Global estimation of finite mixture and misclassification models with an application to multiple equilibria

Yingyao Hu[*] Ruli Xiao[†]

January 20, 2020

Abstract

We show that the identification results of finite mixture and misclassification models are equivalent in a widely used scenario except for an extra ordering assumption. In the misclassification model, an ordering condition is imposed to pin down the precise values of the latent variable, which are also of interest to researchers and need to be identified. In contrast, finite mixture models are usually identified up to permutations of a latent index, which results in local identification. This local identification is satisfactory because the latent index does not convey any economic meaning. However, reaching global identification is important for estimation, especially when researchers use bootstrap to estimate standard errors. This is because standard errors approximated by bootstrap may be incorrect without a global estimator. We demonstrate that games with multiple equilibria fit in our framework and the global estimator with ordering conditions provides more reliable estimates.

Keywords: Finite mixture, misclassification, global estimation, identification, bootstrap, multiple equilibria.

[*] Department of Economics, Johns Hopkins University, 3100 Wyman Park Dr, Baltimore, MD 21211; email: y.hu@jhu.edu.

[†] Department of Economics, Indiana University, 100 S Woodlawn Ave, Bloomington, IN 47405; email: rulixiao@iu.edu.

1 Introduction

Mixture structures arise with the presence of a latent variable, which could be a variable measured with error or unobserved heterogeneity of different sources such as heterogeneous preferences, unobserved heterogeneity within/across markets, different types of beliefs, and multiple equilibria in games. Both finite mixture and misclassification models can be reformulated into similar mixture structures and are widely used in economic applications such as labor economics, industrial organization, and so forth. For example, (Keane and Wolpin, 1997) consider unobserved type-specific endowments; (Hu et al., 2013) control for auction-level unobserved heterogeneity; and (Xiao, 2018) controls for the presence of multiple equilibria in games. See (Hu, 2017) for a survey of applications using measurement error and (Compiani and Kitamura, 2016) for a review of finite mixture models.

This paper shows that the identification results of the two models are equivalent in a widely used scenario without an ordering assumption. Note that both literatures of finite mixture and misclassification models recover the unobserved component-specific distributions through joint distributions of observables, but they rely on different conditions. We provide a unified identification result for the two literatures. Specifically, we build a bridge connecting the conditions used in the two literatures and provide an intuitive understanding of those conditions.

A vast literature studies identification and estimation in the two areas. The finite mixture literature initially focuses on identifying the latent distribution from the observed distribution by imposing restrictions on the component distribution. For example, the identification is feasible when the component distribution belongs to a parametric family (Everitt and David, 1981) or is symmetric ((Bordes et al., 2006) and (Hunter et al., 2007)). Arguably, because these restrictions are implausible in empirical applications, the conditional independence assumption was introduced later in the finite mixture model with a multi-covariate observable. Such a setup is equivalent to the long-existing misclassification model with multiple measurements. In that sense, one may either interpret the misclassification model as an example of a finite mixture model, or observe that the finite mixture setup is merging into the misclassification model. More importantly, this connection means that the existing results for misclassification models are also applicable to finite mixture models. Therefore, in the development of this research area, it is important to clarify the connections and to connect the dots, which is where the contributions of this paper lies.

Both literatures share the same prevalent label swapping issue, but they address the issue differently in accordance with their respective interpretations of the latent variable. In particular, since the latent variable in misclassification models usually carries economic implications, additional conditions are imposed to pin down the precise value of the latent variable. In contrast, the unobserved component in finite mixture models does not convey any economic meaning, so precise location of the unobserved component is not necessarily required. Consequently, misclassification models reach global identification while finite mixture models reach local identification.

A problem arises with local identification when researchers attempt to use bootstrap to estimate the standard errors of the estimators. Without an appropriate ordering condition, the estimator would be a local one in the sense that multiple estimators can generate the same values for the chosen criteria function; thus, it is not straightforward which local estimator should be chosen for each bootstrap resampling. The existing literature on finite mixture models has realized the importance and necessity of pinning down the component order when the standard error is estimated through resampling. For instance, (Kasahara and Shimotsu, 2009) propose determining this component ordering by using the marginal distribution of the component to uniquely pin down the order. (Hall et al., 2003) also suggest similar treatment. (Bonhomme et al., 2016) note that the label swapping issue presents a challenge for inference methods based on resampling algorithms such as bootstrap. In line with this literature, we advocate imposing a condition to pin down the order of the latent components by which a global estimator may be obtained, as in misclassification models. To this end, finite mixture models are very similar to misclassification models.

We apply the proposed global estimator to games with multiple equilibria. Games generally admit multiple equilibria, which is sometimes important for explaining various aspects of economic data. Thus, allowing multiple equilibria in game applications is important. Since the labeling of equilibria does not convey any economic meaning, we can label them in any order without affecting the estimation and interpretation of game payoffs. As a result, imposing the ordering condition is harmless, nonrestrictive, and useful in estimation. In our empirical illustration, we investigate radio stations strategically choosing the timing of commercial breaks, wherein having multiple equilibria is important for rationalizing the clustering patterns of commercial timing in the data. We indeed see that imposing the ordering condition improves standard error estimation via bootstrap.

The remainder of this paper is organized as follows. Section 2 lays out the common framework and shows that the identification results of finite mixture and misclassification models are equivalent in a widely used scenario except for an extra ordering assumption. Section 3 proposes a global estimator for the model. Section 4 provides an empirical illustration using games with multiple equilibria. Section 5 concludes.

2 A Common Framework

Both finite mixture and misclassification models can be represented through an equation associating observables with unknowns, as follows:

$$f_X = \sum_T f_{X|T} f_T, \tag{1}$$

where f is a probability density or mass function, X represents the observables in the data, and $T \in \{t_1, t_2, \ldots, t_K\}$ can represent either the unobserved component of the finite mixture model or the latent true variable of the misclassification model. Several studies (see (Hu, 2008) and (Allman et al., 2009)) focus on the case where there are multiple measurements, i.e., $X = \{X_1, X_2, X_3\}$, which satisfy the following conditional independence condition:

$$X_1 \perp X_2 \perp X_3 \,|\, T. \tag{2}$$

This conditional independence assumption leads to the following representation:

$$f_{X_1 X_2 X_3|T} = f_{X_1|T} f_{X_2|T} f_{X_3|T}. \tag{3}$$

For simplicity, we assume that the cardinality of the unobserved component, K, is known and is the same as the cardinality of X_i, $i = 1, 2$. We allow the cardinality of X_3, denoted as Q, to differ from that of T; that is, Q is allowed to be different from K. The following identification argument applies as long as X_3 provides some variation, i.e., $Q \geq 2$.

Identification of Finite Mixture models In the finite mixture model, the unobserved component is finite, while the observables in the data can be discrete or continuous. Identification is similar for both continuous and finite observable scenarios by using a three-way array and relying on a rank condition. For example, (Allman et al., 2009) follow the fundamental algebraic result in (Kruskal, 1977) to provide conditions for identifying the mixture structures. In particular, in the scenario where X_i has finite state space,[1] they first define a three-dimensional array (tensor) $[\tilde{M_1}, M_2, M_3]$ whose (u, v, w) element is

$$\begin{aligned}
[\tilde{M_1}, M_2, M_3]_{u,v,w} &\equiv \sum_j \pi_j p_j^1(u) p_j^2(v) p_j^3(w) \\
&= \Pr(X_1 = u, X_2 = v, X_3 = w),
\end{aligned}$$

where M_i, for $i = 1, 2$, is of size $K \times K$ with the jth row defined as $p_j^i = \Pr(X_i = \cdot | T = t_j)$. That is,

$$M_i \equiv \begin{pmatrix} p_1^i \\ p_2^i \\ ... \\ p_K^i \end{pmatrix} \equiv \begin{pmatrix} f_{X_i|T}(t_1|t_1) & f_{X_i|T}(t_2|t_1) & ... & f_{X_i|T}(t_K|t_1) \\ f_{X_i|T}(t_1|t_2) & f_{X_i|T}(t_2|t_2) & ... & f_{X_i|T}(t_K|t_2) \\ ... & ... & ... & ... \\ f_{X_i|T}(t_1|t_K) & f_{X_i|T}(t_2|t_K) & ... & f_{X_i|T}(t_K|t_K) \end{pmatrix}.$$

M_3, defined analogously, is of size $K \times Q$. π is the marginal probability distribution of T such that $\pi \equiv (\pi_j) \in (0, 1)^K$ with $\sum_j \pi_j = 1$, and $\tilde{M_1} \equiv diag(\pi) M_1$. Therefore the identification boils down to whether we can recover M_is and π using information on tensor $[\tilde{M_1}, M_2, M_3]$. Note that $[\tilde{M_1}, M_2, M_3]$ is invariant to simultaneously permuting the rows of all M_is and π. Thus, the identification is subject to the label swapping problem.

The identification relies on a rank condition associated with the matrix's Kruskal rank, defined as the largest number I such that every set of I rows of the matrix are linearly independent (Kruskal, 1977). Naturally, the Kruskal rank of matrix M is never larger than the regular rank, i.e., $rank_K(M) \leq rank(M)$. Moreover, if matrix M is full row rank, its Kruskal rank is the same as its regular rank, i.e., $rank_K(M) = rank(M)$. We summarize the identification result in (Allman et al., 2009) (Corollary 2) in the following theorem.

Theorem 1 *(Allman et al., 2009) Consider the model described above. Suppose all entries of π are positive. For each $i = 1, 2, 3$, let $I_i = rank_K(M_i)$. If*

$$I_1 + I_2 + I_3 \geq 2K + 2, \tag{4}$$

the tensor $[\tilde{M_1}, M_2, M_3]$ uniquely determines M_1, M_2, M_3, and π, up to label swapping.

[1] Note that (Allman et al., 2009) do not assume that the X_is are identically distributed conditional on the true T or have the same state space. For illustration purpose, we assume that X_1 and X_2 have the same state space but are not necessary identically distributed, conditional on the true T.

That is, the M_is and π are identified up to a permutation of its support $\{t_1, t_2, \ldots, t_K\}$. Since these unobserved components do not convey any economics meaning, the finite mixture literature does not impose additional assumptions for pinning down their orders.

Note that the Kruskal rank condition required for the identification in Theorem 1 is less restrictive than a full row rank condition. It does not require the M_is to be full rank. However, it does not provide a closed-form expression for the identified components. Consequently, we cannot follow the identification procedure to recover the identified mixture components. As a matter of fact, the identification using the Kruskal rank condition is comparable to a traditional identification argument that a local identification is feasible if the number of restrictions is larger than or equal to that of unknowns. (Allman et al., 2009) further apply this identification result to the scenario of continuous measurement X_i, where the mixture structure of Equation (1) also applies to the corresponding cumulative density function.

Identification of Misclassification Models In the misclassification or the measurement error literature, T represents the latent true variable so it conveys economic meaning itself. For example, in the literature on the returns to education, self-reported education levels may contain measurement error such as those who have not gone to college may report that they have college degrees. In this case, T represents different levels of education. Consequently, pinning down the precise value of T is very important as to evaluate the returns to education.

For identification, we first introduce the following matrix representation.

$$\begin{aligned}
M_i &= \left[\Pr(X_i = t_k | T = t_j)\right]_{j,k}, \qquad i = 1, 2 \\
A(x_3) &\equiv \left[\Pr(X_1 = t_j, X_2 = t_k, X_3 = x_3)\right]_{j,k}, \\
A &\equiv \left[\Pr(X_1 = t_j, X_2 = t_k)\right]_{j,k}, \\
\Omega &\equiv diag\big(\pi_1, ..., \pi_K\big), \\
D(x_3) &\equiv diag\big(Pr(X_3 = x_3 | T = t_1), ..., \Pr(X_3 = x_3 | T = t_K)\big).
\end{aligned}$$

We have the following two matrix representations:

$$\begin{aligned}
A &= M_1^T \Omega M_2, \\
A(x_3) &= M_1^T D(x_3) \Omega M_2.
\end{aligned}$$

With the full rank condition, i.e., M_1 and M_2 are invertible, we have the following key equation summarizing the connection between observables and unknowns.

$$A(x_3)A^{-1} \;\;=\;\; M_1^T D(x_3)(M_1^T)^{-1}. \tag{5}$$

The eigenvalue-eigenvector representation in Equation (5) holds for any value of x_3, and the eigenvector matrix does not change with x_3. Consequently, we can construct a similar eigenvalue-eigenvector expression through aggregating the information of Equation (5) associated with different values of x_3 using some function $\omega(\cdot)$. Specifically, the eigenvalue-eigenvector decomposition still holds with eigenvalue matrix $D(x_3)$ being replaced with the following eigenvalue matrix:

$$D(\omega) = diag\big(E[\omega(X_3)|T = t_1], ..., E[\omega(X_3)|T = t_K]\big),$$

leading to the following eigenvalue-eigenvector expression

$$A(\omega)A^{-1} = M_1^T D(\omega)(M_1^T)^{-1}, \tag{6}$$

where $A(\omega) \equiv \big[E[\omega(X_3)|X_1 = t_j, X_2 = t_k] \Pr(X_1 = t_j, X_2 = t_k)\big]_{j,k}$.

We impose the following condition for the decomposition to be unique.

Assumption 1 *(Distinct eigenvalues) there exist a function $\omega(.)$ such that,*

$$E[\omega(X_3)|T = t_j] \neq E[\omega(X_3)|T = t_k],$$

for any $t_j \neq t_k$.

Intuitively, the conditional distribution of X_3 needs to vary across different values of the latent variable in order to distinguish among them. The introduction of this function is to capture these differences. The function $\omega(\cdot)$ maybe be context-specific, and we impose no functional restrictions on it. Possible examples of this function are: $\omega(X_3) = X_3$; $\omega(X_3) = a(X_3 - b)$, where a and b are some constants; and $\omega(X_3) = I(X_3 = x_3)$, where $I(.)$ is the indicator function. When X_3 is binary, we can further represent the eigenvalue as $E[\omega(X_3)|T = t_j] = \omega(0) + (\omega(1) - \omega(0)) \Pr(X_3 = 1|T = t_j)$, so any function satisfying $\omega(0) \neq \omega(1)$ works. Assumption 1 guarantees that the eigenvalues $E[\omega(X_3)|T = t_j]$ in $D(\omega)$ are distinct. Therefore, the eigenvector matrix M_1 corresponding to each t_j is uniquely identified up to the label t_j.

To address the issue of identification up to relabeling, we provide a set of flexible ordering conditions to pin down the value of the latent true variable. Note that the ordering for eigenvalues and eigenvectors are always consistent, indicating that we can either impose conditions to pin down the ordering of the eigenvalues or that of the eigenvectors.

Assumption 2 *(Ordering) one of the following conditions holds:*

1) there exists t_i such that $f_{X_3|T}(t_i|t)$ is increasing or decreasing in t;

2) there exists a function $\omega(\cdot)$ such that $E[\omega(X_3)|T = t]$ is increasing in t;

3) there exists t_i such that $f_{X_1|T}(t_i|t)$ is decreasing in t;

4) $f_{X_1|T}(\cdot|t)$ has a unique mode at t;

5) $f_{X_1|T}(\cdot|t)$ has a median (min, max, or a known quantile) at t.

Condition 2.1-2.2 are imposed to pin down the order of the eigenvalues while condition 2.3 - 2.5 are imposed to pin down the order of the eigenvectors. Conditions 2.1-2.4 are consistent with Assumptions 2.4-2.7 in (Hu, 2008). Condition 2.5 is a general version of condition 2.4. Understanding the empirical contexts and the economic theory behind the study is very important as to adopt the appropriate ordering condition to pin down the order. We use a few specific empirical examples below to explain these ordering conditions.

1. For studying the impact of education on labor supply, where X_3 is an indicator for whether or not the individual has a job, and X_1 and X_2 are two measurements of the true latent education T, which can come from different waves of the survey, respectively. Economic theory tends to predict that the proportion of people with higher education are more likely to be working than lower educated people. That is, $f_{X_3|T}(1|t)$ is increasing in t (condition 1).

2. For studying returns to education, one intuitive condition we can use to pin down the order of the latent education level is that mean income increases with true education, i.e., $E(X_3|T = t)$ is increasing in t. Condition 2 is satisfied with $\omega(X_3) = X_3$.

3. In the context of latent education with self-reported information, existing validation studies find evidence that the probability of reporting the true education is higher than that of reporting any other values (condition 4). For instance, (Kane et al., 1999) report the proportion of differential educational attainment by self-reported and transcript-reported sources (Table 1). The probability of "correct" self-reported

education attainment is higher than that of the mis-reported one, if we assume that transcripts provide the "true" education attainment.

We summarize the global identification result of misclassification models as follows:

Theorem 2 *(Hu, 2008) Consider a structure described as Equation (1). Suppose that matrix M_1 and M_2 have full rank, and that Assumptions 1 and 2 are satisfied. Then, M_1, M_2, M_3, and π are uniquely identified.*

We next prove the equivalence of the identification condition from the two strands of literature for a general discrete X_3. We first present a set of sufficient and necessary conditions for identification of the two models. We then provide detailed discussions. Note that we focus on the case where M_1 and M_2 have full rank. This important assumption enables us to compare the identification conditions from the two literature. Our main results are summarized as follows:

Theorem 3 *Consider a structure described as Equation (1). Suppose that M_1 and M_2 have full rank, i.e., $I_1 = I_2 = K$. The following four statements are equivalent:*

1. *(Nontrivial Kruskal rank) The Kruskal rank of M_3 is at least 2, i.e., $I_3 \geq 2$;*

2. *(Distinct eigenvalues) Assumption 1 holds. That is, there exists a function $\omega(.)$ such that, for any $t_j \neq t_k$,*

$$E[\omega(X_3)|T = t_j] \neq E[\omega(X_3)|T = t_k].$$

3. *(Non-redundant) For any $t_j \neq t_k$, there exist an x_3 such that*

$$\Pr(X_3 = x_3|T = t_j) \neq \Pr(X_3 = x_3|T = t_k).$$

   Note that x_3 is chosen separately for each pair (t_j, t_k).

4. *(Unique) M_1, M_2, M_3, and π are uniquely identified, up to label swapping.*

Theorem 4 states that the Kruskal rank condition is not only the sufficient condition for identification (Allman et al., 2009), but also the necessary condition; the distinct eigenvalue condition is not only the sufficient condition for identification (Hu, 2008), but also the necessary condition. It is intuitively hard to understand the identification condition in both literatures. It is also difficult to prove the necessity of these identification conditions directly. Thus, we introduce the non-redundant condition, which provides a transparent explanation for the nontrivial Kruskal rank condition and the distinct eigenvalues condition. This non-redundant condition is empirically testable as these components can be estimated directly from the eigen-decomposition in Equation 5. We prove Theorem 3 in the following procedures: the equivalence between the Kruskal rank condition and the non-redundant condition, the equivalence between the distinct eigenvalue condition and the non-redundant condition, and the equivalence between the non-redundant condition and identification.

Note that the non-redundant condition is important as it builds a bridge between the two sets of identification results. This condition indicates that, for any two rows of the M_3 matrix, there exists at least one column, i.e., a value x_3, wherein the probabilities in the two rows differ from each other. It rules out the scenario where two rows of matrix M_3 are the same, which means that the information associated with the two levels of the latent variable is redundant. However, it does not require that there exists an x_3 such that all elements in this column differ from each other; that is, there exists an x_3 such that, for any $t_j \neq t_k$, $\Pr(X_3 = x_3|T = t_j) \neq \Pr(X_3 = x_3|T = t_k)$.

We illustrate the non-redundant condition with two examples. For illustration purposes, assume that $Q = K = 3$ and the matrix M_3 is specified as follows.

$$M_3^1 = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.2 & 0.3 & 0.5 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}; \qquad M_3^2 = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.2 & 0.3 & 0.5 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

Matrix M_3^1 satisfies the non-redundant condition because, for pair $T = t_1$ and $T = t_2$, there exits $X_3 = t_1$ such that $f_{X_3|T}(t_1|t_1) = 0.1 \neq 0.2 = f_{X_3|T}(t_1|t_2)$; for pair $T = t_1$ and $T = t_3$, there exits $X_3 = t_2$ such that $f_{X_3|T}(t_2|t_1) = 0.4 \neq 0.3 = f_{X_3|T}(t_2|t_3)$; for pair $T = t_2$ and $T = t_3$, there exits $X_3 = t_1$ such that $f_{X_3|T}(t_1|t_2) = 0.2 \neq 0.1 = f_{X_3|T}(t_1|t_3)$. Matrix M_3^2 does not satisfy the non-redundant condition because, for pair $T = t_1$ and $T = t_2$, and for any value x_3, $f_{X_3|T}(t_1|t_1) = f_{X_3|T}(t_1|t_2)$. The information provided by $T = t_1$ and $T = t_2$ is redundant so that it is impossible to disentangle t_1 and t_2.

It is important to compare the identification results for the two models. First, we focus on the case where M_1 and M_2 are full rank, while (Allman et al., 2009) provide identification results for a more general setting, which includes the case where each measurement has support smaller than that of the latent variable. For example, when the latent true variable has five possible values, identification is still feasible even if each of the three measurements can only take four possible values. This scenario, however, is not considered in (Hu, 2008) or the current paper. Second, conditions in (Allman et al., 2009) are based on the abstract Kruskal rank, which may be difficult to test. In contrast, the regular rank condition in the current paper is directly testable from the data. Last, it is not clear how to extend the Kruskal rank condition to the case of continuous latent variables while the regular rank condition in (Hu, 2008) can be intuitively extended to the injectivity condition of the continuous case as in (Hu and Schennach, 2008).

With the equivalence of the two identification conditions, the only difference between the two literatures lies in the ordering condition. It seems that the identification condition in the misclassification literature is more restrictive than that of the finite mixture models. However, the additional condition (Assumption 2) is innocuous and very intuitive. It transforms a local identification into a global identification, which is very helpful in estimation. Thus, to address the problem caused by local identification in estimation of finite mixture models, one should impose some version of the ordering condition as Assumption 2 in the estimation.

If the latent true variable contains economic meaning, we need to be careful when introducing the ordering conditions, as those conditions may impose restrictions to the economic models. However, if the latent true variable contains no economic meaning, which is the case in finite mixture models, it is very flexible to introducing ordering conditions. In an example of an entry game with two equilibria, the indexing of equilibria has no economic meaning. One natural ordering condition is that the probability of entry associated with equilibrium 1 is higher than that of equilibrium 2. This would not impose any extra restrictions to the payoff because the equilibrium labels have no economic meaning.

3 Global Estimation

With the model identified, one can use a minimum Hellinger distance estimator (MHD), defined by minimizing the distance of the joint distribution directly from the data $\hat{f}$ and predicted by the models f, respectively. The MHD estimator for finite mixture models can be represented as:

$$\big(f_{M_1|T}, f_{M_2|T}, f_{M_3|T}, f_T\big) = \arg \min_{f_{M_1|T}, f_{M_2|T}, f_{M_3|T}, f_T} \|\hat{f}^{1/2}_{M_1M_2M_3} - (\Sigma f_{M_1|T} f_{M_2|T} f_{M_3|T} f_T)^{1/2}\|, \tag{7}$$

and the MHD estimator for misclassification models can be represented as:

$$\begin{aligned}\big(f_{M_1|T}, f_{M_2|T}, f_{M_3|T}, f_T\big) &= \arg \min_{f_{M_1|T}, f_{M_2|T}, f_{M_3|T}, f_T} \|\hat{f}^{1/2}_{M_1M_2M_3} - (\Sigma f_{M_1|T} f_{M_2|T} f_{M_3|T} f_T)^{1/2}\|, \\ & s.t. \;\; \text{Assumption 2 holds,} \end{aligned} \tag{8}$$

where $\| \cdot \|$ represents the L_2 norm. Since the finite mixture model is identified up to a permutation of T the estimator is a local estimation in the sense that there are $K!$ minima of the criterion function and these minima all lead to the same value for the chosen criteria function. The estimator for the misclassification model is a global one because it directly pins down which minima is the correct one. This may not seem to be a problem because the permutations of the T types do not matter economically. However, such identification up to a permutation makes the bootstrap method invalid because it is unclear which local minimum the estimator reaches in each bootstrap draw. Therefore, we argue that it is still better off to impose Assumption 2 in the estimation of the finite mixture model, i.e., treating it as a misclassification model.

This label swapping issue is a problem for more than just the minimum distance estimation. It is a prevalent problem due to the identification strategy, and thereby affects every estimator. Some may argue that we do not need to worry about this problem if one can derive the variance-covariance matrix for the estimator theoretically. However, some applications, especially applications such as dynamic discrete choice models or dynamic games, rely heavily on a sequential estimation approach to estimate structural parameters while also requiring controls for unobserved heterogeneity. In those applications, deriving the variance matrix is very challenging and may be infeasible. Thus, bootstrap is a popular alternative for standard deviation approximation. The label swapping issue again causes similar problems in these scenarios.

4 Empirical Illustration

This section illustrates the importance of the ordering condition in the estimation of games with multiple equilibria and/or unobserved market-level factors. The application uses a simultaneous move game to characterize the timing decisions for broadcasting commercials by radio stations with contemporary music formats (Contemporary Hit Radio (CHR)/Top 40, Country, Rock, etc.).

4.1 Data and Model

Radio listeners seek to avoid commercials by switching to other stations or opting out with alternatives, such as tapes or CDs. If stations air commercials at the same time, listeners would not be able to skip them, which would be beneficial for advertisers. However, hosting commercial breaks at the same time may not be optimal for radio stations because it may push some listeners to opt out, which might harm the station's overall popularity, measured by averaging audience tune-ins over both commercial and noncommercial programming.

In reality, stations indeed tend to cluster commercials timings (figure 1). There are three peaks in the distributions of commercial timing, which are measured by the average proportion of stations playing commercials in each minute during two different hours of the day. Moreover, the peak pattern varies across markets (Figure 2) even though Arbitron uses the same methodology to compute listenership. One explanation is that some time slots in each hour are particularly desirable for commercials but these desirable time slots differ in different markets. That is, some market-level factors are driving the observation that different markets display different peaks. Another possible explanation is the presence of multiple equilibria. Stations coordinate to take commercial breaks at the same time to avoid listener switching, and different markets coordinate at different times, indicating that they employ different equilibria. Both unobserved market-level factors and multiple equilibria rationalize the clustering pattern in general and the different clustering patterns across markets. In this empirical illustration, we assume away market-level unobserved factors and assume that the presence of multiple equilibria is the only rationale for the data pattern.

To study radio station decisions regarding the timing of commercials, we model the decision process in every hour as choosing from two time blocks to air their commercials simultaneously, as in (Sweeting, 2009) and (Xiao, 2018). Specifically, we construct two exclusive time intervals in every hour using the clustering time peaks and refer to them as option 0 (:48-:52) and option 1 (:53-:57), respectively. The data used in this paper are constructed using hourly airplay logs collected by Medabase 24/7 and extracted from airplay logs that stations play on a minute-by-minute basis[2]. In summary, there are 144 markets in total; the number of stations in each market varies from 3 to 15 with a mean of 5.7; each station has 236 observations over the course of 59 days (Table 1).

In every hour, each station's decision is to choose between the two options to air their commercials. We rule out the possibility of airing commercials in both intervals. For illustrative purposes, we assume that markets differ only in the number of radio stations and that stations are homogeneous so there is not need to keep track of radio station identity. An individual station i's payoff for placing a commercial in time block $t \in \{0, 1\}$ is defined as follows:

$$\begin{aligned} \pi(a_i = 1, a_{-i}) &= \alpha + \delta \frac{\sum_{j \neq i} I(a_j = 1)}{n - 1} + \epsilon_{i1}, \\ \pi(a_i = 0, a_{-i}) &= \delta \frac{\sum_{j \neq i} I(a_j = 0)}{n - 1} + \epsilon_{i0}, \end{aligned}$$

where α captures the gap of the average profit of airing commercials between timing 0 and 1, δ captures the coordination incentives, and ϵs represent station i's idiosyncratic private profit shocks. The ϵs captures the potential that a station may air commercials at different times every day. This introduces variation due to the length of other programming, such as songs or travel news, and can be unpredictable. We assume ϵ_{it} to be independent across actions, players, and markets. Furthermore, ϵ_{it} follows a type-I extreme value distribution.

Following the existing literature, we use the probability that firms choose time slot 0 and 1, denoted as p_0 and p_1, respectively, to characterize the equilibrium. We focus on symmetric equilibria. The equilibrium condition then can be characterized as:

$$\begin{aligned} p_1 &= \int I(\alpha + \delta p_1 + \epsilon_{i1} > \delta p_0 + \epsilon_{i0}) dF(\epsilon_i) \\ &= \frac{exp(\alpha + \delta p_1)}{exp(\delta p_0) + exp(\alpha + \delta p_1)}. \end{aligned} \tag{9}$$

The first equality holds because, radio stations are homogeneous and they have correct beliefs regarding rivals' behaviors in equilibrium. Specifically, the fraction of rivals airing commercials in time slot 1 is consistent with the probability of each radio station airing commercials in time slot 1, i.e., $\frac{\sum_{j \neq i} I(a_j = 1)}{n-1} = p_1$. This feature indicates that each station's strategic decision regarding commercial timing is not affected by the number of rivals in

[2] We thank Andew Sweeting for sharing the data. Please refer to (Sweeting, 2009) for a detailed description of the data.

the market. As a result, we can pool data from markets with different numbers of radio stations for estimation. The second equality holds due to the assumption that ϵ_i follows a type-I extreme value distribution. We further rewrite the equilibrium condition as

$$\log p_1 - \log p_0 \quad = \quad \alpha + \delta(p_1 - p_0). \tag{10}$$

If the data are generated by a single equilibrium, despite the potential existence of multiple equilibria in theory, we can consistently estimate the equilibrium p_1 and p_0 using the fraction of radio stations airing commercials in the two time slots in the data. We then can estimate an individual radio station's payoff using the equilibrium condition characterized above. However, when the data are generated by multiple equilibria, the fraction of radio stations airing commercials in the two time slots in the data are a finite mixture of the equilibrium counterparts. That is,

$$\Pr(a_1, a_2, .., a_n) \quad = \quad \sum_k \lambda(k) \Pr(a_1, a_2, .., a_n|k), \tag{11}$$

where k indexes the equilibrium, $\lambda(k)$ is the proportion of markets adopting equilibrium k, i.e., the equilibrium selection probability, and $\Pr(a_1, a_2, .., a_n|k)$ denotes the joint probability of airing commercials, $\{a_1, a_2, .., a_n\}$, associated with equilibrium k. This finite mixture structure fits our identification framework exactly, so it serves well as an illustration. Thus, we use the identification result described above to identify the probability of airing commercials in time slots 0 and 1 associated with different equilibria. With the equilibrium probability being identified, the payoff parameters can be identified using the equilibrium condition (Xiao, 2018).

4.2 Estimation and Results

Note that the number of equilibria is unknown and needs to be identified and estimated.[3] In this empirical illustration, we use the result from (Xiao, 2018) that the number of equilibria is estimated to be two. We then estimate the equilibrium probabilities using the proposed minimum distance estimation. To illustrate the problem of local estimators, we estimate the probability of airing commercials in time slot 1 associated with the two equilibria with and without imposing an ordering condition. We pool markets with different numbers of radio stations for estimation and use markets with at least three radio stations for estimation.

[3] Please refer to (Xiao, 2018) for identifying and estimating the number of equilibria.

Since the labeling of equilibria does not convey any economic meaning, we can arbitrarily label one equilibrium as equilibrium 1, denoted as $eqm1$, and the other as equilibrium 2, denoted as $eqm2$. One natural ordering we can impose in this context can be that the probability of airing commercial in time slot 1 is greater in equilibrium 1 than in equilibrium 2; that is, $p_1(eqm1) > p_1(eqm2)$. As long as the probability of airing commercials in time slot 1 differs for both equilibria, we can label the one with a higher probability as $eqm1$ and the one with a lower probability as $eqm2$. We should emphasize that this ordering condition imposes no additional restrictions on payoff primitives as both equilibrium probabilities satisfy the same equilibrium condition. That is,

$$\begin{aligned}\log p_1(eqm1) - \log p_0(eqm1) &= \alpha + \delta\left[p_1(eqm1) - p_0(eqm1)\right] \\ \log p_1(eqm2) - \log p_0(eqm2) &= \alpha + \delta\left[p_1(eqm2) - p_0(eqm2)\right].\end{aligned}$$

We present the estimation results regarding the equilibrium probability of airing commercials associated with both equilibria in Table 2. The standard deviation estimated from imposing the ordering condition is significantly smaller than that without the ordering condition. The estimation of payoff primitives are the same with or without imposing the ordering condition. We skip the estimation of the payoff primitives here since it is not the focus of this paper.

5 Conclusion

This paper connects the identification results of finite mixture models and misclassification models in a widely used scenario in empirical research. While existing studies provide sufficient identification conditions for a more general case, we present sufficient and necessary conditions for the identification of this widely used case. In the misclassification model, an ordering condition is usually imposed to pin down the precise value of the latent variable, which are also of interest to researchers and need to be identified. In contrast, the identification of finite mixture models is usually up to label swapping. We argue that the ordering condition in misclassification models leads to global identification and should be imposed in estimation, especially, when researchers use bootstrap to estimate standard errors. As an empirical illustration, games with multiple equilibria fit in our framework well, and we show that the global estimator with ordering assumptions provides reliable estimates with real data.

References

Allman, Elizabeth S, Catherine Matias, and John A Rhodes, "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, 2009, pp. 3099–3132. 2, 2, 1, 1, 2, 2, B

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin, "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016, *78* (1), 211–229. 1

Bordes, Laurent, Stéphane Mottelet, Pierre Vandekerkhove et al., "Semiparametric estimation of a two-component mixture model," *The Annals of Statistics*, 2006, *34* (3), 1204–1232. 1

Compiani, Giovanni and Yuichi Kitamura, "Using mixtures in econometric models: a brief review and some new results," *The Econometrics Journal*, 2016, *19* (3). 1

Everitt, Brian S and J Hand David, "Finite mixture distributions," in "Monogr. Appl. Probab. Stat.," Chapman Hall, 1981. 1

Hall, Peter, Xiao-Hua Zhou et al., "Nonparametric estimation of component distributions in a multivariate mixture," *The annals of statistics*, 2003, *31* (1), 201–224. 1

Hu, Yingyao, "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 2008, *144* (1), 27–61. 2, 2, 2, 2, B

_ , "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of Econometrics*, 2017, *200* (2), 154–168. 1

_ **and Susanne M Schennach**, "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 2008, *76* (1), 195–216. 2

_ **, D McAdams, and M Shum**, "Nonparametric identification of first-price auctions with non-separable unobserved heterogeneity," *Journal of Econometrics*, 2013, *174* (2), 186–193. 1

Hunter, David R, Shaoli Wang, and Thomas P Hettmansperger, "Inference for mixtures of symmetric distributions," *The Annals of Statistics*, 2007, pp. 224–251. 1

Kane, Thomas J, Cecilia Elena Rouse, and Douglas Staiger, "Estimating returns to schooling when schooling is misreported," Technical Report, National Bureau of Economic Research 1999. 3

Kasahara, Hiroyuki and Katsumi Shimotsu, "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 2009, *77* (1), 135–175. 1

Keane, Michael P and Kenneth I Wolpin, "The career decisions of young men," *Journal of political Economy*, 1997, *105* (3), 473–522. 1

Kruskal, Joseph B, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, 1977, *18* (2), 95–138. 2

Sweeting, A., "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria," *The RAND Journal of Economics*, 2009, *40* (4), 710–742. 4.1, 2, 1, 2

Xiao, Ruli, "Identification and estimation of incomplete information games with multiple equilibria," *Journal of Econometrics*, 2018, *203* (2), 328–343. 1, 4.1, 4.1, 4.2, 3

Table 1: Descriptive Statistics

Variable	Obs	Mean	Std. Dev	Min	Max
No. of Players	92766	5.641	2.054	3	15
Timing	92766	.499	.489	0	1
Day	92766	31.745	17.723	1	59

Table 2: Estimation of eqm Strategy

	Ordering		No Ordering	
	estimates	std (bootstrap)	estimates	std (bootstrap)
p_1 (eqm1)	0.602	0.065	0.602	0.120
p_1 (eqm2)	0.420	0.056	0.420	0.123
λ(Prob of eqm1)	0.451	0.229	0.451	0.232

Appendix

A Graphs and Tables

Figure 1: Timing Patterns for Commercials across Markets (Sweeting (2009))

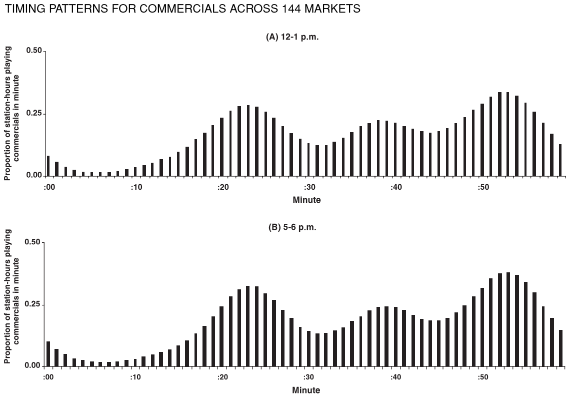
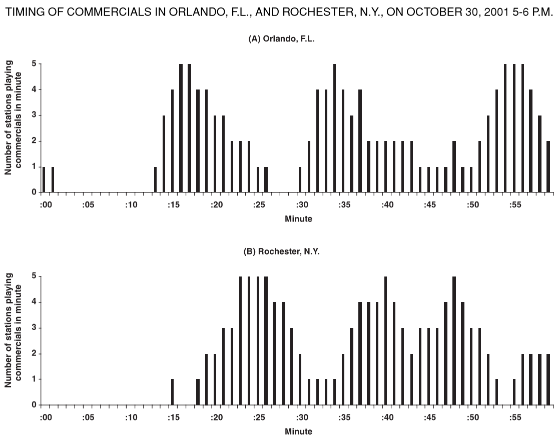


Figure 2: Timing Patterns for Commercials in Different Markets (Sweeting (2009))



TIMING OF COMMERCIALS IN ORLANDO, F.L., AND ROCHESTER, N.Y., ON OCTOBER 30, 2001 5-6 P.M.

(A) Orlando, F.L.

(B) Rochester, N.Y.

B Proof

Proof of Theorem 3 We show that the four statements are equivalent in the following three steps.

Step 1: We show that the Kruskal rank of M_3 is at least 2 if and only if, for any $t_j \neq t_k$, there exists an $x_3 \in \{t_1, t_2, ..., t_Q\}$ such that $f_{X_3|T}(x_3|t_j) - f_{X_3|T}(x_3|t_k) \neq 0$. First, we show that the non-redundant condition implies the Kruskal rank condition. Recall that M_3 is defined as

$$M_3 \quad = \quad \begin{pmatrix} f_{X_3|T}(t_1|t_1) & f_{X_3|T}(t_2|t_1) & ... & f_{X_3|T}(t_Q|t_1) \\ f_{X_3|T}(t_1|t_2) & f_{X_3|T}(t_2|t_2) & ... & f_{X_3|T}(t_Q|t_2) \\ ... & ... & ... & ... \\ f_{X_3|T}(t_1|t_K) & f_{X_3|T}(t_2|t_K) & ... & f_{X_3|T}(t_Q|t_K) \end{pmatrix}.$$

We consider the following reduced-size matrix of dimension $2 \times K$ constructed by any two rows t_j and t_k of matrix M_3, with $t_j \neq t_k$,

$$M_{3.2} \equiv \begin{pmatrix} f_{X_3|T}(t_1|t_j) & f_{X_3|T}(t_2|t_j) & ... & f_{X_3|T}(t_Q|t_j) \\ f_{X_3|T}(t_1|t_k) & f_{X_3|T}(t_2|t_k) & ... & f_{X_3|T}(t_Q|t_k) \end{pmatrix}.$$

Without loss of generality, let $x_3 = t_m, m = 1, ..., Q$, such that $f_{X_3|T}(t_m|t_j) - f_{X_3|T}(t_m|t_k) \neq 0$. Define $\mathbf{1} = (1, 1, ..., 1)^T$ and $e_m = (0, ..., 0, 1, 0, ..., 0)^T$, where 1 is at the m-th coordinate. We consider

$$M_{3.2} \times (e_m \quad \mathbf{1}) = \begin{pmatrix} f_{X_3|T}(t_m|t_j) & 1 \\ f_{X_3|T}(t_m|t_k) & 1 \end{pmatrix}.$$

Therefore, the rank of $M_{3.2}$ is 2 if $f_{X_3|T}(x_3|t_j) - f_{X_3|T}(x_3|t_k) \neq 0$ for any two rows. That means the Kruskal rank of M_3 is greater than or equals to 2.

Second, we show that Kruskal rank condition ($I_3 \geq 2$) implies the non-redundant condition. If the Kruskal rank of M_3 is at least 2, the regular rank of matrix $M_{3.2}$ for any $t_j \neq t_k$ is 2, meaning that any two rows of matrix M_3 are not the same.

Step 2: The distinct eigenvalue condition holds if and only if the non-redundant condition holds. First, we show the distinct eigenvalue condition implies the non-redundant condition, which is demonstrated by contradiction. Suppose non-redundant condition does not hold. That means there may exist at least one pair $T \in \{t_j, t_k\}$ such that distribution $f_{X_3|T}(.|t_j)$ is the same as $f_{X_3|T}(.|t_k)$. Then $E[\omega(X_3)|T = t_j] = E[\omega(X_3)|T = t_k]$ for any function ω, which is contradictory to the distinct eigenvalue condition.

Next, we show that the non-redundant condition implies the distinct eigenvalue condition. For any pair $T = \{t_j, t_k\}$, we define a $Q \times 1$ column vector $Df_{j,k}$

$$Df_{j,k} \equiv \begin{pmatrix} f_{X_3|T}(t_1|t_j) - f_{X_3|T}(t_1|t_k) \\ f_{X_3|T}(t_2|t_j) - f_{X_3|T}(t_2|t_k) \\ ... \\ f_{X_3|T}(t_Q|t_j) - f_{X_3|T}(t_Q|t_k) \end{pmatrix}.$$

The non-redundant condition guarantees that $Df_{j,k} \neq 0$ for all combinations of row j and k if $j \neq k$. Therefore, there exists a vector $W = (w_1, w_2, ..., w_Q)'$ such that W is not orthogonal to $Df_{j,k}$ for any $j \neq k$. That is

$$W' \times Df_{j,k} \neq 0, \quad j \neq k, j, k = 1, 2, ..., K.$$

Therefore, we can define a function of X_3 as

$$\omega(X_3) = \sum_{i=1,2,...,Q} w_i \times I(X_3 = t_i),$$

which satisfies, for any $t_j \neq t_k$,

$$E[\omega(X_3)|T = t_j] - E[\omega(X_3)|T = t_k] = W' \times Df_{j,k} \neq 0.$$

The existence of such a vector W can be shown by contradiction. Suppose such a vector W does not exist. Then, for any $W \in R^Q$, there exists a pair (j, k) with $k \neq j$ such that $W' \times Df_{j,k} = 0$. Note that the function $W' \times Df_{j,k}$ is continuous in W. Thus, for this given pair of (j,k), there exist K linearly independent vectors W^1,...,W^K such that $[(W^1)',...,(W^K)']' \times Df_{j,k} = 0$. As a result, we have $Df_{j,k} = 0$, a contradiction.

Step 3: The non-redundant condition holds if and only if the model is identified. First, we show the non-redundant condition implies identification. We have shown that the non-redundant condition implies the Kurskal rank condition, i.e., $I_3 \geq 2$. Since $I_1 + I_2 + I_3 \geq 2K + 2$, Allman et al. (2009) shows the unique identification up to label swapping.

Next, we show that if the non-redundant condition does not hold, the structure is not identified. The failure of the non-redundant condition indicates that at least two different rows of matrix M_3 are the same. This means for any function of X_3, the distinct eigenvalue assumption fails, which means that any convex combination of the eigenvectors, e.g., $f_{X_1|T}(.|t_j)$ and $f_{X_1|T}(.|t_k)$, corresponding to the same eigenvalue is an eigenvector (Hu, 2008). Therefore, the eigenvectors in M_1 are not uniquely identified. This suggests that we cannot use the eigenvalue-eigenvector decomposition to identify the eigenvector matrix.

To show that the model is non-identified, we construct two sets of matrices both satisfying the underlying structure when the non-redundant condition fails. For illustrative purposes, assume that X_3 is a binary variable, i.e., $X_3 \in \{0,1\}$. Suppose the data generating process is as follows, with matrices M_1, Ω, M_2, and M_3 being the true components,

$$\begin{aligned} A &= M_1^T \Omega M_2, \\ A(x_3) &= M_1^T D(x_3) \Omega M_2, \quad \text{where} \quad x_3 = 0,1 \\ M_i &= \begin{pmatrix} M_{i1} \\ ... \\ M_{iK} \end{pmatrix}, i = 1,2. \quad M_3 \equiv \begin{pmatrix} M_{31} \\ ... \\ M_{3K} \end{pmatrix} \end{aligned}$$

The failure of non-redundant condition indicates that the two rows of M_3 are the same. Without loss of generality, assume $M_{31} = M_{32}$, i.e., $f_{X_3|T}(0|t_1) = f_{X_3|T}(0|t_2)$ and $f_{X_3|T}(1|t_1) = f_{X_3|T}(1|t_2)$. We then show that the matrices constructed below can also explain the data.

$$\tilde{M}_1 \equiv \begin{pmatrix} \tilde{M}_{11} \\ \tilde{M}_{12} \\ ... \\ \tilde{M}_{1K} \end{pmatrix} \equiv \begin{pmatrix} b \times M_{11} + (1-b) \times M_{12} \\ M_{12} \\ ... \\ M_{1K} \end{pmatrix}, \tilde{\Omega}\tilde{M}_2 = (\tilde{M}_1^T)^{-1} A \ \& \ \tilde{M}_3 = M_3,$$

where b is within 0 and 1. The observable equivalence is equivalent to show that

- $A = \tilde{M}_1^T \tilde{\Omega} \tilde{M}_2$, which holds by construction.

- For any x_3, $A(x_3) = \tilde{M}_1^T \tilde{D}(x_3)\tilde{\Omega}\tilde{M}_2$, which is shown in the following. In particular,

$$\begin{aligned}
& A(x_3) = \tilde{M}_1^T D(x_3)\tilde{\Omega}\tilde{M}_2 = \tilde{M}_1^T D(x_3)(\tilde{M}_1^T)^{-1}A \\
\Leftrightarrow\ & A(x_3)A^{-1}\tilde{M}_1^T = \tilde{M}_1^T D(x_3) \\
\Leftrightarrow\ & \begin{cases} A(x_3)A^{-1}(b \times M_{11}^T + (1-b) \times M_{12}^T) = f_{X_3|T}(x_3|t_1)(b \times M_{11}^T + (1-b) \times M_{12}^T) & \text{if } k = 1 \\ A(x_3)A^{-1}M_{1k}^T = f_{X_3|T}(x_3|t_k)M_{1k}^T & \text{if } k \geq 2 \end{cases}
\end{aligned}$$

Note that $A(x_3)A^{-1}M_{1k}^T = f_{X_3|T}(x_3|t_k)M_{1k}^T$, for $k = 1, ..., K$. Consequently, to show that $A(x_3) = \tilde{M}_1^T \tilde{D}(x_3)\tilde{\Omega}\tilde{M}_2$, we just need to show that $A(x_3)A^{-1}(b \times M_{11}^T + (1-b) \times M_{12}^T) = f_{X_3|T}(x_3|t_1)(b \times M_{11}^T + (1-b) \times M_{12}^T)$. Specifically,

$$\begin{aligned}
A(x_3)A^{-1}(b \times M_{11}^T + (1-b) \times M_{12}^T) &= b \times A(x_3)A^{-1}M_{11}^T + (1-b) \times A(x_3)A^{-1}M_{12}^T \\
&= b \times f_{X_3|T}(x_3|t_1)M_{11}^T + (1-b) \times f_{X_3|T}(x_3|t_2)M_{12}^T \\
&= b \times f_{X_3|T}(x_3|t_1)M_{11}^T + (1-b) \times f_{X_3|T}(x_3|t_1)M_{12}^T \\
&= f_{X_3|T}(x_3|t_1)(b \times M_{11}^T + (1-b) \times M_{12}^T).
\end{aligned}$$

The second equality holds because $A(x_3)A^{-1}M_{1k}^T = f_{X_3|T}(x_3|t_k)M_{1k}^T$, for $k = 1, ..., K$. The third equality holds because $f_{X_3|T}(x_3|t_1) = f_{X_3|T}(x_3|t_2)$. Thus, the data can be also explained by the matrices $\tilde{M}_1, \tilde{\Omega}, \tilde{M}_2$, and $\tilde{M}_3$, indicating that the structure is not identified if the non-redundant condition fails.

■