduction (Bankier 1990), ridge regression (Bardsley and Chambers 1984), and various variable selection procedures (Silva and Skinner 1997), have to be used to improve the poor prediction performance of the model. Third, in the presence of outliers, the least squares estimate may be severely affected by the outliers.

There are attempts to lessen the dependence of estimators on the linear model (1). Firth and Bennett (1998) identify a sufficient "internal bias calibration" condition under which a model-based estimator is automatically design consistent, regardless of how well the underlying model fits the population. The condition is met by certain estimators based on linear models, certain canonical link generalized linear models and nonparametric regression estimators constructed from them by a particular style of local likelihood fitting.

Bias can also be calibrated externally, if not internally. Chambers, Dorfman and Wehrly (1993) start with a predictor of the population mean based on a heteroscedastic linear model and adjust for its bias using nonparametric regression. Kuk and Welsh (2001) propose a robustified model-based approach whereby a working model is first fitted using robust methods and subsequently the conditional distributions of the residuals given  $\mathbf{x}$  are estimated nonparametrically to account for local model departure or outliers in localized regions.

Another way of incorporating auxiliary information into an estimator into an estimator in a design consistent manner is the model-calibrated approach first proposed by Deville and Särndal (1992). The basic idea is to choose weights that satisfy certain calibration equations and are closest to the normal Horvitz-Thompson design weights according to some distance measure. Theberge (1999) applies the calibration technique to estimate population parameters other than the means. More recently, Wu and Sitter (2001) extends the calibration approach to deal with nonlinear as well as generalized linear models by using the fitted values under these working models to set up the calibration equations. The model-calibration approach can be classified as "model-assisted" because while the efficiency of the model-calibrated estimator depends on the validity of the model, consistency does not.

There is certainly a growing trend in the survey literature in using nonlinear and nonparametric regression. Instead of model (1), one considers,

$$y_t = g(\mathbf{x}_t) + \epsilon_t,$$

where the regression function  $g(\cdot)$  can be any arbitrary smooth function. Dorfman (1992) estimates  $g$  using the Nadaraya-Watson kernel estimator  $\hat{g}$  to result in the

<sup>1</sup> Faming Liang, Department of Statistics, Texas A&M University, College Station, TX77843-3143. E-mail: fliang@stat.tamu.edu; Anthony Yung Cheung Kuk, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543. E-mail: stakuka@nus.edu.sg.

following model-based estimator or predictor of the finite population mean,

$$\hat{y}_K = N^{-1} \left\{ \sum_{i=1}^n y_i + \sum_{i=n+1}^N \hat{g}(x_i) \right\},$$

where it is assumed without loss of generality that the sample consists of the first  $n$  elements of the population. Kuk (1993) makes use of kernel method to estimate the conditional distribution of  $y$  given  $x$  as a way of incorporating auxiliary information in the estimation of the finite population distribution of  $y$ . For the case of scalar  $x$ , Breidt and Opsomer (2000) estimates  $g$  using local polynomial regression with design weights incorporated to account for the sampling design used and propose a generalized difference estimator,

$$\hat{y}_{LP} = N^{-1} \left\{ \sum_{i=1}^n \frac{y_i - \hat{g}(x_i)}{\pi_i} + \sum_{i=1}^N \hat{g}(x_i) \right\} = N^{-1} \left\{ \sum_{i=1}^n w_i y_i \right\},$$

where  $\pi_i$  is the sample inclusion probability. It can be shown that the weights  $w_i$  are calibrated to match the totals of  $x$  up to the  $q^{\text{th}}$  order, where  $q$  is the order of the local polynomial. As a consequence,  $\hat{y}_{LP}$  is exactly model-unbiased if the true regression function is a polynomial of degree  $q$  or less. Breidt and Opsomer (2000) also show that  $\hat{y}_{LP}$  is asymptotically design-unbiased and consistent under mild conditions. For more discussions on nonlinear and nonparametric methods, see Valliant, Dorfman and Royall (2000) (chapter 11).

In this paper, another nonlinear regression method, Bayesian neural network (BNN), is applied to the problem. BNN has an important advantage of being able to handle multivariate auxiliary variables and model selection with ease, which is not the case for many other nonlinear and nonparametric techniques. BNNs were first introduced by Buntine and Weigend (1991) and MacKay (1992), and were further developed by Neal (1996), Müller and Insua (1998), Marrs (1998), Holmes and Mallick (1998), and Liang and Wong (2001). But the BNN proposed in this paper is different from those cited above in one important respect: A prior is put on each network connection, instead of only on the number of hidden units as done in the literature. This allows us to treat the selection of network structure and the selection of input variables (auxiliary variables) uniformly. The network is trained by sampling from the joint posterior of the network structure and connection weights. The sampled network has often a sparse structure, which effectively prevents the data from being overfitted. A heavy tail distribution, such as the student  $t$ -distribution, is proposed to model the disturbances of the data with outliers. Numerical results show that BNN models have offered a significant improvement over the linear regression based models in finite population estimation.

The remaining part of this article is organized as follows. In section 2, we describe the BNN models and the associated estimators for finite populations. In section 3, we present our numerical results for one finite population example with two choices of auxiliary variables and comparisons with various linear regression based models. In section 4, we present our numerical results for another finite population example demonstrate how a cross-validation procedure can be applied to determine the parameter setting for BNN models. In section 5, we conclude the paper with a brief discussion.

## 2. FINITE POPULATION ESTIMATION WITH BAYESIAN NEURAL NETWORKS

### 2.1 Bayesian Neural Network Models

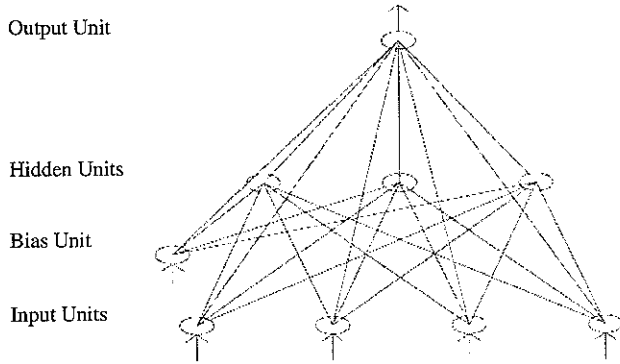
Suppose we have data pairs  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ , which were generated from the relationship

$$y_i = g(x_i) + \epsilon_i, \quad (2)$$

where  $y_i \in R^1$ ,  $x_i = (x_{i1}, \dots, x_{ip}) \in R^p$ ,  $g(\cdot)$  is the true regression function of unknown form, and  $\epsilon_i / \sigma \sim t(\nu)$  with  $\nu > 2$  being a known degree of freedom of the  $t$ -distribution. Here  $g(\cdot)$  may be highly nonlinear, and  $\sigma$  is an unknown scale parameter. We use the student  $t$ -distribution, instead of the Gaussian distribution as usual, to model the disturbances in order to accommodate extreme observations that are often present in the data from social sample surveys.

Before describing our BNN model, we first give a brief description for feed-forward neural networks. Figure 1 illustrates a one-hidden layer feed-forward neural network. It consists of four types of units, bias units, input units, hidden units, and output units. The unit to which the input features are presented is referred to as the input unit. The bias unit is a special type of input units with a constant input, say, 1. The unit where the network output is formed is referred to as the output unit. The hidden unit is so called because its input and output are only used for internal connections and are unavailable to the outside world. In a feed-forward neural network, each hidden unit independently processes the values fed to it by the units in the preceding layer and then presents its output to the units in the next layer for further processing. It has been shown by several authors (Cybenko 1989; Funahashi 1989; Hornik, Stinchcombe and White 1989) that neural networks are universal approximators in that a one-hidden layer feed-forward neural network with linear output units can approximate any continuous functions arbitrarily well on compact sets by increasing the number of hidden units. To survey regression, this is an important advantage of neural network models over other regression models. In the survey regression literature, whether model-assisted or model-based,

there is usually considerable attention paid to the consequences of model misspecification. The neural network model avoids this consideration partially due to its specific property of universal approximation. In section 2.2.1, we show that as the sample size is large, the unknown regression function  $g(\cdot)$  in (2) can be well approximated by BNN models, regardless of the true function form of  $g(\cdot)$ . Essentially, BNN falls into the class of data-driven methods.



**Figure 1.** A fully connected one hidden layer feed-forward neural network with 4 input units, 3 hidden units and 1 output unit. The arrows indicate the direction of data feeding.

In our BNN model, the function  $g(\cdot)$  in model (2) is approximated by a function of the form

$$\hat{g}(x_t, \alpha, \beta, \gamma) = \alpha_0 I_{\alpha_0} + \sum_{i=1}^p x_{ti} \alpha_i I_{\alpha_i} + \sum_{j=1}^M \beta_j I_{\beta_j} \psi \left( \sum_{i=1}^p x_{ti} \gamma_{ji} I_{\gamma_{ji}} + \gamma_{j0} I_{\gamma_{j0}} \right), \quad (3)$$

where  $I_\zeta$  is an indicator function which indicates the effectiveness of the connection  $\zeta$ ;  $M$  denotes the maximum number of hidden units which is specified by users;  $\alpha_0$  denotes the bias term of the output unit,  $\alpha_1, \dots, \alpha_p$  denote the weights on the connections from the input units to the output unit;  $\beta_1, \dots, \beta_M$  denote the weights on the connections from hidden units to the output unit;  $\gamma_{j0}$  denotes the bias term of the  $j^{\text{th}}$  hidden unit,  $\gamma_{j1}, \dots, \gamma_{jp}$  denote the weights on the connections from the input units to the  $j^{\text{th}}$  hidden unit; and  $\psi(\cdot)$  denotes the activation function. Sigmoid and hyperbolic tangent functions are two popular choices for the activation function. We set  $\psi(z) = \tanh(z)$  for all examples of this paper.

Let  $\Lambda$  be the vector consisting of all indicators of model (3). Note that  $\Lambda$  specifies the structure of the corresponding network. Let  $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)$ ,  $\beta = (\beta_1, \dots, \beta_M)$ ,  $\gamma_j = (\gamma_{j0}, \dots, \gamma_{jp})$ ,  $\gamma = (\gamma_1, \dots, \gamma_M)$ , and  $\theta = (\alpha_\Lambda, \beta_\Lambda, \gamma_\Lambda, \sigma^2)$ , where  $\alpha_\Lambda, \beta_\Lambda$  and  $\gamma_\Lambda$  denote the non-zero subsets of  $\alpha, \beta$  and  $\gamma$ , respectively. Thus, the model (3) is completely

specified by the tuple  $(\theta, \Lambda)$ . For simplicity, in the following we will use  $\theta_\Lambda$  to denote a BNN model and use  $\hat{g}(x_t, \theta_\Lambda)$  to re-denote the function  $\hat{g}(x_t, \alpha, \beta, \gamma)$ . Also, we let  $\theta_\Lambda = (\theta, \Lambda)$ , and use  $\theta_\Lambda$  and  $(\theta, \Lambda)$  exchangeably. To conduct a Bayesian analysis for model (3), we have the following prior distributions:  $\alpha_i \sim N(0, \sigma_\alpha^2)$  for  $\alpha_i \in \alpha_\Lambda$ ;  $\beta_j \sim N(0, \sigma_\beta^2)$  for  $\beta_j \in \beta_\Lambda$ ;  $\gamma_{ji} \sim N(0, \sigma_\gamma^2)$  for  $\gamma_{ji} \in \gamma_\Lambda$ ; and  $f(\sigma^2) \sim 1/\sigma^2$ . The total number of effective connections in  $\Lambda$  is  $m = \sum_{i=0}^p I_{\alpha_i} + \sum_{j=1}^M I_{\beta_j} \delta(\sum_{i=0}^p I_{\gamma_{ji}}) + \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}}$ , where  $\delta(z) = 1$  if  $z > 0$  and 0 otherwise. The model  $\Lambda$  is subject to a prior probability that is proportional to the mass put on  $m$  by a truncated Poisson ( $\lambda$ ) with rate  $\lambda$ ,

$$P(\Lambda) = \begin{cases} \frac{1}{Z} \frac{\lambda^m}{m!}, & m = 3, 4, \dots, U \\ 0, & \text{otherwise} \end{cases}$$

where  $U = (M+1)(p+1) + M$  is the number of connections of the full model in which all  $I_\zeta = 1$ ; and  $Z = \sum_{\Lambda \in \Omega} \lambda^m / m!$ . Here we let  $\Omega$  denote the set of all possible models with  $3 \leq m \leq U$ . We set the minimum number of  $m$  to three based on our views: neural networks are usually used for complex problems, and three has been a small enough number as a limiting network size. In these prior distributions,  $\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2$  and  $\lambda$  are hyper-parameters to be specified by users (discussed below). Furthermore, we assume that these prior distributions are independent *a priori*. Thus, we have the following log-posterior (up to an additive constant),

$$\begin{aligned} \log \pi(\theta_\Lambda | D) = & \text{Constant} - \left( \frac{n}{2} + 1 \right) \log \sigma^2 - \frac{v+1}{2} \sum_{t=1}^n \log \left( 1 + \frac{(y_t - \hat{g}(x_t, \theta_\Lambda))^2}{v \sigma^2} \right) \\ & - \frac{1}{2} \sum_{i=0}^p I_{\alpha_i} \left( \log \sigma_\alpha^2 + \frac{\alpha_i^2}{\sigma_\alpha^2} \right) - \frac{1}{2} \sum_{j=1}^M I_{\beta_j} \delta \left( \sum_{i=0}^p I_{\gamma_{ji}} \right) \\ & \left( \log \sigma_\beta^2 + \frac{\beta_j^2}{\sigma_\beta^2} \right) \\ & - \frac{1}{2} \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}} \left( \log \sigma_\gamma^2 + \frac{\gamma_{ji}^2}{\sigma_\gamma^2} \right) - \frac{m}{2} \log(2\pi) \\ & + m \log \lambda - \log(m!). \end{aligned} \quad (4)$$

Our BNN model is different from other BNN models existing in the literature in two important respects. First, the input variables of our BNN model are selected automatically by sampling from the joint posterior of the network structure and weights. Second, the structure of our BNN model is usually sparse and its performance less depends on

the initial specification for the input patterns and the number of hidden units. The sparse is in the sense that only a small number of connections are active in the network. So our BNN model avoids the problem of overfitting in a more natural way.

For data preparation and hyperparameter setting, we have the following suggestions. To avoid some weights that are trained to be extremely large or small (in absolute value) to accommodate different scales of input and output variables, we suggest that all input and output variables be normalized before feeding to the networks. In all examples of this article, the data is normalized by  $(y_i - \bar{y})/S_y$ , where  $\bar{y}$  and  $S_y$  denote the mean and standard deviation of the training data, respectively. Based on the belief that a network with a large weight variation usually has a poor generalization performance, we suggest that  $\sigma_\alpha^2, \sigma_\beta^2$  and  $\sigma_\gamma^2$  are chosen for moderate values to penalize a large weight variation. For example, we set  $\sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\gamma^2 = 5$  for all examples of this article. The setting should also be fine for the other problems. The value of  $\lambda$  reflects our belief on the network size needed for the data under consideration. Here we follow the suggestion of Weigend, Huberman and Rumelhart (1990) to choose  $\lambda$  such that the number of connection weights is about one tenth of the size of the training sample. In one simulation, we assessed the influence of  $\lambda$  on BNN model size and predictionability. The numerical results suggest that the prediction ability of BNN models is rather robust to the variation of  $\lambda$ , although the BNN model size increases slowly as  $\lambda$  increases.

To sample from the posterior (4), a Monte Carlo algorithm, so called the reversible jump evolutionary Monte Carlo (RJEMC) algorithm, is developed. This algorithm extends the evolutionary Monte Carlo algorithm (Liang and Wong 2001) to sample from a variable dimensional space by incorporating some reversible jump moves proposed in Green (1995). For details of the algorithm, please refer to the support documents and software for the paper. They are available at <http://www.stat.tamu.edu/~fliang>.

## 2.2 Finite Population Estimation with Bayesian Neural Networks

### 2.2.1 Bayesian Model Averaging

In this subsection, we review some basic results of Bayesian model averaging and show one theorem for BNN models, which form the theoretical basis for the estimators described in section 2.2.2. Suppose that we are interested in estimating the quantity  $\rho(\theta_\Lambda)$ , which is a function of both  $\Lambda$  and  $\theta$ . The Bayesian estimator of  $\rho(\theta_\Lambda)$  can be written as

$$E_\pi \rho(\theta_\Lambda) = \sum_{k=0}^K P(\Lambda_k | D) \int \rho(\theta_k, \Lambda_k) \pi(\theta_k | \Lambda_k, D) d\theta_k, \quad (5)$$

where  $K$  denotes the total number of models under consideration,  $\theta_k$  denotes the parameters associated with model

$\Lambda_k$ , and  $\pi(\theta_k | \Lambda_k, D)$  denotes the posterior density of  $\theta_k$  conditional on model  $\Lambda_k$ . Madigan and Raftery (1994) argued for this estimator that Bayesian model averaging (averaging over all the models in this fashion) accounts for the model uncertainty, and provides better predictive ability, as measured by the logarithmic scoring rule, than using any single model  $\Lambda_k$ . See Hoeting, Madigan, Raftery and Volinsky (1999) for a tutorial on Bayesian model averaging.

Suppose that samples  $(\theta_1, \Lambda_1), \dots, (\theta_M, \Lambda_M)$  have been drawn from the posterior distribution  $\pi(\theta_\Lambda | D)$  by a MCMC algorithm, then  $\rho(\theta_\Lambda)$  can be estimated by

$$\hat{\rho}(\theta_\Lambda) = \frac{1}{M} \sum_{i=1}^M \rho(\theta_{\Lambda_i}), \quad (6)$$

where  $\theta_{\Lambda_i} = (\theta_i, \Lambda_i)$ . Applying the standard Markov chain theory (Tierney 1994; Roberts and Casella 1999), under regularity conditions we have the following results. If  $E_\pi |\rho(\theta_\Lambda)| < \infty$ , then

$$\frac{1}{M} \sum_{i=1}^M \rho(\theta_{\Lambda_i}) \rightarrow E_\pi \rho(\theta_\Lambda), \quad \text{a.s.}, \quad (7)$$

as  $M \rightarrow \infty$ . Furthermore, if  $E_\pi |\rho(\theta_\Lambda)|^{2+\delta} < \infty$  for some  $\delta > 0$ , then

$$M^{1/2} \left\{ \frac{1}{M} \sum_{i=1}^M \rho(\theta_{\Lambda_i}) - E_\pi \rho(\theta_\Lambda) \right\} \rightarrow N(0, \tau^2), \quad (8)$$

for some positive constant  $\tau^2$  as  $M \rightarrow \infty$ , and the convergence is in distribution.

Similar to (7) and (8), we have the following theorem for BNN models, of which proof is presented in Appendix.

**Theorem 2.1** *Let  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  denote a simple random sample drawn from a population which can be modeled by model (2). Let  $(\theta_1, \Lambda_1), \dots, (\theta_M, \Lambda_M)$  denote the sample drawn from the posterior distribution  $\pi(\theta_\Lambda | D)$ , given in (4), by a MCMC method. Then, for any  $x_0$  drawn from the same distribution with the observations  $D$ , we have*

(a) 
$$E_\pi |\hat{g}(x_0, \theta_\Lambda)|^{2+\delta} < \infty, \quad (9)$$

for some  $\delta > 0$ , as  $n \rightarrow \infty$ .

(b) 
$$\frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) \rightarrow g(x_0), \quad \text{a.s.}, \quad (10)$$

as  $n \rightarrow \infty$  and  $M \rightarrow \infty$ .

(c) 
$$M^{1/2} \left[ \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - g(x_0) \right] \rightarrow N(0, \tau_*^2), \quad (11)$$

for some positive constant  $\tau_*^2$  as  $n \rightarrow \infty$  and  $M \rightarrow \infty$ , and the convergence is in distribution.

To show some properties of moments of  $1/M \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i})$ , we need the following theorem (Billingsley 1986, page 348, Corollary),

**Theorem 2.2** *Let  $r$  be a positive integer. If  $X_M \rightarrow X$  in distribution and  $\sup_m E|X_m|^{r+\delta} < \infty$ , where  $\delta > 0$ , then  $E|X|^r < \infty$  and  $EX_m^r \rightarrow EX^r$ .*

Following from (9), (11) and Theorem 2.2, we know

$$ME \left[ \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - g(x_0) \right]^2 \rightarrow \tau_*^2, \quad (12)$$

as  $n \rightarrow \infty$  and  $M \rightarrow \infty$ . It implies that

$$E \left[ \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - g(x_0) \right]^2 = \frac{\tau_*^2}{M} + o\left(\frac{1}{M}\right) \quad (13)$$

holds as  $n$  and  $M$  are both large.

Note we have shown that (11) and (13) hold as the sample size  $n \rightarrow \infty$ . In the context of finite population, especially for a small finite population, a more precise expression for (11) and (13) would be

$$M^{1/2} \left[ \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - E(y_0|D, x_0) \right] \rightarrow N(0, \tau_*^2), \quad (14)$$

and

$$E \left[ \frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) - E(y_0|D, x_0) \right]^2 = \frac{\tau_*^2}{M} + o\left(\frac{1}{M}\right), \quad (15)$$

where  $E(y_0|D, x_0)$  denotes the prediction of  $y_0$  which is the survey variable corresponding to  $x_0$ . The equations (14) and (15) take into accounts the possible bias of the sample  $D$ . In the case that the population constitutes many exact copies of the sample  $D$ ,  $E(y_0|D, x_0) = g(x_0)$  holds, and equations (14) and (15) are reduced to (11) and (13), respectively.

### 2.2.2 BMA Estimators in Finite Populations

Consider a finite population of  $N$  distinguishable elements. Associated with the  $i^{\text{th}}$  elements are the survey variable  $y_i$  and the auxiliary variables  $x_i$ . The values  $x_1, \dots, x_N$  are known for the entire population, while  $y_i$  is known only if the  $i^{\text{th}}$  unit is selected in the sample. Suppose a simple random sample  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  has been drawn from the finite population, a BNN model has been built for the sample, and  $(\theta_1, \Lambda_1), \dots, (\theta_M, \Lambda_M)$  have been drawn from the posterior distribution of the BNN model, the BMA estimator for the mean of the finite population is

$$\bar{y}_{\text{BNN}} = f \bar{y} + \frac{1}{MN} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(x_t, \theta_{\Lambda_i}),$$

where  $\bar{y}$  is the sample mean of  $y_1, \dots, y_n$ , and  $f = n/N$  is the sample fraction. About this estimator, we have the

following comments. First,  $\bar{y}_{\text{BNN}}$  is a model-based estimator, so that all the inference is with respect to the model for the  $y_i$ 's, not the survey design. As long as the model holds, the BNN estimator will have the mean squared error properties described below for any ignorable sampling design. Second, this estimator is identical to that proposed in Dorfman (1992), except that the BNN is replaced by a kernel-based regression. Third, this estimator can be used to estimate the mean of a finite population as long as each of the unsampled elements has the same distribution as the sample  $D$ .

The accuracy of an estimate can be measured by its mean squared error  $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$ , where  $\bar{Y}$  denotes the true population mean. To estimate  $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$ , we first consider

$$\begin{aligned} & E \left[ (\bar{y}_{\text{BNN}} - \bar{Y})^2 \mid D, \mathbf{X}_{n+1}^N \right] \\ &= E \left[ \left\{ \frac{1}{MN} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(x_t, \theta_{\Lambda_i}) - \frac{1}{N} \sum_{t=n+1}^N (g(x_t) + \epsilon_t) \right\}^2 \mid D, \mathbf{X}_{n+1}^N \right] \\ &= \frac{(N-n)^2}{N^2} E \left[ \left\{ \frac{1}{M(N-n)} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(x_t, \theta_{\Lambda_i}) - \frac{1}{N-n} \sum_{t=n+1}^N g(x_t) \right\}^2 \mid D, \mathbf{X}_{n+1}^N \right] \\ &\quad + \frac{N-n}{N^2} \text{var}(\epsilon_t) \\ &= \frac{(N-n)^2}{N^2} E \left[ \left\{ \begin{aligned} & \frac{1}{M(N-n)} \sum_{i=1}^M \sum_{t=n+1}^N \hat{g}(x_t, \theta_{\Lambda_i}) \\ & - E(\bar{y}_u \mid D, \mathbf{X}_{n+1}^N) \\ & + E(\bar{y}_u \mid D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u) \end{aligned} \right\}^2 \mid D, \mathbf{X}_{n+1}^N \right] \\ &\quad + \frac{N-n}{N^2} \text{var}(\epsilon_t) \\ &\approx \frac{\tau_D^2}{M} + (1-f)^2 \left\{ E(\bar{y}_u \mid D, \mathbf{X}_{n+1}^N) - E(\bar{y}_u) \right\}^2 \\ &\quad + \frac{1-f}{N} \text{var}(\epsilon_t), \quad (16) \end{aligned}$$

where  $\mathbf{X}_{n+1}^N = (x_{n+1}, \dots, x_N)$  denotes the set of auxiliary vectors of the unsampled elements;  $\bar{y}_u$  denotes the averaged survey value of the unsampled elements, and

$$E(\bar{y}_u) = \frac{1}{N-n} \sum_{t=n+1}^N g(x_t).$$

The last approximation of (16) follows from (15), that is, as  $M$  is large,

$$E\left\{\frac{1}{MN}\sum_{i=1}^M\sum_{t=n+1}^M\hat{g}(x_t, \theta_{\Lambda_i}) - (1-f)E(\bar{y}_u|D, X_{n+1}^N)\right\}^2 \approx \frac{\tau_D^2}{M},$$

for some positive constant  $\tau_D^2$ . The term  $E(\bar{y}_u|D, X_{n+1}^N) - E(\bar{y}_u)$  is the prediction bias due to the randomness or sampling bias of  $D$ . Following from (16), we have

$$E(\bar{y}_{\text{BNN}} - \bar{Y})^2 \approx \frac{E\tau_D^2}{M} + (1-f)^2$$

$$E\left\{E(\bar{y}_u|D, X_{n+1}^N) - E(\bar{y}_u)\right\}^2 + \frac{1-f}{N}\text{var}(\epsilon_t). \quad (17)$$

The quantity  $\tau_D^2$  can be estimated by the batch means method (Roberts 1996) as follows. Run the Markov chain for  $M = rs$  iterations, where  $s$  is the batch size and is assumed sufficiently large such that

$$\bar{y}_{\text{BNN},k} = f\bar{y} + \frac{1}{sN}\sum_{i=(k-1)s+1}^{ks}\sum_{t=n+1}^N\hat{g}(x_t, \theta_{\Lambda_i}),$$

is approximately independently  $N(f\bar{y} + (1-f)E(\bar{y}_u|D, X_{n+1}^N), \tau_D^2/s)$ . Therefore  $\tau_D^2$  can be approximated by

$$\hat{\tau}_D^2 = \frac{s}{r-1}\sum_{k=1}^r(\bar{y}_{\text{BNN},k} - \bar{y}_{\text{BNN}})^2, \quad (18)$$

which can be substituted into (17) in lieu of  $E\tau_D^2$ . Under the assumption  $\epsilon_t/\sigma \sim t(\nu)$ , the BMA estimator  $\text{var}(\epsilon_t)$  is

$$\hat{\text{var}}(\epsilon_t) = \frac{\nu}{\nu-2}\frac{1}{M}\sum_{i=1}^M\hat{\sigma}_i^2. \quad (19)$$

Under the assumption that the population is made up of exact copies of the training data, we have  $E(\bar{y}_u|D, X_{n+1}^N) - E(\bar{y}_u) \approx \hat{y} - \bar{y}$ , where  $\hat{y}$  denotes the fitted sample mean, and

$$E(\hat{y} - \bar{y})^2 = E\left\{\frac{1}{n}\sum_{t=1}^n\hat{\epsilon}_t\right\}^2 = \frac{1}{N}\text{var}(\hat{\epsilon}_t), \quad (20)$$

where  $\hat{\epsilon}_t = \sum_{i=1}^M\hat{g}(x_t, \theta_{\Lambda_i})/M - y_t$  is the residual of the  $t^{\text{th}}$  element of  $D$ , and  $\hat{\epsilon}_t$ 's are assumed to be iid and  $E(\hat{\epsilon}_t) = 0$ . Under the true model, we have  $\text{var}(\hat{\epsilon}_t) \approx \text{var}(\epsilon_t)$ . Hence, we suggest  $E\{E(\bar{y}_u|D, X_{n+1}^N) - E(\bar{y}_u)\}^2$  be estimated by

$$\hat{\text{Bias}}^2 = \frac{1}{n}\hat{\text{var}}(\epsilon_t). \quad (21)$$

In summary,  $E(\bar{y}_{\text{BNN}} - \bar{Y})^2$  can be estimated by

$$\hat{E}(\bar{y}_{\text{BNN}} - \bar{Y})^2 = \frac{\hat{\tau}_D^2}{M} + (1-f)^2\hat{\text{Bias}}^2$$

$$+ \frac{1-f}{N}\hat{\text{var}}(\epsilon_t) = \frac{\hat{\tau}_D^2}{M} + \frac{1-f}{n}\hat{\text{var}}(\epsilon_t). \quad (22)$$

As  $M \rightarrow \infty$  we have

$$\hat{E}(\bar{y}_{\text{BNN}} - \bar{Y})^2 = \frac{1-f}{n}\hat{\text{var}}(\epsilon_t). \quad (23)$$

We note that this estimate is identical in form to that given by Cochran (1977) for the linear regression estimator.

### 3. FIRST SIMULATION STUDY

#### 3.1 The Data

Our simulation population comprises 426 records for heads of household surveyed using the sample (long) questionnaire during the 1988 Test Population Census of Limeira, in São Paulo state, Brasil. This test was carried out as a pilot survey during the preparation for the 1991 Brazilian Population Census. For a detailed description for the test census, see Silva and Skinner (1997). We followed Silva and Skinner (1997) to consider the total monthly income as the main survey variable ( $y$ ) together with 11 potential auxiliary variables, namely,

$x_1$	indicator of sex of head of household equal male;
$x_2$	indicator of age of head of household less than or equal to 35;
$x_3$	indicator of age of head of household greater than 35 and less than or equal to 55;
$x_4$	total number of rooms in household;
$x_5$	total number of bathrooms in household;
$x_6$	indicator of ownership of household;
$x_7$	indicator that household type is house;
$x_8$	indicator of ownership of at least one car in household;
$x_9$	indicator of ownership of color TV in household;
$x_{10}$	years of study of head of household;
$x_{11}$	proxy of total monthly income of head of household.

Figure 2, the scatter plots of  $y$  versus the 11 auxiliary variables, shows that a linear regression model is not appropriate for the data. Although  $y$  and  $x_{11}$  are strongly linearly correlated, the scatter plots of  $y$  versus some other auxiliary variables, say  $x_4, x_5$  and  $x_{10}$ , suggest that their relationships can not be well modeled by a linear regression. In addition, if the data is modeled by a linear regression, the outlier, the 53<sup>th</sup> element, may have a high influence on fitting and prediction of the model. More precisely, if the element is included in the training data, the fitted response curve will have a up-drift comparing to the true curve and as a result the finite population mean will be overestimated; if

the element is not included in the training data, prediction will proceed as though there were not outliers and as a result the finite population mean will be underestimated. The presence of the strong influence element also mounts a great challenge on BNN models and other data analysis strategies.

We followed Silva and Skinner (1997) to construct two alternative sets of auxiliary variables for simulations. The first set contains  $x_1, \dots, x_4$  and  $x_{11}$ , which includes the proxy variable  $x_{11}$  and has a reasonable explanatory power in predicting  $y$ . The second set contains  $x_1, \dots, x_{10}$ , which has a weaker explanatory power than the first one due to the exclusion of  $x_{11}$ . So these two sets illustrate the predictive performances of BNN models with strong and weak auxiliary variables, respectively. As in Silva and Skinner (1997), 1,000 sample replicates of size 100 from this simulation population are selected by simple random sampling without replacement. The following computation were performed on the 1,000 replicates.

For each replicate, say  $k$ , it was analyzed by BNN models and various linear regression based strategies (reviewed below). For any strategy, the population mean estimate and its estimated mean squared error for the replicate  $k$  are denoted by  $\bar{y}(k)$  and  $V(\bar{y}(k))$ , respectively.

The computational results were summarized by computing the mean (MEAN), bias (BIAS), mean square error (MSE) and average of mean squared error estimates (AVMSE) from the set of the 1,000 replicates, given respectively by

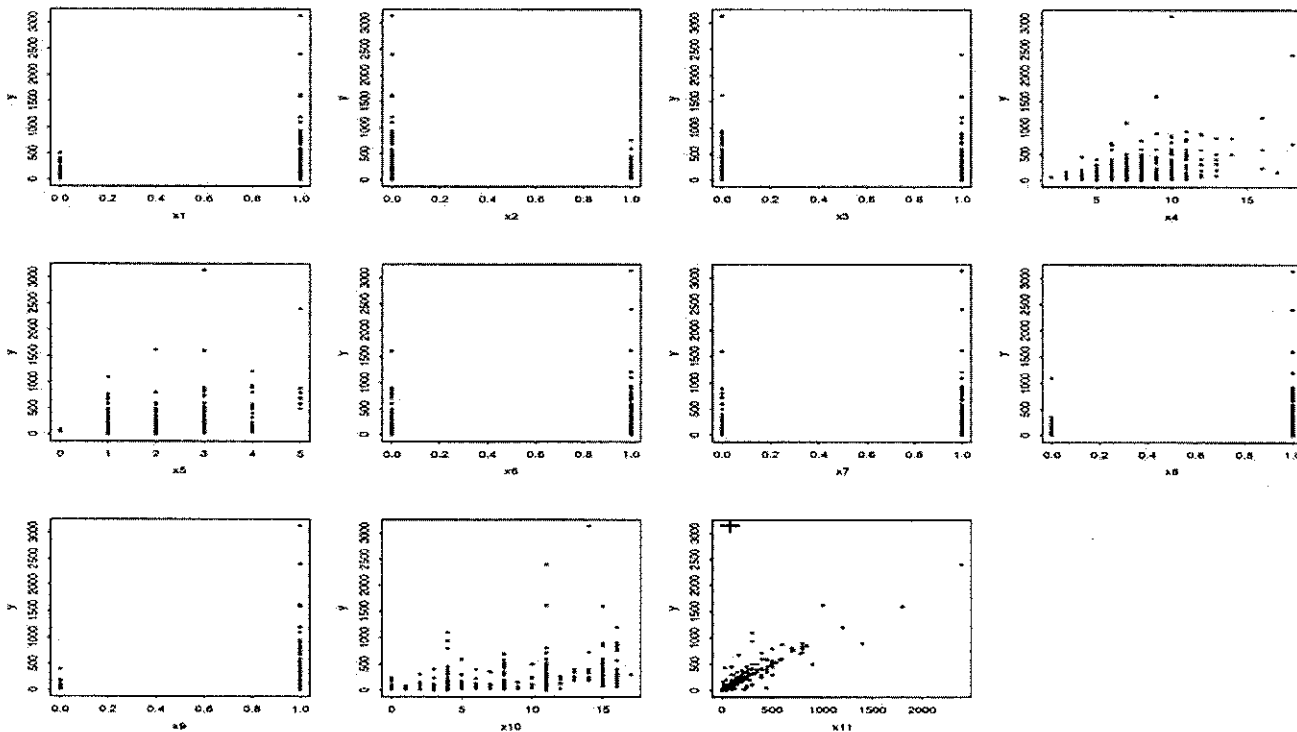
$$MEAN = \sum_{k=1}^S \bar{y}(k) / S;$$

$$BIAS = MEAN - \bar{Y};$$

$$MSE = \sum_{k=1}^S [\bar{y}(k) - \bar{Y}]^2 / S;$$

$$AVMSE = \sum_{k=1}^S V(\bar{y}(k)) / S,$$

where  $S$  is the total number of sample replicates under consideration, and  $\bar{Y} = 194.34$  for the simulation population. Empirical coverage rates for 95% confidence intervals based on asymptotic normal theory were also computed for each strategy and these rates, expressed as percentages, are presented in the last columns of Tables 1 and 3.



**Figure 2.** Scatter plots of the response variable  $y$  versus the auxiliary variables. In the plot of  $y$  versus  $x_{11}$  the “+” represents the 53<sup>rd</sup> element of the population.

### 3.2 Review of the Linear Regression Based Strategies

The linear regression based strategies that have been considered by Silva and Skinner (1997) are listed as follows.

---

SM)	Sample mean estimator, with no auxiliary variables ( $\bar{y}, V_s$ ).
Fs)	Forward selection of auxiliary variables with ( $\bar{y}_r, V_d$ ).
Fd)	Forward selection of auxiliary variables with ( $\bar{y}_r, V_d$ ).
Fg)	Forward selection of auxiliary variables with ( $\bar{y}_r, V_g$ ).
Bs)	Best subset selection from all subsets of auxiliary variables with ( $\bar{y}_r, V_s$ ).
Bd)	Best subset selection from all subsets of auxiliary variables with ( $\bar{y}_r, V_d$ ).
Bg)	Best subset selection from all subsets of auxiliary variables with ( $\bar{y}_r, V_g$ ).
FI)	Fixed subset of auxiliary variable with ( $\bar{y}_r, V_s$ ).
SS)	Saturated subset of auxiliary variable with ( $\bar{y}_r, V_s$ ).
FR)	Forward subset selection using SAS PROC REG, with ( $\bar{y}_r, V_s$ ).
CN)	Condition number reduction subset selection procedure with ( $\bar{y}, V_s$ ).
RI)	Ridge regression estimator proposed by Dunstan and Chambers (1986).

---

To facilitate the description for the above strategies, we define the following notations. Let  $U = \{1, \dots, N\}$  denote a finite population of  $N$  distinguishable elements,  $D \subset U$  denote a sample replicate of  $n$  elements drawn from  $U$  by simple random sampling without replacement,  $x_i = (x_{i1}, \dots, x_{ip})'$  be the vector of auxiliary variables associated with the  $i^{\text{th}}$  element, and  $\beta = (\beta_1, \dots, \beta_p)$ . Let  $\bar{X} = N^{-1} \sum_{i \in U} x_i$  be the vector of population means,  $\bar{x} = n^{-1} \sum_{i \in D} x_i$  be the vector of sample means,  $\bar{y} = n^{-1} \sum_{i \in D} y_i$  be the sample mean of the response variable,  $\hat{S}_x = n^{-1} \sum_{i \in D} (x_i - \bar{x})(x_i - \bar{x})'$ ,  $\hat{S}_{xy} = n^{-1} \sum_{i \in D} (x_i - \bar{x})(y_i - \bar{y})$ ,  $g_i = 1 + (\bar{X} - \bar{x})' \hat{S}_x^{-1} (x_i - \bar{x})$  the so-called  $g$ -weights (Särndal, Swensson and Wretman 1989), and  $\hat{\beta} = \hat{S}_x^{-1} \hat{S}_{xy}$  the least squares estimator of  $\beta$ . The regression estimator of  $\bar{Y}$  is

$$\bar{y}_r = \bar{y} + (\bar{X} - \bar{x})' \hat{\beta}.$$

The  $V_s, V_d$  is and  $V_g$  are three estimators of the mean squared error of  $\bar{y}_r$ . The  $V_s$  is given by Cochran (1977, page 195),

$$V_s = \frac{1-f}{n(n-p-1)} \sum_{i \in D} \hat{\epsilon}_i^2,$$

where  $\hat{\epsilon}_i = (y_i - \bar{y}) - (x_i - \bar{x})' \hat{\beta}$  and  $f = n/N$  is the sample fraction. The  $V_d$  is generalized (from  $p=1$  to  $p > 1$ ) from one estimator studied by Deng and Wu (1987) and it is expected to have a smaller bias than  $V_s$  (Silva 1996),

$$V_d = \frac{1-f}{n(n-1)} \sum_{i \in D} \alpha_i \hat{\epsilon}_i^2,$$

where

$$\alpha_i = (g_i^2 - 2g_i f + f) / \left\{ (1-f) \left[ 1 - (x_i - \bar{x})' \hat{S}_x^{-1} (x_i - \bar{x}) / (n-1) \right] \right\}.$$

The  $V_g$  is modified from one estimator given by Särndal *et al.* (1989), and it has a similar performance to  $V_d$ ,

$$V_g = \frac{1-f}{n(n-p-1)} \sum_{i \in D} g_i^2 \hat{\epsilon}_i^2.$$

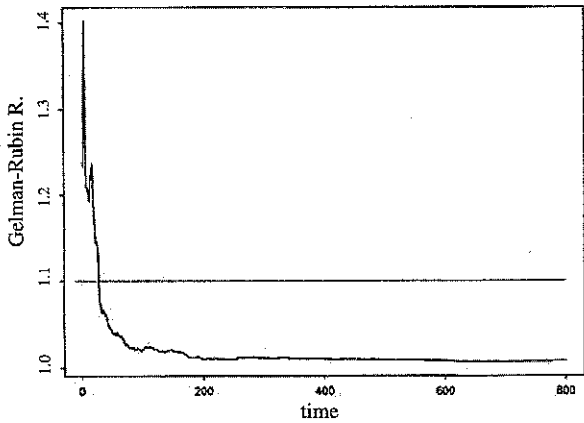
The best subset selection strategy (Bs, Bd and Bg) is to choose one subset which has the smallest mean squared error estimate among all  $2^p$  possible subsets. The forward selection strategy (Fs, Fd and Fg) starts with the sample mean as an estimator, then adds the variable which minimizes the mean squared error estimate, and the procedure is repeated until the mean squared error estimate starts to increase. Refer to Silva and Skinner (1997) for details of the implementations of the strategies CN and RI.

### 3.3 Illustration on One Sample Replicate

To understand the behavior of  $\bar{y}_{\text{BNN}}$  in presence of outliers and the role played by  $\nu$  in robust inference, we focus on one particular sample. The training data comprises the first 100 elements of the population, and the auxiliary variables include  $x_1, \dots, x_4$  and  $x_{11}$  as the first explanatory set. Note that the 53<sup>th</sup> element has been included in the training data.

For BNN models, we set  $\lambda = 5$  and  $M = 8$  which produces 62 connections for the full BNN model, and tried  $\nu = 25, 50, 100, 200$  and  $+\infty$ , where  $\nu = +\infty$  is equivalent to the assumption  $\epsilon_i \sim N(0, \sigma^2)$ . For each setting, RJEMC was run as follows: the network connections were first set to some random numbers drawn from  $N(0, 0.01)$ , and then were updated for 1,000 iterations in the parameter space of the full model, *i.e.*, all indicator variables are set to 1 in those iterations. After the initialization process, 4,000 iterations of RJEMC were run, and 800 samples were collected from these iterations at the lowest temperature level with an equal time space. The convergence of RJEMC can be diagnosed using the Gelman-Rubin statistic  $\hat{R}$  (Gelman and Rubin 1992) based on multiple independent runs. Figure 3 shows  $\hat{R}$  values computed from 10 independent runs. For each sample replicate of the simulation population, RJEMC converges ( $\hat{R} < 1.1$ ) very fast, usually within the first 500 iterations (100 BNN samples). We discarded the first 200 samples for the burn-in process, and used the remaining 600 samples for the further inference.

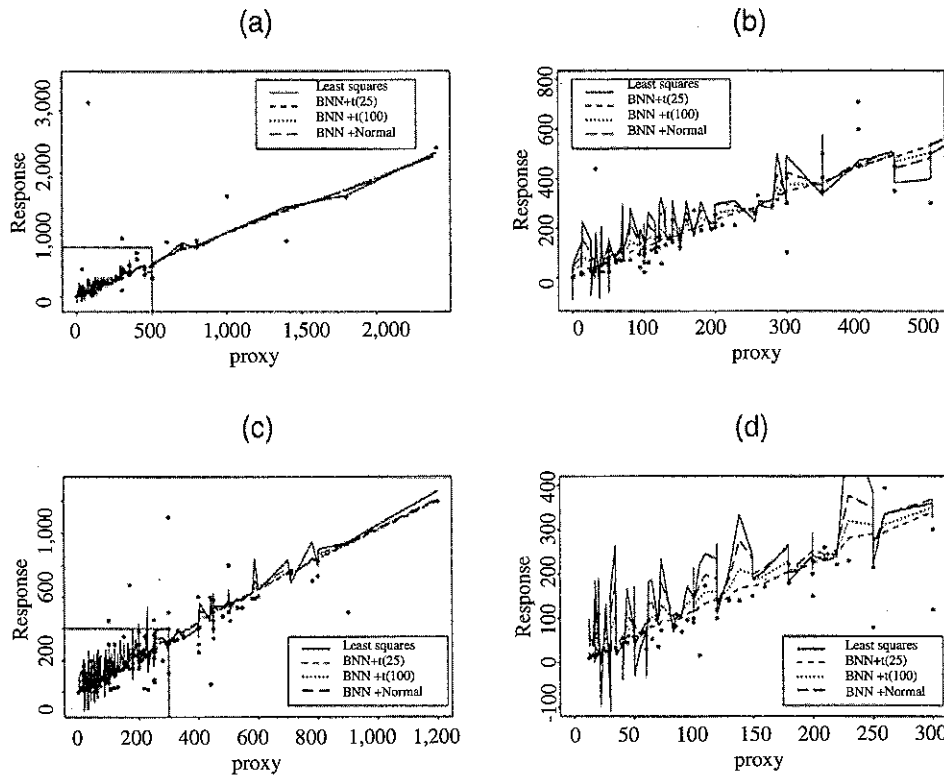
For comparison, the linear regression model (1) was also applied to this sample replicate.



**Figure 3.** Gelman-Rubin statistic  $\hat{R}$ . The curve was computed based on 10 independent runs of RJEMC. The random errors are assumed to be distributed according to  $t(100)$ .

Figure 4 shows the original data together with the fitted and predicted values produced by various models. The BNN

results were all obtained in one run of RJEMC. It can be seen that the linear regression model is not appropriate for this population as some fitted and predicted values produced by the model are negative for this sample replicate. Also, the fitted response curve (the solid curve in Figure 4(a) and 4(b)) is strongly influenced by the 53<sup>th</sup> element and lies above almost two-thirds of the data points. A similar phenomenon occurs for the prediction of unsampled values, see Figure 4(c) and 4(d). As a result, the population mean is overestimated (Figure 5). Comparing to that of the linear regression model, the results of the BNN models are less affected by the 53<sup>th</sup> element, especially for those computed with small values of  $\nu$ . Figure 5 shows that as  $\nu$  decreases, the estimated population mean by BNN models gets closer and closer to the true value, and the estimated 95% confidence interval of the population mean becomes narrower and narrower. It indicates that the influence of the 53<sup>th</sup> element on these estimates becomes weaker and weaker as  $\nu$  decreases. This is not surprising as the use of a heavily tailed error distribution is known to make the inference more robust.



**Figure 4.** Fitted and predicted response curves by various models. The curves are plotted against the proxy variable, and the true response values are shown by points. (a) The fitted response curves for the sampled elements. (b) The amplification of the square region of (a). (c) The predicted response curves for the unsampled elements. (d) The amplification of the square region of (c), and for clearness only every fourth elements are plotted in the order of sorted proxy values.

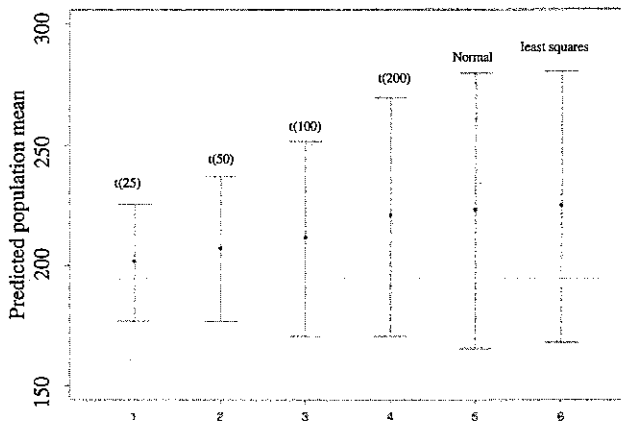


Figure 5. Estimated population mean and the associated 95% confidence interval by various models. The dotted line shows the true population mean which is 194.34.

### 3.4 Numerical Results on More Sample Replicates

BNN models were applied to analyze the 1,000 sample replicates. For each sample replicate of the first explanatory set, we set  $\nu = 100$ ,  $\lambda = 5$  and  $M = 8$  which produces 62 connections for the full BNN model. RJEMC was run as described in section 3.3. In each run 600 BNN samples were obtained for the inference. The computational results were summarized in Table 1. It shows that BNN models have made a significantly improvement over the linear regression based models in population mean estimation for the first explanatory set. Although the BNN estimate is slightly biased (The relative bias is about 2.5% in terms of absolute values and is still acceptable.), it has the smallest MSE value among all estimates in Table 1 and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows). As discussed in the last subsection, we expect  $\bar{y}_{\text{BNN}}$  to behave differently for samples containing and not containing the outlying element 53. When averaged over only those samples that contain element 53,  $\bar{y}_{\text{BNN}}$  with  $\nu = 50$  performs very well with bias 1.51 and 99.6% coverage. The result is obviously not as good as for those samples not containing element 53 due to the inevitable underestimation of the finite population mean. Frankly, there is not much one can do if there are outliers in the population but none in the sample. No statistical method based on sample information alone will be able to predict the occurrence of outliers in the non-sample. We believe that  $\bar{y}_{\text{BNN}}$  will perform very well for populations without outliers due to the universal approximation property of neural networks and the technique of Bayesian model averaging.

Let  $\bar{x}_{11}$  denote the average of proxy values of the elements in one sample replicate. To see how the performance of the BNN models varied with  $\bar{x}_{11}$ , we ordered the 1,000 sample replicates according to their values of  $\bar{x}_{11}$  and

divided them into 20 groups of 50 replicates, the first group containing the 50 replicates whose  $\bar{x}_{11}$  are smallest, and so forth. For each group, we calculated MEAN, MSE and AVMSE. Figure 6 shows these conditional values. From Figure 6(a) it is easy to see that BNN models possess one good property, namely, the population mean estimate is not sensitive to the value of  $\bar{x}_{11}$ . From Figure 6(b) it is easy to see that AVMSE provides an essentially unbiased estimate for MSE regardless of averaged proxy values.

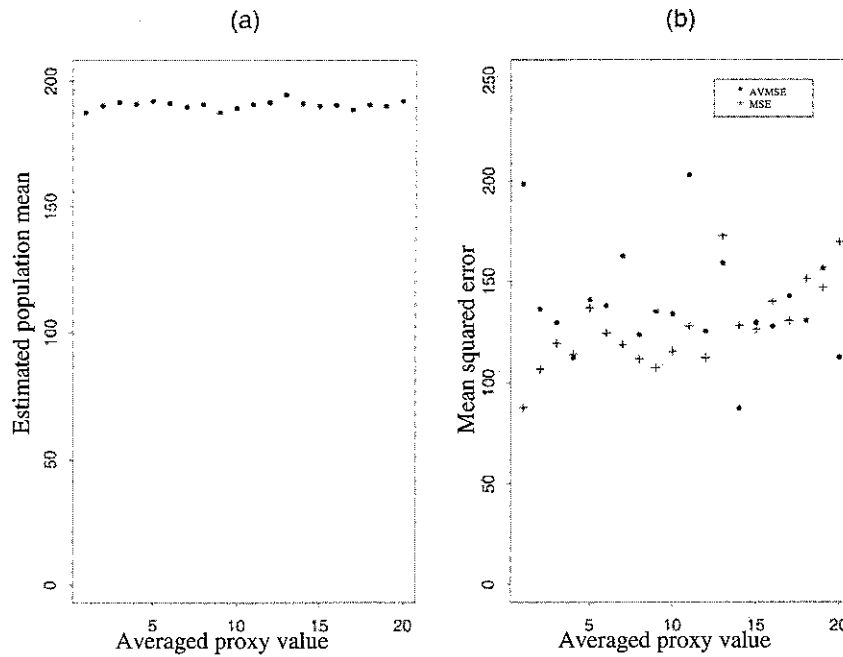
To assess the influence of  $\nu$ ,  $M$  and  $\lambda$  on BNN model size and prediction ability for the first explanatory set, we conducted three groups of experiments. In the first group of experiments, we fixed  $M = 8$  and  $\lambda = 5$ , and varied the value of  $\nu$ ,  $\nu = 50, 100$  and  $150$ . In the second group of experiments, we fixed  $\nu = 100$  and  $\lambda = 5$ , and varied the value of  $M$ ,  $M = 6, 8$  and  $10$ . In the third group of experiments, we fixed  $\nu = 100$  and  $M = 8$ , and varied the value of  $\lambda$ ,  $\lambda = 4, 5$  and  $6$ . For each setting, RJEMC was run as described in section 3.3 for the 1,000 sample replicates. The computational results were summarized in Table 2. It shows that the averaged model size produced by each setting is about the same, although it increases slowly as  $M$  and  $\lambda$  increase. The results of the first group of experiments show clearly that for BNN models there is a trade-off between BIAS and MSE or AVMSE by choosing the value of  $\nu$ . The results of the second and third group of experiments show that BIAS, MSE, AVMSE and the coverage probability are rather stable to the variation of  $M$  and  $\lambda$ , although the latter three statistics have a slow tendency to increase as  $M$  and  $\lambda$  increase. The increasing trend of these statistics is due to the fact that the neural networks tend to be overfitted as  $M$  and  $\lambda$  increase.

Table 1

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using  $x_1, \dots, x_4$  and  $x_{11}$  as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage <sup>a</sup> (%)
SM) Sample mean ( $\bar{y}, V_s$ )	0.25	620.09	619.05	91.8
CN) Cond. num. red. ( $\bar{y}, V_s$ )	0.34	507.33	483.63	89.8
RI) Ridge	2.12	304.95	257.07	82.5
Fs) Forward ( $\bar{y}_r, V_s$ )	0.40	233.78	239.62	82.7
Fd) Forward ( $\bar{y}_r, V_d$ )	-1.25	<b>188.08</b>	<b>196.88</b>	<b>82.0</b>
Fg) Forward ( $\bar{y}_r, V_g$ )	-1.28	<b>188.38</b>	<b>192.73</b>	<b>81.1</b>
Bs) Best ( $\bar{y}_r, V_s$ )	0.44	236.90	239.49	82.7
Bd) Best ( $\bar{y}_r, V_d$ )	-1.22	<b>190.52</b>	<b>196.84</b>	<b>82.0</b>
Bg) Best ( $\bar{y}_r, V_g$ )	-1.24	<b>190.83</b>	<b>192.71</b>	<b>81.1</b>
FI) Fixed ( $\bar{y}_r, V_s$ )	0.29	227.90	241.24	83.3
SS) Saturated ( $\bar{y}_r, V_s$ )	0.30	233.58	242.32	82.5
FR) Proc REG ( $\bar{y}_r, V_s$ )	0.38	235.86	240.26	82.5
BNN) $t(100)$	-4.91	<b>138.11</b>	<b>127.14</b>	<b>84.8</b>

<sup>a</sup> Nominal 95% coverage.



**Figure 6.** MEAN (panel (a)), MSE and AVMSE (Panel (b)) conditional on the averaged proxy values. The 1,000 sample replicates are ordered on  $\bar{x}_{11}$  and divided into 20 groups of 50 samples.

**Table 2**

Assessment of the influence of  $\nu$ ,  $M$  and  $\lambda$  on BNN model size and prediction ability for the first explanatory set. For convenience of comparison, the results of the setting  $\nu = 100$ ,  $M = 8$  and  $\lambda = 5$  were repeated in panels B and C.

Experiment	$\nu$	$M$	$\lambda$	Size <sup>a</sup>	BIAS	MSE	AVMSE	Coverage <sup>b</sup> (%)
A	50	8	5	10.53	-6.78	131.78	90.08	82.0
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	150	8	5	10.79	-3.81	156.55	160.28	85.5
B	100	6	5	9.52	-4.90	136.72	122.58	84.1
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	100	10	5	11.83	-5.14	140.13	132.20	86.4
C	100	8	4	9.42	-4.94	138.04	125.99	85.2
	100	8	5	10.70	-4.91	138.11	127.14	84.8
	100	8	6	11.83	-4.92	139.62	128.64	85.7

<sup>a</sup> Size =  $\sum_{k=1}^{1,000} \sum_{i=1}^M m(\Lambda_i) / M / 1,000$ , where  $m(\Lambda_i)$  is the number of connections of the neural network  $\Lambda_i$ .

<sup>b</sup> Nominal 95% coverage.

The above experiments also address the issue of model misspecification. Note the BNN model proposed in this paper is specified by the three parameters,  $\nu$ ,  $M$  and  $\lambda$ . Table 2 shows that the BNN model can still perform well even when the parameter setting has some departures from the optimal setting. In practice, the setting of  $\nu$ ,  $M$  and  $\lambda$  can be determined by a cross-validation experiment. This will be demonstrated in the second simulation study.

Finally, we consider the weaker set of auxiliary variables  $x_1, \dots, x_{10}$ . For each sample replicate, we set  $\nu = 100$ ,  $\lambda = 5$  and  $M = 8$  which produces 107 connections for the full BNN model. RJEMC was run as in section 3.3. The

computational results were summarized in Table 3. It shows clearly that BNN models continue to provide a significant improvement over the linear regression based models in population mean estimation when the strongest predictor  $x_{11}$  is excluded. The BNN estimate has the smallest MSE value among all estimates in Table 3, and has the smallest bias and the highest nominal coverage probability among the estimates with smaller MSE values (the boldfaced rows).

To assess the influence of  $\nu$ ,  $M$  and  $\lambda$  on BNN model sizes and prediction abilities for the second explanatory set, we conducted the same three groups of experiments as for

the first explanatory set. The computational results were summarized in Table 4. Panel A shows again the trade-off between BIAS and MSE or AVMSE made for BNN models by the value of  $\nu$ . Panels B and C show that BIAS, MSE, AVMSE and the coverage probability have an even more stable performance across different choices of  $M$  and  $\lambda$  than that of the first explanatory set.

**Table 3**

Bias, mean squared error, average of mean squared error estimates and empirical coverage of various estimation strategies for the population mean using  $x_1, \dots, x_{10}$  as auxiliary variables. Figures other than BNN are reproduced from Silva and Skinner (1997).

Estimation strategy	BIAS	MSE	AVMSE	Coverage <sup>a</sup> (%)
SM) Sample mean ( $\bar{y}, V_s$ )	0.25	620.09	619.05	91.8
CN) Cond. num. red. ( $\bar{y}, V_s$ )	3.49	562.91	450.36	87.3
RI) Ridge	1.05	480.18	472.82	89.4
Fs) Forward ( $\bar{y}_r, V_s$ )	0.06	468.46	397.99	86.7
Fd) Forward ( $\bar{y}_r, V_d$ )	-8.12	434.27	338.90	81.7
Fg) Forward ( $\bar{y}_r, V_g$ )	-7.90	433.71	328.46	81.6
Bs) Best ( $\bar{y}_r, V_s$ )	-0.00	466.16	397.59	86.6
Bd) Best ( $\bar{y}_r, V_d$ )	-7.90	434.54	336.88	81.5
Bg) Best ( $\bar{y}_r, V_g$ )	-7.60	433.26	326.05	81.6
FI) Fixed ( $\bar{y}_r, V_s$ )	0.45	490.49	461.86	89.0
SS) Saturated ( $\bar{y}_r, V_s$ )	-0.20	462.71	413.17	86.9
FR) Proc REG ( $\bar{y}_r, V_s$ )	-0.07	466.13	399.34	86.4
BNN) $\lambda(100)$	-5.78	395.25	323.12	86.5

<sup>a</sup> Nominal 95% coverage.

**4. SECOND SIMULATION STUDY**

In the first simulation study, we show that the BNN model works well for the data sets with outliers. In this simulation study, we show that the BNN model works even better for the data sets without outliers. In this study, we also demonstrate how a cross-validation procedure can be applied to determine a setting for the parameters  $\nu$ ,  $M$  and  $\lambda$  of the BNN model.

The simulation population comprises the records of the serious crimes of 141 large standard Metropolitan Statistical Areas (SMSAs) in the United States. A SMSA includes a city (or cities) of specified population size. The data generally pertains to the years 1976 and 1977, and is available in Neter, Kutner, Nachtsheim and Wasserman (1996). We consider the total number of serious crimes in 1977 as the survey variable ( $y$ ) and the following 9 variables as potential auxiliary variables.

$x_1$	Land area (in square miles);
$x_2$	Estimated 1977 total population (in thousands);
$x_3$	Percent of 1976 SMSA population in central city or cities;
$x_4$	Percent of 1976 SMSA population 65 years old or older;
$x_5$	Number of professionally active nonfederal physicians as of December 31, 1977;
$x_6$	Total number of beds, cribs, and bassinets during 1977;
$x_7$	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school, according to the 1970 Census of the Population;
$x_8$	Total number of persons in civilian labor force (persons 16 years old or older classified as employed or unemployed) in 1977 (in thousands);
$x_9$	Total current income received in 1976 by residents of the SMSA from all sources (in millions of dollars).

**Table 4**

Assessment of the influence of  $\nu$ ,  $M$  and  $\lambda$  on BNN model size and prediction ability for the second explanatory set. For convenience of comparison, the results of the setting  $\nu = 100$ ,  $M = 8$  and  $\lambda = 5$  were repeated in panels B and C of the table.

Experiment	$\nu$	$M$	$\lambda$	Size <sup>a</sup>	BIAS	MSE	AVMSE	Coverage <sup>b</sup> (%)
A	50	8	5	14.87	-9.30	394.11	270.09	82.5
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	150	8	5	15.17	-4.38	412.56	346.75	87.1
B	100	6	5	13.90	-5.77	394.79	319.13	86.0
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	100	10	5	16.05	-5.91	396.27	327.86	87.1
C	100	8	4	13.23	-5.62	397.65	323.68	86.4
	100	8	5	15.06	-5.78	395.25	323.12	86.5
	100	8	6	16.76	-5.78	396.45	321.98	86.6

<sup>a</sup> Size =  $\sum_{k=1}^{1,000} \sum_{i=1}^M m(\Lambda_i) / M / 1,000$ , where  $m(\Lambda_i)$  is the number of connections of the neural network  $\Lambda_i$ .

<sup>b</sup> Nominal 95% coverage.

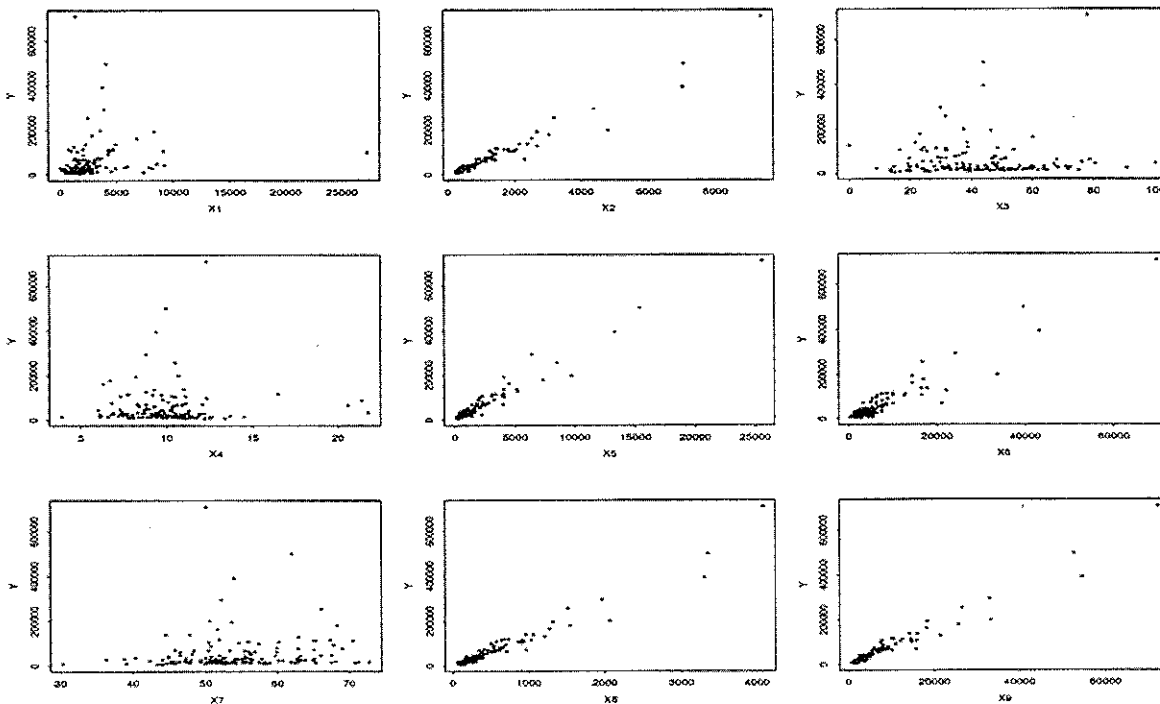


Figure 7: Scatter plots of the response variable  $y$  versus the auxiliary variables for the second simulation study.

**Table 5**  
Cross-validation experiments for the SMSA example. For convenience of comparison, the results of the setting  $\nu = 100$ ,  $M = 3$  and  $\lambda = 5$  were repeated in panels B and C.

Experiment	$\nu$	$M$	$\lambda$	Size	BIAS ( $\times 10^3$ )	MSE ( $\times 10^6$ )	AVMSE ( $\times 10^6$ )	Coverage <sup>a</sup> (%)
A	50	3	5	10.68	-0.472	4.78	4.19	91
	100	3	5	10.74	-0.527	5.04	4.24	92
	$\infty$	3	5	10.74	-0.543	4.76	4.21	92
B	100	1	5	7.29	-0.466	4.63	3.66	89
	100	2	5	9.42	-0.500	4.61	3.91	90
	100	3	5	10.74	-0.527	5.04	4.24	92
	100	4	5	11.66	-0.480	4.74	4.47	91
C	100	3	4	9.56	-0.434	4.68	4.12	92
	100	3	5	10.74	-0.527	5.04	4.24	92
	100	3	6	11.82	-0.455	4.66	4.28	93

<sup>a</sup> Nominal 95% coverage.

Figure 7, the scatter plot of  $y$  versus the 9 auxiliary variables, suggests that a linear regression model may not be appropriate for the data set. There is a strong nonlinear relationship between  $y$  and  $x_1, x_3, x_4$  and  $x_7$ . Also, the explanatory variables  $x_2, x_5, x_6, x_8$  and  $x_9$  are highly correlated. First, we demonstrate how a cross-validation procedure can be applied to determine the setting for the parameters  $\nu$ ,  $M$  and  $\lambda$  of the BNN model. We treated the first 70 records as a small finite population, generated 100 sample replicates of size 50 from these 70 records by the method of simple random sampling without replacement, and then conducted the following experiments. In the first group of experiments, we fixed  $M = 3$  and  $\lambda = 5$ , and varied the value of  $\nu$ ,  $\nu = 50, 100$  and  $\infty$ , where  $\nu = \infty$  is just an

indicator which indicates the normality assumption for the disturbance. Note  $M = 3$  results in a full model of 43 connections, which has been large enough for the data set. In the second group of experiments, we fixed  $\nu = 100$  and  $\lambda = 5$ , and varied the value of  $M$ ,  $M = 1, 2, 3, 4$ . In the third group of experiments, we fixed  $\nu = 100$  and  $M = 3$ , and varied the value of  $\lambda$ ,  $\lambda = 4, 5, 6$ . For each setting, RJEMC was run as in the first simulation study. The computational results were summarized in Table 5. It shows that the performance of the BNN model is rather stable to the variation of the settings. It also suggests that the setting  $\nu = 100, M = 3$  and  $\lambda = 4$  probably be a good setting for this simulation population by a synthetic considerations on all values of BIAS, MSE, AVMSE and coverage probability.

In the further analysis, we generated 500 sample replicates of size 70 from all the 141 records by the method of simple random sampling without replacement. For each replicate, RJEMC was run as in the first simulation study. The computational results were summarized in Table 6. It shows that the BNN model also works well for this population. We also tried the other settings given in Table 5 for the 500 sample replicates. The computational results are all similar.

**Table 6**  
Computational results for the second simulation study with  $v = 100, M = 3$  and  $\lambda = 4$

Size	BIAS ( $\times 10^3$ )	MSE ( $\times 10^6$ )	AVMSE ( $\times 10^6$ )	Coverage <sup>a</sup> (%)
9.20	-0.512	3.36	3.25	92.6

<sup>a</sup> Nominal 95% coverage.

### 5. DISCUSSION

In this article, we studied the use of Bayesian neural networks in finite population estimation. The numerical results show that it has made a significant improvement over the linear regression based methods. The improvement is not from Bayesian model averaging, but mainly from BNN models. We also applied the linear regression based Bayesian model averaging method (Liang, Truong and Wong 2001) to the same problem, and the improvement over Silva and Skinner (1997) is only marginal. Although our implementation for BNN models is not specific to finite populations, we do not think this is a shortcoming of our method. The generality of our method suggests its wide applications, for example, in nonlinear regression and nonlinear time series (the program is available by an request from the first author). Of course, a further research on how to use the known auxiliary variable information for a finite population in BNN training is also of interest.

### APPENDIX

Before proving Theorem 2.1, we give one formula which will be used in the proof.

**Formula 5.1** (Laplace's method)

$$\int b(\theta) \exp\{-nh(\theta)\} d\theta = (2\pi/n)^{p/2} \left| \sum \right|^{1/2} \exp\{-nh(\hat{\theta})\} b(\hat{\theta}) \{1 + O(n^{-1})\}, \quad (24)$$

as  $n \rightarrow \infty$ , where  $b(\cdot)$  is a general function which does not depend on  $n$ ,  $h(\theta)$  is a constant-order function of  $n$  as  $n \rightarrow \infty$ ,  $p$  is the dimension of  $\theta$ ,  $\hat{\theta}$  is the maximizer of  $-h(\theta)$  and  $\Sigma = (D^2h(\hat{\theta}))^{-1}$  is the inverse of the negative Hessian matrix evaluated at  $\hat{\theta}$ .

For the general formulation of Laplace's method, see Kass and Vaidyanathan (1992).

### Proof of Theorem 2.1

**Proof: Part (a)** By definition of expectation,  $E_{\pi} |g(x_0, \theta_{\Lambda})|^{2+\delta}$  can be written as

$$E_{\pi} |g(x_0, \theta_{\Lambda})|^{2+\delta} = \sum_{k=0}^K P(\Lambda_k | D) \int |g(x_0, \theta_{\Lambda})|^{2+\delta} \pi(\theta_k | \Lambda_k, D) d\theta_k.$$

Following from the normality of the posterior distributions  $\pi(\theta_k | \Lambda_k, D)$  (Walker 1969) and the fact that the activation function  $\psi(\cdot)$  in (3) is bounded, we know (9) holds. Walker (1969) showed that the posterior distribution is Gaussian in the limit of infinite training data.

**Part (b).** For a given observation  $x_0, E_{\pi} \hat{g}(x_0, \theta_{\Lambda})$  can be written as

$$E_{\pi} = \hat{g}(x_0, \theta_{\Lambda}) = \frac{\sum_{\Lambda \in \Omega} P(\Lambda) \int \hat{g}(x_0, \theta_{\Lambda}) \exp\{-nh(\theta_{\Lambda})\} \tilde{\pi}(\theta_{\Lambda} | \Lambda) d\theta_{\Lambda}}{\sum_{\Lambda \in \Omega} P(\Lambda) \int \exp\{-nh(\theta_{\Lambda})\} \tilde{\pi}(\theta_{\Lambda} | \Lambda) d\theta_{\Lambda}} \quad (25)$$

where

$$\begin{aligned} \log \tilde{\pi}(\theta_{\Lambda} | \Lambda) = & -\log \sigma^2 - \frac{1}{2} \sum_{i=0}^p I_{\alpha_i} \left( \log \sigma_{\alpha}^2 + \frac{\alpha_i^2}{\sigma_{\alpha}^2} \right) \\ & - \frac{1}{2} \sum_{j=1}^M I_{\beta_j} \delta \left( \sum_{i=0}^p I_{\gamma_{ji}} \right) \left( \log \sigma_{\beta}^2 + \frac{\beta_j^2}{\sigma_{\beta}^2} \right) \\ & - \frac{1}{2} \sum_{j=1}^M \sum_{i=0}^p I_{\beta_j} I_{\gamma_{ji}} \left( \log \sigma_{\gamma}^2 + \frac{\gamma_{ji}^2}{\sigma_{\gamma}^2} \right) \\ & - \frac{m}{2} \log(2\pi) + m \log \lambda - \log(m!), \end{aligned} \quad (26)$$

and

$$\begin{aligned} h(\theta_{\Lambda}) = & \frac{1}{n} \left[ \frac{n}{2} \log \sigma^2 + \frac{v+1}{2} \sum_{i=1}^n \log \left( 1 + \frac{(y_i - \hat{f}(x_i))^2}{v\sigma^2} \right) \right] \\ \approx & \frac{1}{n} \left[ \frac{n}{2} \log \sigma^2 + \frac{v+1}{2} \sum_{i=1}^n \frac{(y_i - \hat{f}(x_i))^2}{v\sigma^2} \right] \\ \approx & \frac{1}{2} \log \sigma^2 + \frac{v+1}{2v\sigma^2} E(y_i - \hat{g}(x_i, \theta_{\Lambda}))^2 \\ = & \frac{1}{2} \log \sigma^2 + \frac{v+1}{2v\sigma^2} \left[ E(y_i - g(x_i))^2 + (g(x_i) - \hat{g}(x_i, \theta_{\Lambda}))^2 \right], \end{aligned} \quad (27)$$

where the first approximation follows from the Taylor expansion,  $\log(1+z) \approx z$ , when  $z$  lies in a neighbourhood of zero; and the second approximation follows from the weak law of large numbers by assuming that  $n$  is large. Note  $\nu$  is often set to a large number, say, a number greater than 30. In the first example of this paper, we set  $\nu = 100$ . The equation (27) implies that the minimum of  $h(\theta_\Lambda)$  is attained when  $g(x_t) = \hat{g}(x_t, \theta_\Lambda)$  holds, that is,  $\hat{g}(x_t, \theta_\Lambda) = g(x_t)$ , where  $\hat{\theta}_\Lambda = \arg \min_{\theta_\Lambda} h(\theta_\Lambda)$ .

By applying Laplace's method to the numerator of (25) with  $b(\cdot) = \hat{g}(x_0, \theta_\Lambda) \tilde{\pi}(\theta_\Lambda | D)$ , we have

$$\begin{aligned} & \sum_{\Lambda \in \Omega} P(\Lambda) \int \hat{g}(x_0, \theta_\Lambda) \exp\{-\tau H(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda \\ & \approx \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} \left| \sum_{\Lambda} \right|^{1/2} \\ & \quad \exp\{-nh(\hat{\theta}_\Lambda)\} \hat{g}(x_0, \hat{\theta}_\Lambda) \tilde{\pi}(\hat{\theta}_\Lambda | D) \\ & \approx g(x_0) \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} \left| \sum_{\Lambda} \right|^{1/2} \\ & \quad \exp\{-nh(\hat{\theta}_\Lambda)\} \tilde{\pi}(\hat{\theta}_\Lambda | D), \end{aligned} \tag{28}$$

where the first approximation follows from the Laplace formula (24), and the second approximation follows from the equality  $\hat{g}(x_t, \hat{\theta}_\Lambda) = g(x_t)$ . Here we assume that the number of hidden units of each  $\Lambda$  is sufficiently large such that  $g(\cdot)$  can be approximated arbitrarily well by the network with properly adjusted weights. Otherwise, that term will take a small value and is negligible in the last approximation of (28).

Similarly, by applying the Laplace's method to the denominator of (25) with  $b(\cdot) = \tilde{\pi}(\theta_\Lambda | D)$ , we have

$$\begin{aligned} & \sum_{\Lambda \in \Omega} P(\Lambda) \int \exp\{-nh(\theta_\Lambda)\} \tilde{\pi}(\theta_\Lambda | \Lambda) d\theta_\Lambda \\ & \approx \sum_{\Lambda \in \Omega} P(\Lambda) (2\pi/n)^{m/2} \left| \sum_{\Lambda} \right|^{1/2} \exp\{-nh(\hat{\theta}_\Lambda)\} \tilde{\pi}(\hat{\theta}_\Lambda | D). \end{aligned} \tag{29}$$

Following from (28), (29), and the approximation accuracy ( $O(n^{-1})$ ) of Laplace's method, we have

$$E_\pi \hat{g}(x_0, \theta_{\Lambda_i}) \rightarrow g(x_0), \tag{30}$$

as  $n \rightarrow \infty$ . Following from (7), (9) and (30), we have

$$\frac{1}{M} \sum_{i=1}^M \hat{g}(x_0, \theta_{\Lambda_i}) \rightarrow g(x_0), \quad a.s.,$$

as  $n \rightarrow \infty$  and  $M \rightarrow \infty$ .

**Part (c).** It follows from (8), (9), (30) and Slutsky's Theorem (Casella and Berger 2002). The proof is completed.

## ACKNOWLEDGEMENTS

The authors would like to thank Chris Skinner for providing the test census data set, and thank the anonymous referees, the associate editor and editor Dr. M.P. Singh for their constructive comments which have led to a significant improvement of this paper.

## REFERENCES

- BANKIER, M.D. (1990). Two step generalized least squares estimation. Ottawa: Statistics Canada, Social Survey Methods Division, Internal reports.
- BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.
- BILLINGSLEY, P. (1986). *Probability and Measure* (Second Edition). New York: John Wiley & Sons, Inc.
- BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- BUNTINE, W.L., and WEIGEND, A.S. (1991). Bayesian back-propagation. *Complex Systems*, 5, 603-643.
- CASELLA, G., and BERGER, R.L. (2002). *Statistical Inference* (Second Edition). United States: Thompson Learning.
- CHAMBERS, R.L., DORFMAN, A.H. and WEHRLY, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- COCHRAN, W.G. (1977). *Sampling techniques* (3<sup>rd</sup> Ed.). New York: John Wiley & Sons, Inc.
- CYBENKO, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303-314.
- DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.
- DEVILLE, J.-C., and SÄRNDALL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DORFMAN, A.H. (1992). Non-parametric regression for estimating totals in finite populations. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA. 622-625.
- DUNSTAN, R., and CHAMBERS, R.L. (1986). Model-based confidence intervals in multipurpose surveys. *Applied Statistics*, 35, 276-280.
- FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society B*, 60, 3-21.
- FUNAHASHI, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2, 183-192.
- GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.

- GREEN, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.
- HOETING, J.A., MADIGAN, D., RAFTERY, A.E. and VOLINSKY, C. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382-417.
- HOLMES, C.C., and MALLICK, B.K. (1998). Bayesian radial basis functions of variable dimension. *Neural Computation*, 10, 1217-1233.
- HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.
- KASS, R.E., and VAIDYANATHAN, S. (1992). Approximate Bayesian factor and orthogonal parameters, with applications to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B*, 54, 129-144.
- KUK, A.Y.C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika*, 80, 385-392.
- KUK, A.Y.C., and WELSH, A.H. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society B*, 63, 277-292.
- LIANG, F., TRUONG, Y.K. and WONG, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statistica Sinica*, 11, 1005-1029.
- LIANG, F., and WONG, W.H. (2001). Real parameter evolutionary Monte Carlo with applications in Bayesian mixture models. *Journal of the American Statistical Association*, 96, 653-666.
- MACKAY, D.J.C. (1992). A practical Bayesian framework for backprop networks. *Neural Computation*, 4, 448-472.
- MADIGAN, D., and RAFTERY, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535-1546.
- MARRS, A.D. (1998). An application of reversible-jump MCMC to multivariate spherical Gaussian mixtures. In *Advances in Neural Information Processing Systems 10*. San Mateo, CA: Morgan Kaufmann. 577-583.
- MÜLLER, P., and INSUA, D.R. (1998). Issues in Bayesian analysis of neural network models. *Neural Computation*, 10, 749-770.
- NEAL, R.M. (1996). *Bayesian Learning For Neural Networks*. New York: Springer-Verlag.
- NETER, J., KUTNER, M.H., NACHTSHEIM, C.J. and WASSERMAN, W. (1996). *Applied Linear Statistical Models* (Fourth Edition). Chicago: Irwin.
- ROBERTS, C.P., and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer.
- ROBERTS, G.O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman & Hall/CRC. 45-57.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.
- SILVA, P.L.D. (1996). Some asymptotic results on the mean squared error of the regression estimator under simple random sampling without replacement. Southampton: University of Southampton, Center for Survey Data Analysis Technical Report 6-2.
- SILVA, P.L.D., and SKINNER, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.
- THEBERGE, A. (1999). Extensions of calibration estimators in survey sampling. *Journal American Statistical Association*, 94, 635-644.
- TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, 22, 1701-1786.
- VALLIANT, R., DORFMAN, A.H. and ROYALL, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley & Sons, Inc.
- WALKER, A.M. (1969). On the asymptotic behaviour of posterior distributions. *Journal Royal Statistics Society, B*, 31, 80-88.
- WEIGEND, A.S., HUBERMAN, B.A. and RUMELHART, D.E. (1990). Predicting the future: A connectionist approach. *Int. J. Neural Syst.* 1, 193-209.
- WU, C., and SITTE, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal American Statistical Association*, 96, 185-193.