Journal of Classification 30 (2013)
DOI: 10.1007/s00357-013-9134-7

Using Neural Network Analysis to Define Methods of DINA Model Estimation for Small Sample Sizes

Zhan Shu

The University of North Carolina at Greensboro

Robert Henson

The University of North Carolina at Greensboro

John Willse

The University of North Carolina at Greensboro

Abstract: The DINA model is a commonly used model for obtaining diagnostic information. Like many other Diagnostic Classification Models (DCMs), it can require a large sample size to obtain reliable item and examinee parameter estimation. Neural Network (NN) analysis is a classification method that uses a training dataset for calibration. As a result, if this training dataset is determined theoretically, as was the case in Gierl's attribute hierarchical method (AHM), the NN analysis does not have any sample size requirements. However, a NN approach does not provide traditional item parameters of a DCM or allow for item responses to influence test calibration. In this paper, the NN approach will be implemented for the DINA model estimation to explore its effectiveness as a classification method beyond its use in AHM. The accuracy of the NN approach across different sample sizes, item quality and Q-matrix complexity is described in the DINA model context. Then, a Markov Chain Monte Carlo (MCMC) estimation algorithm and Joint Maximum Likelihood Estimation is used to extend the NN approach so that item parameters associated with the DINA model are obtained while allowing examinee responses to influence the test calibration. The results derived by the NN, the combination of MCMC and NN (NN MCMC) and the combination of JMLE and NN are compared with that of the well-established Hierarchical MCMC procedure and JMLE with a uniform prior on the attribute profile to illustrate their strength and weakness.

Keywords: Neural network; Diagnostic Classification Model; MCMC.

Corresponding Author's Address: Zhan Shu, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541; e-mail: zshu@ets.org.

1. Introduction

Diagnostic Classification Models (DCMs) are gaining popularity in educational measurement. There has been increasing pressure in educational assessment to make assessments sensitive to specific examinee attributes, knowledge, and other cognitive features (Junker and Sijtsma 2001). DCMs have been developed primarily for the purpose of identifying master or non-master students on fine-grained specific attributes, knowledge or other cognitive attributes. Diagnosis information provided by DCMs enables educators to focus on students' cognitive characteristics as opposed to only on a total score. However, one of the challenges of the application of DCMs is that they typically require a large sample size to reliably estimate model parameters.

An exception of the large sample size requirement is the method proposed by Gierl (2004) for the Attribute Hierarchical Method (AHM). This method uses a Neural Network (NN) analysis to classify students as mastery or non-mastery on a set of specific attributes. Garson (1998) noted that the NN analysis can be used for classification problems that involve mapping a large number of inputs onto a small number of output classes. In the AHM, real response patterns for examinees can be transformed into attribute patterns through the linking functions between the expected response pattern and the attribute pattern, calibrated by the NN analysis. Because of their approach, which involves training the NN using a pre-specified data set, sample size is not a factor affecting the accuracy of the NN classification. As a result, the NN approach could be applied in small sample size cases for relatively reliably estimating students' attribute profile, such as in a formative exam developed for a single classroom.

Like in the AHM, the current research uses the NN approach to estimate mastery or non-mastery profile, without assuming any hierarchical relationships among attributes. Furthermore, the current research shows how the item parameters can be estimated and empirical evidence can be allowed to influence the calibration by combining the NN approach and a Markov Chain Monte Carlo (MCMC) estimation algorithm and Joint Maximum Likelihood Estimation (JMLE) for the DINA estimation. In this study, the NN's properties under different conditions are illustrated at first, especially its features under small sample size conditions. And then the feasibility of combining the NN approach and MCMC (NN MCMC) and JMLE (NN JMLE) are addressed, with a special attention to how the NN interact with the MCMC and JMLE in small sample size conditions.

2. Background

2.1 DINA

The DINA model (Deterministic Input; Noisy "And" Gate, Haertel 1989; Junker and Sijtsma 2001) is a commonly discussed DCM (Junker and Sijtsma 2001; de la Torre and Douglas 2004; Templin and Henson 2006). The probability of a correct response defined by the DINA model is a function of a latent variable η_{ij}:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}} . \tag{1}$$

Here, α_{ik} is the i^{th} examinee's real knowledge status (master or non-master) on the k^{th} attribute and q_{jk} is the element of Q-matrix that defines requirement of the k^{th} attribute by the j^{th} item. If the i^{th} examinee has mastered all the measured attributes in the Q-matrix, then $\eta_{ij} = 1$; otherwise, $\eta_{ij} = 0$.

Given η_{ij}, the probability of a correct response P $(X_{ij} = 1 | \, \eta_{ij})$ is defined by the DINA for the j^{th} item as:

$$\mathrm{P}\left(X_{ij} = 1 \middle| \, \eta_{ij}\right) = \left(1 - s_j\right)^{\eta_{ij}} * g_j{}^{1-\eta_{ij}} , \tag{2}$$

here the slipping parameter s_j is the probability of an incorrect response for the j^{th} item when the i^{th} individual has mastered all the attributes measured by the j^{th} item. The guessing parameter g_j is the probability of a correct response for the j^{th} item when the i^{th} individual has not mastered all of the attributes measured by the j^{th} item. The DINA model is a conjunctive model that uses two parameters (slip and guessing) to define the probability of a correct response. A positive response is most likely when all attributes measured by the item has been mastered. Lacking a single attribute or more measured by that item will reduce an examinee's probability of a correct response to the level of guessing.

2.2 Attribute Hierarchical Method

The Attribute Hierarchical Method (AHM; Leighton, Gierl, and Hunka 2004)) shares some common features with the DINA model. Leighton, Gierl and Hunka (2004) described that "*The basic assumption of the AHM is that the test performance depends on a set of hierarchically ordered competencies called attributes. The test takers must master all* *attributes measured by the item to answer the item correctly*" (p. 207). The attributes hierarchy of the AHM specifies that mastery of the low-level attributes is required before mastery of the higher level attributes. For example, if Attribute 1 is the low-level attribute of Attribute 2, then students cannot master Attribute 2 without firstly mastering Attribute 1. When an item requires Attribute 2, then examinees should master both Attribute 1 and Attribute 2 to obtain a correct response on the item. Essentially, the DINA model and AHM share a common assumption that all of the attributes measured by an item must be mastered for having a correct response to that item.

The attribute hierarchy is among the most important inputs for the AHM, because it is the key factor that defines the attribute patterns and the test's Q-matrix. However, the attributes in a DINA application do not necessarily follow a hierarchy. Thus, if the number of attributes is *K*, the total number of attribute patterns without assuming any attribute hierarchies is 2^K. In the AHM, attribute patterns that do not comply with the attribute hierarchy are removed, while the remaining patterns define the test's Q-matrix (LGierl 2004). The transposed test's Q-matrix represents the possible attribute patterns of examinees under the constraint of the attribute hierarchy (The term "possible attribute patterns" is used to represent the attribute patterns defined by the attribute hierarchical relationship under AHM context). The expected response patterns are defined based on the test's Q-matrix and the possible attribute patterns. The term "expected examinees' response patterns" (Leighton, Gierl, and Hunka 2004) implies that the response pattern should be observed if the attribute hierarchy is true, examinees invoke attributes consistently with attribute hierarchy without slipping/guessing and assuming the Q-matrix is correct. In other words, the AHM is a constrained DINA model, with specified attribute relationship and zero slipping/guessing. That is why the NN can be naturally applied in the DINA for estimating students' attribute profile.

2.3 Neural Network Analysis

The NN can be described broadly as a type of parallel-processing architecture that transforms a stimulus received by an input unit to a signal for the output unit through a series of hidden units (Gierl 2007; Garson 1998). It is formed in three layers: input layer, output layer and hidden layer. A training process is needed to establish the connections between the units in the input layer and those in the output layer. Through the training process, the NN is supposed to identify a linking function f to map the units in the input layers $\boldsymbol{x}$ to the units of the output layers $\mathbf{y}$, $y=f(x)$, through the hidden layers g. The linking function between x and y is defined in Equation 3:

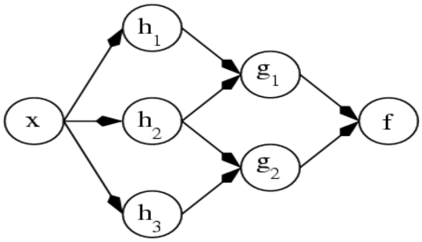

Figure 1. NN with two hidden layers

$$y = f(x) = K(\sum_i^n \omega_{1i} * \boldsymbol{g}(\boldsymbol{x})) \quad (3)$$

Typically, K is a predefined function, such as hyberbolic tangent function. ***g(x)*** represents a collection of units in the hidden layers. An example of typical NN structure is shown in Figure 1. x is the unit of the input layer, h and g represents the two hidden layers and y is the unit of the output layer. In this example, x is mapped to y through two hidden layers g and h, where g is predicted by h. The linking functions of this example are defined as follows:

$$y = f(x) = K(\sum_i^2 \omega_{1i} * g_i(x)), \quad (4)$$

$$g_1(x) = K(\omega_{21} * h_1(x) + \omega_{22} * h_2(x)), \quad (5)$$

$$g_2(x) = K(\omega_{23} * h_3(x) + \omega_{24} * h_2(x)). \quad (6)$$

Essentially, a training process is an optimization process which identify the best linking function f between the input x and output y from a collection of linking functions F. A cost function, such as $C= E[(y-f(x))^2]$, is defined for evaluating the degree of error of the training process. As long as C is less than a predefined level (called "learning rate"), the learning process stops and the optimized function f is saved for transforming a set of inputs to desired output.

In the AHM context, the expected response pattern (η_i) and the corresponding attribute patterns (α_i) for that response pattern are the raw information to train the NN (i.e., supervised NN training). The elements of the expected response patterns are the stimulus of the input layer and the corresponding attribute patterns are the desired signal for the output layers. Notice that in doing this all items are assumed to behave in the same way and so no differentiation is made between items with respect to item quality (i.e., the slipping and guessing level in the DINA model). After the training process, the linking function between α_i and η_i, f is established:

$$\eta_i = f(\alpha_i). \quad (7)$$

The real response patterns X serve as the stimulus of the input layer and then are transformed through the established linking function f in the training process. In this way, the response patterns are transformed into the corresponding attribute patterns S, which is defined as:

$$S = f(X). \tag{8}$$

The elements in S are usually values between 0 and 1, and thus are typically interpreted as "probabilities"—that is the "probability" of each examinee mastering each attribute.

The accuracy of the NN analysis mainly depends on the accuracy of the training process, that is, how accurate the training process can represent the relationship between the input and output. In the AHM context, the linking functions between the expected response patterns and the possible attribute patterns are characterized by the assumption of the attribute hierarchy, non-compensatory/conjunctive property and the test Q-matrix. In addition, factors relative to the NN itself such as its structure, learning rate and predefined functions, can also impact the accuracy of the NN analysis. However, after the linking functions have been calibrated and established in the training, the sample size of examinees does not affect the accuracy of the NN. This may be considered as one of the NN analysis's strengths, that is, it does not require a large sample size for reliably estimating students' skill profile and item parameters. And yet a limitation of the use of a NN analysis is that all "calibration" of the model is based on connections between response patterns and attribute patterns calibrated in the training process, no empirical information (i.e., students' response patterns) is used to influence the procedure.

Given the commons shared between the AHM and DINA model, the application of the NN analysis in the DINA model is a natural extension. However, one should note that when the NN analysis is directly applied in the DINA model, the typical item parameters (slips and guesses) are not calibrated and provided. Thus, item and test development are difficult to be evaluated, which is why this paper proposes combining the NN analysis with the MCMC and JMLE estimation.

3. Methodology

In this section, the NN is firstly applied in the DINA for estimating examinees' attribute profile, and then expanded through its combination with the MCMC and JMLE algorithms. Especially in such expansions, the combinations are supposed to be effective in cases with small sample size (like the NN itself), while also allowing for empirical information to influence skill estimation and evaluate item performance (i.e., providing items' slipping and guessing level).

The NN analysis in the DINA model has a similar process as in the AHM: defining the network (e.g., how many hidden layers), then training the network for calibrating the linking functions, and finally classifying examinees using the real response patterns as the stimulus. However, because no assumed of an attribute hierarchy when estimating the DINA there are more possible attribute patterns to be considered in the training process. Specifically, all possible (2^K) attribute patterns are considered as the target output in the training along with their expected response patterns as the input. Although the NN, as a classification method, may be enough for attribute pattern classification, no item differentiation in terms of item quality (some items may have a higher discrimination between masters and non-masters) is incorporated in the NN's classification process. Nor are any traditional item parameters directly estimated, which could be used to evaluate item quality. Therefore, this approach is expanded using more typical estimation methods for DCM estimation, MCMC and JMLE.

The MCMC is a method used for estimation of item parameters and classification of examinees as mastery or non-mastery on a set of specific attributes in the DINA model (e.g., Henson and Templin 2006). It is a Bayesian theory based estimation and, as a result, parameter estimation is based on the information provided by the prior distributions of the parameters (called "priors") and the empirical information provided by the data. If the priors provide no information (i.e., non-informative), the MCMC estimation can be directly compared to the maximum likelihood estimation because parameter estimates through the MCMC maximize the likelihood of the data. However, if the "priors" are informative then estimates are biased because they are "compromises" between the information in the prior distributions for the parameters and the empirical information. Traditionally, priors have been defined such that they are generally non-informative (e.g., the sampling candidates are sampled from a uniform distribution or standardized normal distribution). However, alternative approaches empirically estimate the priors (commonly referred to as hierarchical Bayesian estimation, not to be confused to the previously discussed attribute hierarchy). Examples of a hierarchical Bayesian approach can be found in the research about higher order latent trait models for cognitive diagnosis by De la Torre and Roussos (2004) and hierarchical MCMC estimation procedure for the DINA by Templin, Henson, Templin, and Roussos (2008). The Hierarchical MCMC algorithm introduces three parameters to model examinees' attribute profile as opposed to using examinee-attribute specific priors: common attribute (i.e., the second-order parameter underlying all the attributes), location and slope parameters for each attribute. The reason why this algorithm is called Hierarchical MCMC is that it makes use of hierarchical parameters to model the attribute profile. Although a hierarchical approach can provide more accurate estimation of attribute mastery classification, because of the increase of estimated model parameters (the priors are also estimated) the demand for sample sizes is accordingly higher.

In addition to the MCMC, the Joint Maximum Likelihood Estimation (JMLE; de la Torre 2009) is another method to estimate item parameters and examinees' attribute profile. The JMLE assumes the conditional independence given examinees' attribute profile, and simultaneously estimates item parameters and examinees' attribute profile (α) by achieving the maximum likelihood which is defined as:

$$\prod_{i=1}^{I} \sum_{j=1}^{J} P(\alpha_i)^{X_{ij}} [1 - P(\alpha_i)]^{1-X_{ij}}, \qquad (9)$$

where i refers to the i^{th} examinee, and j refers to the j^{th} item. The slipping parameter of a particular item is estimated based on the percentage of examinees who master all the required attributes of the item having a wrong answer to that item in the examinees who master all the required attributes. Similarly, the guessing parameter of a particular item is estimated based on the percentage of examinees who lack of at least one of the required attributes by the item have a correct answer to the item in the examinees who should have the item wrong. However, when Q-matrix is complex (resulting in a great number of skill patterns) or number of examinees is small, it is very likely that we do not have any examinee who has certain skill pattern required by certain items, and so no information could be provided to estimate slipping or guessing parameters of those items, consequently, the JMLE does not converge in such cases.

As an alternative to model the priors through hierarchical Bayesian or directly setting the priors (e.g., as a non-informative prior), the information from the NN can be used to set the priors of a MCMC algorithm. Thus, it is possible to obtain parameter estimates (both attribute mastery and item parameters) that are "compromises" between the results provided from a NN and the empirical information provided by the data. Specifically, the examinee attribute classification from the NN is typically interpreted as the probability that a specific attribute has been mastered by that examinee. When an examinee has a predicted value near 1 for a given attribute by NN, that examinee has responded in a way consistent with the training result of the NN and therefore is very likely to be a master. Also values close to 0 imply that an examinee is most likely to be a non-master. More importantly, examinee-attribute values predicted by the NN that are close to 0.5 imply that the examinee has not responded in a way that it is easily classified by the NN analysis. Thus, the training of the NN is not effective in classifying these attributes as having been mastered or not mastered, however, this may be because the NN is not sensitive to differences in item quality (i.e., empirical information). By using the results from the NN as a prior, those values in the NN analysis that are close to the value of 1 or 0 are "strong" priors and therefore output of the MCMC/JMLE will remain very similar to that of the NN analysis (the prior). However, when the NN analysis results are close to 0.5 for a specific attribute and examinee, the prior will resemble an uninformative prior. So the empirical information provided in the data will directly influence the posterior probability of mastery and thus classification of mastery for that examinee's attribute.

It should be noted that when specifying the prior for the MCMC and JMLE algorithm using the results from NN, examinee-attribute specific priors are defined. That is, a probability of mastery for every examinee for each attribute is defined as a prior, where all priors are assumed to be independent. However, even though these priors are assumed to be independent, the association between any two attributes in the population is indirectly incorporated because examinees with high probabilities on one attribute may have high probabilities on the other attributes and vice versa. The impact of the prior is relative to the variation of sample size. If the sample size is small (as in a classroom setting), the impact of the prior on the estimation quality is large, and vice versa. Thus, when there is little or no data available, the NN result works as the prior of the attribute profile and that prior mostly determines the accuracy of the estimation. However, the MCMC/JMLE algorithm and any information with respect to item quality (as indicated by item parameters) can aid in the accuracy of the estimation when the sample size is large.

Given the fact that the sample size of examinees does not critically affect the accuracy of the NN analysis, it is reasonable to believe that the MCMC/JMLE estimation using the NN results as the prior on examinees' attribute profile may probably provide accurate estimates when the sample size is small, and perform similarly as other MCMC methods such as the Hierarchical MCMC when the sample size is large. The combined algorithms of the NN and MCMC/JMLE, on one hand, make use of the beneficial properties of the NN, and on the other hand extend the application of the NN by accounting for the item differentiation information.

4. Simulation Design

A simulation study was designed to demonstrate the performance of the combination of the NN with the MCMC/JMLE algorithms, especially their properties under small sample size conditions. The well-established Hierarchical MCMC and JMLE algorithm with non-informative priors are used in this study as a comparison reference. Specifically, four factors have been taken into consideration: the complexity of the Q-matrix, alpha patterns (i.e., students' skill profile), item quality and sample size.

In the design of the Q-matrix, two types of Q-matrix structure are considered in this research: simple structure and complex structure. In the simple structure cases, items measure one and only one attribute. The attributes are evenly measured by the same number of items; in the complex Q-matrix, each item measures between one and three attributes, and each attribute is measured by at least 10 items. The other factor determining the complexity of the Q-matrix is the number of attributes. Two conditions are considered, the first type of Q-matrix has 4 attributes and the other type of Q-matrix has 6 attributes.

With respect to the alpha pattern (α), it is simulated by using a dichotomized multivariate normal distribution with covariance matrix Ξ and mean vector μ, defined as:

$$\Xi=\begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}, \tag{10}$$

$$\mu = c(\mu_1, \ldots \mu_k),$$

where ρ is randomly selected from U (0.4, 0.6), the elements of the μ vector are set to zero in the simulation process. The cut points for each attribute are set to zero.

$$\alpha_{ij} = \begin{cases} 1, z_{ij} \geq 0 \\ 0, z_{ij} < 0 \end{cases}. \tag{11}$$

Here z_{ij} is the element of the simulated values of the i^{th} examinee mastering the j^{th} attribute, "α_{ij} =1" represents the mastery of attributes and "α_{ij} =0" refers to the non-mastery of attributes. Notice that, this structure is approximately a single factor structure, which also necessarily incorporates attribute associations within a population.

In addition, two types of items are considered in this simulation. In the conditions with well-performed items, the slip and guessing parameters are randomly sampled from U(0.1, 0.2). In the other cases with relatively poorly-performed items, the slips and guessing parameters are randomly generated from U(0.2,0.4). Finally, six different sample sizes are considered: 1000, 500, 200, 150, 50 and 20, for demonstrating the properties of the NN and its combinations with the MCMC/JMLE under various sample sizes.

As a summary, all the variables considered in this simulation study are listed in Table 1. The conditions will be combined together to create 48 joint conditions and each joint condition is sampled 25 times. In the simu-

Table 1.Summary of Considered Variables

Variables	Number of conditions
Q-matrix structure	2
Number of attributes	2
Item quality	2
Sample size	6

lation process, the package "AMORE" (Castejón Limas and Ordieres Meré 2009) built in R is used to complete the NN classification. The software "LDCM.exe" (Henson 2008), written for the Log-Linear Cognitive Diagnostic Model (LCDM) estimation is used to conduct the Hierarchical MCMC of the DINA model. The Joint Maximum Likelihood estimation in R statistic language written by Willse (2010) is used as the JMLE estimation software.

5. Results

The results of the NN and its combinations are presented and discussed in this section. As a note, the chain length for the posterior samples in the MCMC is 5000 where the first 3000 samples serve as the burn-in and the last 2000 samples after the burn-in are used to characterize the posterior distribution of each parameter.

5.1 Robustness of the NN as a Classification Method

The results of the average CCR across all attributes derived by the NN, NN MCMC, NN JMLE (i.e., JMLE with NN as initial values), the hierarchical MCMC and JMLE with non-informative initial values, are reported in Tables 2 to 5. In the four tables, the first column is the average CCR of the NN analysis, the second column represents the results derived by the NN MCMC, and the third column is the average CCR of the hierarchical MCMC, the fourth column is the results by the JMLE using non-informative initial values, and the last column has the results of the NN JMLE. As a note, the JMLE and NN JMLE do not converge in a lot of small sample size cases, which is why some of their CCRs are filled with "NA". The percentage of their non-convergence (the number of non-convergence over the total number of replication for each condition) is presented in Table 6. The eight combined conditions are listed in Table 7 with assigned labels for a convenience description in the following sections.

Table 2. The CCR of the cases with 4 attributes and item parameters within U (0.1, 0.2)

structure	sample Size	NN	NN_MCMC	HI_MCMC	JMLE	NN_JMLE
simple	1000	0.974	0.979	0.983	0.979	0.978
	500	0.973	0.979	0.983	0.978	0.978
	200	0.975	0.978	0.982	0.979	0.980
	150	0.975	0.978	0.981	0.977	0.976
	50	0.975	0.977	0.973	0.973	0.977
	20	0.974	0.977	0.967	0.973	0.976
complex	1000	0.861	0.910	0.941	0.934	0.918
	500	0.861	0.910	0.939	0.934	0.914
	200	0.871	0.919	0.942	0.934	0.914
	150	0.878	0.922	0.943	0.937	0.914
	50	0.971	0.917	0.924	0.931	0.900
	20	0.875	0.916	0.923	NA	NA

NN_MCMC= MCMC procedure with NN as prior; Hi_MCMC=Hierarchical MCMC procedure.

Table 3. The CCR of the cases with 6 attributes and item parameters within U (0.1, 0.2)

structure	sample Size	NN	NN_MCMC	HI_MCMC	JMLE	NN_JMLE
simple	1000	0.942	0.949	0.962	0.946	0.950
	500	0.941	0.948	0.962	0.948	0.950
	200	0.943	0.950	0.960	0.946	0.948
	150	0.943	0.949	0.956	0.946	0.948
	50	0.945	0.948	0.939	0.948	0.944
	20	0.950	0.945	0.926	0.943	0.941
complex	1000	0.768	0.825	0.888	0.880	0.862
	500	0.768	0.822	0.882	0.880	0.863
	200	0.775	0.833	0.883	0.876	0.855
	150	0.771	0.829	0.877	0.875	0.854
	50	0.775	0.833	0.857	0.874	0.842
	20	0.769	0.818	0.814	NA	NA

NN _MCMC= MCMC procedure with NN as prior; Hi_MCMC=Hierarchical MCMC procedure.

Table 4. The CCR of the cases with 4 attributes and item parameters within U (0.2, 0.4)

structure	sample Size	NN	NN_MCMC	HI_MCMC	JMLE	NN_JMLE
simple	1000	0.843	0.858	0.873	0.851	0.854
	500	0.840	0.858	0.870	0.846	0.850
	200	0.838	0.854	0.819	0.844	0.849
	150	0.838	0.853	0.775	0.846	0.851
	50	0.838	0.841	0.706	NA	NA
	20	0.845	0.850	0.710	NA	NA
complex	1000	0.720	0.768	0.828	0.785	0.763
	500	0.715	0.763	0.825	0.787	0.772
	200	0.713	0.761	0.785	0.750	0.773
	150	0.714	0.758	0.746	0.795	0.773
	50	0.723	0.764	0.626	NA	NA
	20	0.720	0.765	0.603	NA	NA

NN _MCMC= MCMC procedure with NN as prior; Hi_MCMC=Hierarchical MCMC procedure

Table 5. The CCR of the cases with 6 attributes and item parameters within U (0.2, 0.4)

structure	sample Size	NN	NN_MCMC	HI_MCMC	JMLE	NN_JMLE
simple	1000	0.787	0.802	0.828	0.789	0.795
	500	0.788	0.800	0.802	0.792	0.792
	200	0.785	0.797	0.727	0.794	0.793
	150	0.786	0.792	0.700	NA	NA
	50	0.790	0.795	0.658	NA	NA
	20	0.788	0.785	0.677	NA	NA
complex	1000	0.643	0.677	0.742	0.714	0.696
	500	0.650	0.680	0.742	0.718	0.697
	200	0.643	0.675	0.702	NA	NA
	150	0.642	0.673	0.649	NA	NA
	50	0.633	0.666	0.563	NA	NA
	20	0.627	0.660	0.538	NA	NA

NN _MCMC= MCMC procedure with NN as prior; Hi_MCMC=Hierarchical MCMC procedure.

Table 6. The percentage of non-convergence

Number of Attributes	Sample Size	U(0.1, 0.2)		U(0.2, 0.4)	
		Complex	Simple	Complex	Simple
4-Attributes	1000	0.00%	0.00%	0.00%	0.00%
	500	0.00%	0.00%	0.00%	0.00%
	200	0.00%	0.00%	0.00%	0.00%
	150	0.00%	0.00%	0.00%	0.00%
	50	0.00%	0.00%	**36.00%**	**44.00%**
	20	**4.00%**	0.00%	**80.00%**	**72.00%**
6-Attributes	1000	0.00%	0.00%	0.00%	0.00%
	500	0.00%	0.00%	0.00%	0.00%
	200	0.00%	0.00%	**8.00%**	0.00%
	150	0.00%	0.00%	**24.00%**	**4.00%**
	50	0.00%	0.00%	**52.00%**	**20.00%**
	20	**12.00%**	0.00%	**68.00%**	**64.00%**

Complex= the Q-matrix with complex structure; Simple= the Q-matrix with simple structure; U(0.1,0.2)= item parameters within the interval between 0.1 and 0.2; U(0.2, 0.4)= item parameters within the interval between 0.2 and 0.4.

Table 7. The Eight Conditions

Condition label	Combination of the conditions
4att_s_h	4 attributes & simple structure & U(0.1,0.2)
4att_c_h	4 attributes & complex structure & U(0.1,0.2)
6att_s_h	6 attributes & simple structure & U(0.1,0.2)
6att_c_h	6 attributes & complex structure & U(0.1,0.2)
4att_s_l	4 attributes & simple structure & U(0.2,0.4)
4att_c_l	4 attributes & complex structure & U(0.2,0.4)
6att_s_l	6 attributes & simple structure & U(0.2,0.4)
6att_c_l	6 attributes & complex structure & U(0.2,0.4)

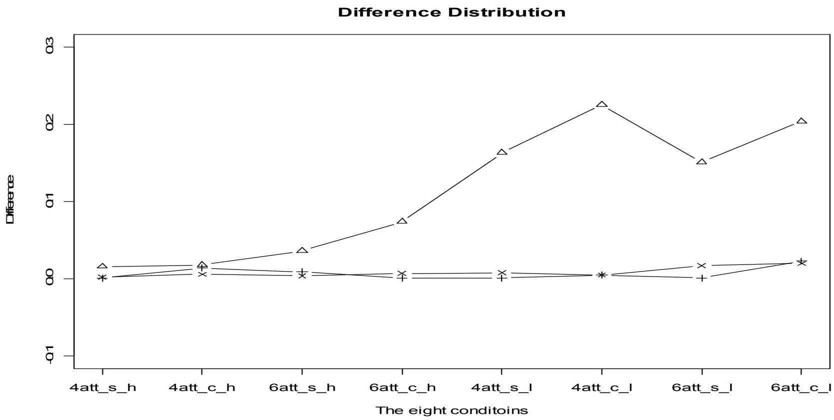

Figure 2. "Δ" = the Hierarchical MCMC's range; "+"= the NN's range; "X"= the NN MCMC's range.

The NN's CCR is stable across different sample sizes. The ranges of the average CCR within the eight conditions (i.e., the greatest difference among the average CCRs across the three sample sizes) are presented in Figure 2. The range of the NN analysis, as well as the NN MCMC, is almost zero in every condition. In comparison, the range of the Hierarchical MCMC is greater than that of the NN and NN MCMC, and such range is even much greater in cases with high level of slips/guessing or a high degree of Q-matrix complexity. The small range is a direct demonstration that both the NN and NN MCMC are robust to the sample size variation, while the Hierarchical MCMC is more sensitive to sample size than the NN and NN MCMC.

Furthermore, well-developed items (i.e., low level of slips and guessing parameters) will improve the NN's CCR. For example in Table 8, the average CCR of the four cases under U (0.1, 0.2) is uniformly greater than that of the corresponding cases under U (0.2, 0.4). Oppositely, an increasing Q-matrix complexity will decrease the NN' CCR. The average CCR under the simple structure cases is uniformly greater than that of the complex structure cases. Similarly, the average CCR under the 4-attribute cases is uniformly greater than that of the 6-attribute cases.

As a illustration in Table 8, the average CCRs under simple structure cases (the first column in Table 8) are uniformly higher than that of the complex structure cases (the second column in Table 8); similarly, the average CCRs under 4-attribute cases (the first and third rows in Table 8) are uniformly higher than those of 6-attribute cases (the second and forth rows in Table 8).

Table 8. The CCR of the NN analysis under different conditions

Conditions	Simple	Complex
U(0.1,0.2)/4	0.974	0.886
U(0.1,0.2)/6	0.944	0.771
U(0.2,0.4)/4	0.840	0.718
U(0.2,0.4)/6	0.787	0.640

U (0.1, 0.2) and U (0.2, 0.4) are the item parameter interval; 4 and 6 represent the number of attribute]

The NN MCMC performs better than the NN, and also shows a higher degree of robustness than the Hierarchical MCMC in the small sample size cases. For example, the NN MCMC has a similar degree of accuracy of classifying the students as the Hierarchical MCMC does in the simple structure cases, but it has a lower CCR than the Hierarchical MCMC in the complex structure cases when the sample size is greater or equal to 200. In contrast, its CCR is uniformly greater than the Hierarchical MCMC in both the simple and complex structure cases when the samples are below 150. Similarly, the NN JMLE performs better than the JMLE (with non-informative initial values) in the simple-structure cases and slightly worse in the cases with a complex-structure Q-matrix.

As shown in Table 6, the JMLE and NN JMLE may not be able to converge and provide analysis results in many small sample size cases, especially in the cases where the Q-matrix is complex or the level of slipping/guessing is high. A smaller sample size will result in a greater percentage of non-convergence. The non-convergence cases of the JMLE or NN JMLE should be due to the methodology itself because there are no enough samples to provide information for certain attribute patterns in the Q-matrix. Even though there is the possibility of non-convergence, the NN JMLE and JMLE with non-informative initial values are computationally efficient. For example, the JMLE/NN JMLE take about 3 seconds to provide estimates with 1000 students in R statistic language on a Windows XP (32-bit) computer with Inter Core P4 processor and 1GB Installed Memory. In contrast the NN MCMC will take about 40 minutes to run 5000 iterations. Therefore, the NN JMLE should be a better choice than the NN MCMC or NN in the converged cases, because the NN JMLE requires much less computational time and yet derives better results, but the NN MCMC should be the alternative when the NN JMLE does not converge.

5.2 Item Estimation Quality of the NN MCMC

The NN MCMC's item estimation quality is discussed in this section. Note, the NN JMLE is not discussed because of its non-convergence in small sample size cases. The correlation between the estimated and true item parameters (COR), the mean difference between the estimated and true parameters (MD), and the root mean square difference (RMSD) are used in Tables 9 and 10 to describe the estimation quality. The recovery quality of the *slips* has a similar pattern as that of the *guessing* in all the conditions, and so the following discussion will just focus on *slips*.

The item estimation of the NN MCMC deteriorates along with the decreasing sample size and increasing Q-matrix complexity. For example, in the simple structure cases with 1000 and 500 samples, the correlations between the true and estimated parameters are all above 0.73. The mean of the MD and RMSD of these conditions are -0.01 and 0.0225. In 200 and 150 sample cases, the correlations drop with a range from 0.50 to 0.68. When sample size drops down to 50 or below, the correlations are between 0.20 and 0.37.

As compared to the Hierarchical MCMC, the NN MCMC performs as well as the Hierarchical MCMC in the simple-structure Q-matrix cases with 1000 students, and outperforms the Hierarchical MCMC when the sample size is 500 or below. Oppositely, in the complex structure cases, the Hierarchical MCMC overall has a better item estimation than the NN MCMC, but the difference of the three indices between the two methods becomes smaller as the sample size decreases. As shown by this comparison, the advantage of the NN MCMC is its capability to obtain relatively more accurate estimation than the hierarchical MCMC in the small sample size cases. However, when the Q-matrix is of complex structure and the sample size is equal/below to 500, the item estimation quality of the NN MCMC is questionable, and should not be applied in practice no matter the Q-matrix is of simple or complex structure.

As a summary, the classification quality of the NN analysis will not be impacted by the sample size, but influenced by the Q-matrix complexity and item quality. The NN is combined with the MCMC/JMLE for defining relatively efficient methods for the DINA in cases with a small sample size. Such combinations provide more accurate examinees' classification and item estimation than the hierarchical MCMC and JMLE with non-informative starting values. However, it is a worthy note that the item estimation in small sample size cases is not reliable.

Table 9. Item estimation quality indices when true item parameters within U (0.1, 0.2)

		NN _MCMC						HI_MCMC					
		Slips			Guessing			Slips			Guessing		
Number	Structure	Size	COR	MD	RMSD	COR	MD	RMSD	COR	MD	RMSD	COR	MD	RMSD

Number	Structure	Size	NN Slips COR	NN Slips MD	NN Slips RMSD	NN Guessing COR	NN Guessing MD	NN Guessing RMSD	HI Slips COR	HI Slips MD	HI Slips RMSD	HI Guessing COR	HI Guessing MD	HI Guessing RMSD
		1000	0.86	0.00	0.01	0.88	0.00	0.01	0.87	0.00	0.01	0.88	0.00	0.01
	simple	500	0.75	0.00	0.02	0.81	0.00	0.02	0.75	0.00	0.02	0.81	0.00	0.02
		200	0.64	0.01	0.03	0.62	0.00	0.03	0.64	0.00	0.03	0.62	-0.01	0.03
		150	0.57	0.01	0.03	0.53	0.01	0.03	0.57	-0.01	0.04	0.51	-0.01	0.04
		50	0.37	0.03	0.06	0.37	0.03	0.06	0.36	-0.01	0.07	0.35	-0.01	0.07
4 Attributes		20	0.22	0.06	0.09	0.21	0.06	0.09	0.20	-0.01	0.10	0.17	-0.01	0.11
		1000	0.73	0.01	0.02	0.63	-0.01	0.02	0.82	0.00	0.02	0.87	0.00	0.01
	complex	500	0.66	0.01	0.03	0.62	-0.01	0.02	0.73	0.00	0.02	0.80	0.00	0.02
		200	0.56	0.02	0.04	0.52	0.00	0.03	0.58	0.00	0.04	0.63	-0.01	0.03
		150	0.51	0.02	0.04	0.51	0.00	0.03	0.51	-0.01	0.04	0.58	-0.01	0.03
		50	0.25	0.05	0.08	0.39	0.01	0.05	0.23	-0.01	0.08	0.39	-0.01	0.06
		20	0.16	0.10	0.11	0.22	0.04	0.07	0.16	-0.02	0.11	0.23	-0.02	0.08
		1000	0.86	0.00	0.01	0.86	0.00	0.01	0.86	0.00	0.01	0.86	0.00	0.01
	simple	500	0.73	0.00	0.02	0.74	0.00	0.02	0.74	0.00	0.02	0.75	0.00	0.02
		200	0.54	0.00	0.03	0.54	0.00	0.03	0.55	-0.01	0.04	0.52	-0.01	0.04
		150	0.50	0.00	0.04	0.53	0.00	0.03	0.47	-0.01	0.04	0.54	-0.01	0.04
		50	0.34	0.02	0.06	0.30	0.02	0.06	0.30	-0.02	0.07	0.29	-0.01	0.08
6 Attributes		20	0.13	0.06	0.09	0.15	0.06	0.08	0.12	-0.02	0.12	0.13	-0.01	0.11
		1000	0.52	0.03	0.04	0.53	-0.02	0.03	0.76	0.00	0.02	0.86	0.00	0.01
	complex	500	0.47	0.03	0.05	0.52	-0.01	0.03	0.68	0.00	0.03	0.78	0.00	0.02
		200	0.40	0.03	0.05	0.45	-0.01	0.04	0.47	-0.01	0.04	0.59	0.00	0.03
		150	0.31	0.05	0.07	0.41	-0.01	0.04	0.40	-0.01	0.05	0.53	0.00	0.04
		50	0.21	0.06	0.09	0.32	0.00	0.05	0.23	-0.02	0.09	0.34	-0.02	0.06
		20	0.13	0.12	0.13	0.26	0.03	0.07	0.17	0.00	0.13	0.16	-0.02	0.09

Table 10. Item estimation quality indices when true item parameters within U (0.2, 0.4)

			NN _MCMC						HI_MCMC					
			Slips			Guessing			Slips			Guessing		
Number	Structure	Size	COR	MD	RMSD	COR	MD	RMSD	COR	MD	RMSD	COR	MD	RMSD
		1000	0.91	-0.01	0.02	0.92	-0.02	0.03	0.91	0.00	0.02	0.91	0.00	0.02
	simple	500	0.86	-0.01	0.03	0.83	-0.02	0.03	0.85	0.00	0.03	0.79	0.00	0.03
		200	0.68	-0.01	0.04	0.66	-0.02	0.05	0.59	0.00	0.07	0.46	-0.03	0.08
		150	0.68	0.00	0.05	0.67	-0.01	0.05	0.52	0.02	0.09	0.39	-0.05	0.11
		50	0.43	0.01	0.08	0.43	-0.01	0.07	0.31	0.05	0.14	0.18	-0.14	0.18
4 Attributes		20	0.31	0.03	0.10	0.34	-0.01	0.09	0.24	-0.01	0.17	0.12	-0.11	0.19
		1000	0.60	0.01	0.06	0.52	-0.02	0.05	0.90	0.00	0.02	0.91	0.00	0.02
	complex	500	0.53	0.02	0.06	0.52	-0.02	0.06	0.80	0.00	0.03	0.79	0.00	0.03
		200	0.46	0.02	0.07	0.43	-0.02	0.07	0.45	-0.01	0.10	0.54	-0.03	0.07
		150	0.39	0.03	0.08	0.42	-0.02	0.06	0.37	-0.02	0.14	0.45	-0.04	0.09
		50	0.34	0.04	0.11	0.35	-0.01	0.08	0.20	0.05	0.20	0.21	-0.12	0.17
		20	0.22	0.07	0.13	0.31	-0.02	0.09	0.24	0.09	0.21	0.18	-0.15	0.20
		1000	0.86	-0.03	0.03	0.85	-0.04	0.04	0.83	-0.01	0.03	0.85	0.00	0.03
	simple	500	0.80	-0.03	0.04	0.80	-0.03	0.04	0.69	0.00	0.05	0.62	-0.02	0.05
		200	0.63	-0.02	0.05	0.60	-0.03	0.05	0.41	0.01	0.11	0.35	-0.07	0.13
		50	0.60	-0.02	0.06	0.58	-0.03	0.06	0.41	0.02	0.12	0.28	-0.09	0.15
		50	0.39	-0.01	0.08	0.34	-0.02	0.08	0.25	0.02	0.16	0.11	-0.14	0.20
6 Attributes		20	0.23	0.02	0.10	0.20	-0.02	0.09	0.15	-0.03	0.19	0.02	-0.12	0.21
		1000	0.31	0.05	0.11	0.36	-0.02	0.08	0.83	0.00	0.03	0.81	0.00	0.02
	complex	500	0.32	0.05	0.11	0.38	-0.03	0.08	0.73	-0.01	0.05	0.73	-0.01	0.03
		200	0.28	0.05	0.11	0.35	-0.03	0.08	0.45	-0.01	0.10	0.54	-0.03	0.07
		150	0.28	0.05	0.11	0.36	-0.03	0.08	0.37	-0.02	0.14	0.45	-0.04	0.09
		50	0.19	0.07	0.14	0.38	-0.03	0.09	0.20	0.05	0.20	0.21	-0.12	0.17
		20	0.13	0.09	0.16	0.27	-0.02	0.09	0.07	0.02	0.23	0.15	-0.13	0.20

6. Discussion

Determined by the NN itself and proved by the simulation study, the accuracy of classifying students as mastery or non-mastery of attributes is not impacted by the variation in sample size in the NN analysis. However, it is affected by the Q-matrix complexity and item quality. An increasing Q-matrix complexity or decreasing item quality (i.e., a greater level of slips and guessing) tends to reduce the accuracy of the NN's classification. The MCMC/JMLE with the NN results as the prior of examinee-attribute specific profile acquires the properties of both the NN and MCMC/JMLE estimation. The information captured by MCMC/JMLE improves the NN classification rate and such improvement become greater when the sample size gets larger. In turn the NN improves the MCMC/JMLE's robustness over the small sample size and the degree of the improvement is determined by the accuracy of the NN analysis.

Of particular importance is the ability of the combination of the NN and MCMC/JMLE to provide estimates of item parameters. These estimates are important for two reasons. First, item parameters may allow for differential contribution of items in attribute mastery estimation. Note that because estimation using only NN assumes that item are determinant then all items with the same Q-matrix are assumed to contribute equally to estimation. Thus, these results are expected to become even more dramatic as the variance of item quality increases. Item parameter estimates may also prove useful in test development, where they can be used to identify items that should be revised or removed from the test. As for the two methods with the NN as the prior of student's attribute profile, the advantage of the NN JMLE lies in its efficient computation and slightly better performance. However, the NN MCMC is able to provide the item estimation results in small sample size cases, while the NN JMLE might not be able to converge.

This study is only a starting point of the application of NN in DCMs, which leaves a lot of new questions open to be investigated. For instance, Q-matrix misspecification has not been considered in this study, where the Q-matrix is assumed to be correctly specified. How the misspecification of the Q-matrix will impact the classification rate of the NN should be further discussed in the future. Furthermore, the correlations among the attributes are set between 0.4 and 0.6 in this article, how the correlations in the other intervals will impact the classification rate of the NN still needs to be clarified. Questions, such as, the application of the NN and the NN MCMC under other DCM contexts (e.g., RUM, DINO) may be interesting for some particular users.

References

CASTEJÓN LIMAS, M., and ORDIERES MERÉ, J. (2009), "The AMORE Package: A MORE Flexible Neural Network Package" (R package), http://cran.r-project.org/.

DE LA TORRE, J. (2008), "An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications", *Journal of Educational Measurement , 45(4),* 343–362.

DE LA TORRE, J. (2009), "DINA Model and Parameter Estimation: A Didactic", *Journal of Educational and Behavioral Statistics, 34(1),* 115–130.

DE LA TORRE, J., and DOUGLAS, J. (2004), "Higher Order Latent Trait Models for Cognitive Diagnosis, *Psychometrika,69(3),* 333–353.

GARSON, G. (1998), *Neural Networks: An Introductory Guide for Social Scientists*, London: SAGE Publication Ltd.

GIERL, M. (2007), "Making Diagnostic Inferences about Cognitive Attributes Using the Rule-Space Model and Attribute Hierarchy Method", *Journal of Educational Measurement, 44(4),* 325–340.

GIERL, M.J., Cui, Y., and HUNKA, S. (2007), "Using Connectionist Models to Evaluate Examinees' Response Patterns on Tests", paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.

GIERL, M., WANG, C., and ZHOU, J. (2008), "Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT", *Journal of Technology, Learning, and Assessment, 6(6).*

GIERL, M., ZHENG, Y., and CUI, Y. (2008), "Using the Attribute Hierarchy Method to Identify and Interpret Cognitive Skills that Produce Group Differences", *Journal of Educational Measurement, 45(1),* 65–89.

HAERTEL, E. (1989), ``Using Restricted Latent Class Models to Map the Sill Structure of Achievement Items", *Journal of Educational Measurement (26(4)*, 301–321.

HENSON, R., and TEMPLIN, J. (2006), *Implications of Q-matrix Misspecification in Cognitive Diagnosis*, manuscript submitted for publication.

HENSON, R. (2008), "Functions of Estimating Log-Linear Cognitive Diagnostic Model," Department of Educational Research Methodology, The University of North Carolina at Greensboro, Greensboro, NC.

JUNKER, B.W., and SIJTSMA, K. (2001), "Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory", *Applied Psychological Measurement, 25,* 258–272.

LEIGHTON, J., GIERL, M., and HUNKA, S. (2004), "The Attribute Hierarchical Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach", *Journal of Educational Measurement, 41(3),* 205–237.

PATZ, R., and JUNKER, B. (1999), "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models", *Journal of Educational and Behavioral Statistics, 24(2),* 146–178.

RUPP, A., TEMPLIN, J., and HENSON, R. (2010), *Diagnostic Measurement: Theory, Methods and Applications*, New York, NY: The Guildford Press.

STERGIOU, C., and SIGANOS, D., "Neural Networks", http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

TEMPLIN, J., HENSON, R., TEMPLIN, S., and ROUSSO, L. (2008), "Robustness of Hierarchical Modeling of Skill Association in Cognitive Diagnosis Models", *Applied Psychological Measurement (OnlineFirst).*

TEMPLIN, J., and HENSON, R. (2006), ``Measurement of Psychological Disorders Using Cognitive Diagnosis Models”, *Psychological Methods 11(3),* 287–305.

WILLSE, J.T. (2010). “Functions for Estimating DINA and DINO Models Using JML or MML”, Department of Educational Research Methodology, The University of North Carolina at Greensboro, Greensboro, NC.