


MAP Estimation using a Possibly Misspecified Parameter Redundant Model

Richard M. Golden 

Abstract In this paper, new theorems are proved which show how in some cases the asymptotic distribution of Maximum A Posteriori (MAP) estimates can be obtained for parameter redundant probability models which are possibly misspecified. The new methods are then empirically investigated in a simulation study investigating confidence interval coverage for Cognitive Diagnostic Models (CDMs). The empirical results are shown to be relevant in the application of CDMs to small sample size situations.

Key words: identifiability, parameter redundancy, model misspecification, cognitive diagnostic model, M-estimation

1 Introduction

A key challenge in probabilistic inference is the construction of an appropriate probabilistic model. If the probability model is too constrained, then the possibility of model misspecification is likely to increase with undesirable consequences. One approach to addressing the misspecification challenge is to implement a nonparametric modeling methodology in which a very flexible model with many free parameters is used to estimate the data generating process. However, there are several major challenges associated with parameter estimation in flexible probability models.

First, highly flexible models with many parameters require larger sample sizes to ensure reliable and unique parameter estimates. Second, the presence of redundant parameters can lead to overfitting phenomena and larger sampling error. Third,

Richard M. Golden
Cognitive Informatics and Statistics Lab, School of Behavioral and Brain Sciences (GR4.1),
University of Texas at Dallas, Richardson, TX. e-mail: golden@utdallas.edu. This project was
partially funded by The University of Texas at Dallas Office of Research and Innovation to Richard
Golden through the SPARK program.

standard asymptotic statistical theory used to characterize the asymptotic behavior of parameter estimates typically makes strong assumptions that the probability model is not parameter redundant and optimal parameters are locally identifiable.

A variety of methods for parameter redundancy detection (Catchpole and Morgan, 1997; Cole et al., 2010; Ran and Hu, 2017, Theorem 17; Cole, 2020) and parameter redundancy correction (Catchpole et al., 1998; Dasgupta et al., 2007; Ran and Hu, 2017, Theorem 18; Cole, 2020) have been proposed. However, many of these discussions have not focused upon the development of methods for parameter redundancy detection and correction which are valid in the presence of possible model misspecification.

In this paper, we theoretically investigate the asymptotic distribution of maximum likelihood estimates and Maximum A Posteriori (MAP) estimates when parameter redundancy is present without requiring the assumption of correct model specification. When parameter redundancy is present, the asymptotic distribution of the maximum likelihood estimates is not necessarily Gaussian. However, it is shown that the asymptotic distribution of linear combinations of model parameters may still be Gaussian under certain conditions. Some aspects of the resulting mathematical theory are then empirically investigated in a series of simulation studies by evaluating the effectiveness of confidence interval estimation in the presence of parameter redundancy for a Deterministic Input Noisy And (DINA) Cognitive Diagnostic Model fit to an extract of the Tatsuoka (1983) Fraction-Subtraction data set.

2 MAP Estimation Theory for Parameter Redundant Models

The goal of this section is to develop a theory of maximum likelihood estimation and Maximum A Posteriori estimation for probability models which are possibly parameter redundant and possibly misspecified. To simplify the exposition, it is assumed the Data Generating Process is a bounded sequence of discrete random vectors but this assumption can be relaxed.

2.1 Assumptions and Definitions

2.1.1 DGP and Modeling Assumptions

The Data Generating Process (DGP) is a mechanism which generates a data set using the DGP probability mass function (DGP pmf).

Assumption (A1. Data Generating Process) Let Ω be a finite subset of \mathcal{R}^d . The data set $\mathcal{D}_n \equiv [\mathbf{x}_1, \dots, \mathbf{x}_n]$ is a realization of the stochastic sequence $\tilde{\mathcal{D}}_n \equiv [\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n]$ of i.i.d. d -dimensional random vectors with common DGP pmf $p_e : \Omega \rightarrow [0, 1]$. \square

Let Θ be a closed, bounded, subset of \mathcal{R}^q . A *model representation* for a DGP pmf $p_e : \Omega \rightarrow [0, 1]$ is a function $p_\Theta : \Omega \times \Theta \rightarrow [0, 1]$ defined such that $p_\Theta(\mathbf{x}|\theta)$ is the probability of observing \mathbf{x} given some θ . A *probability model*, \mathcal{M}_Θ , with *parameter space* Θ is a collection of probability mass functions defined with respect to a model representation such that $\mathcal{M}_\Theta \equiv \{p_\Theta(\cdot|\theta) : \theta \in \Theta\}$.

Assumption (A2. Probability Model Smoothness) Let Θ be a convex, closed, bounded subset of \mathcal{R}^q . Let $p_\Theta : \Omega \times \Theta \rightarrow [0, 1]$ be a model representation. Assume $\log p_\Theta(\mathbf{x}, \cdot)$ is a twice continuously differentiable function on Θ for each $\mathbf{x} \in \Omega$. \square

2.1.2 MAP and ML Estimation Algorithms

Let the observed data $\mathcal{D}_n \equiv [\mathbf{x}_1, \dots, \mathbf{x}_n]$ be defined as in Assumption A1. Assume a smooth probability model $\mathcal{M}_\Theta = \{p_\Theta(\cdot|\theta) : \theta \in \Theta\}$ as in Assumption A2. Let $p_\theta : \Theta \rightarrow (0, \infty)$ be an absolutely continuous density function which is called the *parameter prior*. The goal of MAP estimation (Maris, 1999; Golden, 2020, Chapter 13) is to compute a parameter estimate $\hat{\theta}_n$ such that:

$$\hat{\theta}_n \equiv \arg \min_{\theta \in \Theta} \hat{\ell}_n(\theta), \quad \hat{\ell}_n(\theta) = -(1/n) \sum_{i=1}^n \log p(\mathbf{x}_i|\theta) - (1/n) \log p_\theta(\theta). \quad (1)$$

When the density p_θ is a uniform density on Θ , then the MAP estimation goal specified in (1) is equivalent to Maximum Likelihood (ML) estimation for each sample size n . Moreover, we make the strong assumption there exists a finite number K such that $|\log p_\theta(\theta)| < K$ so that the second term on the right-hand side of (1) converges to zero as $n \rightarrow \infty$. This latter result implies that the MAP estimation goal in (1) is asymptotically equivalent to ML estimation as $n \rightarrow \infty$.

Given the above assumptions, it follows from the Uniform Law of Large Numbers (e.g., Golden, 2020, Theorem 13.1.1) that as $n \rightarrow \infty$, $\hat{\ell}_n \rightarrow \ell$ uniformly with probability one where

$$\ell(\theta) \equiv - \sum_{\mathbf{x} \in \Omega} p_e(\mathbf{x}) \log p(\mathbf{x}|\theta). \quad (2)$$

When model misspecification is present, a *true parameter value* θ^* defined such that $p_\Theta(\mathbf{x}|\theta^*) = p_e(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ does not exist. However, in the special case where θ^* is a true parameter value, then it can be shown that θ^* will correspond to a global minimum of (2). Therefore, to provide a general framework for estimation and inference in the presence of model misspecification, the goal of the parameter estimation process is formulated as seeking a strict local minimizer of $\ell(\theta)$.

2.1.3 Theorem Assumptions and Notation

The Theorems in the remainder of this paper all assume that the observed data $\mathcal{D}_n \equiv [\mathbf{x}_1, \dots, \mathbf{x}_n]$ is defined as in Assumption A1. In addition, all theorems assume a smooth probability model $\mathcal{M}_\Theta = \{p_\Theta(\cdot|\theta) : \theta \in \Theta\}$ as in Assumption A2. The parameter space Θ is assumed to be a closed, convex, and bounded subset of \mathcal{R}^q which contains a strict local minimizer of $\ell(\theta)$ denoted by θ^* . It is further assumed that θ^* is the unique global minimizer of ℓ on Θ and θ^* is located in the interior of Θ . The notation \log refers to the natural logarithm. Let $\mathbf{g}_\Theta(\mathbf{x}, \theta)$ denote the column-vector valued function which is the gradient of $-\log p_\Theta(\mathbf{x}, \theta)$ with respect to θ .

Define $\mathbf{A}(\mathbf{x}, \theta)$ as the Hessian of $-\log p_\Theta(\mathbf{x}|\theta)$. Let $\bar{\mathbf{g}}_\Theta(\theta) = \sum_{\mathbf{x} \in \Omega} \mathbf{g}_\Theta(\mathbf{x}, \theta) p_e(\mathbf{x})$. The Hessian of $\ell(\theta)$, $\bar{\mathbf{A}}_\Theta$, and Outer-Product Gradient (OPG) matrix, $\bar{\mathbf{B}}_\Theta$, defined with respect to model representation p_Θ are given respectively by:

$$\bar{\mathbf{A}}_\Theta(\theta) = \sum_{\mathbf{x} \in \Omega} \mathbf{A}_\Theta(\mathbf{x}, \theta) p_e(\mathbf{x}) \text{ and } \bar{\mathbf{B}}_\Theta(\theta) = \sum_{\mathbf{x} \in \Omega} \mathbf{g}_\Theta(\mathbf{x}, \theta) \mathbf{g}_\Theta(\mathbf{x}, \theta)^T p_e(\mathbf{x}). \quad (3)$$

Let the notation $\mathbf{A}_\Theta^* \equiv \bar{\mathbf{A}}(\theta^*)$ and $\mathbf{B}_\Theta^* \equiv \bar{\mathbf{B}}(\theta^*)$. Let the notation $\mathbf{g}_\Theta^* \equiv \bar{\mathbf{g}}(\theta^*)$. In practice, \mathbf{A}_Θ^* and \mathbf{B}_Θ^* are not directly observable and are typically estimated respectively by $\hat{\mathbf{A}}_n$ which is the second derivative of $\hat{\ell}_n$ evaluated at $\hat{\theta}_n$ and $\hat{\mathbf{B}}_n \equiv (1/n) \sum_{i=1}^n \mathbf{g}_\Theta(\mathbf{x}_i, \hat{\theta}_n) \mathbf{g}_\Theta(\mathbf{x}_i, \hat{\theta}_n)^T$. The notation $\mathbf{0}_q$ refers to the q -dimensional column vector of zeros. The notation \mathbf{I}_q refers to the q -dimensional identity matrix.

2.1.4 Identifiability and Redundancy Definitions

Definition 1 (Observationally Equivalent (Rothenberg, 1971; Bowden, 1973))

Let $\mathcal{M}_\Theta \equiv \{p_\Theta(\cdot|\theta) : \theta \in \Theta\}$ be a probability model. Two distinct parameter vectors θ_1 and θ_2 in the parameter space Θ of model \mathcal{M}_Θ are said to be *observationally equivalent* if $p_\Theta(\mathbf{x}|\theta_1) = p_\Theta(\mathbf{x}|\theta_2)$ for all $\mathbf{x} \in \Omega$.

Definition 2 (Locally Identifiable (Bowden, 1973; Ran and Hu, 2017))

A parameter vector $\theta^* \in \Theta$ is said to be *globally identifiable* on Θ if there is no point other than θ^* in Θ which is observationally equivalent to θ^* . A parameter vector $\theta^* \in \Theta$ is said to be *locally identifiable* if there exists a neighborhood of θ^* such that no point in that neighborhood other than θ^* is observationally equivalent to θ^* .

Definition 3 (Parameter Redundant Model Representation (Ran and Hu, 2017))

Let $\Theta \subseteq \mathcal{R}^q$. Let $\Psi \subseteq \mathcal{R}^k$. Let $p_\Theta : \Omega \times \Theta \rightarrow [0, 1]$ and $p_\Psi : \Omega \times \Psi \rightarrow [0, 1]$ be two model representations for the same DGP pmf. Assume a continuous function $\check{\psi} : \Theta \rightarrow \Psi$ exists such that $p_\Theta(\mathbf{x}|\theta) = p_\Psi(\mathbf{x}|\check{\psi}(\theta))$ for all $\theta \in \Theta$ and for all $\mathbf{x} \in \Omega$. Then $\check{\psi}$ is called a *reparameterization function*. If the dimension of k is strictly less than q then the model representation p_Θ is called *parameter redundant* on Θ . Let Γ be a set of reparameterization functions. If p_Θ is not parameter redundant for every reparameterization function in Γ , then p_Θ is called a *minimal parameterization* with respect to Γ .

2.2 Theorems

2.2.1 Parameter Redundancy and Identifiability

The following theorem is similar to existing theorems in the literature (e.g., Bowden, 1973; Ran and Hu, 2017, Theorem 5, Theorem 6). The Theorem shows that if \mathbf{A}^* is positive definite, then that is a sufficient condition for θ^* to be locally identifiable.

Definition 4 (Hessian Positive Definiteness Condition) The *Hessian Positive Definiteness Condition* holds with respect to model representation p_Θ and DGP pmf p_e when \mathbf{A}^* is positive definite.

Theorem 1 (Sufficient Local Identifiability Condition) If (i) $\mathbf{g}_\Theta^* = \mathbf{0}_q$, and (ii) the eigenvalues of \mathbf{A}_Θ^* are positive, then θ^* is locally identifiable.

Proof By Theorem 5.3.3 of Golden (2020) (a standard optimization theory result) it follows that from conditions (i) and (ii) that θ^* is a strict local minimizer of $\ell(\theta)$. Since θ^* is a strict local minimizer, there exists a closed, bounded, convex neighborhood of θ^* , \mathcal{N}^* , such that θ^* is the unique global minimizer on \mathcal{N}^* . This implies that:

$$\ell(\theta) - \ell(\theta^*) = - \sum_{\mathbf{x} \in \Omega} p_e(\mathbf{x}) \log \frac{p_\Theta(\mathbf{x}|\theta)}{p_\Theta(\mathbf{x}|\theta^*)} \quad (4)$$

is strictly positive when $\theta \neq \theta^*$. Now suppose that θ^* is not globally identifiable on \mathcal{N}^* . This means that there exists a $\theta^+ \in \mathcal{N}^*$ which is not equal to θ^* such that $p_\Theta(\mathbf{x}|\theta^+) = p_\Theta(\mathbf{x}|\theta^*)$ for all $\mathbf{x} \in \Omega$. But this would mean that $\ell(\theta^+) - \ell(\theta^*) = 0$ from (4) when $\theta \neq \theta^*$ resulting in a contradiction. \square

The following theorem and proof have been previously discussed in the literature (e.g., Catchpole and Morgan, 1997; Ran and Hu, 2017, Theorem 18) but more explicit details of the argument are provided here. The Theorem shows that if \mathbf{B}^* is positive definite, then that is a sufficient condition for the model representation p_Θ to be minimal parameterization on a neighborhood of θ^* .

Definition 5 (OPG Positive Definiteness Condition) The *OPG Positive Definiteness Condition* holds with respect to model representation p_Θ and DGP pmf p_e when \mathbf{B}^* is positive definite.

In practice, \mathbf{B}^* is often estimated by $\hat{\mathbf{B}}_n$. A necessary but not sufficient condition for the q -dimensional matrix $\hat{\mathbf{B}}_n$ to be positive definite is that the number of data points, n , is greater than the number of free parameters q .

Theorem 2 (Sufficient Condition for No Parameter Redundancy) Let \mathcal{N} be a closed, bounded subset of Θ . Assume Ψ is a bounded subset of \mathcal{R}^k where $k \leq q$. Let Γ be the set of all continuously differentiable reparameterization functions with domain \mathcal{N} and range Ψ . If the q -dimensional matrix $\mathbf{B}_\Theta(\theta)$ defined with respect to p_Θ in (3) is positive definite on \mathcal{N} , then p_Θ is a minimal parameterization with respect to Γ .

Proof Suppose there exists a continuously differentiable reparameterization function $\psi : \mathcal{R}^q \rightarrow \mathcal{R}^k$ in Ψ and model representation p_Ψ such that $p_\Theta(\mathbf{x}|\theta) = p_\Psi(\mathbf{x}|\psi(\theta))$ where $k < q$. From the matrix chain rule:

$$-\mathbf{g}_\Theta(\mathbf{x}, \theta)^T = \frac{d \log p_\Theta(\mathbf{x}|\theta)}{d\theta} = \frac{d \log p_\Psi(\mathbf{x}|\psi(\theta))}{d\theta} = \frac{d \log p_\Psi(\mathbf{x}|\psi(\theta))}{d\psi} \frac{d\psi}{d\theta}. \quad (5)$$

Using (5),

$$\mathbf{B}_\Theta(\theta) = E\{\mathbf{g}_\Theta(\tilde{\mathbf{x}}, \theta)\mathbf{g}_\Theta(\tilde{\mathbf{x}}, \theta)^T\} = \left(\frac{d\psi}{d\theta}\right)^T E\{\mathbf{g}_\Psi(\tilde{\mathbf{x}}, \psi)\mathbf{g}_\Psi(\tilde{\mathbf{x}}, \psi)^T\} \frac{d\psi}{d\theta} \quad (6)$$

which has rank $k < q$ and is therefore singular on \mathcal{N} . By a contrapositive argument, this implies that if $\mathbf{B}_\Theta(\theta)$ is positive definite on \mathcal{N} , then p_Θ is a minimal parameterization on Γ . \square

Theorem 2 shows that the OPG positive definiteness condition is sufficient to ensure the absence of parameter redundancy near a strict local minimizer θ^* . Theorem 1 shows that the Hessian positive definiteness condition is sufficient to ensure local identifiability at θ^* . However, the Hessian positive definiteness condition does not necessarily imply the OPG positive definiteness condition holds (e.g., Ran and Hu, 2017; Fox and Golden, 2022). Still, in the special case where the probability model is correctly specified at a strict local minimizer θ^* , then the OPG positive definiteness and Hessian positive definiteness conditions are equivalent.

Theorem 3 (Information Matrix Equality) *If there exists a point θ^* in the interior of Θ such that $p(\mathbf{x}|\theta^*) = p_e(\mathbf{x})$ for all $\mathbf{x} \in \Omega$, then $\mathbf{A}_\Theta^* = \mathbf{B}_\Theta^*$.*

Proof See Golden (2020, Theorem 16.3.1) for a proof. \square

Theorem 4 (Equivalence of Redundancy and Identifiability Conditions) *If there exists a point θ^* in the interior of Θ such that $p(\mathbf{x}|\theta^*) = p_e(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ and either \mathbf{A}_Θ^* or \mathbf{B}_Θ^* is positive definite, then both $\mathbf{A}_\Theta(\theta)$ and $\mathbf{B}_\Theta(\theta)$ are positive definite on a neighborhood of θ^* .*

Proof By assumption and the the Information Matrix Equality Theorem, both $\mathbf{A}_\Theta^* = \mathbf{B}_\Theta^*$ have strictly positive eigenvalues at the point θ^* . Those eigenvalues will also be strictly positive in some neighborhood of θ^* since those eigenvalues are continuous functions of \mathbf{A}_Θ and \mathbf{B}_Θ (Franklin, 1968, p. 191) which are, by definition, continuous functions of θ . \square

In many discussions of parameter redundancy and identifiability in the literature, the conclusion of the Information Matrix Equation Theorem (Theorem 4) is assumed to hold (Bowden, 1973, Equation 7; Dasgupta et al., 2007, Assumption A9; Little et al., 2010; Proof of Theorem 3; Ran and Hu, 2014, Equation 18). However, if the model is misspecified, the conclusion of (Theorem 4) may not hold.

2.2.2 MAP Estimate Asymptotic Distribution

The next theorem investigates the asymptotic normality of the MAP estimates. The statement of the following theorem, unlike usual treatments of the asymptotic distribution of maximum likelihood estimates (e.g., White, 1994, Assumption 3.9, Assumption 6.1, Theorem 6.4; Golden, 2020, Chapter 13, Chapter 15) does not assume that \mathbf{A}_Θ^* and \mathbf{B}_Θ^* are positive definite.

The essential idea is to reparameterize the model with a linear transformation of the parameters using methods similar to those discussed by (Catchpole et al., 1998; Dasgupta et al., 2007; Ran and Hu, 2017). Unlike many of these approaches, we focus on a particular linear parameterization strategy specified by the reparameterization function: $\tilde{\theta}(\psi) = \theta^* + \mathbf{P}(\psi - \psi^*)$ with $\psi^* \equiv \mathbf{P}^T \theta^*$. The $q \times k$ -dimensional projection matrix \mathbf{P} consisting of k orthonormal columns is constructed to ensure the reparameterized model is locally identifiable at θ^* and not parameter redundant in a neighborhood of θ^* . This is achieved by choosing \mathbf{P} such that $\mathbf{A}_\Psi^* \equiv \mathbf{P}^T \mathbf{A}_\Theta^* \mathbf{P}$ and $\mathbf{B}_\Psi^* \equiv \mathbf{P}^T \mathbf{B}_\Theta^* \mathbf{P}$ are both positive definite which respectively imply local identifiability by Theorem 1 and no local redundancy by Theorem 2.

In addition, the approach investigated here differs from many previous approaches in which a reparameterized model is constructed to support inference (e.g., Catchpole et al., 1998; Dasgupta et al., 2007; Ran and Hu, 2017; Cole, 2020) since the results are subsequently projected back into the original parameter space. Consequently, the resulting model remains parameter redundant but now it becomes possible to show that some (but not all) linear combinations of parameter estimates have well-defined asymptotic Gaussian distributions despite the presence of parameter redundancy. Formally, a set of linear combinations of elements of the q -dimensional MAP estimate $\hat{\theta}_n$ is represented by $\mathbf{S}\hat{\theta}_n$ where \mathbf{S} is called the *selection matrix*. If \mathbf{S} is the identity matrix, this corresponds to the standard case where the asymptotic distribution of $\hat{\theta}_n$ is of interest. Choosing \mathbf{S} to be the m th row vector of a q -dimensional identity matrix corresponds to the situation where $\mathbf{S}\hat{\theta}_n$ is the m th element of $\hat{\theta}_n$. It is important to emphasize that $\mathbf{S}\hat{\theta}_n$ will only have an asymptotic multivariate Gaussian distribution for certain choices of \mathbf{S} so it is necessary to check that the covariance matrix of $\mathbf{S}\hat{\theta}_n$ denoted as \mathbf{C}_S^* is positive definite. With this introduction, the following new theorem is introduced.

Theorem 5 (Asymptotic Normality with Parameter Redundancy) *Let N^* be a neighborhood of a strict local minimizer θ^* of $\ell(\theta)$ such that: (i) θ^* is the unique global minimizer in the interior of $N^* \subseteq \Theta$, and (ii) $\hat{\theta}_n$ is a minimizer of $\hat{\ell}_n(\theta)$ on N^* for $n = 1, 2, \dots$. Assume the span of \mathbf{A}_Θ^* and \mathbf{B}_Θ^* is specified by k orthonormal vectors corresponding to the columns of the full column rank matrix $\mathbf{P} \in \mathcal{R}^{q \times k}$ such that $\mathbf{A}_\Psi^* \equiv \mathbf{P}^T \mathbf{A}_\Theta^* \mathbf{P}$ and $\mathbf{B}_\Psi^* \equiv \mathbf{P}^T \mathbf{B}_\Theta^* \mathbf{P}$ are positive definite. Let $\mathbf{C}_\Psi^* \equiv (\mathbf{A}_\Psi^*)^{-1} \mathbf{B}_\Psi^* (\mathbf{A}_\Psi^*)^{-1}$. Let $\mathbf{S} \in \mathcal{R}^{q \times s}$ have positive column rank s . Then the following results hold.*

1. As $n \rightarrow \infty$, $\mathbf{S}^T \hat{\theta}_n \rightarrow \mathbf{S}^T \theta^*$ with probability one.
2. As $n \rightarrow \infty$, $\sqrt{n} \mathbf{S}^T (\hat{\theta}_n - \theta^*)$ converges in distribution to a zero mean Gaussian random vector with covariance matrix $\mathbf{C}_S^* \equiv \mathbf{S}^T \mathbf{P} \mathbf{C}_\Psi^* \mathbf{P}^T \mathbf{S}$ provided $\mathbf{S}^T \mathbf{P}$ has positive row rank s .

Proof

Part 1. Define $\hat{\boldsymbol{\psi}}_n \equiv \mathbf{P}^T \hat{\boldsymbol{\theta}}_n$. Consider the empirical risk function $\hat{\ell}_n^{\Psi} : \Psi \rightarrow \mathcal{R}$ defined such that:

$$\hat{\ell}_n^{\Psi}(\boldsymbol{\psi}) \equiv \hat{\ell}_n(\hat{\boldsymbol{\theta}}(\boldsymbol{\psi})) = -(1/n) \sum_{i=1}^n \log p_{\Theta}(\tilde{\mathbf{x}}_i, \hat{\boldsymbol{\theta}}(\boldsymbol{\psi})).$$

Since $\hat{\boldsymbol{\psi}}_n \rightarrow \boldsymbol{\psi}^*$ as $n \rightarrow \infty$ by standard M-estimation theorems (e.g., White, 1994, Theorem 3.4; Golden, 2020, Theorem 13.1.1) and the definition of the continuous function $\hat{\boldsymbol{\theta}}(\boldsymbol{\psi})$ it follows that: $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\psi}}_n) \rightarrow \boldsymbol{\psi}^* = \boldsymbol{\theta}^*$ as $n \rightarrow \infty$.

Part 2. Since, by assumption, $\mathbf{S}^T \mathbf{P}$ has full positive row rank s and \mathbf{C}_{Ψ}^* is positive definite, it follows that $\sqrt{n} \mathbf{S}^T \mathbf{P}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*)$ converges in distribution to a zero-mean multivariate Gaussian random vector with positive definite covariance matrix $\mathbf{C}_{\mathcal{S}}^*$ using standard M-estimation theorems (e.g., White, 1994, Theorems 6.2, 6.4; Golden, 2020, Theorem 15.2.2). Therefore, $\sqrt{n} \mathbf{S}^T \mathbf{P}(\hat{\boldsymbol{\psi}}_n - \boldsymbol{\psi}^*) = \sqrt{n} \mathbf{S}^T(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)$ converges in distribution to a zero-mean multivariate Gaussian random vector with covariance matrix $\mathbf{C}_{\mathcal{S}}^*$. □

3 Simulation Study

3.1 Methods

3.1.1 Data Set, Model, and Estimation Algorithm

This simulation study used an extraction of the Tatsuoka (1983) Fraction-Subtraction data set consisting of 15 questions, 5 skills, and 536 students. In this study, 200 bootstrap data sets each of sample size N were generated by sampling with replacement from the original data set which consisted of 536 data records. Next, each of the 200 bootstrap data sets was fit to a Reparameterized Deterministic Input gate (RDINA) CDM (DeCarlo, 2011). The RDINA CDM was specified by a known independent attribute Bernoulli distribution (Maris, 1999) and 15 two-parameter item models for each of the 15 questions respectively. A MAP estimation methodology was used that incorporated an uninformative Gaussian prior. The uninformative Gaussian parameter prior was designed to specify the most likely values of a guess or slip probability parameter value to be 0.354 with a variance of $1e+8$.

3.1.2 Evaluation of Confidence Interval Estimation Methods

Four different methods were used to evaluate how confidence intervals were calculated. The **No-Selection/No-Projection Method** computed confidence intervals

Table 1 Percentage of times bootstrap parameter estimates were not included in an averaged 95% confidence interval as a function of bootstrap data set sample size N . Numbers in parentheses indicate number of bootstrap data sets used to estimate percent inclusion. Bold-faced values indicate close agreement with expected theoretical value of **0.05**.

Sample Size (30 parameters)	No Selection No Projection	Selection No Projection	No Selection Projection	Selection Projection
$N = 15$	0.56 (200)	0.58 (188)	0.59 (200)	0.06 (26)
$N = 60$	0.20 (200)	0.20 (197)	0.20 (200)	0.07 (121)
$N = 120$	0.11 (200)	0.10 (191)	0.11 (200)	0.07 (157)
$N = 240$	0.08 (200)	0.08 (192)	0.08 (200)	0.07 (176)
$N = 536$	0.06 (200)	0.06 (194)	0.06 (200)	0.05 (180)

for all of the 200 bootstrap data sets and used the averaged confidence interval as the estimator using the identity matrix as the projection matrix \mathbf{P} . The **Selection/No-Projection method** only included bootstrap data sets for estimating the two-dimensional confidence interval item model covariance matrix \mathbf{C}_S^* which had full numerical rank using the identity matrix as the projection matrix \mathbf{P} . The **No-Selection/Projection Method** computed confidence intervals for all 200 bootstrap data sets using the projection matrix \mathbf{P}_B constructed such that its columns were the k orthonormal eigenvectors of $\hat{\mathbf{B}}_n^{\ominus}$ associated with the non-negligible eigenvalues. The **Selection/Projection Method** only included bootstrap data sets where \mathbf{C}_S^* had full numerical rank using projection matrix \mathbf{P}_B .

Results and Discussion

As shown in Table 1, when the sample size of a bootstrap data set was $N = 536$, the simulation results agreed relatively closely with the expected theoretically expected prediction that 5% of the values would not be included in the average confidence interval. The simulation results also agreed closely with the expected theoretical results for the Selection-Projection Method for all sample sizes including the small sample size case of $N=15$ where the number of data records was half the size of the number of free parameters. For the $N=15$ case, the Selection-Projection Method obtained good confidence interval coverage by strategically identifying 26 out of the 200 bootstrap data sets which contained sufficient information for reliable inference. Without the use of the methods introduced here, however, confidence interval estimation performance was ineffective for smaller sample sizes. In addition, using either the Selection Method or the Projection Method by themselves was also ineffective, both of these methods had to be used in combination to obtain reliable inferences across all sample sizes.

References

1. Bowden, R. The Theory of Parametric Identification. *Econometrica*. **41**, 1069-1074 (1973), <http://www.jstor.org/stable/1914036>
2. Catchpole, E. & Morgan, B. Detecting parameter redundancy. *Biometrika*. **84**, 187-196 (1997), <https://doi.org/10.1093/biomet/84.1.187>
3. Catchpole, E., Morgan, J. & Freeman, S. Estimation in parameter-redundant models. *Biometrika*. **85**, 462-468 (1998), <https://doi.org/10.1093/biomet/85.2.462>
4. Cole, D. Parameter Redundancy and Identifiability. (CRC Press,2020)
5. Cole, D., Morgan, B. & Titterton, D. Determining the parametric structure of models. *Mathematical Biosciences*. **228**, 16-30 (2010), 1879-3134 Cole, D J Morgan, B J T Titterton, D M Journal Article United States 2010/08/31 Math Biosci. 2010 Nov;228(1):16-30. doi: 10.1016/j.mbs.2010.08.004. Epub 2010 Aug 25.
6. Dasgupta, A., Self, S. & Das Gupta, S. Non-identifiable parametric probability models and reparametrization. *Journal Of Statistical Planning And Inference*. **137**, 3380-3393 (2007), <https://www.sciencedirect.com/science/article/pii/S0378375807001036>
7. DeCarlo, L. On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*. **35**, 8-26 (2011)
8. Fox, C. & Golden, R. Regularized robust confidence interval estimation in cognitive diagnostic models. *Quantitative Psychology: The 87th Annual Meeting Of The Psychonomic Society*. **422** pp. 233-242 (2022)
9. Franklin, J. Matrix Theory. (Prentice-Hall, Inc.,1968)
10. Golden, R. Statistical Machine Learning (Texts in Statistical Sciences Series). (Chapman-Hall, CRC Press,2020), <https://www.routledge.com/Statistical-Machine-Learning-A-Unified-Framework/Golden/p/book/9781138484696>
11. Little, M., Heidenreich, W. & Guangquan, L. Parameter identifiability and redundancy: Theoretical considerations. *PloS One* **5.1**. **5** (2010)
12. Maris, E. Estimating multiple classification latent class models. *Psychometrika*. **64**, 187-212 (1999), <https://doi.org/10.1007/BF02294535>
13. Ran, Z. & Hu, B. Determining structural identifiability of parameter learning machines. *Neurocomputing*. **127** pp. 88-97 (2014), <https://www.sciencedirect.com/science/article/pii/S0925231213009612>
14. Ran, Z. & Hu, B. Parameter Identifiability in Statistical Machine Learning: A Review. *Neural Computation*. **29**, 1151-1203 (2017), https://doi.org/10.1162/NECO_a.00947
15. Rothenberg, T. Identification in Parametric Models. *Econometrica*. **39**, 577-591 (1971), <http://www.jstor.org/stable/1913267>
16. Tatsuoka, K. Rule-space. An approach for dealing with misconceptions based on item response theory. *Journal Of Educational Measurement*. **20** pp. 345-354 (1983), <https://www.jstor.org/stable/1434951>
17. White, H. Estimation, inference, and specification analysis. (Cambridge University Press,1994), 92030563 Halbert White. 24 cm. Includes bibliographical references and indexes. Econometric Society monographs ; no. 22.