EVALUATING COUNTERFACTUAL QUANTITIES USING CROSS-WORLD FACTOR MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The evaluation of counterfactual queries is a central objective of causal inference. We develop an identification strategy for counterfactual queries using cross-world factor models, which provide a compact representation of causal relationships across K treatment levels using J scalar latent factors. First, we characterize when this representation is saturated, that is, when it can be adopted without loss of generality relative to an unrestricted structural causal model, and show that saturation requires at least as many latent factors as treatment levels ($J \geq K$). Next, under conditional mutual independence and additional regularity and loading-structure assumptions, we derive identification results for the joint distribution of the potential outcomes from their conditional marginal distributions. Our identification results require no more latent factors than treatment levels ($J \leq K$), implying that a model satisfying both saturation and our identification conditions must use the same number of latent factors as treatment levels ($J = K$). We further develop a parametric maximum likelihood estimation procedure and demonstrate the proposed approach using a real-world randomized recommendation dataset.

1 INTRODUCTION

The evaluation of counterfactual queries is a central objective of causal inference (Pearl, 2009). For a multilevel treatment X with K levels, an outcome Y, and subject-level covariates W, counterfactual queries are often expressed as conditional expectations of functions of the potential outcomes:

$$\mathbb{E}\left[g(Y_0, \ldots, Y_{K-1})|W = w\right], \tag{1}$$

where Y_x denotes the potential outcome under treatment level $X = x$ and W denotes the observed covariates. Many causal quantities can be expressed in this form, as summarized in Table 4 in Appendix A. These quantities enable the analysis of a wide range of causal questions related to causal effect heterogeneity and facilitate counterfactual decision-making. However, even when the treatment-specific marginal distributions of the potential outcomes are identified, their joint distribution generally remains unidentified because potential outcomes under different treatment levels are never jointly observed for the same unit. Thus, identifying counterfactual quantities that depend on multiple potential outcomes requires additional structure on their cross-world dependence.

One existing approach to identifying counterfactual quantities is based on monotonicity assumptions (Pearl, 1999; Tian & Pearl, 2000; Mueller & Pearl, 2023; Zhang et al., 2025; Hashimoto et al., 2026) or rank preservation (Heckman et al., 1997; Wu et al., 2025; Kawakami & Tian, 2026). Monotonicity restricts the direction of treatment effects, whereas rank preservation requires individuals to retain their relative ordering across potential outcomes under different treatment conditions.

Another existing approach to identifying counterfactual quantities is based on conditional independence assumptions. For binary treatments, conditional independence assumptions such as $Y_1 \perp\!\!\!\perp Y_0|W$ and $Y_1 - Y_0 \perp\!\!\!\perp Y_0|W$ have been used to identify treatment-benefit or individual-effect distributions (Zhang et al., 2013; Shen et al., 2013; Post & Van Den Heuvel, 2025). These independence conditions arise from two familiar binary-treatment representations. With $(U_0, U_1) = (Y_0, Y_1)$, the observed outcome is $Y = U_0\mathbb{I}(X = 0) + U_1\mathbb{I}(X = 1)$ (Imbens & Rubin, 2015). With $(U_0, U_1) = (Y_1 - Y_0, Y_0)$, it is $Y = U_0\mathbb{I}(X = 1) + U_1$ (Post & Van Den Heuvel, 2025). In each representation, the corresponding independence condition is $U_0 \perp\!\!\!\perp U_1|W$. The first representation uses separate factors for the two treatments, whereas the second includes a baseline factor shared across treatments.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Factor models have also been used to identify joint counterfactual distributions. Carneiro et al. (2003) develop multifactor models with multiple state-contingent outcomes, while Cunha et al. (2005) use a multifactor structure for counterfactual earnings under two schooling states. Their identification analyses use auxiliary information from choice equations, measurements, and repeated outcomes to recover the factor structure and dependence across counterfactuals. For specified factor-sharing structures, we characterize when the representation is saturated relative to an unrestricted SCM and establish conditions under which the treatment-specific conditional marginals identify the joint potential-outcome distribution. As an extension of the binary-treatment representations in Imbens & Rubin (2015) and Post & Van Den Heuvel (2025), we introduce a class of causal models in which the potential outcomes are represented using finitely many scalar latent factors $U_0, \ldots, U_{J-1}$.

Given J subsets $A_0, \ldots, A_{J-1}$ of $\{0, \ldots, K-1\}$, we consider models of the form

$$Y := \sum_{x=0}^{K-1} \left[\mu_x(W) + \sum_{j=0}^{J-1} \alpha_{xj}(W) U_j \mathbb{I}\{x \in A_j\} \right] \mathbb{I}\{X = x\}. \tag{2}$$

For every $x \in \{0, \ldots, K-1\}$, the potential outcome satisfies

$$Y_x = \mu_x(W) + \sum_{j=0}^{J-1} \alpha_{xj}(W) U_j \mathbb{I}\{x \in A_j\}. \tag{3}$$

We call this a *cross-world factor model* because the same latent factor may contribute to potential outcomes under multiple treatment levels. We call $\alpha_{xj}(W)$ the *loading* of U_j in Y_x. The factor U_j contributes to Y_x if and only if $x \in A_j$ and $\alpha_{xj}(W) \neq 0$, so A_j specifies the treatment levels under which U_j may enter the potential outcome. $\mu_x(W)$ represents the treatment-specific baseline component of Y_x.

Depending on the choice of $A_0, \ldots, A_{J-1}$ and the loadings $\alpha_{xj}(W)$, the representation may exclude relationships that are admissible under an unrestricted SCM. We therefore characterize when the representation is saturated, meaning that it can represent the outcome equation of any unrestricted SCM, and show that saturation requires ($J \geq K$).

Then, for specified sets $A_0, \ldots, A_{J-1}$, we establish an identification strategy for the joint distribution of the potential outcomes from their conditional marginal distributions under conditional mutual independence, regularity conditions, and treatment-invariant loadings. Our identification results require $J \leq K$. Combined with the saturation requirement ($J \geq K$), this implies that a model that is both saturated and identifiable under our conditions must have $J = K$.

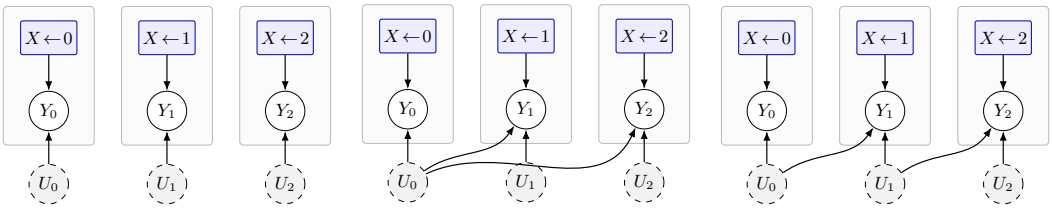
Using our identification results, causal queries can, in principle, be estimated via density deconvolution (Meister, 2008). However, the required operations may be unstable with realistic sample sizes. Instead, we posit a finite-dimensional parametric model compatible with the factor representation, develop a maximum likelihood estimation (MLE) procedure, and demonstrate its application to a real-world dataset.

The main contributions of this paper are:

- We introduce cross-world factor models, which represent causal relationships using scalar latent factors and explicitly encode how these factors are shared across potential-outcome worlds.
- We characterize saturation of these models, showing that they can represent every relationship allowed by an unrestricted SCM if and only if the loading matrix has full row rank.
- Given a specified factor-sharing structure and the conditional marginal distributions of the potential outcomes, we establish sufficient conditions for identifying their conditional joint distribution under conditional mutual independence of the latent factors, regularity of the characteristic functions, and treatment-invariant loadings.
- We develop finite-sample estimation procedures under parametric latent-factor distributions and illustrate the resulting counterfactual analysis using a real-world randomized recommendation dataset.

2 Illustrative Examples with Three Treatment Levels

To motivate our study, we present three illustrative examples of models with three treatment levels and three latent factors. We consider a short-video platform that exposes a user to one of three types



(a) Separate response mechanisms **(b) Shared baseline factor** **(c) Overlapping response mechanisms**

Figure 1: Parallel-world graphs for three latent-factor models in the short-video setting. Treatment levels 0, 1, and 2 correspond to urban short drama, television drama, and animation, respectively. Each shaded box represents the submodel induced by one treatment intervention. The latent factors are shared across worlds; an arrow $U_k \to Y_x$ indicates that $x \in A_k$. Panel (c) corresponds to the overlapping factor structure used in Section 6.

of video content: urban short drama, television drama, or animation. The treatment is defined as

$$X = 0 : \text{ urban short drama}, \quad X = 1 : \text{ television drama}, \quad X = 2 : \text{ animation}. \tag{4}$$

The outcome Y is the user's viewing time. For simplicity, we consider the case without covariates. Figure 1 shows parallel-world graphs for the three latent-factor models. A more detailed formulation, including observed covariates, is given in Appendix B.

Example 1 (Separate Response Mechanisms)**.** The user has three treatment-specific latent characteristics: preference for urban short drama U_0, preference for television drama U_1, and preference for animation U_2. Each latent characteristic affects the viewing time under only one type of video content. The potential outcomes are

$$Y_0 = \mu_0 + \alpha_{00}U_0, \quad Y_1 = \mu_1 + \alpha_{11}U_1, \quad Y_2 = \mu_2 + \alpha_{22}U_2. \tag{5}$$

Equivalently, the observed outcome is generated by the structural equation

$$Y = (\mu_0 + \alpha_{00}U_0)\,\mathbb{I}\{X = 0\} + (\mu_1 + \alpha_{11}U_1)\,\mathbb{I}\{X = 1\} + (\mu_2 + \alpha_{22}U_2)\,\mathbb{I}\{X = 2\}. \tag{6}$$

Example 2 (Shared Baseline Factor)**.** The user has a general viewing tendency U_0, which affects the response to all three content types. In addition, U_1 represents the content-specific appeal of television drama, and U_2 represents the content-specific appeal of animation. The model has potential outcomes

$$Y_0 = \mu_0 + \alpha_{00}U_0, \quad Y_1 = \mu_1 + \alpha_{10}U_0 + \alpha_{11}U_1, \quad Y_2 = \mu_2 + \alpha_{20}U_0 + \alpha_{22}U_2. \tag{7}$$

The loading of U_0 may differ across the three content types.

Example 3 (Overlapping Response Mechanisms)**.** The three content types have partially overlapping characteristics. Let U_0 affect viewing time under urban short drama and television drama, let U_1 affect viewing time under television drama and animation, and let U_2 affect viewing time only under animation. The resulting potential outcomes are

$$Y_0 = \mu_0 + \alpha_{00}U_0, \quad Y_1 = \mu_1 + \alpha_{10}U_0 + \alpha_{11}U_1, \quad Y_2 = \mu_2 + \alpha_{21}U_1 + \alpha_{22}U_2. \tag{8}$$

The loading of each shared latent factor may differ across the content types to which it contributes.

In practice, researchers may specify such models by introducing a collection of latent factors—often one factor for each treatment level—and determining which factors affect each potential outcome.

3 Cross-World Factor Models and Model Saturation

In the following, random variables are denoted by capital letters such as X, and their realizations by the corresponding lowercase letters such as x. Let $\mathbb{I}\{A\}$ denote the indicator of an event A. For a random variable Z, let Ω_Z denote its state space and $\mathbb{E}[Z]$ its expectation whenever the expectation exists.

Structural causal models (SCMs) We use the language of SCMs as our basic semantic and inferential framework (Pearl, 2009). An SCM $\mathcal{M}$ is a tuple $< U, V, \mathcal{F}, \mathbb{P}_U >$, where U is a set of exogenous (unobserved) variables following a joint distribution $\mathbb{P}_U$, and V is a set of endogenous (observable) variables whose values are determined by structural functions $\mathcal{F} = \{f_{V_i}\}_{V_i \in V}$ such that $v_i \coloneqq f_{V_i}(pa_{V_i}, u_{V_i})$, where $PA_{V_i} \subseteq V$ and $U_{V_i} \subseteq U$. An atomic intervention of setting a set of endogenous variables X to constants x, denoted by $do(x)$, replaces the original equations of X by $X \coloneqq x$ and induces a *sub-model* $\mathcal{M}_x$. We denote the potential outcome Y under intervention $do(x)$ by $Y_x(u)$, which is a solution of Y given $U = u$ in the sub-model $\mathcal{M}_x$.

The number of latent factors in f_Y is unrestricted in the definition of SCM (Pearl, 2009). We consider J scalar latent factors for a treatment with K levels and characterize when models are saturated.

Saturation of Models Before introducing our class of representations, we formalize the notion of saturation discussed by Post & Van Den Heuvel (2025) as follows:

Definition 1 (Saturation)**.** *Fix a treatment X with $\Omega_X = \{0, \dots, K-1\}$, observed covariates W, and an outcome $Y \in \Omega_Y \subseteq \mathbb{R}$. Consider a representation of the form $Y = h(X, W, \overline{U})$. We call the representation saturated given $W = w$ if, for every SCM whose structural equation for Y has the form $Y \coloneqq f_Y(X, W, U_Y)$, there exists a random vector $\overline{U}$ such that $h(x, w, \overline{U}) = f_Y(x, w, U_Y)$ holds almost surely for every $x \in \{0, \dots, K-1\}$.*

A model class is called *saturated* if every model in the class is saturated in the sense of Definition 1. Such models can reproduce any collection of potential outcomes generated by an SCM when the latent joint distribution is unrestricted. Saturation ensures compatibility with arbitrary potential-outcome distributions and, in particular, with any collection of observed marginal distributions. This claim is proved in Appendix G.

Cross-World Factor Model We define cross-world factor models as follows:

Definition 2 (Cross-World Factor Model)**.** *For a multilevel treatment $X \in \Omega_X = \{0, \dots, K-1\}$, an outcome $Y \in \Omega_Y$ and subject-level covariates $W \in \Omega_W$, given J subsets $A_0, \dots, A_{J-1}$ of $\{0, \dots, K-1\}$, a* cross-world factor model *is a representation of the form*

$$Y \coloneqq \sum_{x=0}^{K-1} \left[\mu_x(W) + \sum_{j=0}^{J-1} \alpha_{xj}(W) U_j \mathbb{I}\{x \in A_j\} \right] \mathbb{I}\{X = x\}, \tag{9}$$

where $U_0, \dots, U_{J-1}$ are real-valued latent factors, and μ_x and α_{xj} are measurable functions defined on Ω_W.

For each $x \in \{0, \dots, K-1\}$ and $j \in \{0, \dots, J-1\}$, define $m_{xj} \coloneqq \mathbb{I}(x \in A_j)$ and $b_{xj}(w) \coloneqq \alpha_{xj}(w) m_{xj}$. Let $M \coloneqq (m_{xj}) \in \{0,1\}^{K \times J}$ and $B(w) \coloneqq (b_{xj}(w)) \in \mathbb{R}^{K \times J}$. The entries of M indicate whether each latent factor enters each potential outcome, while the entries of $B(w)$ are the corresponding loadings, with zero entries when $x \notin A_j$. Let $\boldsymbol{Y} \coloneqq (Y_0, \dots, Y_{K-1})^\top$, $\boldsymbol{U} \coloneqq (U_0, \dots, U_{J-1})^\top$, and $\boldsymbol{\mu}(w) \coloneqq (\mu_0(w), \dots, \mu_{K-1}(w))^\top$. Thus, $\boldsymbol{Y}$ collects the K potential outcomes, $\boldsymbol{U}$ collects the J latent factors, and $\boldsymbol{\mu}(w)$ collects the treatment-specific baseline components. Then

$$\boldsymbol{Y} = \boldsymbol{\mu}(w) + B(w)\boldsymbol{U} \tag{10}$$

holds given $W = w$. Equation (10) is the vector-matrix form of the K potential-outcome equations implied by (9) given $W = w$.

Theorem 1 (Saturation)**.** *A model of the form* (9) *is saturated given $W = w$ if and only if* $\operatorname{rank}(B(w)) = K$.

Thus, saturation requires the latent factors to be able to generate arbitrary deviations of the K potential outcomes from their treatment-specific baseline components. Since $\operatorname{rank}(B(w)) \leq \min\{K, J\}$, a saturated model must satisfy $J \geq K$. A saturated model therefore requires at least K latent factors ($J \geq K$), with $J = K$ being the smallest number compatible with saturation.

4 IDENTIFICATION OF COUNTERFACTUAL QUERIES WITH CROSS-WORLD FACTOR MODELS

Next, we study identification under the proposed causal model. We provide sufficient conditions for identifying the conditional joint distribution of the potential outcomes and, consequently, joint expectations involving multiple potential outcomes. In the remainder of the paper, we assume that $\mathbb{E}[g(Y_0, \dots, Y_{K-1})|W = w]$ exists for the given function g. We assume access to the conditional marginal distribution $\mathbb{P}_{Y_x|W=w}$ for every treatment level $x \in \{0, \dots, K-1\}$ and every relevant covariate value w.

We impose the following conditional mutual independence assumption on the latent factors:

Assumption 1 (Conditional Mutual Independence of Latent Factors)**.** *$U_0, \dots, U_{J-1}$ are mutually independent given $W = w$.*

In Example 1, this assumption requires the treatment-specific preferences for urban short drama, television drama, and animation to be mutually independent given $W = w$. In Example 2, it requires the general viewing tendency, television-drama-specific appeal, and animation-specific appeal to be mutually independent given $W = w$. In Example 3, it requires the latent response shared by urban short drama and television drama, the latent response shared by television drama and animation, and the animation-specific response to be mutually independent given $W = w$.

We also assume the following condition:

Assumption 2. *The conditional distribution of U_j given $W = w$ has a density for every $j \in \{0, \dots, J-1\}$. For each $x \in \{0, \dots, K-1\}$, the conditional characteristic function $\varphi_x(t|w) := \mathbb{E}[\exp(\mathrm{i}tY_x)|W = w]$ is nonzero for every $t \in \mathbb{R}$.*

Assumption 3 (Cross-World Loading Homogeneity)**.** *For each $j \in \{0, \dots, J-1\}$, the coefficient $\alpha_{xj}(w)$ is constant across all $x \in A_j$; we denote its common value by $\alpha_j(w)$.*

Assumptions 1 and 2 alone do not ensure identification in general; see Appendix C.

Assumption 3 is plausible when a shared factor represents a common additive component whose magnitude does not depend on the treatment level. In Example 1, Assumption 3 holds automatically. In Example 2, it requires $\alpha_{00} = \alpha_{10} = \alpha_{20}$. In Example 3, it requires $\alpha_{00} = \alpha_{10}$ and $\alpha_{11} = \alpha_{21}$.

As the next theorem shows, the conditional joint distribution is not identifiable in general when the number of latent factors exceeds the number of treatment levels ($J > K$).

Theorem 2. *Fix $w \in \Omega_W$. Suppose that $J > K$ and the columns of M are distinct and nonzero. There exist two models of the form (9) with the same $\boldsymbol{\mu}(w)$ and $B(w) = M$ that satisfy Assumptions 1, 2, and 3 and induce the same conditional marginal distributions of $Y_0, \dots, Y_{K-1}$ given $W = w$ but different conditional joint distributions of $\boldsymbol{Y}$ given $W = w$.*

In the remainder of the identification analysis, we restrict attention to $J \leq K$. A saturated model in this range must have $J = K$, because saturation requires $J \geq K$ (Figure 2).

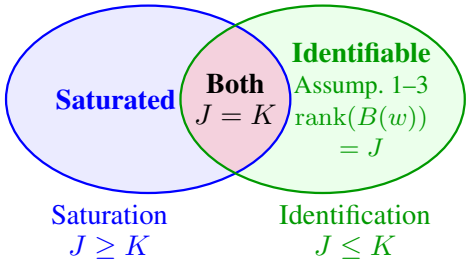


Figure 2: **Saturation and identification.** Saturation requires $J \geq K$, while our identification results consider $J \leq K$; hence models satisfying both have $J = K$.

We then establish identification under treatment-invariant loadings.

Theorem 3 (Identification under treatment-invariant loadings)**.** *Fix $w \in \Omega_W$. Consider a model determined by (9) with specified M and $J \leq K$ that satisfies Assumptions 1, 2, and 3. Suppose that $\mathrm{rank}(B(w)) = J$ so that M has full column rank. Then the conditional marginal distributions of $Y_0, \dots, Y_{K-1}$ given $W = w$ identify the conditional joint distribution of $\boldsymbol{Y}$ given $W = w$. More precisely, the following statements hold.*

(i): For a matrix $L \in \mathbb{R}^{J \times K}$ with $LM = I_J$, define

$$\boldsymbol{V} := L\boldsymbol{Y}, \qquad \boldsymbol{\delta}(w) := (I_K - ML)\boldsymbol{\mu}(w). \tag{11}$$

For each $j \in \{0, \ldots, J-1\}$, let $q_j(\cdot|w)$ be the conditional density of V_j given $W = w$. Let $\boldsymbol{\ell}(t|w) \coloneqq (\ell_0(t|w), \ldots, \ell_{K-1}(t|w))^\top$, where $\ell_x(\cdot|w)$ is the continuous logarithm of $\varphi_x(\cdot|w)$ satisfying $\ell_x(0|w) = 0$. Then $\boldsymbol{\delta}(w)$ is identified as

$$\boldsymbol{\delta}(w) = \frac{1}{\mathrm{i}t}(I_K - ML)\boldsymbol{\ell}(t|w), \qquad \text{for any } t \neq 0. \tag{12}$$

$q_j(\cdot|w)$ is also identified and, if $\psi_j(\cdot|w) \coloneqq \exp((L\boldsymbol{\ell}(\cdot|w))_j) \in L^1(\mathbb{R})$, it is represented as

$$q_j(v|w) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp(-\mathrm{i}tv)\psi_j(t|w)dt \qquad \text{for Lebesgue-a.e. } v \in \mathbb{R}. \tag{13}$$

(ii): $V_0, \ldots, V_{J-1}$ are mutually independent given $W = w$, and

$$\boldsymbol{Y} = \boldsymbol{\delta}(w) + M\boldsymbol{V} \tag{14}$$

holds given $W = w$[1]. Consequently,

$$\mathbb{E}[g(Y_0, \ldots, Y_{K-1})|W = w] = \int_{\mathbb{R}^J} g(\boldsymbol{\delta}(w) + M\boldsymbol{v}) \prod_{j=0}^{J-1} q_j(v_j|w)d\boldsymbol{v}. \tag{15}$$

(iii): If $J = K$, then $L = M^{-1}$, $\boldsymbol{\delta}(w) = \mathbf{0}$, and the conditional joint density of $\boldsymbol{Y}$ is identified as

$$p_{\boldsymbol{Y}|W=w}(\boldsymbol{y}) = |\det(M^{-1})| \prod_{j=0}^{K-1} q_j((M^{-1}\boldsymbol{y})_j|w) \qquad \text{for } \boldsymbol{y} \in \mathbb{R}^K. \tag{16}$$

Identification in Examples 1–3 In all three examples, $J = K = 3$, $\det(M) = 1$, $\boldsymbol{\delta} = \mathbf{0}$, and $\boldsymbol{V} = M^{-1}\boldsymbol{Y}$. Let $\psi_j(t) \coloneqq \mathbb{E}[\exp(\mathrm{i}tV_j)]$ denote the characteristic function of V_j. Suppose that $\psi_j \in L^1(\mathbb{R})$ for every $j \in \{0, 1, 2\}$. Then

$$q_j(v) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp(-\mathrm{i}tv)\psi_j(t)dt \qquad \text{for a.e. } v \text{ and every } j \in \{0, 1, 2\}. \tag{17}$$

In Example 1, $\boldsymbol{V} = (Y_0, Y_1, Y_2)^\top$. Thus $\psi_0(t) = \varphi_0(t)$, $\psi_1(t) = \varphi_1(t)$, and $\psi_2(t) = \varphi_2(t)$, and $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(v_0, v_1, v_2) \prod_{j=0}^2 q_j(v_j)d\boldsymbol{v}$. In Example 2, $\boldsymbol{V} = (Y_0, Y_1 - Y_0, Y_2 - Y_0)^\top$. In this case, $\psi_0(t) = \varphi_0(t)$, $\psi_1(t) = \varphi_1(t)/\varphi_0(t)$, and $\psi_2(t) = \varphi_2(t)/\varphi_0(t)$, so $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(v_0, v_0 + v_1, v_0 + v_2) \prod_{j=0}^2 q_j(v_j)d\boldsymbol{v}$. Finally, in Example 3, $\boldsymbol{V} = (Y_0, Y_1 - Y_0, Y_2 - Y_1 + Y_0)^\top$. We have $\psi_0(t) = \varphi_0(t)$, $\psi_1(t) = \varphi_1(t)/\varphi_0(t)$, and $\psi_2(t) = \varphi_0(t)\varphi_2(t)/\varphi_1(t)$, and hence $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(v_0, v_0 + v_1, v_1 + v_2) \prod_{j=0}^2 q_j(v_j)d\boldsymbol{v}$.

Testability of Assumption 1 Violations of Assumption 1 are sometimes testable. Consider a model of the form (9) satisfying Assumption 3 with $\mathrm{rank}(B(w)) = J$, and fix L satisfying $LM = I_J$. Suppose that the components of $\boldsymbol{V} = L\boldsymbol{Y}$ have finite conditional variances given $W = w$. Define $\boldsymbol{v}_Y(w) \coloneqq (\mathrm{Var}(Y_0|W = w), \ldots, \mathrm{Var}(Y_{K-1}|W = w))^\top$ and $\boldsymbol{v}_V(w) \coloneqq (\mathrm{Var}(V_0|W = w), \ldots, \mathrm{Var}(V_{J-1}|W = w))^\top$. If Assumption 1 holds, then

$$\boldsymbol{v}_Y(w) = M\boldsymbol{v}_V(w),\ \boldsymbol{v}_V(w) = L\boldsymbol{v}_Y(w),\ (I_K - ML)\boldsymbol{v}_Y(w) = \mathbf{0}. \tag{18}$$

Thus every component of $L\boldsymbol{v}_Y(w)$ must be nonnegative and $(I_K - ML)\boldsymbol{v}_Y(w) = \mathbf{0}$ must hold. These conditions are necessary, but not sufficient, consequences of Assumption 1.

Interpretation of Assumption 1 in Terms of Potential Outcomes Consider a model of the form (9) satisfying Assumption 3 with $\mathrm{rank}(B(w)) = J$, and fix L satisfying $LM = I_J$. Then Assumption 1 is equivalent to conditional mutual independence of the components of $L\boldsymbol{Y}$, a vector of linear combinations of the potential outcomes, given $W = w$. When $J < K$, the left inverse L is not unique, but the resulting independence assumption is the same for every left inverse of M.

In Examples 1–3, Assumption 1 is therefore equivalent, respectively, to conditional mutual independence of the components of $(Y_0, Y_1, Y_2)^\top$, $(Y_0, Y_1 - Y_0, Y_2 - Y_0)^\top$, and $(Y_0, Y_1 - Y_0, Y_2 - Y_1 + Y_0)^\top$ given $W = w$.

[1]The value on the right-hand side of (15) is the same for every L satisfying $LM = I_J$. The values $\mu_x(w)$, the loadings $\alpha_j(w)$, and the conditional distributions of U_j given $W = w$ are not identified separately.

Sensitivity to the Loading Homogeneity Assumption Assumption 3 requires each latent factor to have the same loading in every potential outcome to which it contributes. We allow the loadings to vary across treatment levels as

$$\alpha_{xj}(w) = \lambda_x(w)\alpha_j(w) \tag{19}$$

for every $j \in \{0, \dots, J-1\}$ and $x \in A_j$, where $\lambda_x(w) \neq 0$ and $\alpha_j(w) \neq 0$. Let $\boldsymbol{\lambda}(w) \coloneqq (\lambda_0(w), \dots, \lambda_{K-1}(w))^\top$. Every $\boldsymbol{\lambda}(w) = (a, \dots, a)^\top$ with $a \neq 0$ satisfies Assumption 3. Let $\Lambda(w) \subset (\mathbb{R}\setminus\{0\})^K$ consist of values of $\boldsymbol{\lambda}(w)$ for which the given conditional marginal distributions are induced by a model satisfying Assumptions 1, 2, and (19).

Fix $\boldsymbol{\lambda}(w) \in \Lambda(w)$ and define $D_{\boldsymbol{\lambda}}(w) \coloneqq \operatorname{diag}(\lambda_0(w), \dots, \lambda_{K-1}(w))$ and $\boldsymbol{Z}_{\boldsymbol{\lambda}} \coloneqq D_{\boldsymbol{\lambda}}(w)^{-1}\boldsymbol{Y}$. Then

$$Z_{\boldsymbol{\lambda},x} = \frac{\mu_x(w)}{\lambda_x(w)} + \sum\nolimits_{j=0}^{J-1} \alpha_j(w) U_j \mathbb{I}\{x \in A_j\} \tag{20}$$

for every $x \in \{0, \dots, K-1\}$. Thus $\boldsymbol{Z}_{\boldsymbol{\lambda}}$ has treatment-invariant loadings. Assumptions 1 and 2 also hold for (20): the latent factors are unchanged, and the conditional characteristic function of $Z_{\boldsymbol{\lambda},x}$ is $\varphi_x(t/\lambda_x(w)|w)$. If $\operatorname{rank}(M) = J$, then the coefficient matrix $M \operatorname{diag}(\alpha_0(w), \dots, \alpha_{J-1}(w))$ has rank J. Therefore Theorem 3 applies to $\boldsymbol{Z}_{\boldsymbol{\lambda}}$ for every $\boldsymbol{\lambda}(w) \in \Lambda(w)$.

Let $\boldsymbol{\delta}_{\boldsymbol{\lambda}}(w)$ and $q_{\boldsymbol{\lambda},0}(\cdot|w), \dots, q_{\boldsymbol{\lambda},J-1}(\cdot|w)$ be the quantities identified by Theorem 3. Since $\boldsymbol{Y} = D_{\boldsymbol{\lambda}}(w)\boldsymbol{Z}_{\boldsymbol{\lambda}}$,

$$\mathbb{E}[g(Y_0, \dots, Y_{K-1})|W = w] = \int_{\mathbb{R}^J} g(D_{\boldsymbol{\lambda}}(w)(\boldsymbol{\delta}_{\boldsymbol{\lambda}}(w) + M\boldsymbol{v}))\prod\nolimits_{j=0}^{J-1} q_{\boldsymbol{\lambda},j}(v_j|w)d\boldsymbol{v}. \tag{21}$$

The sensitivity analysis evaluates (21) over $\boldsymbol{\lambda}(w) \in \Lambda(w)$.

Illustration of the Sensitivity Analysis with Examples 1–3 In all three examples, $J = K = 3$, $\det(M) = 1$, $\boldsymbol{\delta}_{\boldsymbol{\lambda}} = \mathbf{0}$, and $\boldsymbol{V}_{\boldsymbol{\lambda}} \coloneqq M^{-1}\boldsymbol{Z}_{\boldsymbol{\lambda}}$. Write $\lambda_x \coloneqq \lambda_x(w)$ for $x \in \{0,1,2\}$. Let $\psi_{\boldsymbol{\lambda},j}(t) \coloneqq \mathbb{E}[\exp(\mathrm{i}tV_{\boldsymbol{\lambda},j})]$ denote the characteristic function of $V_{\boldsymbol{\lambda},j}$.

Suppose that $\psi_{\boldsymbol{\lambda},j}(\cdot) \in L^1(\mathbb{R})$ for every $j \in \{0,1,2\}$. Then the Fourier inversion in (17) determines $q_{\boldsymbol{\lambda},j}(\cdot)$ from $\psi_{\boldsymbol{\lambda},j}(\cdot)$ for every $j \in \{0,1,2\}$.

More explicitly, in Example 1, Assumption 3 holds for every $\boldsymbol{\lambda}(w)$ because each latent factor contributes to only one potential outcome. Moreover, $\boldsymbol{V}_{\boldsymbol{\lambda}} = (\frac{Y_0}{\lambda_0}, \frac{Y_1}{\lambda_1}, \frac{Y_2}{\lambda_2})^\top$, and the characteristic functions satisfy $\psi_{\boldsymbol{\lambda},j}(t) = \varphi_j(\frac{t}{\lambda_j})$ for every $j \in \{0,1,2\}$. Therefore, $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(\lambda_0 v_0, \lambda_1 v_1, \lambda_2 v_2) \times \prod_{j=0}^2 q_{\boldsymbol{\lambda},j}(v_j)d\boldsymbol{v}$. For Example 2, $\boldsymbol{V}_{\boldsymbol{\lambda}} = (\frac{Y_0}{\lambda_0}, \frac{Y_1}{\lambda_1} - \frac{Y_0}{\lambda_0}, \frac{Y_2}{\lambda_2} - \frac{Y_0}{\lambda_0})^\top$. In this case, we see $\psi_{\boldsymbol{\lambda},0}(t) = \varphi_0(\frac{t}{\lambda_0})$, $\psi_{\boldsymbol{\lambda},1}(t) = \varphi_1(\frac{t}{\lambda_1})/\varphi_0(\frac{t}{\lambda_0})$, and $\psi_{\boldsymbol{\lambda},2}(t) = \varphi_2(\frac{t}{\lambda_2})/\varphi_0(\frac{t}{\lambda_0})$. Hence $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(\lambda_0 v_0, \lambda_1(v_0+v_1), \lambda_2(v_0+v_2)) \times \prod_{j=0}^2 q_{\boldsymbol{\lambda},j}(v_j)d\boldsymbol{v}$. In Example 3, $\boldsymbol{V}_{\boldsymbol{\lambda}} = (\frac{Y_0}{\lambda_0}, \frac{Y_1}{\lambda_1} - \frac{Y_0}{\lambda_0}, \frac{Y_2}{\lambda_2} - \frac{Y_1}{\lambda_1} + \frac{Y_0}{\lambda_0})^\top$. Then $\psi_{\boldsymbol{\lambda},0}(t) = \varphi_0(\frac{t}{\lambda_0})$, $\psi_{\boldsymbol{\lambda},1}(t) = \varphi_1(\frac{t}{\lambda_1})/\varphi_0(\frac{t}{\lambda_0})$, and $\psi_{\boldsymbol{\lambda},2}(t) = \varphi_0(\frac{t}{\lambda_0})\varphi_2(\frac{t}{\lambda_2})/\varphi_1(\frac{t}{\lambda_1})$. Consequently, we see $\mathbb{E}[g(Y_0, Y_1, Y_2)] = \int_{\mathbb{R}^3} g(\lambda_0 v_0, \lambda_1(v_0 + v_1), \lambda_2(v_1 + v_2)) \times \prod_{j=0}^2 q_{\boldsymbol{\lambda},j}(v_j)d\boldsymbol{v}$.

5 Maximum Likelihood Estimation of Counterfactual Queries

Theorem 3 identifies $q_0, \dots, q_{J-1}$ at the population level from the conditional marginal distributions of $Y_0, \dots, Y_{K-1}$. A direct finite-sample implementation would estimate these marginal distributions nonparametrically and recover $q_0, \dots, q_{J-1}$ through characteristic-function deconvolution and Fourier inversion. These inverse steps can amplify estimation error, and direct nonparametric recovery may be unstable with realistic sample sizes (Meister, 2008). To avoid these difficulties, we restrict the treatment-specific marginal densities to a compatible finite-dimensional parametric family, estimate its parameter by maximum likelihood, and evaluate the counterfactual query using the implied densities $q_0, \dots, q_{J-1}$.

Let $\Theta_K \subset \mathbb{R}^{d_K}$ be a parameter space, and let $p_x(y, w|\boldsymbol{\theta})$ denote a model for the conditional density of Y_x given $W = w$. The parameter $\boldsymbol{\theta}$ is common to all covariate values. Let $\boldsymbol{\theta}^0$ be its true value. For every $\boldsymbol{\theta} \in \Theta_K$, we assume that the marginal densities are induced by the independent-factor representation in Theorem 3, with $\boldsymbol{\delta}(w, \boldsymbol{\theta})$ and conditional factor densities $q_0(\cdot|w, \boldsymbol{\theta}), \dots, q_{J-1}(\cdot|w, \boldsymbol{\theta})$.

At $\boldsymbol{\theta}^0$, this representation is assumed to describe the conditional joint distribution of the potential outcomes. The full model specification is given in Appendix D.

For each treatment level $x \in \{0, \dots, K-1\}$, let $\{(Y^{(x,i)}, W^{(x,i)})\}_{i=1}^{n_x}$ be i.i.d. experimental observations of (Y_x, W) from the same population. The samples from different treatment groups are independent, and $n \coloneqq \sum_{x=0}^{K-1} n_x$. Let $\widehat{\boldsymbol{\theta}} \in \Theta_K$ maximize $\prod_{x=0}^{K-1} \prod_{i=1}^{n_x} p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta})$. For every measurable function g for which the integral exists, define

$$\tau_g(w, \boldsymbol{\theta}) \coloneqq \int_{\mathbb{R}^J} g(\boldsymbol{\delta}(w, \boldsymbol{\theta}) + M\boldsymbol{v}) \prod_{j=0}^{J-1} q_j(v_j|w, \boldsymbol{\theta}) d\boldsymbol{v}, \qquad \widehat{\tau}_{g,n}(w) \coloneqq \tau_g(w, \widehat{\boldsymbol{\theta}}). \tag{22}$$

By Theorem 3, $\tau_g(w, \boldsymbol{\theta}^0) = \mathbb{E}[g(Y_0, \dots, Y_{K-1})|W = w]$. Then, the MLE-based estimators are asymptotically normal:

Theorem 4 (Asymptotic normality)**.** *Suppose that Assumption 4 in Appendix D holds. Let $\mathcal{I}_K(\boldsymbol{\theta}^0)$ denote the Fisher information defined by* (48) *in Appendix D.*

(i): For a fixed K, we have

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^0, \qquad \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}), \qquad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_{\mathbb{P}}(n^{-1/2}). \tag{23}$$

Fix w such that $\boldsymbol{\theta} \mapsto \tau_g(w, \boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}^0$, and let $\widehat{\tau}_{g,n}(w) = \tau_g(w, \widehat{\boldsymbol{\theta}})$. Then we have

$$\sqrt{n}\{\widehat{\tau}_{g,n}(w) - \tau_g(w, \boldsymbol{\theta}^0)\} \xrightarrow{d} N\left(0, \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)\right). \tag{24}$$

(ii): Consider a family of these models indexed by K, and let d_K denote the dimension of $\boldsymbol{\theta}$. For the fixed w above, suppose there exist constants $d_0, C_g, c_I > 0$, independent of K, such that

$$d_K \le d_0 K, \qquad \|\nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)\|^2 \le d_K C_g^2, \qquad \lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0)) \ge \frac{c_I}{K}, \tag{25}$$

where $\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))$ is the smallest eigenvalue of $\mathcal{I}_K(\boldsymbol{\theta}^0)$. Then

$$\frac{1}{n} \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0) \le \frac{d_0 C_g^2}{c_I} \frac{K^2}{n}. \tag{26}$$

Consequently, the corresponding asymptotic standard deviation is $O(K/\sqrt{n})$.

Part (ii) bounds, across models indexed by K, the asymptotic variance obtained from the fixed-K limit in part (i).

Synthetic Experiments For treatment r, we consider the probability of achieving the best potential outcome (PoB) given by $\mathbb{P}(Y_r > \max_{x \neq r} Y_x)$ (Kawakami & Tian, 2026).

We use synthetic data to examine the finite-sample behavior of the PoB estimator for $r = 1$ as the sample size and the number of treatment levels vary, with only the marginal location parameters estimated. Table 1 focuses on the dependence on K with the sample size per treatment group fixed. The first relation in the table has one factor shared by all treatments and additional treatment-specific factors, whereas in the second relation each potential outcome contains all or all but one of the latent factors. The variance decreases slightly with K under the first relation. Under the second, the variance increases sharply and the mean estimate falls below 0.25 for large K. Appendix E gives the details.

Table 1: Behavior of the PoB estimator as K increases. The true PoB is 0.25, each treatment group contains $1{,}000$ observations, and each entry is based on 300 independent datasets.

	$Y_0 = V_0,\ Y_x = V_0 + V_x\ (x \neq 0)$		$Y_1 = \sum_{j=0}^{K-1} V_j,\ Y_x = \sum_{j \neq x} V_j\ (x \neq 1)$	
K	Mean	Variance	Mean	Variance
3	0.251	1.884×10^{-4}	0.251	7.629×10^{-4}
10	0.251	1.311×10^{-4}	0.247	5.750×10^{-3}
50	0.250	1.180×10^{-4}	0.233	2.593×10^{-2}
100	0.249	1.115×10^{-4}	0.222	4.267×10^{-2}

6 Application to a Real-World Dataset

Setting We use KuaiRand-Pure, which records videos displayed by a randomized recommendation procedure on the Kuaishou short-video platform (Gao et al., 2022). We consider recommendations of urban short drama ($X = 0$), television drama ($X = 1$), and animation ($X = 2$) as the

three treatments and use viewing time in milliseconds (ms) as the outcome Y. We use the overlapping tag structure as the factor-sharing pattern in Example 3 and do not condition on covariates. Assumption 1 requires mutual independence of the tag-associated response components, and Assumption 3 requires each shared component to have the same loading under both content types. We use a normal inverse Gaussian specification satisfying Assumption 2 and estimate its parameters by maximum likelihood as in Section 5, using one observation from each of $17{,}036$ users. We study the probability of achieving the best potential outcome (PoB) (Kawakami & Tian, 2026). For comparison, we also estimate PoB and the pairwise probabilities under the rank preservation assumption (Kawakami & Tian, 2026). Data construction and estimation details are given in Appendix F.

Table 2: Sample mean viewing times and estimated PoB in KuaiRand-Pure. The intervals are 95% bootstrap percentile intervals.

Video content	Users	Mean (ms)	PoB [95% interval]	
			Our model	Rank preservation
Urban short drama	3,939	9,288	0.304 [0.243, 0.358]	0.000 [0.000, 0.084]
Television drama	8,069	**11,061**	0.327 [0.273, 0.388]	**0.998** [0.906, 0.998]
Animation	5,028	6,128	**0.369** [0.352, 0.384]	0.000 [0.000, 0.009]

Table 3: Estimated pairwise probabilities.

Quantity	Our model	Rank preservation
$\mathbb{P}(Y_1 > Y_0)$	0.528	0.998
$\mathbb{P}(Y_2 > Y_0)$	0.429	0.003
$\mathbb{P}(Y_2 > Y_1)$	0.381	0.000

Results Tables 2 and 3 report the sample mean viewing times, PoB estimates, and pairwise probabilities. Television drama has the largest sample mean, whereas animation has the smallest sample mean but the largest estimated PoB under our model. Animation is less likely than either alternative to produce a longer viewing time in the pairwise comparisons, yet it has the largest probability of exceeding both alternatives simultaneously. This illustrates that pairwise comparisons do not determine a three-way counterfactual ranking, since the PoB depends on how the potential outcomes vary jointly across users. In addition, the bootstrap intervals for the PoB differences distinguish animation from urban short drama more clearly than from television drama (Appendix F).

By contrast, under rank preservation, television drama has an estimated PoB close to one, while the pairwise point estimates indicate the same treatment ordering under both methods. Rank preservation requires each user's potential outcomes to have the same percentile rank in the treatment-specific distributions, and the empirical quantiles for television drama exceed those for both alternatives at nearly all percentiles. On the other hand, our factor model allows users' relative ranks to vary across treatments, and this may partly explain the difference in the estimated PoB.

Sensitivity to Assumption 3 We examine sensitivity to violations of Assumption 3 using the treatment-level scaling model in Section 4. The details and results are in Appendix F.

7 Conclusion

We introduced cross-world factor models, which represent causal relationships using finitely many scalar latent factors, and established identification of counterfactual queries under conditional mutual independence and loading homogeneity. The framework provides an alternative route to identification when these assumptions are credible, even if monotonicity or rank preservation is not.

Furthermore, our framework may open new directions for causal inference, including bounding and sensitivity analyses when the conditional mutual independence assumption is violated. When the identification assumptions fail and the causal quantities of interest are no longer identifiable, bounding methods have been developed in a variety of settings (Balke & Pearl, 1997; Tian & Pearl, 2000; Fan & Park, 2010; Duarte et al., 2024; Sachs et al., 2023; Kawakami & Tian, 2025; 2026). Likewise, sensitivity analysis is commonly used to assess the robustness of causal conclusions to violations of identification assumptions (Gadbury et al., 2001; Zhang et al., 2013; Ciocănea-Teodorescu et al., 2025). Developing sensitivity analyses for the conditional mutual independence assumptions considered in this paper, together with corresponding bounding methods based on the proposed saturated models, is an interesting direction for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, language refinement, organization, formatting, and code development. AI-assisted code was reviewed and tested by the authors before being used in the experiments. Generative AI tools were also used to assist with mathematical exposition and the presentation of experimental results. We did not use generative AI tools to autonomously generate experimental findings or determine the scientific conclusions of the work. All AI-assisted text, code, and mathematical content were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Pedro Carneiro, Karsten T Hansen, and James J Heckman. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college. Working Paper 9546, National Bureau of Economic Research, March 2003. URL `http://www.nber.org/papers/w9546`.

Iuliana Ciocănea-Teodorescu, Erin E Gabriel, and Arvid Sjölander. Sensitivity analysis for the probability of benefit in randomized controlled trials with a binary treatment and a binary outcome. *Biostatistics*, 26(1):kxaf011, 2025.

Flavio Cunha, James Heckman, and Salvador Navarro. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers*, 57(2):191–261, 2005.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024.

R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. ISBN 9781108473682.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010.

G L Gadbury, H K Iyer, and D B Allison. Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat*, 11(4):313–333, Nov 2001.

Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 3953–3957, 2022.

Naoya Hashimoto, Yuta Kawakami, and Jin Tian. Bounds and identification of joint probabilities of potential outcomes and observed variables under monotonicity assumptions. In *The 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

James J. Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 10 1997.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81 (396):945–960, 1986.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Yuta Kawakami and Jin Tian. Moments of causal effects. *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI-2025)*, 2025.

Yuta Kawakami and Jin Tian. Potential outcome rankings for counterfactual decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36679–36687, 2026.

Yuta Kawakami, Manabu Kuroki, and Jin Tian. Probabilities of causation for continuous and vector variables. *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI-2024)*, 2024.

Alexander Meister. Density deconvolution. In *Deconvolution Problems in Nonparametric Statistics*, pp. 5–105. Springer, 2008.

Scott Mueller and Judea Pearl. Monotonicity: Detection, refutation, and ramification, 2023.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Richard AJ Post and Edwin R Van Den Heuvel. Beyond conditional averages: Estimating the individual causal effect distribution. *Journal of Causal Inference*, 13(1):20240007, 2025.

Michael C Sachs, Gustav Jonzon, Arvid Sjölander, and Erin E Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023.

Changyu Shen, Jaesik Jeong, Xiaochun Li, Peng-Sheng Chen, and Alfred Buxton. Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics*, 69(3):724–731, Sep 2013.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Peng Wu, Haoxuan Li, Chunyuan Zheng, Yan Zeng, Jiawei Chen, Yang Liu, Ruocheng Guo, and Kun Zhang. Learning counterfactual outcomes under rank preservation. *Advances in Neural Information Processing Systems*, 38:148262–148289, 2025.

Chao Zhang, Zhi Geng, Wei Li, and Peng Ding. Identifying and bounding the probability of necessity for causes of effects with ordinal outcomes. *Biometrika*, pp. asaf049, 2025.

Zhiwei Zhang, Chenguang Wang, Lei Nie, and Guoxing Soon. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *J R Stat Soc Ser C Appl Stat*, 62 (5):687–704, Nov 2013.

APPENDIX TO "EVALUATING COUNTERFACTUAL QUANTITIES USING CROSS-WORLD FACTOR MODELS"

A SUMMARY OF CAUSAL QUANTITIES

Table 4 summarizes the causal quantities of interest. For a binary treatment $X \in \{0, 1\}$, it includes the average causal effect (ACE) (Holland, 1986), the distribution of causal effects (Fan & Park, 2010), the probability of necessity and sufficiency (PNS) (Kawakami et al., 2024), and the (central) moments of causal effects (Kawakami & Tian, 2025). For a discrete treatment $X \in \{0, 1, \ldots, K-1\}$, it includes the PNS with multi-hypothetical terms (Kawakami et al., 2024), the covariance of causal effects (central product moment of causal effects) (Kawakami & Tian, 2025), the probabilities of potential outcome ranking (PoR) and the probability of action achieving the best potential outcome (PoB) (Kawakami & Tian, 2026).

B DETAILED FORMULATION OF THE ILLUSTRATIVE EXAMPLES

We provide a more detailed description of the examples in Section 2. Here we write the same three models with W included and relate them explicitly to the general formulation in Section 3. For each $x \in \{0, 1, 2\}$, $\mu_x(W)$ represents the baseline component under treatment level x, and $\alpha_{xj}(W)$ represents the loading of U_j on Y_x. Both may vary with W. The conditional joint distribution of (U_0, U_1, U_2) given W may also vary with W.

Within each model, the same variables W, U_0, U_1, U_2 are used in the equations for Y_0, Y_1, Y_2. Thus, when the same latent factor appears in two potential-outcome equations, the same random variable contributes to both potential outcomes. The definitions of the three examples themselves do not impose independence among U_0, U_1, U_2. The observed outcome satisfies

$$Y = Y_0\mathbb{I}\{X = 0\} + Y_1\mathbb{I}\{X = 1\} + Y_2\mathbb{I}\{X = 2\}. \tag{27}$$

Example 1 (Separate Response Mechanisms). The three latent factors represent treatment-specific characteristics. With the observed covariates included, the potential outcomes are

$$\begin{aligned} Y_0 &= \mu_0(W) + \alpha_{00}(W) \underbrace{U_0}_{\text{urban short drama preference}}, \\ Y_1 &= \mu_1(W) + \alpha_{11}(W) \underbrace{U_1}_{\text{television drama preference}}, \\ Y_2 &= \mu_2(W) + \alpha_{22}(W) \underbrace{U_2}_{\text{animation preference}}. \end{aligned} \tag{28}$$

The corresponding sets are

$$A_0 = \{0\}, \qquad A_1 = \{1\}, \qquad A_2 = \{2\}, \tag{29}$$

and hence

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{30}$$

The factor U_0 enters only Y_0, U_1 enters only Y_1, and U_2 enters only Y_2. In particular, no latent factor is shared across treatment levels.

Substituting (28) into (27), the observed outcome satisfies

$$\begin{aligned} Y = {} & (\mu_0(W) + \alpha_{00}(W)U_0)\,\mathbb{I}\{X = 0\} + (\mu_1(W) + \alpha_{11}(W)U_1)\,\mathbb{I}\{X = 1\} \\ & + (\mu_2(W) + \alpha_{22}(W)U_2)\,\mathbb{I}\{X = 2\}. \end{aligned} \tag{31}$$

Treatment	Causal Quantities	Expression	Reference
Binary	Average causal effect (ACE)	$\mathbb{E}[Y_1 - Y_0]$	Holland (1986)
	Distribution of causal effects	$\mathbb{P}(Y_1 - Y_0 < y)$	Fan & Park (2010); Post & Van Den Heuvel (2025)
	Prob. of necessity and sufficiency (PNS)	$\mathbb{P}(Y_0 < y \leq Y_1)$	Kawakami et al. (2024)
	Moments of causal effects	$\mathbb{E}[(Y_1 - Y_0)^m],\ \mathbb{E}\left[\left(Y_1 - Y_0 - \mathbb{E}[Y_1 - Y_0]\right)^m\right]$	Kawakami & Tian (2025)
Discrete	PNS with multi-hypothetical terms	$\mathbb{P}(Y_{x_0} < y_1 \leq Y_{x_1}, Y_{x_1} < y_2 \leq Y_{x_2}, \ldots, Y_{x_{P-1}} < y_P \leq Y_{x_P})$	Kawakami et al. (2024)
	Covariance of causal effects	$\mathbb{E}[\{(Y_i - Y_j) - (\mathbb{E}[Y_i] - \mathbb{E}[Y_j])\}\{(Y_k - Y_h) - (\mathbb{E}[Y_k] - \mathbb{E}[Y_h])\}]$	Kawakami & Tian (2025)
	Prob. of potential outcome ranking (PoR)	$\mathbb{P}(Y_{r_1} > Y_{r_2} > \cdots > Y_{r_K})$	Kawakami & Tian (2026)
	Prob. of best outcome (PoB)	$\mathbb{P}(Y_r > \max_{k \neq r} Y_k)$	

Table 4: Causal quantities for a continuous outcome expressible as joint expectations over potential outcomes. Here $i, j, k, h \in \{0, 1, \ldots, K-1\}$ for covariance of causal effects, $\boldsymbol{r} = (r_1, \ldots, r_K)$ denotes a ranking of treatment levels for PoR, and $r \in \{0, 1, \ldots, K-1\}$ denotes a treatment level for PoB. The values $y_1, \ldots, y_P$ are outcome thresholds, and $x_0, x_1, \ldots, x_P$ are treatment levels for PNS with multi-hypothetical terms.

Example 2 (Shared Baseline Factor). The factor U_0 represents a general viewing tendency and enters the potential outcome under every treatment level. The factors U_1 and U_2 represent the content-specific appeal of television drama and animation, respectively, and each enters only one potential-outcome equation. The potential outcomes are

$$\begin{aligned} Y_0 &= \mu_0(W) + \alpha_{00}(W) \underbrace{U_0}_{\text{general viewing tendency}}, \\ Y_1 &= \mu_1(W) + \alpha_{10}(W) \underbrace{U_0}_{\text{general viewing tendency}} + \alpha_{11}(W) \underbrace{U_1}_{\text{television drama appeal}}, \\ Y_2 &= \mu_2(W) + \alpha_{20}(W) \underbrace{U_0}_{\text{general viewing tendency}} + \alpha_{22}(W) \underbrace{U_2}_{\text{animation appeal}}. \end{aligned} \tag{32}$$

The corresponding sets are

$$A_0 = \{0, 1, 2\}, \qquad A_1 = \{1\}, \qquad A_2 = \{2\}, \tag{33}$$

so

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}. \tag{34}$$

The same realization of U_0 therefore contributes to Y_0, Y_1, and Y_2. The three loadings $\alpha_{00}(W)$, $\alpha_{10}(W)$, and $\alpha_{20}(W)$ may differ. Sharing the same latent factor across treatment levels does not require equal loadings for that factor.

The observed outcome satisfies

$$\begin{aligned} Y = {} & (\mu_0(W) + \alpha_{00}(W)U_0)\,\mathbb{I}\{X = 0\} \\ & + (\mu_1(W) + \alpha_{10}(W)U_0 + \alpha_{11}(W)U_1)\,\mathbb{I}\{X = 1\} \\ & + (\mu_2(W) + \alpha_{20}(W)U_0 + \alpha_{22}(W)U_2)\,\mathbb{I}\{X = 2\}. \end{aligned} \tag{35}$$

Example 3 (Overlapping Response Mechanisms). The factor U_0 enters the potential outcomes under urban short drama and television drama, while U_1 enters the potential outcomes under television drama and animation. The factor U_2 enters only the potential outcome under animation. The potential outcomes are

$$\begin{aligned} Y_0 &= \mu_0(W) + \alpha_{00}(W) \underbrace{U_0}_{\substack{\text{shared by urban short drama} \\ \text{and television drama}}}, \\ Y_1 &= \mu_1(W) + \alpha_{10}(W) \underbrace{U_0}_{\substack{\text{shared by urban short drama} \\ \text{and television drama}}} + \alpha_{11}(W) \underbrace{U_1}_{\substack{\text{shared by television drama} \\ \text{and animation}}}, \\ Y_2 &= \mu_2(W) + \alpha_{21}(W) \underbrace{U_1}_{\substack{\text{shared by television drama} \\ \text{and animation}}} + \alpha_{22}(W) \underbrace{U_2}_{\text{animation-specific}}. \end{aligned} \tag{36}$$

The corresponding sets are

$$A_0 = \{0, 1\}, \qquad A_1 = \{1, 2\}, \qquad A_2 = \{2\}, \tag{37}$$

and

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}. \tag{38}$$

Thus, the same U_0 appears in Y_0 and Y_1, and the same U_1 appears in Y_1 and Y_2. No latent factor enters both Y_0 and Y_2 directly. As in Example 2, sharing a latent factor does not require its loading to be the same in the potential outcomes in which it appears.

The observed outcome satisfies

$$\begin{aligned} Y = {} & (\mu_0(W) + \alpha_{00}(W)U_0)\,\mathbb{I}\{X = 0\} \\ & + (\mu_1(W) + \alpha_{10}(W)U_0 + \alpha_{11}(W)U_1)\,\mathbb{I}\{X = 1\} \\ & + (\mu_2(W) + \alpha_{21}(W)U_1 + \alpha_{22}(W)U_2)\,\mathbb{I}\{X = 2\}. \end{aligned} \tag{39}$$

Relation to saturation and the identification assumptions. The three membership matrices displayed above are invertible and satisfy $\det(M) = 1$. For Example 1,

$$B(w) = \begin{pmatrix} \alpha_{00}(w) & 0 & 0 \\ 0 & \alpha_{11}(w) & 0 \\ 0 & 0 & \alpha_{22}(w) \end{pmatrix}. \tag{40}$$

For Example 2,

$$B(w) = \begin{pmatrix} \alpha_{00}(w) & 0 & 0 \\ \alpha_{10}(w) & \alpha_{11}(w) & 0 \\ \alpha_{20}(w) & 0 & \alpha_{22}(w) \end{pmatrix}. \tag{41}$$

For Example 3,

$$B(w) = \begin{pmatrix} \alpha_{00}(w) & 0 & 0 \\ \alpha_{10}(w) & \alpha_{11}(w) & 0 \\ 0 & \alpha_{21}(w) & \alpha_{22}(w) \end{pmatrix}. \tag{42}$$

In every case,

$$\det(B(w)) = \alpha_{00}(w)\alpha_{11}(w)\alpha_{22}(w). \tag{43}$$

Hence, by Theorem 1, each representation is saturated given $W = w$ if and only if $\alpha_{00}(w)$, $\alpha_{11}(w)$, and $\alpha_{22}(w)$ are all nonzero.

The definitions of Examples 1–3 do not impose Assumptions 1–3. Assumption 1 additionally requires U_0, U_1, U_2 to be mutually independent given $W = w$. Assumption 2 imposes the density and characteristic-function conditions stated in Section 4. Assumption 3 holds automatically in Example 1. In Example 2, it requires $\alpha_{00}(w) = \alpha_{10}(w) = \alpha_{20}(w)$. In Example 3, it requires $\alpha_{00}(w) = \alpha_{10}(w)$ and $\alpha_{11}(w) = \alpha_{21}(w)$. These are additional restrictions used for identification.

C Non-identifiability under Assumptions 1 and 2 alone

As the following theorem shows, Assumptions 1 and 2 alone do not ensure identification in general.

Theorem 5 (Non-identification)**.** *Fix $w \in \Omega_W$, $\boldsymbol{\mu}(w)$, and $B(w)$, and suppose that $J = K$. Suppose that $B(w)$ is invertible and that $(b_{xk}(w)^2)_{x,k=0}^{K-1}$ is singular. Then there exist two models of the form* (9) *with the same $\boldsymbol{\mu}(w)$ and $B(w)$ that are saturated given $W = w$ and satisfy Assumptions 1 and 2. Let $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ denote their potential-outcome vectors. For every $x \in \{0, \ldots, K-1\}$, the conditional distributions of Y_x and Y'_x given $W = w$ coincide, whereas $\mathrm{Cov}(\boldsymbol{Y}|W = w) \neq \mathrm{Cov}(\boldsymbol{Y}'|W = w)$. In particular, the conditional joint distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ given $W = w$ differ.*

D Details of Maximum Likelihood Estimation

Models. Suppose that $\mathrm{rank}(M) = J$ and fix $L \in \mathbb{R}^{J \times K}$ satisfying $LM = I_J$. Let P_W denote the distribution of the pretreatment covariates W in the population. We leave P_W unspecified and assume that it does not involve $\boldsymbol{\theta}$. We assume that the marginal distribution of W is P_W for every treatment level and every $\boldsymbol{\theta} \in \Theta_K$. For every $\boldsymbol{\theta} \in \Theta_K$, suppose that there exist $\boldsymbol{\delta}(w, \boldsymbol{\theta}) \in \mathbb{R}^K$ and random variables $V_0, \ldots, V_{J-1}$ that are mutually independent given $W = w$ and have conditional densities $q_0(\cdot|w, \boldsymbol{\theta}), \ldots, q_{J-1}(\cdot|w, \boldsymbol{\theta})$ for P_W-almost every w. Suppose that $L\boldsymbol{\delta}(w, \boldsymbol{\theta}) = \mathbf{0}$ and that $\delta_x(w, \boldsymbol{\theta}) + \sum_{j=0}^{J-1} m_{xj} V_j$ has conditional density $p_x(\cdot, w|\boldsymbol{\theta})$ given $W = w$ for every $x \in \{0, \ldots, K-1\}$. For every $x \in \{0, \ldots, K-1\}$ and P_W-almost every w, we assume that $\int_{\mathbb{R}} \exp(\mathrm{i}ty) p_x(y, w|\boldsymbol{\theta}) dy \neq 0$ for every $t \in \mathbb{R}$. We also assume that these functions are jointly measurable in their arguments. By Theorem 3, $\boldsymbol{\delta}(w, \boldsymbol{\theta})$ and the densities $q_0(\cdot|w, \boldsymbol{\theta}), \ldots, q_{J-1}(\cdot|w, \boldsymbol{\theta})$ are uniquely determined for P_W-almost every w by the joint distributions of $(Y_0, W), \ldots, (Y_{K-1}, W)$.

Estimation. For the treatment-group samples in Section 5, we obtain $\widehat{\boldsymbol{\theta}}$ by maximizing $\sum_{x=0}^{K-1} \sum_{i=1}^{n_x} \log p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta_K$. We then substitute $\widehat{\boldsymbol{\theta}}$ into $\boldsymbol{\delta}(w, \boldsymbol{\theta})$ and $q_j(\cdot|w, \boldsymbol{\theta})$ and evaluate the integral in (22) to obtain $\widehat{\tau}_{g,n}(w)$. For PoB of treatment r, we use $g(\boldsymbol{y}) = \mathbb{I}\{y_r > \max_{x \neq r} y_x\}$.

Regularity Conditions. We state the condition assumed in Theorem 4 for the treatment-group samples in Section 5. Expectations involving (Y_x, W) below are taken under the joint density $p_x(\cdot,\cdot|\boldsymbol{\theta}^0)$ with respect to $dy\,P_W(dw)$. For every x and $\boldsymbol{\theta}$, we have $\int_{\mathbb{R}} p_x(y,w|\boldsymbol{\theta})dy = 1$ for P_W-almost every w. Write $\boldsymbol{s}_x(y,w,\boldsymbol{\theta}) \coloneqq \nabla_{\boldsymbol{\theta}} \log p_x(y,w|\boldsymbol{\theta})$.

Assumption 4 (Regularity for Maximum Likelihood Estimation)**.** *We assume the following conditions.*

(i) The set Θ_K is bounded and closed, $\boldsymbol{\theta}^0$ is an interior point of Θ_K, and $p_x(\cdot,\cdot|\boldsymbol{\theta}) = p_x(\cdot,\cdot|\boldsymbol{\theta}^0)$ almost everywhere for every $x \in \{0,\dots,K-1\}$ only if $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

(ii) For every $x \in \{0,\dots,K-1\}$, the observations $(Y^{(x,i)}, W^{(x,i)})$ are independent and identically distributed with joint density $p_x(\cdot,\cdot|\boldsymbol{\theta}^0)$, and the samples from different treatment groups are independent. The treatment-group sizes are fixed and satisfy

$$n = \sum\nolimits_{x=0}^{K-1} n_x, \qquad \frac{n_x}{n} \longrightarrow \pi_x \in (0,1) \tag{44}$$

for every $x \in \{0,\dots,K-1\}$ as $n \to \infty$.

(iii) For every $x \in \{0,\dots,K-1\}$ and P_W-almost every w, there exists a measurable set $\mathcal{S}_x(w)$, independent of $\boldsymbol{\theta}$, such that $p_x(y,w|\boldsymbol{\theta}) > 0$ for a.e. $y \in \mathcal{S}_x(w)$ and $p_x(y,w|\boldsymbol{\theta}) = 0$ for a.e. $y \notin \mathcal{S}_x(w)$ for every $\boldsymbol{\theta} \in \Theta_K$. For a.e. $y \in \mathcal{S}_x(w)$, the function $\log p_x(y,w|\cdot)$ is continuous on Θ_K and satisfies

$$\sup_{\boldsymbol{\theta}\in\Theta_K} |\log p_x(y,w|\boldsymbol{\theta})| \le G_x^{(0)}(y,w), \qquad \mathbb{E}[G_x^{(0)}(Y_x,W)] < \infty, \tag{45}$$

where $G_x^{(0)}$ is measurable and nonnegative.

(iv) For every $x \in \{0,\dots,K-1\}$, P_W-almost every w, and a.e. $y \in \mathcal{S}_x(w)$, the function $\log p_x(y,w|\cdot)$ is C^2 on a neighborhood $\mathcal{N} \subset \Theta_K$ of $\boldsymbol{\theta}^0$. For $r \in \{1,2\}$,

$$\nabla_{\boldsymbol{\theta}}^r \int_{\mathcal{S}_x(w)} p_x(y,w|\boldsymbol{\theta})dy = \int_{\mathcal{S}_x(w)} \nabla_{\boldsymbol{\theta}}^r p_x(y,w|\boldsymbol{\theta})dy \tag{46}$$

holds for every $\boldsymbol{\theta} \in \mathcal{N}$. For $r \in \{1,2\}$, the norm of $\nabla_{\boldsymbol{\theta}}^r \log p_x(y,w|\boldsymbol{\theta})$ is bounded on $\mathcal{N}$ by a measurable nonnegative function $G_x^{(r)}(y,w)$ satisfying

$$\mathbb{E}[G_x^{(1)}(Y_x,W)^2] < \infty, \qquad \mathbb{E}[G_x^{(2)}(Y_x,W)] < \infty. \tag{47}$$

(v) The Fisher information

$$\mathcal{I}_K(\boldsymbol{\theta}^0) \coloneqq \sum\nolimits_{x=0}^{K-1} \pi_x \int_{\Omega_W} \int_{\mathcal{S}_x(w)} \boldsymbol{s}_x(y,w,\boldsymbol{\theta}^0)\boldsymbol{s}_x(y,w,\boldsymbol{\theta}^0)^\top p_x(y,w|\boldsymbol{\theta}^0)dy\,P_W(dw) \tag{48}$$

is positive definite.

Remark 1. *Suppose that the parameter vector can be partitioned as $\boldsymbol{\theta} = ((\boldsymbol{\theta}_0)^\top,\dots,(\boldsymbol{\theta}_{K-1})^\top)^\top$, with $\Theta_K = \Theta_0 \times \cdots \times \Theta_{K-1}$, and that $p_x(\cdot,\cdot|\boldsymbol{\theta}) = p_x(\cdot,\cdot|\boldsymbol{\theta}_x)$ for every $x \in \{0,\dots,K-1\}$. Then Assumption 4(i) holds if every Θ_x is bounded and closed, $\boldsymbol{\theta}_x^0$ is an interior point of Θ_x, and $p_x(\cdot,\cdot|\boldsymbol{\theta}_x) = p_x(\cdot,\cdot|\boldsymbol{\theta}_x^0)$ almost everywhere only if $\boldsymbol{\theta}_x = \boldsymbol{\theta}_x^0$ for every $x \in \{0,\dots,K-1\}$. Under this parameterization, Assumption 4(iii) and (iv) can be verified separately for every $x \in \{0,\dots,K-1\}$. The Fisher information in* (48) *is block diagonal. Since $\pi_x > 0$ for every $x \in \{0,\dots,K-1\}$, Assumption 4(v) holds if and only if every diagonal block is positive definite.*

Remark 2. *When W is constant, the integrals over P_W reduce to evaluation at that constant, so the formulas include the no-covariate models used in the numerical analyses. For randomized experiments with random group sizes, the same asymptotic conclusions hold if, conditional on those sizes, the samples have the laws and independence specified in Assumption 4(ii), and $n_x/n \xrightarrow{p} \pi_x \in (0,1)$. For every deterministic sequence of group sizes with these limits, we may apply the laws of large numbers and central limit theorems within each group as in the proof. For random group sizes satisfying the stated condition, we first condition on those sizes and then average over their distribution.*

E SYNTHETIC EXPERIMENTS

We examine the finite-sample behavior of the maximum likelihood plug-in estimator in Section 5 under the normal inverse Gaussian family used in the real-data analysis. The first experiment varies the sample size in each treatment group. The second fixes the true PoB and the sample size in each treatment group and examines how the variance changes with the number of treatment levels under two relations between the potential outcomes and the latent factors.

For $d > 0$ and $a > |b|$, the density of the normal inverse Gaussian distribution $\mathrm{NIG}(\xi, d, a, b)$ is

$$f(v|\xi, d, a, b) = \frac{ad}{\pi\{d^2 + (v-\xi)^2\}^{1/2}} K_1(a\{d^2 + (v-\xi)^2\}^{1/2}) \exp\Big(d\sqrt{a^2 - b^2} + b(v-\xi)\Big), \tag{49}$$

where K_1 is the modified Bessel function of the second kind. Its characteristic function is

$$\varphi(t) = \exp\Big(\mathrm{i}\xi t + d\{\sqrt{a^2 - b^2} - \sqrt{a^2 - (b + \mathrm{i}t)^2}\}\Big), \tag{50}$$

which is nonzero for every $t \in \mathbb{R}$.

In this section, we set $d = 1$, $a = 2$, and $b = 0.75$ unless stated otherwise. Only the marginal location parameters are estimated, while d, a, and b are treated as known. The location parameter space is $[-100, 100]$, and all true parameter values used below are interior points. For these parameter values, a mean-zero normal inverse Gaussian variable has skewness 0.826 and kurtosis 5.528.

Let $f(v|\xi, d, a, b)$ and $F(v|\xi, d, a, b)$ denote the density and distribution function of $\mathrm{NIG}(\xi, d, a, b)$.

Finite-sample behavior. For each $K \geq 3$, let $V_0, \ldots, V_{K-1}$ be mutually independent and suppose that $V_j \sim \mathrm{NIG}(\xi_j, d, a, b)$. We assume

$$Y_0 = V_0, \qquad Y_x = V_0 + V_x, \quad x \in \{1, \ldots, K-1\}. \tag{51}$$

The relation in (51) is invertible because $V_0 = Y_0$ and $V_x = Y_x - Y_0$ for every $x \in \{1, \ldots, K-1\}$. Treatment 1 has the largest potential outcome if and only if $V_1 > 0$ and $V_1 > V_x$ for every $x \in \{2, \ldots, K-1\}$. Conditioning on V_1 therefore implies

$$\tau_K(\boldsymbol{\xi}) = \int_0^\infty f(v|\xi_1, d, a, b) \prod_{x=2}^{K-1} F(v|\xi_x, d, a, b) dv, \tag{52}$$

where $\boldsymbol{\xi} \coloneqq (\xi_1, \ldots, \xi_{K-1})^\top$. We set $\xi_0 = \xi_2 = \cdots = \xi_{K-1} = -db/\sqrt{a^2 - b^2}$, so that $V_0, V_2, \ldots, V_{K-1}$ have mean zero, and choose ξ_1 so that $\tau_K(\boldsymbol{\xi}) = 0.25$.

By closure of the normal inverse Gaussian family under convolution,

$$Y_0 \sim \mathrm{NIG}(\eta_0, d, a, b), \qquad Y_x \sim \mathrm{NIG}(\eta_x, 2d, a, b), \quad x \in \{1, \ldots, K-1\}, \tag{53}$$

where $\eta_0 = \xi_0$ and $\eta_x = \xi_0 + \xi_x$. Let $\widehat{\eta}_x$ be the maximum likelihood estimator of η_x based on treatment group x and define

$$\widehat{\xi}_x \coloneqq \widehat{\eta}_x - \widehat{\eta}_0, \qquad \widehat{\tau}_K \coloneqq \tau_K(\widehat{\xi}_1, \ldots, \widehat{\xi}_{K-1}). \tag{54}$$

This is the maximum likelihood plug-in estimator in Section 5 for the parametric model in (53).

For the normal inverse Gaussian location family, let

$$\mathcal{I}_{\mathrm{loc}}(d, a, b) \coloneqq \int_{\mathbb{R}} \left\{ \frac{\partial}{\partial \xi} \log f(v|\xi, d, a, b) \right\}^2 f(v|\xi, d, a, b) dv. \tag{55}$$

This quantity does not depend on ξ. Let $g_j(\boldsymbol{\xi}) \coloneqq \partial \tau_K(\boldsymbol{\xi})/\partial \xi_j$. Since $\xi_j = \eta_j - \eta_0$, we have $\partial \tau_K/\partial \eta_0 = -\sum_{j=1}^{K-1} g_j(\boldsymbol{\xi})$ and $\partial \tau_K/\partial \eta_j = g_j(\boldsymbol{\xi})$. The maximum likelihood estimators of the marginal locations are independent across treatment groups. Hence, by Theorem 4, the asymptotic variance of (54), when every treatment group contains N observations, is

$$\frac{(\sum_{j=1}^{K-1} g_j(\boldsymbol{\xi}))^2}{N\mathcal{I}_{\mathrm{loc}}(d, a, b)} + \frac{\sum_{j=1}^{K-1} g_j(\boldsymbol{\xi})^2}{N\mathcal{I}_{\mathrm{loc}}(2d, a, b)}, \tag{56}$$

where the derivatives are evaluated at the true parameter.

We use $K \in \{3, 5, 10\}$ and $N \in \{20, 50, 100, 1{,}000, 10{,}000\}$. The total sample size is $n = KN$. For each setting, we generate $1{,}000$ independent datasets. The mean, variance, skewness, kurtosis, and the 2.5% and 97.5% quantiles in Table 5 are computed from the resulting $1{,}000$ estimates.

Table 5: Finite-sample behavior of the PoB estimator in (54). The true PoB is 0.25, and N is the sample size in each treatment group.

K	N	Mean	Variance	Skewness	Kurtosis	Central 95% interval
3	20	0.256	9.321×10^{-3}	0.731	3.522	[0.106, 0.479]
3	50	0.256	4.091×10^{-3}	0.441	2.866	[0.149, 0.394]
3	100	0.253	2.057×10^{-3}	0.385	3.288	[0.170, 0.354]
3	1,000	0.250	2.018×10^{-4}	0.178	3.007	[0.224, 0.280]
3	10,000	0.250	1.956×10^{-5}	-0.042	3.263	[0.241, 0.258]
5	20	0.244	7.345×10^{-3}	0.683	3.345	[0.112, 0.435]
5	50	0.248	3.211×10^{-3}	0.483	3.217	[0.152, 0.371]
5	100	0.249	1.701×10^{-3}	0.307	3.146	[0.173, 0.333]
5	1,000	0.250	1.565×10^{-4}	0.230	3.501	[0.227, 0.277]
5	10,000	0.250	1.869×10^{-5}	0.075	2.871	[0.242, 0.259]
10	20	0.243	6.312×10^{-3}	0.688	3.593	[0.114, 0.425]
10	50	0.247	2.870×10^{-3}	0.441	3.188	[0.152, 0.368]
10	100	0.250	1.287×10^{-3}	0.302	3.325	[0.185, 0.323]
10	1,000	0.250	1.325×10^{-4}	0.055	3.313	[0.227, 0.272]
10	10,000	0.250	1.272×10^{-5}	0.039	3.041	[0.243, 0.257]

For $N \geq 1{,}000$, the absolute difference between the mean estimate and 0.25 is at most 0.0003. Across the three values of K, the variance ranges from 6.312×10^{-3} to 9.321×10^{-3} when $N = 20$ and from 1.272×10^{-5} to 1.956×10^{-5} when $N = 10{,}000$. The decrease is close to the $1/N$ rate suggested by the asymptotic variance in (56), and the empirical variance is between 0.916 and 1.166 times the asymptotic variance over the 15 settings. Correspondingly, the widths of the central 95% intervals decrease from 0.310–0.373 when $N = 20$ to 0.014–0.017 when $N = 10{,}000$. The skewness ranges from 0.683 to 0.731 when $N = 20$ and from -0.042 to 0.075 when $N = 10{,}000$. For every value of N considered, the variance does not increase as K changes from 3 to 10. Figure 3 shows the corresponding distributions.

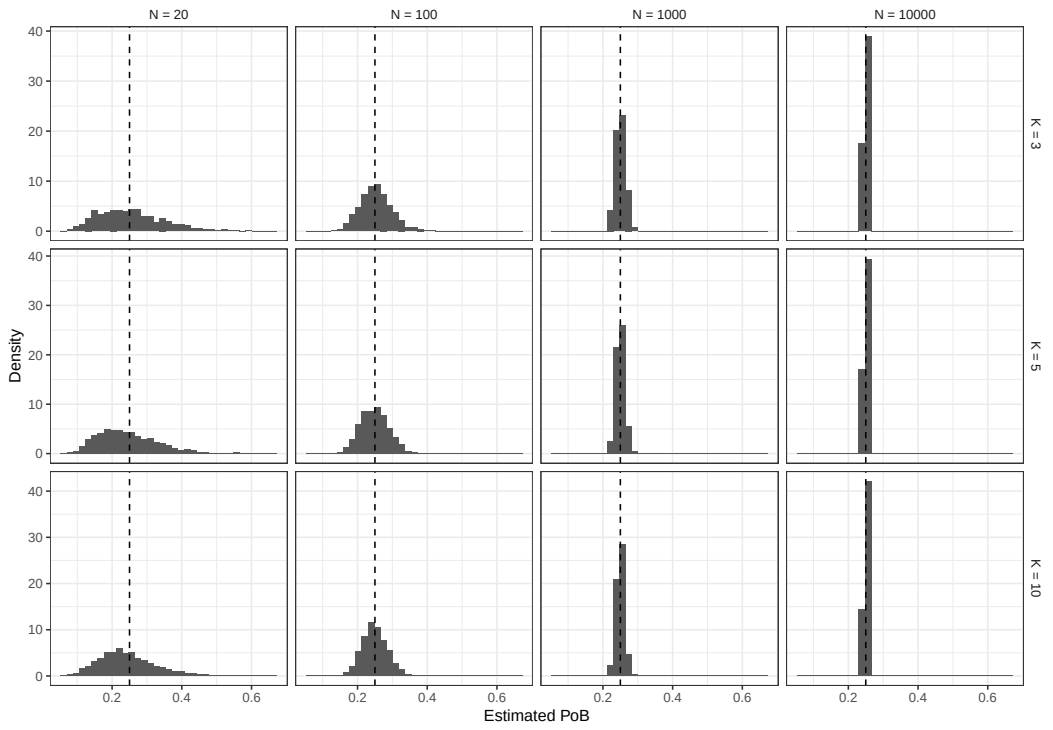


Figure 3: Distributions of the PoB estimator in (54). The dashed line is the true PoB.

Dependence on the number of treatment levels. We next examine whether the variance can increase with K. We retain the same values of d, a, and b and compare (51) with the following relation using mutually independent latent factors. For every $x \in \{0, \dots, K-1\} \setminus \{1\}$, we assume

$$Y_1 = \sum_{j=0}^{K-1} V_j, \qquad Y_x = \sum_{j=0,\, j\neq x}^{K-1} V_j. \tag{57}$$

The relation in (57) is invertible because $V_x = Y_1 - Y_x$ for every $x \neq 1$ and $V_1 = \sum_{x\neq 1} Y_x - (K-2)Y_1$. Set

$$d_K \coloneqq \frac{d}{\sqrt{K}}, \qquad a_K \coloneqq a\sqrt{K}, \qquad b_K \coloneqq b\sqrt{K}. \tag{58}$$

If $Z \sim \mathrm{NIG}(\xi, d, a, b)$, then $Z/\sqrt{K} \sim \mathrm{NIG}(\xi/\sqrt{K}, d_K, a_K, b_K)$. Hence, the variance of each latent factor is proportional to $1/K$. Let $f_K(v) \coloneqq f(v|0, d_K, a_K, b_K)$ and $F_K(v) \coloneqq F(v|0, d_K, a_K, b_K)$. For

$$p_K \coloneqq 0.25^{1/(K-1)}, \qquad \zeta_K \coloneqq -F_K^{-1}(1-p_K), \tag{59}$$

we assume $V_x \sim \mathrm{NIG}(\zeta_K, d_K, a_K, b_K)$ for every $x \neq 1$ and $V_1 \sim \mathrm{NIG}(\xi_1, d_K, a_K, b_K)$, where ξ_1 is chosen so that $\mathbb{E}[Y_1] = 0$. Then $\mathbb{P}(V_x > 0) = p_K$ for every $x \neq 1$. Since $Y_1 > Y_x$ if and only if $V_x > 0$, mutual independence implies

$$\mathbb{P}(Y_1 > Y_x \text{ for every } x \neq 1) = \prod_{x\neq 1} \mathbb{P}(V_x > 0) = p_K^{K-1} = 0.25. \tag{60}$$

Let η_1 and η_x be the location parameters of Y_1 and Y_x. By closure under convolution,

$$Y_1 \sim \mathrm{NIG}(\eta_1, Kd_K, a_K, b_K), \qquad Y_x \sim \mathrm{NIG}(\eta_x, (K-1)d_K, a_K, b_K), \quad x \neq 1. \tag{61}$$

Let $\widehat{\eta}_1$ and $\widehat{\eta}_x$ be the maximum likelihood estimators of the marginal locations. The maximum likelihood plug-in estimator is

$$\widehat{\tau}_K = \prod_{x\neq 1} \{1 - F_K(-(\widehat{\eta}_1 - \widehat{\eta}_x))\}. \tag{62}$$

Define $r_K := 0.25 f_K(-\zeta_K)/p_K$. At the true parameter, $\eta_1 - \eta_x = \zeta_K$ for every $x \neq 1$, so the derivative of the PoB with respect to η_1 is $(K-1)r_K$, and the derivative with respect to each η_x, $x \neq 1$, is $-r_K$. By a first-order Taylor expansion, for fixed K, as $N \to \infty$,

$$\widehat{\tau}_K - 0.25 = (K-1)r_K(\widehat{\eta}_1 - \eta_1) - r_K \sum_{x \neq 1} (\widehat{\eta}_x - \eta_x) + o_{\mathbb{P}}(N^{-1/2}). \tag{63}$$

The maximum likelihood estimators of the marginal locations are independent across treatment groups. By Theorem 4, the asymptotic variance of (62) is

$$\frac{(K-1)^2 r_K^2}{N\mathcal{I}_{\mathrm{loc}}(Kd_K, a_K, b_K)} + \frac{(K-1)r_K^2}{N\mathcal{I}_{\mathrm{loc}}((K-1)d_K, a_K, b_K)}. \tag{64}$$

For both (51) and (57), we use $K \in \{3, 5, 10, 20, 50, 100\}$, set $N = 1{,}000$, and generate 300 independent datasets for each setting. For (51), ξ_1 is recalibrated for every K so that the true PoB remains 0.25. Table 6 reports the full results corresponding to Table 1 in the main text, including the ratios of the empirical variances to the asymptotic variances.

Table 6: Behavior of the PoB estimator as K increases. The true PoB is 0.25, each treatment group contains $N = 1{,}000$ observations, and each entry is based on 300 independent datasets. Each ratio is the empirical variance divided by the corresponding asymptotic variance in (56) or (64).

	(51)			(57)		
K	Mean	Variance	Ratio	Mean	Variance	Ratio
3	0.251	1.884×10^{-4}	0.942	0.251	7.629×10^{-4}	0.927
5	0.250	1.511×10^{-4}	0.942	0.253	2.366×10^{-3}	1.067
10	0.251	1.311×10^{-4}	0.994	0.247	5.750×10^{-3}	0.972
20	0.249	1.262×10^{-4}	1.040	0.240	1.039×10^{-2}	0.771
50	0.250	1.180×10^{-4}	1.027	0.233	2.593×10^{-2}	0.712
100	0.249	1.115×10^{-4}	0.995	0.222	4.267×10^{-2}	0.571

The two relations show different behavior as K increases. Under (51), the variance decreases from 1.884×10^{-4} at $K = 3$ to 1.115×10^{-4} at $K = 100$. Under (57), it increases from 7.629×10^{-4} to 4.267×10^{-2}. The corresponding asymptotic variances show the same contrast, decreasing from 1.999×10^{-4} to 1.120×10^{-4} under (51) and increasing from 8.226×10^{-4} to 7.472×10^{-2} under (57).

For $K = 3$, 5, and 10, the variance ratios in Table 6 are close to 1 under both relations. At $K = 10$, the empirical variance under (57) is approximately 44 times that under (51). Thus, the asymptotic variances accurately describe substantially different levels of estimation precision at these values of K. Under (57), the larger values of K are accompanied by increasing variance, mean estimates below 0.25, and decreasing agreement with (64). Estimation accuracy and the accuracy of the asymptotic approximation therefore deteriorate together, even though the total sample size $n = KN$ increases. Figure 4 shows the empirical variances in Table 6.

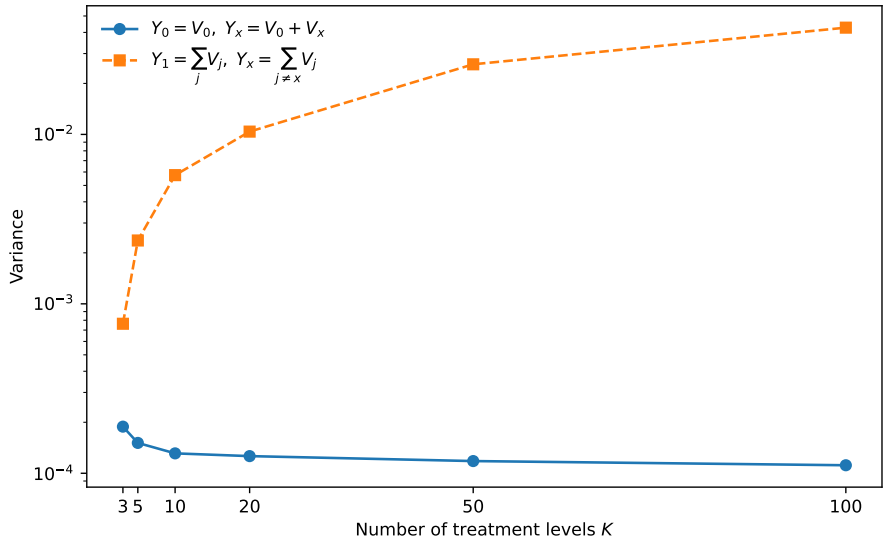


Figure 4: Variance of the PoB estimator as K increases. The solid line corresponds to (51), and the dashed line corresponds to (57). The vertical axis is logarithmic.

F Details of the Real-World Application

Data and Treatment Construction. KuaiRand-Pure records videos displayed by a randomized recommendation procedure on Kuaishou. The exact tag sets and the official content labels determine the three treatments in Table 7. Treatment $X = 0$, 1, or 2 means that the randomized recommendation procedure displays a video with the exact tag set $\{39\}$, $\{39, 43\}$, or $\{20, 43\}$, respectively. The potential outcome Y_x is the viewing time in milliseconds under treatment $X = x$. Tag 39 is shared by urban short drama and television drama, tag 43 is shared by television drama and animation, and tag 20 is specific to animation. We associate tags 39, 43, and 20 with the latent contributions U_0, U_1, and U_2 in Example 3, respectively. We use records from one recommendation interface and require positive video duration and nonnegative viewing time. Among the 17,036 users, 9,416 have more than one display from the selected video sets. We retain the first display for each user, so each user contributes one observation.

Under Assumption 3, define $V_0 := Y_0$, $V_1 := Y_1 - Y_0$, and $V_2 := Y_2 - Y_1 + Y_0$. These variables are affine transformations of the latent factors in Example 3, and

$$Y_0 = V_0, \quad Y_1 = V_0 + V_1, \quad Y_2 = V_1 + V_2. \tag{65}$$

Table 8 summarizes the untransformed viewing time. The sample distributions are strongly right-skewed, so we use a continuous distribution that permits a long upper tail.

Normal Inverse Gaussian Model. We parameterize a compatible model for the treatment-specific marginal distributions through mutually independent V_0, V_1, V_2.

Using the NIG parameterization in (49), let $V_j \sim \mathrm{NIG}(\xi_j, d_j, a, b)$ for $j \in \{0, 1, 2\}$, where $d_j > 0$ and a, b are common across j. Under Assumption 3, each V_j is a nonconstant affine transformation of U_j, so each U_j has a density. Then (65) implies

$$\begin{aligned} Y_0 &\sim \mathrm{NIG}(\xi_0, d_0, a, b), \\ Y_1 &\sim \mathrm{NIG}(\xi_0 + \xi_1, d_0 + d_1, a, b), \\ Y_2 &\sim \mathrm{NIG}(\xi_1 + \xi_2, d_1 + d_2, a, b). \end{aligned} \tag{66}$$

The marginal characteristic functions in (66) are nonzero, so Assumption 2 holds. We maximize the likelihood over $(a, b, \xi_0, \xi_1, \xi_2, d_0, d_1, d_2)$.

Table 7: Treatment definitions and analysis sample.

X	Video content	Tag set	Candidate videos	Users
0	Urban short drama	$\{39\}$	66	3,939
1	Television drama	$\{39, 43\}$	152	8,069
2	Animation	$\{20, 43\}$	89	5,028
	Total		307	17,036

Table 8: Empirical distribution of viewing time in milliseconds. The variance is the sample variance. The fraction with zero viewing time is the proportion of observations with viewing time exactly equal to zero.

Video content	Median	First and third quartiles	Variance	Fraction with zero viewing time
Urban short drama	2,187	[1,413, 4,539]	5.889×10^8	0.0033
Television drama	2,561	[1,485, 6,187]	8.253×10^8	0.0024
Animation	1,760	[1,240, 3,386]	2.866×10^8	0.0034

PoB and Bootstrap. Table 2 reports the sample mean viewing times, PoB estimates, and 95% bootstrap percentile intervals. We evaluate PoB and the pairwise probabilities using $500{,}000$ Monte Carlo draws from the fitted model. By (65), the pairwise probabilities satisfy

$$\begin{aligned} &\mathbb{P}(Y_1 > Y_0) = \mathbb{P}(V_1 > 0), \\ &\mathbb{P}(Y_2 > Y_0) = \mathbb{P}(V_1 + V_2 > V_0), \\ &\mathbb{P}(Y_2 > Y_1) = \mathbb{P}(V_2 > V_0). \end{aligned} \tag{67}$$

The estimates are reported in Table 3. Each probability in (67) compares two potential outcomes, whereas the PoB requires one potential outcome to exceed both alternatives. Since the joint distribution has a density,

$$\begin{aligned} &\mathbb{P}(\min(Y_0, Y_1) < Y_2 < \max(Y_0, Y_1)) \\ &= \mathbb{P}(Y_2 > Y_0) + \mathbb{P}(Y_2 > Y_1) - 2\mathbb{P}(Y_2 > \max(Y_0, Y_1)). \end{aligned} \tag{68}$$

The estimated probabilities of the best, intermediate, and worst outcomes are reported in Table 9. Animation has the largest estimated probabilities of both the best and the worst outcomes and the smallest probability of an intermediate outcome.

We use $1{,}000$ nonparametric bootstrap samples within the three treatment groups. For each bootstrap sample, all normal inverse Gaussian parameters and the three PoB values are recomputed. All $1{,}000$ estimates satisfy the parameter constraints and are included in the percentile intervals.

Comparison under Rank Preservation. We use the same observations to estimate PoB and pairwise probabilities by reference-group sample averages based on the rank-preservation identification result of Kawakami & Tian (2026). Let $\widehat{F}_x$ denote the empirical cumulative distribution function in treatment group x, and let $\widehat{Q}_x(u) \coloneqq \inf\{y \in \mathbb{R} | \widehat{F}_x(y) \geq u\}$ for $0 < u \leq 1$. For each treatment r, we use

$$\widehat{p}_r \coloneqq \frac{1}{n_r} \sum_{i=1}^{n_r} \mathbb{I}\{Y^{(r,i)} > \max_{x \neq r} \widehat{Q}_x(\widehat{F}_r(Y^{(r,i)}))\}. \tag{69}$$

For $\mathbb{P}(Y_r > Y_s)$, the indicator compares $Y^{(r,i)}$ with $\widehat{Q}_s(\widehat{F}_r(Y^{(r,i)}))$. We evaluate the empirical CDF at each observed value, including repeated values, and retain strict inequalities in the comparisons. The 95% PoB intervals in Table 2 use the 2.5% and 97.5% quantiles of $1{,}000$ bootstrap estimates, with resampling performed independently within each treatment group and the empirical distributions reestimated in every resample.

Sensitivity Analysis. Write α_0, α_1, and α_2 for the reference loadings of U_0, U_1, and U_2. With $\lambda_0 = 1$, we use

$$\begin{aligned} Y_0 &= \mu_0 + \alpha_0 U_0, \\ Y_1 &= \mu_1 + \lambda_1(\alpha_0 U_0 + \alpha_1 U_1), \\ Y_2 &= \mu_2 + \lambda_2(\alpha_1 U_1 + \alpha_2 U_2). \end{aligned} \tag{70}$$

Table 9: Estimated probabilities of the best, intermediate, and worst potential outcomes.

Video content	Best	Intermediate	Worst
Urban short drama	0.304	0.436	0.261
Television drama	0.327	0.492	0.180
Animation	0.369	0.072	0.559

Table 10: Differences between the estimated PoB values. The intervals are 95% bootstrap percentile intervals.

Difference	Estimate	95% interval
Animation minus urban short drama	0.065	[0.008, 0.130]
Animation minus television drama	0.042	[-0.021, 0.103]
Television drama minus urban short drama	0.023	[-0.082, 0.146]

The loading ratio for U_0 between Y_1 and Y_0 is λ_1, and the loading ratio for U_1 between Y_2 and Y_1 is λ_2/λ_1. The value $(\lambda_1, \lambda_2) = (1, 1)$ corresponds to Assumption 3.

For each pair $(\lambda_1, \lambda_2) \in \{0.80, 0.85, \ldots, 1.20\}^2$, we set $Z^{(i)} := Y^{(i)}/\lambda_{X^{(i)}}$ and reestimate the normal inverse Gaussian parameters from the transformed observations by maximum likelihood as in Section 5. The observed sample is the same at every grid point, and the fitted marginal distributions may vary with (λ_1, λ_2).

Under (70) and Assumption 1, the component variances satisfy

$$\mathrm{Var}(\alpha_0 U_0) = v_0, \qquad \mathrm{Var}(\alpha_1 U_1) = \frac{v_1}{\lambda_1^2} - v_0, \qquad \mathrm{Var}(\alpha_2 U_2) = \frac{v_2}{\lambda_2^2} - \frac{v_1}{\lambda_1^2} + v_0, \tag{71}$$

where $v_x := \mathrm{Var}(Y_x)$. For

$$\widehat{v}_x := \frac{1}{n_x} \sum_{i:X^{(i)}=x} (Y^{(i)} - \overline{Y}_x)^2, \tag{72}$$

we record whether substituting the sample variances into (71) yields positive component variances, that is, whether

$$\widehat{v}_0 > 0, \qquad \frac{\widehat{v}_1}{\lambda_1^2} - \widehat{v}_0 > 0, \qquad \frac{\widehat{v}_2}{\lambda_2^2} - \frac{\widehat{v}_1}{\lambda_1^2} + \widehat{v}_0 > 0. \tag{73}$$

We reestimate the NIG model and evaluate PoB at all 81 grid points regardless of these sample-variance inequalities, and numerical optimization succeeds at every grid point.

Among the 81 pairs, 38 satisfy (73), as shown in Table 11, and among these 38 grid points, urban short drama, animation, and television drama have the largest estimated PoB at 4, 7, and 27 points, respectively.

Tables 12 and 13 report the PoB estimates at every grid point with successful numerical optimization.

Table 14 summarizes the range of the PoB estimates over the grid points with successful numerical optimization.

Figure 5 shows the treatment with the largest estimated PoB. For each fixed value of λ_1, this treatment is unchanged over the evaluated values of λ_2: urban short drama has the largest PoB for $\lambda_1 \in \{0.80, 0.85, 0.90, 0.95\}$, animation for $\lambda_1 = 1$, and television drama for $\lambda_1 \in \{1.05, 1.10, 1.15, 1.20\}$. At $\lambda_2 = 1$, for example, the treatment with the largest PoB changes from urban short drama to animation and then to television drama as λ_1 changes from 0.95 to 1.00 and then to 1.05, whereas the estimated PoB of animation remains between 0.369 and 0.371. Across all 81 grid points, the estimated PoB of animation ranges from 0.329 to 0.397, compared with 0.012 to 0.519 for urban short drama and 0.091 to 0.596 for television drama (Table 14). Therefore, the changes in the treatment with the largest PoB are driven mainly by the much larger variation in the estimated PoB values of urban short drama and television drama. Among the shared-factor loadings, λ_1 changes both the loading ratio for U_0 and the ratio λ_2/λ_1 for U_1, whereas changing λ_2 changes only the latter.

Table 11: Grid points satisfying the sample-variance inequalities in (73).

$\lambda_1 \backslash \lambda_2$	0.80	0.85	0.90	0.95	1.00	1.05	1.10	1.15	1.20
0.80	–	–	–	–	–	–	–	–	–
0.85	–	–	–	–	–	–	–	–	–
0.90	✓	–	–	–	–	–	–	–	–
0.95	✓	✓	✓	–	–	–	–	–	–
1.00	✓	✓	✓	✓	✓	✓	✓	–	–
1.05	✓	✓	✓	✓	✓	✓	✓	✓	✓
1.10	✓	✓	✓	✓	✓	✓	✓	✓	✓
1.15	✓	✓	✓	✓	✓	✓	✓	✓	✓
1.20	–	–	–	–	–	–	–	–	–

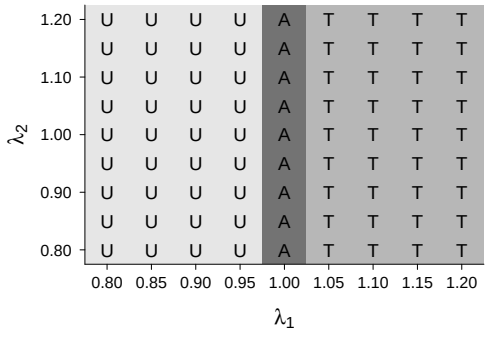


Figure 5: Treatment with the largest estimated PoB. U, T, and A denote urban short drama, television drama, and animation.

Table 12: PoB estimates for $\lambda_1 \in \{0.80, 0.85, 0.90, 0.95, 1.00\}$. The third column reports λ_2/λ_1.

λ_1	λ_2	λ_2/λ_1	Urban short drama	Television drama	Animation
0.80	0.80	1.000	0.488	0.177	0.335
0.80	0.85	1.062	0.494	0.150	0.357
0.80	0.90	1.125	0.496	0.139	0.365
0.80	0.95	1.187	0.502	0.127	0.371
0.80	1.00	1.250	0.503	0.118	0.378
0.80	1.05	1.312	0.508	0.110	0.382
0.80	1.10	1.375	0.511	0.103	0.386
0.80	1.15	1.437	0.516	0.098	0.386
0.80	1.20	1.500	0.519	0.091	0.390
0.85	0.80	0.941	0.477	0.194	0.329
0.85	0.85	1.000	0.479	0.178	0.343
0.85	0.90	1.059	0.486	0.156	0.358
0.85	0.95	1.118	0.487	0.144	0.369
0.85	1.00	1.176	0.491	0.135	0.373
0.85	1.05	1.235	0.493	0.131	0.376
0.85	1.10	1.294	0.499	0.122	0.379
0.85	1.15	1.353	0.502	0.116	0.382
0.85	1.20	1.412	0.504	0.111	0.384
0.90	0.80	0.889	0.459	0.203	0.337
0.90	0.85	0.944	0.460	0.195	0.345
0.90	0.90	1.000	0.465	0.185	0.350
0.90	0.95	1.056	0.467	0.166	0.366
0.90	1.00	1.111	0.474	0.157	0.369
0.90	1.05	1.167	0.473	0.147	0.380
0.90	1.10	1.222	0.475	0.145	0.380
0.90	1.15	1.278	0.482	0.137	0.381
0.90	1.20	1.333	0.483	0.133	0.384
0.95	0.80	0.842	0.429	0.222	0.349
0.95	0.85	0.895	0.433	0.220	0.347
0.95	0.90	0.947	0.436	0.213	0.350
0.95	0.95	1.000	0.440	0.203	0.357
0.95	1.00	1.053	0.443	0.186	0.371
0.95	1.05	1.105	0.445	0.179	0.376
0.95	1.10	1.158	0.447	0.174	0.379
0.95	1.15	1.211	0.447	0.171	0.381
0.95	1.20	1.263	0.451	0.166	0.384
1.00	0.80	0.800	0.298	0.344	0.358
1.00	0.85	0.850	0.298	0.341	0.361
1.00	0.90	0.900	0.299	0.340	0.361
1.00	0.95	0.950	0.303	0.335	0.361
1.00	1.00	1.000	0.304	0.327	0.369
1.00	1.05	1.050	0.307	0.316	0.377
1.00	1.10	1.100	0.309	0.312	0.379
1.00	1.15	1.150	0.309	0.307	0.384
1.00	1.20	1.200	0.312	0.301	0.388

Table 13: PoB estimates for $\lambda_1 \in \{1.05, 1.10, 1.15, 1.20\}$. The third column reports λ_2/λ_1.

λ_1	λ_2	λ_2/λ_1	Urban short drama	Television drama	Animation
1.05	0.80	0.762	0.111	0.520	0.369
1.05	0.85	0.810	0.112	0.516	0.372
1.05	0.90	0.857	0.110	0.519	0.371
1.05	0.95	0.905	0.112	0.518	0.370
1.05	1.00	0.952	0.114	0.515	0.371
1.05	1.05	1.000	0.112	0.512	0.377
1.05	1.10	1.048	0.113	0.506	0.380
1.05	1.15	1.095	0.114	0.503	0.383
1.05	1.20	1.143	0.113	0.501	0.386
1.10	0.80	0.727	0.059	0.563	0.378
1.10	0.85	0.773	0.060	0.562	0.378
1.10	0.90	0.818	0.059	0.564	0.377
1.10	0.95	0.864	0.058	0.560	0.381
1.10	1.00	0.909	0.059	0.563	0.378
1.10	1.05	0.955	0.059	0.563	0.378
1.10	1.10	1.000	0.059	0.559	0.382
1.10	1.15	1.045	0.058	0.556	0.386
1.10	1.20	1.091	0.059	0.555	0.385
1.15	0.80	0.696	0.032	0.580	0.388
1.15	0.85	0.739	0.031	0.580	0.389
1.15	0.90	0.783	0.031	0.584	0.384
1.15	0.95	0.826	0.029	0.583	0.388
1.15	1.00	0.870	0.030	0.584	0.386
1.15	1.05	0.913	0.030	0.584	0.386
1.15	1.10	0.957	0.029	0.582	0.389
1.15	1.15	1.000	0.029	0.586	0.384
1.15	1.20	1.043	0.029	0.581	0.391
1.20	0.80	0.667	0.014	0.590	0.397
1.20	0.85	0.708	0.014	0.591	0.395
1.20	0.90	0.750	0.014	0.590	0.395
1.20	0.95	0.792	0.015	0.595	0.390
1.20	1.00	0.833	0.013	0.594	0.393
1.20	1.05	0.875	0.013	0.596	0.392
1.20	1.10	0.917	0.012	0.596	0.392
1.20	1.15	0.958	0.013	0.596	0.391
1.20	1.20	1.000	0.013	0.596	0.391

Table 14: Minimum and maximum PoB over all 81 grid points.

Video content	Minimum	(λ_1, λ_2)	Maximum	(λ_1, λ_2)
Urban short drama	0.012	$(1.20, 1.10)$	0.519	$(0.80, 1.20)$
Television drama	0.091	$(0.80, 1.20)$	0.596	$(1.20, 1.15)$
Animation	0.329	$(0.85, 0.80)$	0.397	$(1.20, 0.80)$

G PROOFS

We provide proofs of all results stated in the main text.

Justification of the statements following Definition 1. Fix $w \in \Omega_W$ and a representation $Y = h(X, W, \overline{U})$ that is saturated given $W = w$. In this proof, we condition on $W = w$. For any SCM with outcome equation $Y \coloneqq f_Y(X, W, U_Y)$, by Definition 1, there exists a random vector $\overline{U}$ such that $h(x, w, \overline{U}) = f_Y(x, w, U_Y)$ almost surely for every $x \in \{0, \dots, K-1\}$. Since there are finitely many treatment levels, these equalities hold simultaneously with probability one. Consequently,

$$(h(0, w, \overline{U}), \dots, h(K-1, w, \overline{U}))^\top = (Y_0, \dots, Y_{K-1})^\top \tag{74}$$

holds almost surely, so the joint distribution of all potential outcomes is reproduced.

Every probability distribution Q on Ω_Y^K can occur as a joint potential-outcome distribution in an SCM. To see this, take $U_Y = (Z_0, \dots, Z_{K-1})^\top$ with conditional distribution Q given $W = w$ and define $f_Y(x, W, U_Y) = Z_x$ for every treatment level x. The potential outcomes are then $Y_x = Z_x$, so their conditional joint distribution is Q. By the preceding argument, the saturated representation reproduces Q.

Now let $Q_0, \dots, Q_{K-1}$ be arbitrary probability distributions on Ω_Y. In the same construction, choose $Q = Q_0 \otimes \cdots \otimes Q_{K-1}$ and take X to be independent of U_Y given $W = w$, with $\mathbb{P}(X = x | W = w) > 0$ for every x. Since $h(x, w, \overline{U}) = f_Y(x, w, U_Y)$ holds almost surely for every $x \in \{0, \dots, K-1\}$, we have $h(X, w, \overline{U}) = f_Y(X, w, U_Y)$ almost surely. Thus, for every Borel set $C \subseteq \Omega_Y$,

$$\begin{aligned} \mathbb{P}(h(X, w, \overline{U}) \in C | X = x, W = w) &= \mathbb{P}(Z_x \in C | X = x, W = w) \\ &= Q_x(C). \end{aligned} \tag{75}$$

Hence the saturated representation is compatible with arbitrary treatment-specific observed marginal distributions. □

Theorem 1 (Saturation)**.** *A model of the form* (9) *is saturated given $W = w$ if and only if* $\operatorname{rank}(B(w)) = K$.

Proof. Fix $w \in \Omega_W$ and write $B \coloneqq B(w)$ and $\boldsymbol{\mu} \coloneqq \boldsymbol{\mu}(w)$. Suppose first that $\operatorname{rank}(B) = K$. Then there exists $R \in \mathbb{R}^{J \times K}$ such that $BR = I_K$. For an arbitrary SCM $Y \coloneqq f_Y(X, W, U_Y)$, define

$$\boldsymbol{Y}^f(w) \coloneqq (f_Y(0, w, U_Y), \dots, f_Y(K-1, w, U_Y))^\top \tag{76}$$

and set

$$\boldsymbol{U} \coloneqq R(\boldsymbol{Y}^f(w) - \boldsymbol{\mu}). \tag{77}$$

Since $BR = I_K$,

$$\boldsymbol{\mu} + B\boldsymbol{U} = \boldsymbol{\mu} + BR(\boldsymbol{Y}^f(w) - \boldsymbol{\mu}) = \boldsymbol{Y}^f(w) \tag{78}$$

holds almost surely. Therefore, for every $x \in \{0, \dots, K-1\}$,

$$\mu_x(w) + \sum\nolimits_{j=0}^{J-1} b_{xj}(w) U_j = f_Y(x, w, U_Y) \tag{79}$$

holds almost surely. Hence the model is saturated given $W = w$ by Definition 1.

Conversely, suppose that $\operatorname{rank}(B) < K$. Then $\boldsymbol{\mu} + \{B\boldsymbol{u} \mid \boldsymbol{u} \in \mathbb{R}^J\}$ is a proper affine subspace of $\mathbb{R}^K$. Choose $\boldsymbol{y}^* \in \mathbb{R}^K \setminus (\boldsymbol{\mu} + \{B\boldsymbol{u} \mid \boldsymbol{u} \in \mathbb{R}^J\})$ and consider the SCM defined by $f_Y(x, W, U_Y) = y_x^*$ for every $x \in \{0, \dots, K-1\}$. For every random vector $\boldsymbol{U} \in \mathbb{R}^J$,

$$\boldsymbol{\mu} + B\boldsymbol{U} \in \boldsymbol{\mu} + \{B\boldsymbol{u} \mid \boldsymbol{u} \in \mathbb{R}^J\} \tag{80}$$

holds almost surely. Hence $\boldsymbol{\mu} + B\boldsymbol{U} \neq \boldsymbol{y}^*$ almost surely. Therefore the model is not saturated given $W = w$. □

Theorem 2. *Fix $w \in \Omega_W$. Suppose that $J > K$ and the columns of M are distinct and nonzero. There exist two models of the form* (9) *with the same $\boldsymbol{\mu}(w)$ and $B(w) = M$ that satisfy Assumptions 1, 2, and 3 and induce the same conditional marginal distributions of $Y_0, \dots, Y_{K-1}$ given $W = w$ but different conditional joint distributions of $\boldsymbol{Y}$ given $W = w$.*

Proof. We prove the conclusion under the weaker condition $\operatorname{rank}(M) < J$, which holds whenever $J > K$ because $\operatorname{rank}(M) \leq K$. Let $\odot$ denote componentwise multiplication, let $\boldsymbol{r}_x = (m_{x0}, m_{x1}, \ldots, m_{x,J-1}) \in \mathbb{R}^J$ be the transpose of the xth row of M, let $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^J$ be the vector of ones, and define

$$\mathcal{R} \coloneqq \{M^\top \boldsymbol{a} | \boldsymbol{a} \in \mathbb{R}^K\}. \tag{81}$$

We first show that there exists $\boldsymbol{h} \in \ker(M)$ such that

$$M \operatorname{diag}(\boldsymbol{h}) M^\top \neq O. \tag{82}$$

Suppose to the contrary that $M \operatorname{diag}(\boldsymbol{h}) M^\top = O$ for every $\boldsymbol{h} \in \ker(M)$. Then, for every $x, y \in \{0, \ldots, K-1\}$ and $\boldsymbol{h} \in \ker(M)$,

$$0 = (M \operatorname{diag}(\boldsymbol{h}) M^\top)_{xy} = \sum\nolimits_{j=0}^{J-1} m_{xj} m_{yj} h_j = \langle \boldsymbol{r}_x \odot \boldsymbol{r}_y, \boldsymbol{h} \rangle. \tag{83}$$

Hence $\boldsymbol{r}_x \odot \boldsymbol{r}_y \in (\ker(M))^\perp = \mathcal{R}$ for every $x, y \in \{0, \ldots, K-1\}$. For $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{R}$, take $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{R}^K$ satisfying $\boldsymbol{a} = M^\top \boldsymbol{c}$ and $\boldsymbol{b} = M^\top \boldsymbol{d}$. Then

$$\boldsymbol{a} \odot \boldsymbol{b} = \sum\nolimits_{x=0}^{K-1} \sum\nolimits_{y=0}^{K-1} c_x d_y (\boldsymbol{r}_x \odot \boldsymbol{r}_y) \in \mathcal{R}. \tag{84}$$

Thus $\mathcal{R}$ is closed under componentwise multiplication.

We show $\mathbf{1} \in \mathcal{R}$. For each $j \in \{0, \ldots, J-1\}$, let $S_j \coloneqq \{x \in \{0, \ldots, K-1\} | m_{xj} = 1\}$. Since every column of M is nonzero, $S_j \neq \emptyset$. By inclusion-exclusion, the jth component of

$$\sum\nolimits_{\emptyset \neq S \subseteq \{0,\ldots,K-1\}} (-1)^{|S|+1} \bigodot\nolimits_{x \in S} \boldsymbol{r}_x \tag{85}$$

is

$$\sum\nolimits_{\emptyset \neq S \subseteq S_j} (-1)^{|S|+1} = 1. \tag{86}$$

Every term in (85) belongs to $\mathcal{R}$, so $\mathbf{1} \in \mathcal{R}$.

For each $j \in \{0, \ldots, J-1\}$, define

$$\boldsymbol{q}_j \coloneqq \bigodot\nolimits_{x=0}^{K-1} \{m_{xj} \boldsymbol{r}_x + (1 - m_{xj})(\mathbf{1} - \boldsymbol{r}_x)\}. \tag{87}$$

Each factor in (87) belongs to $\mathcal{R}$. Hence $\boldsymbol{q}_j \in \mathcal{R}$. For every $\ell \in \{0, \ldots, J-1\}$, $(\boldsymbol{q}_j)_\ell = 1$ if $m_{x\ell} = m_{xj}$ for every $x \in \{0, \ldots, K-1\}$ and $(\boldsymbol{q}_j)_\ell = 0$ otherwise. The columns of M are distinct, so $\boldsymbol{q}_j$ is the jth standard basis vector of $\mathbb{R}^J$. Therefore $\mathcal{R} = \mathbb{R}^J$, which implies

$$J = \dim(\mathcal{R}) = \operatorname{rank}(M) \leq K. \tag{88}$$

This contradicts $\operatorname{rank}(M) < J$. Thus a vector $\boldsymbol{h} \in \ker(M)$ satisfying (82) exists.

Choose $c > \max_{0 \leq j \leq J-1} |h_j|$ and define $v_j \coloneqq c$ and $v'_j \coloneqq c + h_j$ for every $j \in \{0, \ldots, J-1\}$. Then $v_j > 0$ and $v'_j > 0$ for every j. Consider two models with the same $\boldsymbol{\mu}(w)$ and M, with $\alpha_{xj}(w) = 1$ for every $x \in A_j$, in which the latent factors are mutually independent and satisfy $U_j | W = w \sim N(0, v_j)$ and $U'_j | W = w \sim N(0, v'_j)$. Let

$$\begin{aligned} \boldsymbol{Y} &\coloneqq \boldsymbol{\mu}(w) + M\boldsymbol{U}, \\ \boldsymbol{Y}' &\coloneqq \boldsymbol{\mu}(w) + M\boldsymbol{U}'. \end{aligned} \tag{89}$$

The latent factors have densities, the conditional characteristic functions of all potential outcomes are nonzero everywhere, and the loadings are treatment-invariant. Thus both models satisfy Assumptions 1, 2, and 3.

For every $x \in \{0, \ldots, K-1\}$, $Y_x | W = w$ and $Y'_x | W = w$ are Gaussian with the same mean $\mu_x(w)$. Since $m_{xj}^2 = m_{xj}$,

$$\begin{aligned} \operatorname{Var}(Y'_x | W = w) - \operatorname{Var}(Y_x | W = w) &= \sum\nolimits_{j=0}^{J-1} m_{xj}^2 (v'_j - v_j) \\ &= \sum\nolimits_{j=0}^{J-1} m_{xj} h_j \\ &= (M\boldsymbol{h})_x \\ &= 0. \end{aligned} \tag{90}$$

Thus the conditional marginal distributions of Y_x and Y'_x given $W = w$ coincide for every $x \in \{0, \dots, K-1\}$. On the other hand,

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{Y}'|W=w) - \mathrm{Cov}(\boldsymbol{Y}|W=w) &= M\,\mathrm{diag}(v'_0 - v_0, \dots, v'_{J-1} - v_{J-1})M^\top \\ &= M\,\mathrm{diag}(\boldsymbol{h})M^\top \\ &\neq O\end{aligned} \tag{91}$$

by (82). Therefore the conditional joint distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ given $W = w$ differ. □

Theorem 3 (Identification under treatment-invariant loadings)**.** *Fix $w \in \Omega_W$. Consider a model determined by* (9) *with specified M and $J \leq K$ that satisfies Assumptions 1, 2, and 3. Suppose that $\mathrm{rank}(B(w)) = J$ so that M has full column rank. Then the conditional marginal distributions of $Y_0, \dots, Y_{K-1}$ given $W = w$ identify the conditional joint distribution of $\boldsymbol{Y}$ given $W = w$. More precisely, the following statements hold.*

(i): For a matrix $L \in \mathbb{R}^{J\times K}$ with $LM = I_J$, define

$$\boldsymbol{V} \coloneqq L\boldsymbol{Y}, \qquad \boldsymbol{\delta}(w) \coloneqq (I_K - ML)\boldsymbol{\mu}(w). \tag{11}$$

For each $j \in \{0, \dots, J-1\}$, let $q_j(\cdot|w)$ be the conditional density of V_j given $W = w$. Let $\boldsymbol{\ell}(t|w) \coloneqq (\ell_0(t|w), \dots, \ell_{K-1}(t|w))^\top$, where $\ell_x(\cdot|w)$ is the continuous logarithm of $\varphi_x(\cdot|w)$ satisfying $\ell_x(0|w) = 0$. Then $\boldsymbol{\delta}(w)$ is identified as

$$\boldsymbol{\delta}(w) = \frac{1}{\mathrm{i}t}(I_K - ML)\boldsymbol{\ell}(t|w), \qquad \text{for any } t \neq 0. \tag{12}$$

$q_j(\cdot|w)$ is also identified and, if $\psi_j(\cdot|w) \coloneqq \exp((L\boldsymbol{\ell}(\cdot|w))_j) \in L^1(\mathbb{R})$, it is represented as

$$q_j(v|w) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp(-\mathrm{i}tv)\psi_j(t|w)dt \qquad \text{for Lebesgue-a.e. } v \in \mathbb{R}. \tag{13}$$

(ii): $V_0, \dots, V_{J-1}$ are mutually independent given $W = w$, and

$$\boldsymbol{Y} = \boldsymbol{\delta}(w) + M\boldsymbol{V} \tag{14}$$

holds given $W = w$[2]. Consequently,

$$\mathbb{E}[g(Y_0, \dots, Y_{K-1})|W = w] = \int_{\mathbb{R}^J} g(\boldsymbol{\delta}(w) + M\boldsymbol{v}) \prod_{j=0}^{J-1} q_j(v_j|w)d\boldsymbol{v}. \tag{15}$$

(iii): If $J = K$, then $L = M^{-1}$, $\boldsymbol{\delta}(w) = \boldsymbol{0}$, and the conditional joint density of $\boldsymbol{Y}$ is identified as

$$p_{\boldsymbol{Y}|W=w}(\boldsymbol{y}) = |\det(M^{-1})| \prod_{j=0}^{K-1} q_j((M^{-1}\boldsymbol{y})_j|w) \qquad \text{for } \boldsymbol{y} \in \mathbb{R}^K. \tag{16}$$

Proof. In this proof, we suppose that all distributions and characteristic functions are conditional on $W = w$.

Let $D_\alpha(w) \coloneqq \mathrm{diag}(\alpha_0(w), \dots, \alpha_{J-1}(w))$. By Assumption 3,

$$B(w) = MD_\alpha(w). \tag{92}$$

Then $J = \mathrm{rank}(B(w)) \leq \min\{\mathrm{rank}(M), \mathrm{rank}(D_\alpha(w))\} \leq J$. Hence $\mathrm{rank}(M) = \mathrm{rank}(D_\alpha(w)) = J$. Since $J \leq K$, the equality $\mathrm{rank}(M) = J$ means that M has full column rank. Moreover, $\alpha_j(w) \neq 0$ for every $j \in \{0, \dots, J-1\}$. Therefore, there exists $L \in \mathbb{R}^{J\times K}$ satisfying $LM = I_J$. By (10), (92), and (11),

$$\boldsymbol{V} = L\boldsymbol{\mu}(w) + D_\alpha(w)\boldsymbol{U}. \tag{93}$$

Thus, $V_j = (L\boldsymbol{\mu}(w))_j + \alpha_j(w)U_j$ for every $j \in \{0, \dots, J-1\}$. Each V_j is a nonconstant affine transformation of U_j. Then, Assumption 1 implies that $V_0, \dots, V_{J-1}$ are mutually independent, and Assumption 2 implies that each V_j has a density. Moreover,

$$\begin{aligned}\boldsymbol{\delta}(w) + M\boldsymbol{V} &= (I_K - ML)\boldsymbol{\mu}(w) + ML\boldsymbol{\mu}(w) + MD_\alpha(w)\boldsymbol{U} \\ &= \boldsymbol{\mu}(w) + B(w)\boldsymbol{U} \\ &= \boldsymbol{Y},\end{aligned} \tag{94}$$

[2]The value on the right-hand side of (15) is the same for every L satisfying $LM = I_J$. The values $\mu_x(w)$, the loadings $\alpha_j(w)$, and the conditional distributions of U_j given $W = w$ are not identified separately.

which proves (14).

Let $\widetilde{L} \in \mathbb{R}^{J\times K}$ be another matrix satisfying $\widetilde{L}M = I_J$, and define $\widetilde{\boldsymbol{V}} \coloneqq \widetilde{L}\boldsymbol{Y}$ and $\widetilde{\boldsymbol{\delta}}(w) \coloneqq (I_K - M\widetilde{L})\boldsymbol{\mu}(w)$. Set $\boldsymbol{c} \coloneqq (\widetilde{L} - L)\boldsymbol{\mu}(w)$. Since $\boldsymbol{Y} = \boldsymbol{\mu}(w) + MD_\alpha(w)\boldsymbol{U}$ and $(\widetilde{L} - L)M = O$, we see

$$\widetilde{\boldsymbol{V}} = \widetilde{L}\boldsymbol{Y} = L\boldsymbol{Y} + (\widetilde{L} - L)\boldsymbol{\mu}(w) = \boldsymbol{V} + \boldsymbol{c}. \tag{95}$$

and

$$\widetilde{\boldsymbol{\delta}}(w) = (I_K - M\widetilde{L})\boldsymbol{\mu}(w) = \boldsymbol{\delta}(w) - M\boldsymbol{c}. \tag{96}$$

Then it follows that

$$\widetilde{\boldsymbol{\delta}}(w) + M\widetilde{\boldsymbol{V}} = \boldsymbol{\delta}(w) - M\boldsymbol{c} + M(\boldsymbol{V} + \boldsymbol{c}) = \boldsymbol{\delta}(w) + M\boldsymbol{V}. \tag{97}$$

Thus $\widetilde{\boldsymbol{\delta}}(w) + M\widetilde{\boldsymbol{V}} = \boldsymbol{\delta}(w) + M\boldsymbol{V} = \boldsymbol{Y}$, and the value in (15) does not depend on the choice of L.

Let $\psi_j(t|w) \coloneqq \mathbb{E}[\exp(\mathrm{i}tV_j)|W = w]$. By (14), $Y_x = \delta_x(w) + \sum_{j=0}^{J-1} m_{xj}V_j$ for every $x \in \{0, \dots, K-1\}$. Since $V_0, \dots, V_{J-1}$ are mutually independent given $W = w$,

$$\varphi_x(t|w) = \exp(\mathrm{i}t\delta_x(w)) \prod\nolimits_{j=0}^{J-1} \psi_j(t|w)^{m_{xj}} \tag{98}$$

holds for every $x \in \{0, \dots, K-1\}$ and $t \in \mathbb{R}$.

Every column of M contains a nonzero entry. Suppose that $\psi_j(t|w) = 0$ for some j and t. Since $\mathrm{rank}(M) = J$, there exists x such that $m_{xj} = 1$. Then the right-hand side of (98) is zero, which contradicts Assumption 2. Therefore $\psi_j(t|w) \neq 0$ for every j and t.

For each x, let $\ell_x(\cdot|w)$ be the continuous logarithm of $\varphi_x(\cdot|w)$ satisfying $\ell_x(0|w) = 0$. For each j, let $\eta_j(\cdot|w)$ be the continuous logarithm of $\psi_j(\cdot|w)$ satisfying $\eta_j(0|w) = 0$. Define $\boldsymbol{\ell}(t|w) \coloneqq (\ell_0(t|w), \dots, \ell_{K-1}(t|w))^\top$ and $\boldsymbol{\eta}(t|w) \coloneqq (\eta_0(t|w), \dots, \eta_{J-1}(t|w))^\top$. Taking the continuous logarithms in (98), for every x and t we have

$$\ell_x(t|w) - \mathrm{i}t\delta_x(w) - \sum\nolimits_{j=0}^{J-1} m_{xj}\eta_j(t|w) \in 2\pi\mathrm{i}\mathbb{Z}. \tag{99}$$

The left-hand side is continuous in t and equals zero at $t = 0$. Hence

$$\boldsymbol{\ell}(t|w) = \mathrm{i}t\boldsymbol{\delta}(w) + M\boldsymbol{\eta}(t|w) \tag{100}$$

holds for every $t \in \mathbb{R}$. Since $LM = I_J$ and $L\boldsymbol{\delta}(w) = \boldsymbol{0}$,

$$\psi_j(t|w) = \exp((L\boldsymbol{\ell}(t|w))_j) \tag{101}$$

holds for every j and t.

The conditional marginal distributions of $Y_0, \dots, Y_{K-1}$ identify $\varphi_x(\cdot|w)$ and hence the continuous logarithm $\ell_x(\cdot|w)$ satisfying $\ell_x(0|w) = 0$ for every $x \in \{0, \dots, K-1\}$. Therefore (101) identifies $\psi_j(\cdot|w)$ for every $j \in \{0, \dots, J-1\}$. Since a characteristic function uniquely determines its distribution, the conditional distribution of every V_j is identified. $q_j(\cdot|w)$ is identified for every $j \in \{0, \dots, J-1\}$ because every V_j has a conditional density.

If $\psi_j(\cdot|w) \in L^1(\mathbb{R})$, then the integral in (13) is absolutely convergent because $|\exp(-\mathrm{i}tv)\psi_j(t|w)| = |\psi_j(t|w)|$. By the Fourier inversion theorem (Durrett, 2019, Theorem 3.3.14), this integral divided by 2π is a continuous version of the conditional density of V_j given $W = w$. Hence (13) holds for Lebesgue-almost every $v \in \mathbb{R}$.

Moreover, substituting $\boldsymbol{\eta}(t|w) = L\boldsymbol{\ell}(t|w)$ into (100), we have $(I_K - ML)\boldsymbol{\ell}(t|w) = \mathrm{i}t\boldsymbol{\delta}(w)$. For every $t \in \mathbb{R} \setminus \{0\}$, this implies

$$\boldsymbol{\delta}(w) = \frac{1}{\mathrm{i}t}(I_K - ML)\boldsymbol{\ell}(t|w). \tag{102}$$

Thus $\boldsymbol{\delta}(w)$ is identified from the conditional marginal distributions of $Y_0, \dots, Y_{K-1}$.

By Assumption 1, the conditional joint density of $\boldsymbol{V}$ satisfies $p_{\boldsymbol{V}|W=w}(\boldsymbol{v}) = \prod_{j=0}^{J-1} q_j(v_j|w)$. By (14),

$$\begin{aligned}\mathbb{E}[g(Y_0, \dots, Y_{K-1})|W = w] &= \mathbb{E}[g(\boldsymbol{\delta}(w) + M\boldsymbol{V})|W = w] \\ &= \int_{\mathbb{R}^J} g(\boldsymbol{\delta}(w) + M\boldsymbol{v}) \prod\nolimits_{j=0}^{J-1} q_j(v_j|w)d\boldsymbol{v},\end{aligned} \tag{103}$$

which proves (15).

If $J = K$, then $\mathrm{rank}(M) = K$, so M is invertible and $L = M^{-1}$. Therefore $ML = I_K$, and (11) implies $\boldsymbol{\delta}(w) = \mathbf{0}$. Hence (14) becomes $\boldsymbol{Y} = M\boldsymbol{V}$. The inverse transformation is $\boldsymbol{v} = M^{-1}\boldsymbol{y}$, whose Jacobian matrix is M^{-1}. By the change-of-variables formula and the conditional joint density of $\boldsymbol{V}$,

$$\begin{aligned} p_{\boldsymbol{Y}|W=w}(\boldsymbol{y}) &= |\det\big(M^{-1}\big)| p_{\boldsymbol{V}|W=w}(M^{-1}\boldsymbol{y}) \\ &= |\det\big(M^{-1}\big)| \prod_{j=0}^{K-1} q_j((M^{-1}\boldsymbol{y})_j|w), \end{aligned} \tag{104}$$

which proves (16). □

Theorem 4 (Asymptotic normality)**.** *Suppose that Assumption 4 in Appendix D holds. Let $\mathcal{I}_K(\boldsymbol{\theta}^0)$ denote the Fisher information defined by* (48) *in Appendix D.*

(i): For a fixed K, we have

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^0, \qquad \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}), \qquad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_{\mathbb{P}}(n^{-1/2}). \tag{23}$$

Fix w such that $\boldsymbol{\theta} \mapsto \tau_g(w, \boldsymbol{\theta})$ is differentiable at $\boldsymbol{\theta}^0$, and let $\widehat{\tau}_{g,n}(w) = \tau_g(w, \widehat{\boldsymbol{\theta}})$. Then we have

$$\sqrt{n}\{\widehat{\tau}_{g,n}(w) - \tau_g(w, \boldsymbol{\theta}^0)\} \xrightarrow{d} N\left(0, \nabla_{\boldsymbol{\theta}}\tau_g(w, \boldsymbol{\theta}^0)^\top \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \nabla_{\boldsymbol{\theta}}\tau_g(w, \boldsymbol{\theta}^0)\right). \tag{24}$$

(ii): Consider a family of these models indexed by K, and let d_K denote the dimension of $\boldsymbol{\theta}$. For the fixed w above, suppose there exist constants $d_0, C_g, c_I > 0$, independent of K, such that

$$d_K \le d_0 K, \qquad \|\nabla_{\boldsymbol{\theta}}\tau_g(w, \boldsymbol{\theta}^0)\|^2 \le d_K C_g^2, \qquad \lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0)) \ge \frac{c_I}{K}, \tag{25}$$

where $\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))$ is the smallest eigenvalue of $\mathcal{I}_K(\boldsymbol{\theta}^0)$. Then

$$\frac{1}{n}\nabla_{\boldsymbol{\theta}}\tau_g(w, \boldsymbol{\theta}^0)^\top \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \nabla_{\boldsymbol{\theta}}\tau_g(w, \boldsymbol{\theta}^0) \le \frac{d_0 C_g^2}{c_I} \frac{K^2}{n}. \tag{26}$$

Consequently, the corresponding asymptotic standard deviation is $O(K/\sqrt{n})$.

Proof. Define

$$\ell_n(\boldsymbol{\theta}) \coloneqq \sum_{x=0}^{K-1} \sum_{i=1}^{n_x} \log p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta}). \tag{105}$$

By Assumption 4(iii), $\prod_{x=0}^{K-1}\prod_{i=1}^{n_x} p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta}) > 0$ for every $\boldsymbol{\theta} \in \Theta_K$ with probability one. Since the logarithm is strictly increasing, $\widehat{\boldsymbol{\theta}}$ also maximizes ℓ_n.

Proof of (i). Fix K.

$\underline{\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^0.}$ For each $x \in \{0, \dots, K-1\}$, define

$$\begin{aligned} Q_x(\boldsymbol{\theta}) &\coloneqq \int_{\Omega_W} \int_{\mathcal{S}_x(w)} \log p_x(y, w|\boldsymbol{\theta}) p_x(y, w|\boldsymbol{\theta}^0) dy\, P_W(dw), \\ Q(\boldsymbol{\theta}) &\coloneqq \sum_{x=0}^{K-1} \pi_x Q_x(\boldsymbol{\theta}). \end{aligned} \tag{106}$$

For every $\boldsymbol{\theta} \in \Theta_K$, apply $\log u \le u - 1$ to $u = p_x(y, w|\boldsymbol{\theta})/p_x(y, w|\boldsymbol{\theta}^0)$ on the common support in Assumption 4(iii). Since the marginal distribution of W is P_W under every $\boldsymbol{\theta}$, we have $\int_{\mathcal{S}_x(w)} p_x(y, w|\boldsymbol{\theta}) dy = 1$ for P_W-almost every w. Therefore,

$$\begin{aligned} Q(\boldsymbol{\theta}^0) - Q(\boldsymbol{\theta}) &= \sum_{x=0}^{K-1} \pi_x \int_{\Omega_W} \int_{\mathcal{S}_x(w)} p_x(y, w|\boldsymbol{\theta}^0) \log \frac{p_x(y, w|\boldsymbol{\theta}^0)}{p_x(y, w|\boldsymbol{\theta})} dy\, P_W(dw) \\ &\ge \sum_{x=0}^{K-1} \pi_x \int_{\Omega_W} \int_{\mathcal{S}_x(w)} \{p_x(y, w|\boldsymbol{\theta}^0) - p_x(y, w|\boldsymbol{\theta})\} dy\, P_W(dw) \\ &= 0. \end{aligned} \tag{107}$$

Equality in $\log u \le u - 1$ holds only for $u = 1$. Since $\pi_x > 0$ for every x, equality in (107) holds only if $p_x(\cdot, \cdot|\boldsymbol{\theta}) = p_x(\cdot, \cdot|\boldsymbol{\theta}^0)$ almost everywhere for every x. By Assumption 4(i), we then

have $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. Hence $\boldsymbol{\theta}^0$ is the unique maximizer of Q. By Assumption 4(iii) and the dominated convergence theorem, each Q_x is continuous on Θ_K. Moreover,

$$\sup_{\boldsymbol{\theta}\in\Theta_K} |Q_x(\boldsymbol{\theta})| \le \mathbb{E}[G_x^{(0)}(Y_x, W)] < \infty. \tag{108}$$

For every x, we have $n_x \longrightarrow \infty$ by (44). The observations $(Y^{(x,i)}, W^{(x,i)})$ within group x are i.i.d. by Assumption 4(ii). By Assumption 4(iii), $\log p_x(y, w|\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ almost everywhere and is bounded uniformly on Θ_K by $G_x^{(0)}(y, w)$, whose expectation is finite. The set Θ_K is compact by Assumption 4(i). Thus, by Lemma 2.4 of Newey & McFadden (1994),

$$\sup_{\boldsymbol{\theta}\in\Theta_K} \left| \frac{1}{n_x} \sum\nolimits_{i=1}^{n_x} \log p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta}) - Q_x(\boldsymbol{\theta}) \right| \xrightarrow{p} 0. \tag{109}$$

By (105) and (106),

$$\begin{aligned} \frac{1}{n}\ell_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}) = \sum\nolimits_{x=0}^{K-1} \frac{n_x}{n} \left\{ \frac{1}{n_x} \sum\nolimits_{i=1}^{n_x} \log p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta}) - Q_x(\boldsymbol{\theta}) \right\} \\ + \sum\nolimits_{x=0}^{K-1} \left(\frac{n_x}{n} - \pi_x \right) Q_x(\boldsymbol{\theta}). \end{aligned} \tag{110}$$

By the triangle inequality,

$$\begin{aligned} &\sup_{\boldsymbol{\theta}\in\Theta_K} \left| \frac{1}{n}\ell_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta}) \right| \\ &\le \sum\nolimits_{x=0}^{K-1} \frac{n_x}{n} \sup_{\boldsymbol{\theta}\in\Theta_K} \left| \frac{1}{n_x} \sum\nolimits_{i=1}^{n_x} \log p_x(Y^{(x,i)}, W^{(x,i)}|\boldsymbol{\theta}) - Q_x(\boldsymbol{\theta}) \right| \\ &\quad + \sum\nolimits_{x=0}^{K-1} \left| \frac{n_x}{n} - \pi_x \right| \sup_{\boldsymbol{\theta}\in\Theta_K} |Q_x(\boldsymbol{\theta})|. \end{aligned} \tag{111}$$

For each x, the first summand converges in probability to zero by (109) and $0 \le n_x/n \le 1$. The second summand converges to zero by (44) and (108). The sums have finitely many terms, so

$$\sup_{\boldsymbol{\theta}\in\Theta_K} |n^{-1}\ell_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \xrightarrow{p} 0. \tag{112}$$

The set Θ_K is compact, Q is continuous, and $\boldsymbol{\theta}^0$ is its unique maximizer. Since $\widehat{\boldsymbol{\theta}}$ maximizes ℓ_n, by Theorem 2.1 of Newey & McFadden (1994) and (112), we have

$$\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^0. \tag{113}$$

Asymptotic normality of $n^{-1/2}\nabla_{\boldsymbol{\theta}}\ell_n(\boldsymbol{\theta}^0)$. For a twice differentiable real-valued function f on a subset of $\mathbb{R}^{d_K}$, write

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \coloneqq \left(\frac{\partial f(\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial f(\boldsymbol{\theta})}{\partial\theta_{d_K}} \right)^\top, \qquad \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \coloneqq \left(\frac{\partial^2 f(\boldsymbol{\theta})}{\partial\theta_r \partial\theta_s} \right)_{r,s=1}^{d_K}. \tag{114}$$

Recall the definition of $\boldsymbol{s}_x$ in Appendix D and define the Hessian by

$$\boldsymbol{s}_x(y, w, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log p_x(y, w|\boldsymbol{\theta}), \qquad H_x(y, w, \boldsymbol{\theta}) \coloneqq \nabla^2_{\boldsymbol{\theta}} \log p_x(y, w|\boldsymbol{\theta}). \tag{115}$$

By differentiation of the logarithm,

$$\boldsymbol{s}_x(y, w, \boldsymbol{\theta}) p_x(y, w|\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} p_x(y, w|\boldsymbol{\theta}). \tag{116}$$

For P_W-almost every w, we have $\int_{\mathcal{S}_x(w)} p_x(y, w|\boldsymbol{\theta}) dy = 1$. By Assumption 4(iv) and (116),

$$\begin{aligned} \mathbb{E}[\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)|W = w] &= \int_{\mathcal{S}_x(w)} \boldsymbol{s}_x(y, w, \boldsymbol{\theta}^0) p_x(y, w|\boldsymbol{\theta}^0) dy \\ &= \int_{\mathcal{S}_x(w)} \nabla_{\boldsymbol{\theta}} p_x(y, w|\boldsymbol{\theta}) dy \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\ &= \nabla_{\boldsymbol{\theta}} \int_{\mathcal{S}_x(w)} p_x(y, w|\boldsymbol{\theta}) dy \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\ &= \mathbf{0}. \end{aligned} \tag{117}$$

By Assumption 4(iv),

$$\mathbb{E}[\|\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)\|^2] \leq \mathbb{E}[G_x^{(1)}(Y_x, W)^2] < \infty. \tag{118}$$

Thus $\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)$ is integrable, and by the law of iterated expectations and (117),

$$\mathbb{E}[\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)] = \mathbb{E}[\mathbb{E}[\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)|W]] = \mathbf{0}. \tag{119}$$

Differentiating (116) once more, we have

$$\begin{aligned} \nabla^2_{\boldsymbol{\theta}} p_x(y, w|\boldsymbol{\theta}) &= p_x(y, w|\boldsymbol{\theta}) H_x(y, w, \boldsymbol{\theta}) + \boldsymbol{s}_x(y, w, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} p_x(y, w|\boldsymbol{\theta})^\top \\ &= p_x(y, w|\boldsymbol{\theta}) \{H_x(y, w, \boldsymbol{\theta}) + \boldsymbol{s}_x(y, w, \boldsymbol{\theta}) \boldsymbol{s}_x(y, w, \boldsymbol{\theta})^\top\}. \end{aligned} \tag{120}$$

Again by Assumption 4(iv),

$$\int_{\mathcal{S}_x(w)} \nabla^2_{\boldsymbol{\theta}} p_x(y, w|\boldsymbol{\theta}) dy \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \nabla^2_{\boldsymbol{\theta}} \int_{\mathcal{S}_x(w)} p_x(y, w|\boldsymbol{\theta}) dy \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = O. \tag{121}$$

It follows that

$$\int_{\mathcal{S}_x(w)} H_x(y, w, \boldsymbol{\theta}^0) p_x(y, w|\boldsymbol{\theta}^0) dy = -\int_{\mathcal{S}_x(w)} \boldsymbol{s}_x(y, w, \boldsymbol{\theta}^0) \boldsymbol{s}_x(y, w, \boldsymbol{\theta}^0)^\top p_x(y, w|\boldsymbol{\theta}^0) dy. \tag{122}$$

By (118) and $\mathbb{E}[G_x^{(2)}(Y_x, W)] < \infty$, we may integrate both sides with respect to P_W. Define

$$\mathcal{I}_x(\boldsymbol{\theta}^0) \coloneqq \mathbb{E}[\boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0) \boldsymbol{s}_x(Y_x, W, \boldsymbol{\theta}^0)^\top]. \tag{123}$$

Then

$$\begin{aligned} -\mathbb{E}[H_x(Y_x, W, \boldsymbol{\theta}^0)] &= \mathcal{I}_x(\boldsymbol{\theta}^0), \\ -\sum\nolimits_{x=0}^{K-1} \pi_x \mathbb{E}[H_x(Y_x, W, \boldsymbol{\theta}^0)] &= \sum\nolimits_{x=0}^{K-1} \pi_x \mathcal{I}_x(\boldsymbol{\theta}^0) = \mathcal{I}_K(\boldsymbol{\theta}^0). \end{aligned} \tag{124}$$

For each x, the vectors $\boldsymbol{s}_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}^0)$ are i.i.d. with mean zero and covariance $\mathcal{I}_x(\boldsymbol{\theta}^0)$. By (118) and the multivariate central limit theorem,

$$\frac{1}{\sqrt{n_x}} \sum\nolimits_{i=1}^{n_x} \boldsymbol{s}_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}^0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_x(\boldsymbol{\theta}^0)). \tag{125}$$

By Assumption 4(ii), these sums are independent across treatment groups. Their joint limiting distribution is therefore that of independent normal vectors $\boldsymbol{Z}_x \sim N(\mathbf{0}, \mathcal{I}_x(\boldsymbol{\theta}^0))$, $x \in \{0, \ldots, K-1\}$. Since $n_x/n \longrightarrow \pi_x$, by Slutsky's theorem,

$$\begin{aligned} \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) &= \sum\nolimits_{x=0}^{K-1} \sqrt{\frac{n_x}{n}} \frac{1}{\sqrt{n_x}} \sum\nolimits_{i=1}^{n_x} \boldsymbol{s}_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}^0) \\ &\xrightarrow{d} \sum\nolimits_{x=0}^{K-1} \sqrt{\pi_x} \boldsymbol{Z}_x. \end{aligned} \tag{126}$$

By independence of the vectors $\boldsymbol{Z}_x$, their weighted sum is normal with mean zero and covariance $\sum_{x=0}^{K-1} \pi_x \mathcal{I}_x(\boldsymbol{\theta}^0) = \mathcal{I}_K(\boldsymbol{\theta}^0)$. Hence

$$\frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)). \tag{127}$$

Convergence of the average Hessian. Choose a closed ball $\mathcal{B}$ centered at $\boldsymbol{\theta}^0$ and contained in the interior of the neighborhood $\mathcal{N}$ in Assumption 4(iv). By (113), $\mathbb{P}(\widehat{\boldsymbol{\theta}} \in \mathcal{B}) \longrightarrow 1$. For $t \in [0, 1]$, define $\boldsymbol{\theta}_{n,t} \coloneqq \boldsymbol{\theta}^0 + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)$. On the event $\{\widehat{\boldsymbol{\theta}} \in \mathcal{B}\}$, we have $\boldsymbol{\theta}_{n,t} \in \mathcal{B}$ for every t because $\mathcal{B}$ is convex. Moreover, $\widehat{\boldsymbol{\theta}}$ is an interior maximizer, so $\nabla_{\boldsymbol{\theta}} \ell_n(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. By the chain rule,

$$\frac{d}{dt} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}_{n,t}) = \nabla^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}_{n,t})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \tag{128}$$

By the fundamental theorem of calculus,

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}} \ell_n(\widehat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) + \int_0^1 \nabla^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}_{n,t})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) dt. \tag{129}$$

Define

$$A_n \coloneqq \int_0^1 \frac{1}{n} \sum\nolimits_{x=0}^{K-1} \sum\nolimits_{i=1}^{n_x} H_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}_{n,t}) dt. \tag{130}$$

After division by $\sqrt{n}$, we have

$$\mathbf{0} = \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) + A_n \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \tag{131}$$

For a matrix $C = (c_{rs}) \in \mathbb{R}^{d_K \times d_K}$, write $\|C\| \coloneqq (\sum_{r=1}^{d_K} \sum_{s=1}^{d_K} c_{rs}^2)^{1/2}$. For each treatment group, define

$$\begin{aligned} \widehat{H}_{x,n}(\boldsymbol{\theta}) &\coloneqq \frac{1}{n_x} \sum\nolimits_{i=1}^{n_x} H_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}), \\ \overline{H}_x(\boldsymbol{\theta}) &\coloneqq \mathbb{E}[H_x(Y_x, W, \boldsymbol{\theta})]. \end{aligned} \tag{132}$$

By Assumption 4(iv), each entry of $H_x(y, w, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ on $\mathcal{B}$ and is bounded in absolute value by $G_x^{(2)}(y, w)$, whose expectation is finite. By Lemma 2.4 of Newey & McFadden (1994), for every $r, s \in \{1, \ldots, d_K\}$,

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}} |\widehat{H}_{x,n}(\boldsymbol{\theta})_{rs} - \overline{H}_x(\boldsymbol{\theta})_{rs}| \xrightarrow{p} 0. \tag{133}$$

Since d_K is fixed and

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}} \|\widehat{H}_{x,n}(\boldsymbol{\theta}) - \overline{H}_x(\boldsymbol{\theta})\|^2 \le \sum\nolimits_{r=1}^{d_K} \sum\nolimits_{s=1}^{d_K} \sup_{\boldsymbol{\theta} \in \mathcal{B}} |\widehat{H}_{x,n}(\boldsymbol{\theta})_{rs} - \overline{H}_x(\boldsymbol{\theta})_{rs}|^2, \tag{134}$$

we have

$$\sup_{\boldsymbol{\theta} \in \mathcal{B}} \|\widehat{H}_{x,n}(\boldsymbol{\theta}) - \overline{H}_x(\boldsymbol{\theta})\| \xrightarrow{p} 0. \tag{135}$$

By Assumption 4(iv) and the dominated convergence theorem, $\overline{H}_x$ is continuous at $\boldsymbol{\theta}^0$. By (113),

$$\sup_{t \in [0,1]} \|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}^0\| = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| \xrightarrow{p} 0. \tag{136}$$

Therefore, by continuity of $\overline{H}_x$,

$$\sup_{t \in [0,1]} \|\overline{H}_x(\boldsymbol{\theta}_{n,t}) - \overline{H}_x(\boldsymbol{\theta}^0)\| \xrightarrow{p} 0. \tag{137}$$

On the event $\{\widehat{\boldsymbol{\theta}} \in \mathcal{B}\}$, by the triangle inequality,

$$\begin{aligned} &\left\| \int_0^1 \widehat{H}_{x,n}(\boldsymbol{\theta}_{n,t}) dt - \overline{H}_x(\boldsymbol{\theta}^0) \right\| \\ &\le \int_0^1 \|\widehat{H}_{x,n}(\boldsymbol{\theta}_{n,t}) - \overline{H}_x(\boldsymbol{\theta}_{n,t})\| dt + \int_0^1 \|\overline{H}_x(\boldsymbol{\theta}_{n,t}) - \overline{H}_x(\boldsymbol{\theta}^0)\| dt \\ &\le \sup_{\boldsymbol{\theta} \in \mathcal{B}} \|\widehat{H}_{x,n}(\boldsymbol{\theta}) - \overline{H}_x(\boldsymbol{\theta})\| + \sup_{t \in [0,1]} \|\overline{H}_x(\boldsymbol{\theta}_{n,t}) - \overline{H}_x(\boldsymbol{\theta}^0)\|. \end{aligned} \tag{138}$$

By (135) and (137), both terms in the final line converge in probability to zero. Since $\mathbb{P}(\widehat{\boldsymbol{\theta}} \in \mathcal{B}) \longrightarrow 1$, we obtain

$$\int_0^1 \widehat{H}_{x,n}(\boldsymbol{\theta}_{n,t}) dt \xrightarrow{p} \overline{H}_x(\boldsymbol{\theta}^0). \tag{139}$$

By (130),

$$A_n = \sum\nolimits_{x=0}^{K-1} \frac{n_x}{n} \int_0^1 \widehat{H}_{x,n}(\boldsymbol{\theta}_{n,t}) dt. \tag{140}$$

Consequently,

$$\begin{aligned} &\left\| A_n - \sum\nolimits_{x=0}^{K-1} \pi_x \overline{H}_x(\boldsymbol{\theta}^0) \right\| \\ &\le \sum\nolimits_{x=0}^{K-1} \frac{n_x}{n} \left\| \int_0^1 \widehat{H}_{x,n}(\boldsymbol{\theta}_{n,t}) dt - \overline{H}_x(\boldsymbol{\theta}^0) \right\| + \sum\nolimits_{x=0}^{K-1} \left| \frac{n_x}{n} - \pi_x \right| \|\overline{H}_x(\boldsymbol{\theta}^0)\|. \end{aligned} \tag{141}$$

By (139), (44), and (124),

$$A_n \xrightarrow{p} \sum_{x=0}^{K-1} \pi_x \overline{H}_x(\boldsymbol{\theta}^0) = -\mathcal{I}_K(\boldsymbol{\theta}^0). \tag{142}$$

Asymptotic normality of $\widehat{\boldsymbol{\theta}}$. By Assumption 4(v), $\mathcal{I}_K(\boldsymbol{\theta}^0)$ is invertible. Since the determinant is continuous, $\det(A_n) \xrightarrow{p} \det\big(-\mathcal{I}_K(\boldsymbol{\theta}^0)\big) \neq 0$ by (142). Hence $\mathbb{P}(A_n \text{ is invertible}) \longrightarrow 1$. On the event that $\widehat{\boldsymbol{\theta}} \in \mathcal{B}$ and A_n is invertible, we may rearrange (131) as

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -A_n^{-1} \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0). \tag{143}$$

By (142) and continuity of matrix inversion,

$$-A_n^{-1} \xrightarrow{p} \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}. \tag{144}$$

By (127), $n^{-1/2}\nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) = O_{\mathbb{P}}(1)$. Hence

$$\{-A_n^{-1} - \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}\} \frac{1}{\sqrt{n}} \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0) = o_{\mathbb{P}}(1). \tag{145}$$

By (105), $\nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}^0)$ is the sum of $\boldsymbol{s}_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}^0)$ over x and i. Since $\mathbb{P}(\widehat{\boldsymbol{\theta}} \in \mathcal{B}, A_n \text{ is invertible}) \longrightarrow 1$, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \frac{1}{\sqrt{n}} \sum_{x=0}^{K-1} \sum_{i=1}^{n_x} \boldsymbol{s}_x(Y^{(x,i)}, W^{(x,i)}, \boldsymbol{\theta}^0) + o_{\mathbb{P}}(1). \tag{146}$$

For $\boldsymbol{Z} \sim N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0))$, we have

$$\begin{aligned} \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}\boldsymbol{Z} &\sim N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}\mathcal{I}_K(\boldsymbol{\theta}^0)\mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}) \\ &= N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}). \end{aligned} \tag{147}$$

By (127), (146), and Slutsky's theorem,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}). \tag{148}$$

Since a sequence that converges in distribution is bounded in probability,

$$\sqrt{n}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_{\mathbb{P}}(1), \qquad \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| = O_{\mathbb{P}}(n^{-1/2}). \tag{149}$$

Proof of (24). Since $\tau_g(w, \cdot)$ is differentiable at $\boldsymbol{\theta}^0$, there exists a function r_g such that

$$\tau_g(w, \boldsymbol{\theta}^0 + \boldsymbol{h}) - \tau_g(w, \boldsymbol{\theta}^0) = \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \boldsymbol{h} + r_g(\boldsymbol{h}), \qquad \frac{|r_g(\boldsymbol{h})|}{\|\boldsymbol{h}\|} \longrightarrow 0 \text{ as } \boldsymbol{h} \longrightarrow \mathbf{0}. \tag{150}$$

Substituting $\boldsymbol{h} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$,

$$\tau_g(w, \widehat{\boldsymbol{\theta}}) - \tau_g(w, \boldsymbol{\theta}^0) = \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + r_g(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \tag{151}$$

By (113),

$$\frac{|r_g(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)|}{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|} \xrightarrow{p} 0, \tag{152}$$

where the ratio is taken to be zero when $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^0$. Therefore, together with (149), we see

$$\sqrt{n}|r_g(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)| = \sqrt{n}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\| \frac{|r_g(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)|}{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|} \xrightarrow{p} 0. \tag{153}$$

Multiplying (151) by $\sqrt{n}$,

$$\sqrt{n}\{\widehat{\tau}_{g,n}(w) - \tau_g(w, \boldsymbol{\theta}^0)\} = \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) + \sqrt{n} r_g(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0). \tag{154}$$

By (148) and the continuous mapping theorem,

$$\nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) \xrightarrow{d} N\left(0, \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)^\top \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} \nabla_{\boldsymbol{\theta}} \tau_g(w, \boldsymbol{\theta}^0)\right). \tag{155}$$

By Slutsky's theorem and (154), (24) holds.

Proof of (ii). By (25),

$$\|\nabla_{\boldsymbol{\theta}}\tau_g(w,\boldsymbol{\theta}^0)\|^2 = \sum\nolimits_{r=1}^{d_K}\left|\frac{\partial\tau_g(w,\boldsymbol{\theta}^0)}{\partial\theta_r}\right|^2 \le d_K C_g^2 \le d_0 K C_g^2. \tag{156}$$

Since $\mathcal{I}_K(\boldsymbol{\theta}^0)$ is real, symmetric, and positive definite, there exist an orthogonal matrix P and positive eigenvalues $\lambda_1,\dots,\lambda_{d_K}$ such that

$$\begin{aligned}\mathcal{I}_K(\boldsymbol{\theta}^0) &= P\,\mathrm{diag}(\lambda_1,\dots,\lambda_{d_K})P^\top,\\ \mathcal{I}_K(\boldsymbol{\theta}^0)^{-1} &= P\,\mathrm{diag}(\lambda_1^{-1},\dots,\lambda_{d_K}^{-1})P^\top.\end{aligned} \tag{157}$$

For $\boldsymbol{a}\in\mathbb{R}^{d_K}$, let $\boldsymbol{z}\coloneqq P^\top\boldsymbol{a}$. Since P is orthogonal, $\|\boldsymbol{z}\|=\|\boldsymbol{a}\|$. Therefore

$$\boldsymbol{a}^\top\mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}\boldsymbol{a} = \sum\nolimits_{r=1}^{d_K}\frac{z_r^2}{\lambda_r} \le \frac{1}{\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))}\sum\nolimits_{r=1}^{d_K} z_r^2 = \frac{\|\boldsymbol{a}\|^2}{\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))}. \tag{158}$$

By $\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0)) \ge c_I/K$ in (25), we see

$$\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))^{-1} \le \frac{K}{c_I}. \tag{159}$$

Apply (158) with $\boldsymbol{a}=\nabla_{\boldsymbol{\theta}}\tau_g(w,\boldsymbol{\theta}^0)$ and then use (156). Then we have

$$\frac{1}{n}\nabla_{\boldsymbol{\theta}}\tau_g(w,\boldsymbol{\theta}^0)^\top\mathcal{I}_K(\boldsymbol{\theta}^0)^{-1}\nabla_{\boldsymbol{\theta}}\tau_g(w,\boldsymbol{\theta}^0) \le \frac{\|\nabla_{\boldsymbol{\theta}}\tau_g(w,\boldsymbol{\theta}^0)\|^2}{n\lambda_{\min}(\mathcal{I}_K(\boldsymbol{\theta}^0))} \le \frac{d_0 C_g^2}{c_I}\frac{K^2}{n}. \tag{160}$$

This proves the final assertion. □

Theorem 5 (Non-identification)**.** *Fix $w\in\Omega_W$, $\boldsymbol{\mu}(w)$, and $B(w)$, and suppose that $J=K$. Suppose that $B(w)$ is invertible and that $(b_{xk}(w)^2)_{x,k=0}^{K-1}$ is singular. Then there exist two models of the form* (9) *with the same $\boldsymbol{\mu}(w)$ and $B(w)$ that are saturated given $W=w$ and satisfy Assumptions 1 and 2. Let $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ denote their potential-outcome vectors. For every $x\in\{0,\dots,K-1\}$, the conditional distributions of Y_x and Y_x' given $W=w$ coincide, whereas $\mathrm{Cov}(\boldsymbol{Y}|W=w)\neq\mathrm{Cov}(\boldsymbol{Y}'|W=w)$. In particular, the conditional joint distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ given $W=w$ differ.*

Proof. Since $(b_{xk}(w)^2)_{x,k=0}^{K-1}$ is singular, there exists $\boldsymbol{h}=(h_0,\dots,h_{K-1})^\top\neq\boldsymbol{0}$ such that

$$\sum\nolimits_{k=0}^{K-1} b_{xk}(w)^2 h_k = 0 \tag{161}$$

for every $x\in\{0,\dots,K-1\}$. Choose $c>\max_{0\le k\le K-1}|h_k|$ and define $v_k\coloneqq c$ and $v_k'\coloneqq c+h_k$. Then $v_k>0$ and $v_k'>0$ for every $k\in\{0,\dots,K-1\}$. Consider two models with the same $\boldsymbol{\mu}(w)$ and $B(w)$ and with conditional distributions

$$\boldsymbol{U}|W=w\sim N(\boldsymbol{0},\mathrm{diag}(v_0,\dots,v_{K-1})) \tag{162}$$

and

$$\boldsymbol{U}'|W=w\sim N(\boldsymbol{0},\mathrm{diag}(v_0',\dots,v_{K-1}')). \tag{163}$$

The components of each latent vector are mutually independent and have densities. Since $B(w)$ is invertible, both models are saturated given $W=w$ by Theorem 1. Define $\boldsymbol{Y}\coloneqq\boldsymbol{\mu}(w)+B(w)\boldsymbol{U}$ and $\boldsymbol{Y}'\coloneqq\boldsymbol{\mu}(w)+B(w)\boldsymbol{U}'$. For every $x\in\{0,\dots,K-1\}$, both $Y_x|W=w$ and $Y_x'|W=w$ are Gaussian and

$$\mathbb{E}[Y_x|W=w]=\mathbb{E}[Y_x'|W=w]=\mu_x(w). \tag{164}$$

Moreover,

$$\begin{aligned}\mathrm{Var}(Y_x'|W=w)-\mathrm{Var}(Y_x|W=w) &= \sum\nolimits_{k=0}^{K-1} b_{xk}(w)^2(v_k'-v_k)\\ &= \sum\nolimits_{k=0}^{K-1} b_{xk}(w)^2 h_k\\ &= 0,\end{aligned} \tag{165}$$

where the last equality follows from (161). Thus the conditional distributions of Y_x and Y'_x given $W = w$ coincide for every $x \in \{0, \dots, K-1\}$. Their characteristic functions are nonzero everywhere, so both models satisfy Assumptions 1 and 2.

On the other hand, mutual independence gives

$$\mathrm{Cov}(\boldsymbol{Y}|W = w) = B(w)\mathrm{diag}(v_0, \dots, v_{K-1})B(w)^\top \tag{166}$$

and

$$\mathrm{Cov}(\boldsymbol{Y}'|W = w) = B(w)\mathrm{diag}(v'_0, \dots, v'_{K-1})B(w)^\top. \tag{167}$$

Therefore,

$$\mathrm{Cov}(\boldsymbol{Y}'|W = w) - \mathrm{Cov}(\boldsymbol{Y}|W = w) = B(w)\mathrm{diag}(\boldsymbol{h})B(w)^\top. \tag{168}$$

If the right-hand side were zero, multiplication by $B(w)^{-1}$ from the left and $(B(w)^\top)^{-1}$ from the right would give $\mathrm{diag}(\boldsymbol{h}) = O$, contradicting $\boldsymbol{h} \neq \boldsymbol{0}$. Therefore $\mathrm{Cov}(\boldsymbol{Y}|W = w) \neq \mathrm{Cov}(\boldsymbol{Y}'|W = w)$, and the conditional joint distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ given $W = w$ differ. □