arXiv:2603.24705v1 [stat.ME] 25 Mar 2026

通过等变神经网络对相关误差离散选择模型进行摊销推断

Easton Huch* Michael Keane†

2026年3月

摘要

离散选择模型是《管理科学》、经济学和市场营销领域的重要工具，可用于理解和预测决策行为。基于Logit的模型在应用研究中占据主导地位，主要得益于其选择概率具有便于计算的闭式表达式。然而，这类模型对随机效用成分作出了严格假设，限制了我们刻画现实且具有理论依据的选择行为的能力，尤其是替代模式。本文提出一种摊销推断方法，利用神经网络模拟器近似一般误差分布下的选择概率，包括具有相关误差的分布。我们提出了专门的神经网络架构及配套训练流程，旨在遵循离散选择模型的不变性。我们为该架构提供了群论基础，包括在给定最小一组不变特征的条件下通用逼近的证明。训练完成后，该模拟器可快速进行似然评估和梯度计算。我们采用索博列夫训练，在似然损失中加入梯度匹配惩罚项，使模拟器能够同时学习选择概率及其导数。我们证明，在温和的近似条件下，基于模拟器的极大似然估计量具有一致性和渐近正态性，并给出即使似然近似不完美也仍然有效的三明治标准误。模拟结果表明，与GHK模拟器相比，该方法在准确性和速度方面均有显著提升。

关键词： 摊销推断；DeepSet；离散选择；不变理论；多项式概率单位；神经网络模拟器；置换等变性；索博列夫训练

*约翰斯·霍普金斯大学凯瑞商学院博士后研究员，ehuch@jhu.edu。
†约翰斯·霍普金斯大学凯瑞商学院凯瑞杰出教授，mkeane14@jhu.edu。

1 引言

离散选择模型是管理科学、经济学、市场营销等领域广泛使用的工具，用于理解个人和组织如何在有限备选集合中做出决策（McFadden，1974）。这些模型提供了一个结构化框架，能够得出可解释的参数，例如支付意愿和需求弹性；同时也能预测假设性干预措施的影响，包括定价变化、产品引入和政策调整。

应用研究中占主导地位的离散选择模型是多项Logit模型（MNL）。该模型假设消费者 i 从备选项 j 获得的效用形式为 $U_j = v_j + \epsilon_j$，其中 v_j 是效用的确定性部分，通常写成备选项 j的属性 $\boldsymbol{x}_j$的线性函数$\boldsymbol{x}_j^\top \boldsymbol{\beta}$ ；而 ϵ_j 是I型极值分布（Gumbel）误差，在各备选项 $j = 1$、 ..... 、 K 之间独立同分布（ iid ）。这一设定得出了简单的闭式选择概率：选择备选项j 的概率由“softmax函数” $\exp(v_j)/\sum_{k=1}^{K}\exp(v_k)$给出。由此带来的计算便利，使 MNL 成为无数应用中的默认选择。

然而，这种便利也付出了众所周知的代价。iid Gumbel误差假设意味着一个严格的无关选项独立性（IIA）性质：任意两个备选项之间的胜算比不受其他备选项的存在或属性影响，这可能导致不现实的替代模式。MNL 的一种常见推广是混合 Logit 模型（混合多项Logit模型），其中效用权重 $\boldsymbol{\beta}$ 允许因消费者而异。通过模拟方法进行估计十分简便（Train，2009）。混合多项Logit模型在总体层面放宽了 IIA，但在给定相应随机效用权重i 的情况下，消费者 $\boldsymbol{\beta}_i$的个体选择仍然满足 IIA。

多项Probit模型（MNP）是多项Logit模型（MNL）的知名替代方案。它通过假设 ϵ_1、.....、 ϵ_K服从具有跨选项相关性的多元正态分布，放宽了独立无关替代项假设，并允许灵活的替代模式。尽管具有这一优势，MNP 的应用仍相对有限，主要原因是其计算开销较大。MNP 的选择概率没有闭式解，因此需要计算多元正态矩形概率。其估计所依赖的模拟方法也比混合多项Logit模型所需的方法更复杂：在经典框架下使用 GHK 模拟器（Geweke，1989年；Hajivassiliou 和 McFadden，1998年；Keane，1994 年），或在贝叶斯框架下使用 MCMC（McCulloch 和 Rossi，1994 年）。

我们提出一种截然不同的方法：不再针对每次似然评估重新模拟选择概率，而是训练一个神经网络模拟器，直接近似

选择概率函数。这种策略称为摊销推断，它将计算负担从推断阶段转移到一次性的训练阶段。训练完成后，模拟器只需通过简单的函数调用，即可快速、确定性地近似选择概率。摊销推断框架已在多个科学领域取得显著成效，可用于近似计算密集型仿真模型（Lueck-mann etal.，2019年；Cranmer et al.，2020年）。在此，我们将其应用于离散选择模型，将适用范围从多项Logit模型和多项Probit模型扩展到一般的相关误差分布。

在计量经济学文献中，Norets (2012年) 采用了一种相关策略，通过神经网络近似动态离散选择模型中的期望价值函数。他在一个特定参数模型的背景下应用这一策略，因此，动态模型发生变化时，需要修改并重新训练神经网络。相比之下，我们的模拟器仅根据确定性效用和误差分布的参数处理通用选择问题，因此无需更改或重新训练模拟器，即可修改参数模型（例如，确定性效用的函数形式）。

我们的方法依赖于若干项新的贡献。首先，我们开发了一种专为选择模型设计的神经网络架构。该架构遵循这些模型的基本不变性，并将其形式化为选择概率映射的对称性（群作用）。[1]选择概率对于位置平移（对所有效用加上同一个常数）和尺度变换（将效用乘以一个正数）是不变的，并且对于选择项的置换具有等变性。我们的架构采用预处理变换，通过强制实现位置不变性和尺度不变性来缩减特征空间的维度。处理后的特征随后被传入一个基于 DeepSet 架构的逐选项编码器模块（Zaheer等人，2017年）。然后，将各选择项的输出拼接起来，并通过等变层处理，从而对输出概率施加和为一约束，同时遵循置换等变性。

其次，我们确立了该架构的理论基础。我们证明，除一个零测例外集外，该架构能够在参数空间的紧子集上通用逼近选择概率。这一结果将我们的架构与群作用下轨道分离的理论联系起来，并将 Blum- Smith 等人(2025) 关于对称矩阵的通用逼近结果扩展至效用向量与中心化协方差矩阵构成的联合空间。 Blum-Smith 等人(2025) 运用伽罗瓦理论证明了泛型轨道分离，而我们则基于不变量重构论证给出了直接证明。

[1]如果这些特性未嵌入架构中，神经网络就必须在训练过程中学习它们。由于一系列连续的模型会生成等价的选择概率，这会显著减慢学习速度。

第三，我们确立了基于真实离散选择概率的模拟器近似而构造的极大似然估计量（MLE）的统计性质。我们证明，如果模拟器对真实对数似然的逼近足够好——具体而言，平均逼近误差为 $o_p(n^{-1})$——那么基于模拟器的估计量便继承精确最大似然估计量的一致性和渐近正态性。当这一条件不满足时，我们证明，仍可通过三明治标准误获得有效推断，即在准最大似然框架中将模拟器视为工作模型。

通过精心设计，我们的预处理变换和架构确保选择概率相对于输入——确定性效用及其对应的尺度（协方差）矩阵——具有平滑性。我们采用索博列夫训练（Czarnecki 等人，2017年），使模拟器梯度与对数选择概率的梯度相匹配。这些设计选择结合起来，使自动微分能够可靠地用于下游的模型拟合和推断任务。

借助预训练模拟器，从逻辑斯蒂模型推广到概率单位模型（或任何其他误差分布），在构造似然时只需将闭式 softmax 概率替换为模拟器对选择概率的评估。在概率单位模型的情形下，我们通过模拟表明，在给定固定计算预算时，使用我们的摊销推断程序的最大似然估计量，在估计误差和覆盖率方面，表现不逊于使用 GHK 算法的估计量。此外，我们的方法还能轻松处理多项式正态概率模型以外的模型，在这些模型中，GHK 这类高效模拟算法可能并不存在。

为便于说明，我们的论述大多聚焦于多元正态概率模型。不过，模拟器方法及其支撑理论基本不受所假设误差参数形式的限制。推广到其他误差分布（例如相关 Gumbel 分布）十分直接，只需对训练数据生成和模拟器输入做少量改动。

本文其余部分安排如下。第2 节总结选择建模的相关文献。第3 节介绍多项Probit模型和我们的推断目标。第4 节介绍模拟器架构和训练流程。第5 节阐明该架构及基于模拟器的估计量的理论性质。第6 节展示仿真结果，第7节作总结。

2 相关文献

将机器学习方法应用于离散选择的文献大致可分为两大流派。第一派保留经典的随机效用模型（RUM）选择框架，其中选择概率由一群理性消费者产生

类型构成，这些消费者对选择对象全集具有不同（但稳定）的偏好排序，正如Bloch 和 Marschak (1960年) 以及 McFadden 和 Richter (1990年) 所阐述的那样。[2] 多项Logit模型、多项Probit模型和混合多项Logit模型都属于随机效用模型（RUM）范畴，前提是保持随机效用模型假设，即某个选项的效用仅取决于其自身特征。该研究方向的关键论文以基本的多项Logit模型结构为起点，利用神经网络对效用的函数形式进行推广：Bentz 和 Merunka(2000年)、Sifringer 等人(2020年)、Wang 等人(2020年)、Han 等人 (2022年)，以及Singh 等人(2023年)。近期有两篇重要论文拓展了这项研究：Aouad 和 Desir (2025) 提出了一种架构，可通过灵活的偏好异质性分布实现混合 Logit 模型（RUMnet）；而Bagheri 等人(2025) 则提出另一种架构，用以推广 Gumbel 误差假设（RUM-NN）。两篇论文都在损失函数中使用经 softmax 平滑的样本均值来近似选择概率，因此与纯粹的基于Logit的模型相比，计算时间有所增加。

Aouad 和 Desir (2025) 基于 McFadden 和 Train (2000年所证明的混合多项Logit模型的一般逼近性质进一步开展研究。这一结果依赖于对效用函数进行灵活的基函数展开，以及可能采用非参数形式的混合分布。在实践中，分析人员并不了解这些模型属性。正如 Aouad 和 Desir(2025) 和 McFadden 和 Train (2000年) 都强调的那样，这种逼近性质并非 Gumbel 误差所独有。更一般地，它适用于更大一类层级离散选择模型，包括混合多项正态概率模型。 McFadden 和 Train (2000年) 利用这一逼近性质，论证采用混合多项Logit模型是合理的，因为与多项正态概率模型等缺乏闭式选择概率的其他模型相比，它在计算上更为便利。但鉴于我们在此提出的仿真器这类实用、通用的替代方案，这一采用混合多项Logit模型的理由就不那么有说服力了。

第二类研究摒弃了随机效用模型（RUM）结构，通常旨在放宽多项Logit模型（MNL）中独立无关替代项假设（IIA）等限制性假设，并允许更灵活的替代模式。一些此类研究将离散选择视为适用于机器学习方法的一般分类问题。相关例子包括Wang 和 Ross （2018年）、Lh´eritier 等人（2019年）、Rosenfeld 等人 （2020年）、Chen 和 Mi˘si´c（2022年），以及Chen 等人 （2025）。另一些研究则在顶层保留多项Logit模型（MNL）结构——也就是说，选择概率由一组备选项特定效用构成，这些效用进入 softmax 函数——但使用神经网络以灵活的方式构造备选项特定效用函数，使其偏离随机效用模型（RUM）的假设。例如，备选项的效用 j 可以取决于

[2]假设全集中有 $\bar{K}$ 个对象，消费者面临一个包含 $K \leq \bar{K}$ 个元素的选择集（或“商品组合”）。随机效用模型的一个关键推论是，消费者从产品 k 中获得的效用不随选择集而变化，因此排除了情境效应或商品组合效应。

其他备选项的属性，从而产生情境效应。相关例子包括Wang 等人 （2021年）、Wong 和 Farooq （2021年）、Cai 等人 （2022年）、Pfannschmidt 等人 （2022年），以及Berbeglia 和 Venkataraman（2025）。

这两条研究脉络之间存在着根本性的张力：尽管随机效用模型因其可解释性以及扎根于经济理论而备受青睐，但实证研究中占主导地位的随机效用模型（线性多项Logit模型和混合多项Logit模型）对确定性效用和替代模式施加了约束，在某些应用中可能过于严格；尤其是，它们假设个体选择层面满足独立无关替代项假设。第一条研究脉络旨在通过提高随机效用模型的灵活性来应对这一挑战。然而，许多此类方案仍保留误差项相互独立且服从逻辑斯蒂分布的假设，从而限制了替代模式，并可能导致效用估计产生偏差。例外方案——即 RUMnet 和 RUM-NN——虽能处理替代问题，却要以牺牲可解释性和／或计算效率为代价，原因在于它们（a）采用（可能是）非线性的效用函数，以及（b）对似然函数进行计算成本高昂的样本平均近似。第二条研究脉络则彻底放弃随机效用模型结构，因此难以或无法对许多具有经济意义的量作出可靠推断，例如消费者福利、支付意愿度量、需求弹性和替代效应。总之，这种张力造成了模型灵活性与可解释性之间的权衡：一端是多项Logit模型这类僵化但可解释的模型，另一端则是灵活但缺乏经济理论基础的分类方法。

多元正态概率模型是上述权衡中的一个显著例外，因为它们能够通过可解释的协方差关系实现灵活的替代模式。此外，如果需要更高的灵活性，还可以扩展模型，使确定性效用遵循神经网络的函数形式（Hruschka，2007年）。反之，如果研究者希望模型更加简约或易于解释，分析人员可以像Jiang 等人（2025）那样，通过惩罚方法约束协方差结构，也可以采用我们在第6节中展示的因子结构。尽管具有这些优点，文献中的多项Probit模型实证应用仍相对较少，主要原因在于评估多项Probit模型的选择概率颇具难度。

更广泛地说，多元正态概率模型只是具有相关误差项的一大类随机效用模型中的一种。与多元正态概率模型类似，其他模型也可以通过假设误差具有以下形式来构造： $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*$，其中 $\boldsymbol{\Sigma}^{1/2}$ 是平方根尺度矩阵， $\boldsymbol{\epsilon}^*$ 是可交换误差向量。例如， $\epsilon_j^* \overset{iid}{\sim} \text{Gumbel}(0,1)$ 会产生尺度矩阵为 $\boldsymbol{\Sigma}$的相关 Gumbel 分布。另一个例子是，我们可以假设 $\boldsymbol{\epsilon}^* \sim \text{Multivariate-}t(\mathbf{0}, \mathbf{I}\ \nu)$，通过自由度参数 $\nu > 0$($\mathbf{I}$表示单位矩阵)来捕捉厚尾行为。 Bagheri 等人 (2025) 考虑了以这种方式构造的相关误差分布，

由于只有在高斯情形下才有高效算法（GHK），他们不得不采用计算成本高昂的模拟方法来近似选择概率。

本研究的主要贡献是提出了一种摊销推断框架，可针对具有一般误差分布（包括具有非平凡相关结构的误差分布）的随机效用模型，生成准确、可复用且计算高效的选择概率估计。该框架为解决上述灵活性与可解释性之间的权衡提供了可能的方案，使灵活选择模型的实际估计成为可能，同时不牺牲随机效用模型所具备的经济学意义洞见和简约性。该框架具有坚实的理论依据，包括通用逼近保证，以及在温和近似条件下成立的渐近推断结果（见第 5 节）。此外，该框架与近期机器学习方法的进展大体上具有互补性，包括上述讨论的两类方法。在许多情况下，这些方法可以借助基于模拟器的似然评估得到增强，从而形成具有可解释替代模式且计算需求可控的灵活模型。

3 问题设定

本节概述问题设定，包括离散选择模型的不变性以及我们的推断目标。

3.1 离散选择模型

我们考虑一个决策者，从 K 互斥选项中进行选择。决策者为每个备选项 $j \in \{1, \dots, K\}$ 赋予一个潜在效用 U_j，并选择效用最高的备选项：

$$Y = \underset{j \in \{1,\dots,K\}}{\arg\max}\, U_j. \tag{1}$$

Th潜在效用分解为确定性部分和随机部分 ：

$$U_j = v_j + \epsilon_j, \quad j = 1, \dots, K, \tag{2}$$

其中， v_j 是取决于可观测特征的确定性（或系统性）效用，而 ϵ_j 是一个均值为零的随机误差，用于捕捉不可观测因素。如第1节所述，确定性效用通常采用线性形式 $v_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$，其中 $\boldsymbol{x}_j$ 是备选项特定属性向量， $\boldsymbol{\beta}$ 是待估计的参数向量。

对误差向量 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots\ldots, \epsilon_K)^\top$ 采用不同的分布假设，会得到不同的选择模型。多项 Logit 模型假设这些 ϵ_j 相互独立且服从相同的 Gumbel(0，1)分布。相比之下，多项式正态概率模型假设 $\boldsymbol{\epsilon}$服从均值为零、协方差矩阵为 $\boldsymbol{\Sigma}$的多元正态分布。下文将考虑相关模型，其中 $\boldsymbol{\epsilon} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\epsilon}^*$，且 $\boldsymbol{\Sigma}^{1/2}$ 是第2节所述的平方根尺度矩阵。我们对 $\boldsymbol{\epsilon}^*$作出如下假设。

假设 1。 误差向量 $\boldsymbol{\epsilon}^*$ 具有密度 $f_{\boldsymbol{\epsilon}^*} : \mathbb{R}^K \to [0, \infty)$，相对于勒贝格测度而言。此外， $\boldsymbol{\epsilon}^*$ 的各个元素可交换，因此对于所有 $\boldsymbol{\epsilon}' \in \mathbb{R}^K$ 以及任意置换矩阵$\mathbf{P}_\pi$， $f_{\boldsymbol{\epsilon}^*}(\mathbf{P}_\pi\boldsymbol{\epsilon}') = f_{\boldsymbol{\epsilon}^*}(\boldsymbol{\epsilon}')$。

这一设定和假设涵盖了多项Probit模型这一典型示例，其中 $f_{\boldsymbol{\epsilon}^*}$ 等于 $\mathcal{N}(\mathbf{0}_K, \mathbf{I}_K)$ 分布的概率密度函数。不过，这一设定足够通用，还涵盖许多其他可能性，例如第 t 2 节所述的相关 Gumbel 分布和多元分布。我们将在第 5 节中讨论其他正则条件，并在假定参数模型及其参数向量 $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$的背景下，推导我们框架的理论性质。

我们将备选项 j 的选择概率记为

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}) = \Pr(U_j \geq U_k \text{ for all } k \neq j) = \Pr(\epsilon_k - \epsilon_j \leq v_j - v_k \text{ for all } k \neq j), \tag{3}$$

其中 $\boldsymbol{v} = (v_1, \ldots, v_K)^\top$。在实际应用中，我们也可以选择纳入控制 $\boldsymbol{\epsilon}^*$分布的其他参数，例如第 ν2 节所述的自由度参数，但为简洁起见，记号中省略了这些参数。选择概率可表示为以下积分：

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}) = \int_{\mathcal{R}_j} f_{\boldsymbol{\epsilon}^*}(\boldsymbol{\epsilon}^*)\, d\boldsymbol{\epsilon}^*, \tag{4}$$

$$\mathcal{R}_j = \left\{\boldsymbol{\epsilon}^* \in \mathbb{R}^K : v_j + \left[\boldsymbol{\Sigma}^{1/2}\right]_{j\cdot}\boldsymbol{\epsilon}^* \geq v_k + \left[\boldsymbol{\Sigma}^{1/2}\right]_{k\cdot}\boldsymbol{\epsilon}^* \text{ for all } k \neq j\right\}, \tag{5}$$

其中， $\mathcal{R}_j$ 是选择备选项 j 的误差空间区域， $\left[\boldsymbol{\Sigma}^{1/2}\right]_{j\cdot}$ 表示 $\boldsymbol{\Sigma}^{1/2}$ 的第j行，因此 $\left[\boldsymbol{\Sigma}^{1/2}\right]_{j\cdot}\boldsymbol{\epsilon}^* = \epsilon_j$。一般来说，这个积分无法通过解析方法求解。尤其是在多元正态概率模型中， $K \geq 3$没有闭式解，因此需要采用数值方法，例如 GHK 算法，通过递归条件化来模拟这些概率。每次评估时，GHK 需要 R 次模拟抽样，其近似误差以 $O(R^{-1/2})$的速率递减。

3.2 不变性

选择概率满足三种基本的不变性。

位置不变性。 给所有效用加上一个常数，不会改变选择概率：

$$P_j(\boldsymbol{v} + C\mathbf{1}_K, \boldsymbol{\Sigma}) = P_j(\boldsymbol{v}, \boldsymbol{\Sigma}) \tag{6}$$

其中， $\mathbf{1}_K$ 是由 K个 $\mathbf{1}$ 组成的向量， $C \in \mathbb{R}$ 可能是随机变量。这是因为选择仅取决于效用差。

尺度不变性 。 对所有效用和尺度矩阵进行缩放，可保持选择概率不变：

$$P_j(\alpha\boldsymbol{v}, \alpha^2\boldsymbol{\Sigma}) = P_j(\boldsymbol{v}, \boldsymbol{\Sigma}) \quad \text{for all } \alpha > 0. \tag{7}$$

这是因为，缩放后的模型等价于将原模型中的 $\boldsymbol{v}$ 和 $\boldsymbol{\epsilon}$分别替换为 $\alpha\boldsymbol{v}$ 和 $\alpha\boldsymbol{\epsilon}$。

置换等变性。 重新标记备选项会相应地置换选择概率。对于任意一个 $\{1, \ldots, K\}$ 的置换 π，其对应的置换矩阵为 $\mathbf{P}_\pi$：

$$P_j(\boldsymbol{v}, \boldsymbol{\Sigma}) = P_{\pi^{-1}(j)}(\mathbf{P}_\pi\boldsymbol{v}, \mathbf{P}_\pi\boldsymbol{\Sigma}\mathbf{P}_\pi^\top). \tag{8}$$

构建模拟器的朴素方法可能无法满足这些性质。平移、缩放或置换备选项都可能导致模拟器预测不同，从而造成学习缓慢、泛化能力差，因为模拟器需要手动学习这些性质。在第 4 节中，我们提出了一种解决方案，通过两个设计组件自动满足这些性质：用于固定位置和尺度的预处理步骤，以及在不牺牲表达能力的前提下确保置换等变性的等变架构。

上述不变性意味着，未经归一化的选择模型无法识别。我们始终假设模型已进行适当归一化，使参数向量 $\boldsymbol{\theta}$ 可以识别。实践中，通常通过将一个备选项的效用确定性地设为零，并固定 $\boldsymbol{\Sigma}$ 中的一个非零对角元素来实现（参见Keane (1992年）。同样，我们假设数据具有足够的变异性，能够识别其余参数；这通常要求备选项特定协变量（如价格）存在变异。

3.3 推断目标

给定观测值 $\{(y_i,\ \mathbf{X}_i)\}_{i=1}^n$，其中 $y_i \in \{1,\dots,K\}$ 是观测选择，$\mathbf{X}_i$ 表示协变量，我们旨在估计参数向量 $\boldsymbol{\theta}$。该参数向量既包含影响确定性效用的系数，也包含用于控制尺度矩阵的其他参数。

（平均）对数似然函数为

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log P_{y_i}\left\{\boldsymbol{v}_i(\mathbf{X}_i,\boldsymbol{\theta}), \boldsymbol{\Sigma}_i(\mathbf{X}_i,\boldsymbol{\theta})\right\}, \tag{9}$$

其中，$\boldsymbol{v}_i(\mathbf{X}_i,\boldsymbol{\theta})$ 和 $\boldsymbol{\Sigma}_i(\mathbf{X}_i,\boldsymbol{\theta})$ 分别表示观测i 的效用向量和尺度矩阵，它们是协变量和参数的函数。最大似然估计需要对 Θ 上的 $\ell_n(\boldsymbol{\theta})$ 进行最大化，这需要反复计算 $\ell_n(\boldsymbol{\theta})$ 及其梯度$\nabla_{\boldsymbol{\theta}} \log\ P_j(\boldsymbol{v},\boldsymbol{\Sigma})$。类似地，哈密顿蒙特卡洛等贝叶斯方法需要反复计算 $\ell_n(\boldsymbol{\theta})$ 和 $\nabla_{\boldsymbol{\theta}}$ $\log\ P_j(\boldsymbol{v},\boldsymbol{\Sigma})$，在整个估计过程中通常要计算数千次。

ˆ 我们的目标是构建一个神经网络模拟器 $P_j(\boldsymbol{v},\boldsymbol{\Sigma};\boldsymbol{\phi})$，以快速、准确地近似真实的选择概率，并通过自动微分提供解析梯度。该模拟器只需在覆盖相关输入空间的模拟数据上训练一次，之后即可用于对多个数据集和模型进行快速推断。

4 模拟器设计

本节概述模拟器的预处理变换、神经网络架构、训练流程和推断过程。

4.1 设计目标

模拟器必须满足以下四项要求：

1. **遵循不变性**。 模拟器应遵循第 3.2 节所述的不变性：对位置平移和尺度变换具有不变性，对排列具有等变性。

2. **提供平滑的近似**。 模拟器应能够对其输入求导，从而支持基于梯度的优化和自动微分。

3. **适用于不同设定**。 单个训练后的模拟器应能适用于多种效用和协方差结构。

4. **提高计算效率**。 模拟器评估应足够快速，以便在估计中常规使用。

我们通过结合预处理变换与精心设计的神经网络架构来实现这些目标。

4.2 预处理：中心化与缩放

我们通过将输入投影到规范子空间的预处理变换，处理位置不变性和尺度不变性。

中心化。 定义中心化矩阵

$$\mathbf{M} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top, \tag{10}$$

我们使用该矩阵将效用向量 $\boldsymbol{U} \in \mathbb{R}^K$ 投影到与 $\mathbf{1}_K$ 正交的子空间，从而得到变换后的效用 $\boldsymbol{U} = \mathbf{M}(\boldsymbol{v} + \boldsymbol{\epsilon}) = \boldsymbol{v} + \boldsymbol{\epsilon} - \mathbf{1}_K(\bar{v} + \bar{\epsilon})$，其中$\bar{v} = \frac{1}{K}\sum_{j=1}^K v_j$ 且 $\bar{\epsilon} = \frac{1}{K}\sum_{j=1}^K \epsilon_j$。 这些变换后的效用在每次实现中之和为零： $\sum_{j=1}^K \tilde{U}_j = 0$。实际上，该变换会得到经过修改的模拟器输入：

$$\tilde{\boldsymbol{v}} = \mathbf{M}\boldsymbol{v} = \boldsymbol{v} - \mathbf{1}_K\bar{v}, \quad \tilde{\boldsymbol{\Sigma}} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top. \tag{11}$$

缩放。我们用中心化尺度矩阵的迹进行归一化，得到变换后的效用 $\boldsymbol{U}^* = \sqrt{K/\mathrm{tr}(\tilde{\boldsymbol{\Sigma}})}\tilde{\boldsymbol{U}}$。最终变换后的参数为

$$\boldsymbol{v}^* = \sqrt{\frac{K}{\mathrm{tr}(\tilde{\boldsymbol{\Sigma}})}}\,\tilde{\boldsymbol{v}}, \quad \boldsymbol{\Sigma}^* = \frac{K}{\mathrm{tr}(\tilde{\boldsymbol{\Sigma}})}\,\tilde{\boldsymbol{\Sigma}}. \tag{12}$$

~我们可以省略 $\mathrm{tr}(\boldsymbol{\Sigma}) = 0$ 的情况，因为这会产生已知的确定性选择。完整的变换得到归一化输入 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$，满足：

1. $\sum_{j=1}^K v_j^* = 0$ （中心化的确定性效用），

2. $\boldsymbol{\Sigma}^*\mathbf{1}_K = \mathbf{0}$ （中心化尺度矩阵），

3. $\text{tr}(\boldsymbol{\Sigma}^*) = K$（尺度归一化），

4. $\boldsymbol{\Sigma}^*$ 是半正定的，秩至多为 $K-1$。

这种归一化缩小了模拟器的输入空间，从而加快学习并改善泛化能力。重要的是，这种变换与置换矩阵可交换：对于任意置换矩阵 $\mathbf{P}_\pi$，将该变换应用于 $(\mathbf{P}_\pi \boldsymbol{v}, \mathbf{P}_\pi \boldsymbol{\Sigma} \mathbf{P}_\pi^\top)$，会得到 $(\mathbf{P}_\pi \boldsymbol{v}^*, \mathbf{P}_\pi \boldsymbol{\Sigma}^* \mathbf{P}_\pi^\top)$。这确保预处理保留置换等变性。

4.3 神经网络架构

预处理后，我们构建一个神经网络，将 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 映射为选择概率，同时保持置换等变性。该架构由三个组件组成：逐备选项编码器、置换等变层和输出层。下文为简化说明，将 $\boldsymbol{\Sigma}$ 视为协方差矩阵，并分别将其对角元素和非对角元素称为方差和协方差。这种解释适用于多元正态概率模型。对于其他模型，这仅是一种示意性解释，因为这些参数可能与方差和协方差并不完全对应。

逐备选项编码器。对于每个备选项 $j \in \{1, \ldots, K\}$，我们构建一个表示 $\boldsymbol{z}_j \in \mathbb{R}^{d_z}$，用于捕捉备选项 j 与所有其他备选项之间的关系。该表示由两个互补的 DeepSet 网络构成——一个对角 DeepSet 和一个非对角 DeepSet——并将它们的输出与备选项 j自身的特征相结合。对角 DeepSet 处理 j 与每个其他备选项之间的成对关系，而非对角 DeepSet 则概括除 j之外的备选项之间的协方差结构。在下文中，标记为“基础输入”的特征足以实现选择概率的通用逼近（见第5.1节）。我们还加入了其他特征，以增强模拟器的表达能力，因为我们发现这样可以改善其预测效果。

对角 DeepSet。对于每一对 $(j,\ k)$（满足 $k \neq j$），我们构建一个特征向量 $\boldsymbol{d}_{jk}$，其中包含与备选项 j 和 k 相关的效用和协方差的可微函数。基础输入为效用 v_j^* 和 v_k^*、方差 Σ_{jj}^* 和 Σ_{kk}^*，以及协方差 Σ_{jk}^*。基于这些输入，我们推导出有助于学习的额外特征变换，包括标准差 $\sigma_j = \sqrt{\Sigma_{jj}^*}$ 和 $\sigma_k = \sqrt{\Sigma_{kk}^*}$、相关性 $\rho_{jk} = \Sigma_{jk}^* / (\sigma_j \sigma_k)$，以及标准化效用差（z分数）：

$$z_{jk} = \frac{v_j^* - v_k^*}{\sqrt{\Sigma_{jj}^* + \Sigma_{kk}^* - 2\Sigma_{jk}^*}}. \tag{13}$$

对角 DeepSet 使用 Zaheer等人 (2017年) 提出的架构处理这些特征：

$$\boldsymbol{h}_j^{\text{diag}} = \rho_{\text{diag}}\left(\sum_{k\neq j}\phi_{\text{diag}}(\boldsymbol{d}_{jk})\right), \tag{14}$$

其中 ϕ_{diag} 是应用于每一对对象的多层感知机（多层感知机），求和操作对所有 $k \neq j$进行聚合，而 ρ_{diag} 是另一个多层感知机，用于生成求和结果的学习得到的非线性表示。我们采用 ϕ_{diag} 和 ρ_{diag} 的记法，遵循Zaheer等人 (2017年)；多层感知机 ρ_{diag} 和ρ_{off} （如下文所定义）可根据其下标与相关性 ρ_{jk} 区分开来。

Singh等人(2023年)也采用了 DeepSet 架构的一种变体。不过，他们的方法假设误差独立同分布，并基于特定问题特征运行，因此需要修改并重新训练其神经网络以适应新的数据结构。相比之下，我们的架构依赖通用的随机效用模型输入，例如确定性效用。

非对角 DeepSet。 对于每一对 $(k,\ l)$，其中 $k < l$ 且 $k,\ l \neq j$，我们构造特征向量 $\boldsymbol{o}_{kl}$，其中包含用于比较备选项 k 和 l的可微对称函数；由于这些配对没有天然的顺序，因此我们要求函数具有对称性。基础输入为 Σ^*_{kl}，我们还纳入其他派生特征，例如相关性 ρ_{kl}、效用差平方（$v_k^* - v_l^*$）^{2},、z 分数平方 z_{kl}^2以及方差和 $\Sigma^*_{kk} + \Sigma^*_{ll}$。

非对角 DeepSet 的结构与对角 DeepSet 类似：

$$\boldsymbol{h}_j^{\text{off}} = \rho_{\text{off}}\left(\sum_{\substack{k<l\\k,l\neq j}}\phi_{\text{off}}(\boldsymbol{o}_{kl})\right). \tag{15}$$

组合网络。 两个 DeepSets 的输出与额外的“直通特征” $\boldsymbol{s}_j$拼接在一起，例如 v_j^*、 σ_j，以及备选项 j 与其他备选项之间的协方差汇总统计量。与上述额外特征一样，直通特征并非实现通用逼近所必需，但我们发现它们在经验上能提升性能。随后，拼接结果由组合多层感知机 ζ 处理，以生成逐备选项表示：

$$\boldsymbol{z}_j = \zeta\left(\boldsymbol{h}_j^{\text{diag}}, \boldsymbol{h}_j^{\text{off}}, \boldsymbol{s}_j\right). \tag{16}$$

置换等变层。将逐备选项表示堆叠成矩阵$\mathbf{Z} \in \mathbb{R}^{K\times d_z}$，并通过线性置换等变层（Zaheer

等人，2017年），从而实现备选项之间的信息交换。这些层的形式为

$$L(\mathbf{Z}) = \sigma\left(\mathbf{Z}\mathbf{A} + \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\mathbf{Z}\mathbf{B} + \mathbf{1}_K\boldsymbol{c}^\top\right), \tag{17}$$

其中 $\mathbf{A}$、$\mathbf{B}$ $\in \mathbb{R}^{d_{\text{in}}\times d_{\text{out}}}$ 和 $\boldsymbol{c} \in \mathbb{R}^{d_{\text{out}}}$ 是可学习参数，而 σ 是平滑激活函数。第一项对每个备选项分别进行变换，第二项汇总备选项之间的信息，第三项则添加共享偏差（截距）。

输出层。 最后一层为每个备选项生成一个逻辑斯蒂模型：

$$\mathbf{logit} = \mathbf{Z}^{(L)}\boldsymbol{a}, \tag{18}$$

其中 $\mathbf{Z}^{(L)}$ 是最后一个等变层的输出， $\boldsymbol{a} \in \mathbb{R}^{d_L}$ 使得输出**逻辑斯蒂模型**是 $\mathbb{R}^K$中的一个向量。我们省略其他项，因为它们只会产生常数平移，不会影响输出概率。

选择概率　　通过对输出层应用 softmax 函数获得。　　er:

$$\hat{P}_j = \frac{\exp(\text{logit}_j)}{\sum_{k=1}^K \exp(\text{logit}_k)}. \tag{19}$$

对于概率单纯形 Δ^{K-1}上的任意有效选择概率向量（P_1，……， P_K），都存在相应的逻辑值向量，使得通过 softmax函数可得到（P_1，……， P_K），因此这种参数化不失一般性。图 1展示了整体架构。

激活函数。 在整个架构中，我们使用平滑（无限可微）激活函数。这确保模拟器输出是输入的平滑函数，从而能够为下游模型拟合准确地计算梯度。经验表明，Swish 激活函数 Swish(x) $= x/(1+e^{-x})$ 效果良好（ Ramachandran 等人，2018 年），不过也可以使用其他平滑激活函数，例如双曲正切或 softplus（Dugas 等人，2000年）。

计算复杂度。 由于需要计算非对角 DeepSet，逐备选项编码器的朴素实现需要 $O(K^3)$ 次前向传播运算。不过，通过预先计算所有非对角配对的总和，并通过减法排除涉及备选项 j的配对，我们可以将复杂度降至 $O(K^2)$。由于 $\boldsymbol{\Sigma}$有 $O(K^2)$ 个影响选择概率的不同元素，因此这种缩放对于一般的 $\boldsymbol{\Sigma}$而言是最优的。

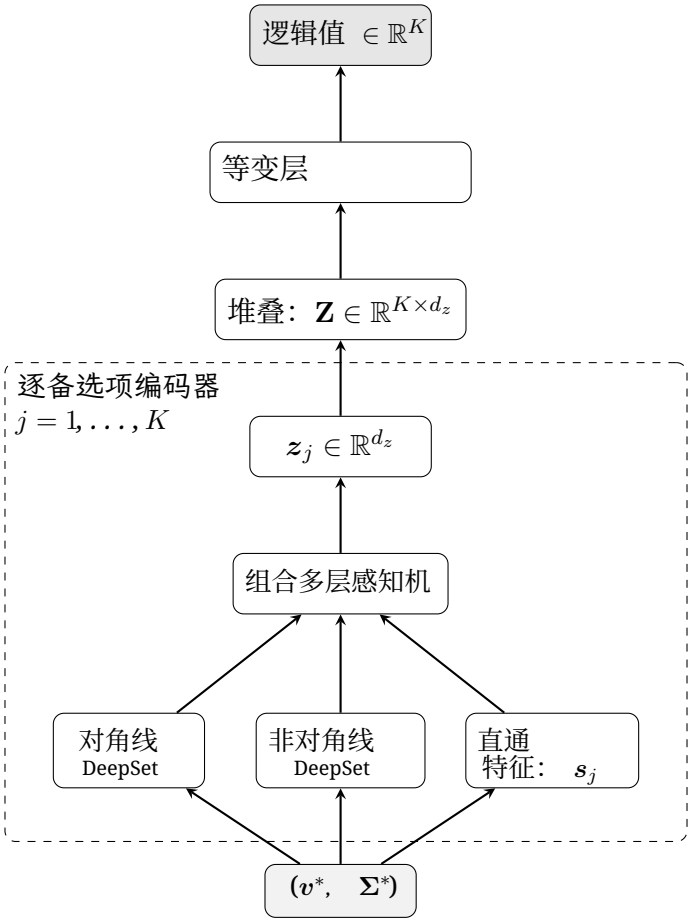


图1：神经网络模拟器的架构。对于每个备选项 j，逐备选项编码器会处理对角特征（描述 j 与其他备选项之间关系）、非对角特征（除 j 之外的备选项之间的成对特征），以及备选项 j 自身的特征。通过组合多层感知机将这些特征合并，生成表示 $\boldsymbol{z}_j$。所有备选项的表示会堆叠起来，并经过置换等变层处理，最终生成逻辑值。

网络规模。 该架构可以很紧凑。在我们的模拟中，每个组件使用包含 8–64 个隐藏单元和 0–2 层隐藏层的网络。通常， K 的值越大，要实现准确的近似就需要更大的网络。

与 Blum-Smith 等人 (2025) 的关系。 我们提出的架构可视为对 Blum-Smith 等人 (2025) 所提出的 DS-CI 架构的扩展，二者有三项关键区别。首先，Blum-Smith 等人 (2025) 需要一个结合对角元素与非对角元素的"连接特征"，而我们的中心化过程使这一特征变得多余，从而简化了所需输入的集合。其次，我们的问题包含一个额外的向量输入（$\boldsymbol{v}$），这是 Blum-Smith 等人 (2025) 未考虑的；这一额外复杂性要求我们将确定性效用值与对角方差元素耦合起来。第三，由于我们的问题具有等变性，我们需要分别对每个备选项进行编码，因此在逐备选项编码器中将一个指定备选项（j）与其他竞争备选项（$k \neq j$）区分开来。

4.4 训练流程

我们使用涵盖多种效用与协方差配置的模拟数据来训练模拟器，并采用索博列夫训练，以确保对选择概率及其梯度进行准确近似。

数据生成。为了生成第6节中涉及多项式正态概率模型的模拟所需的每个训练样本，我们遵循以下流程：

1. **误差协方差**。 从 Wishart($\mathbf{I}_K/(K+2)$, $K+2$) 分布中抽取一个协方差矩阵 $\boldsymbol{\Sigma}$。2. **原始确定性效用**。 抽取独立效用 $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}_K)$。3. **确定性效用协方差**。 （独立地从与 $\boldsymbol{\Sigma}$ 相同的分布中）抽取第二个协方差矩阵 $\boldsymbol{\Sigma}'$，抽取 $\omega \sim \text{Uniform}(0, 1)$，并构造凸组合 $\boldsymbol{\Sigma}'' = \omega\boldsymbol{\Sigma} + (1-\omega)\boldsymbol{\Sigma}'$。4. **变换后的确定性效用**。 变换效用，使其协方差结构与 $\boldsymbol{\Sigma}''$ 相关： $\boldsymbol{v}' = (\boldsymbol{\Sigma}'')^{1/2}\boldsymbol{v}$。由此得到 $\text{Cov}(\boldsymbol{v}') = \boldsymbol{\Sigma}''$。5. **预处理**。 应用预处理变换，从 $(\boldsymbol{v}', \boldsymbol{\Sigma})$ 得到 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$。6. **模拟选择**。 使用效用 $\boldsymbol{v}^*$ 和尺度矩阵 $\boldsymbol{\Sigma}^*$（或等价地，使用 $\boldsymbol{v}'$ 和 $\boldsymbol{\Sigma}$），从多项式正态概率模型中模拟 10^6 个选择。

步骤 3 和 4 确保确定性效用与误差协方差之间的关系具有多样性。当 $\omega = 1$ 时，效用与协方差结构完全一致；当 $\omega = 0$ 时，它们相互独立。中间值则会产生部分相关性。

通过在最后一步中将多项式正态概率模型替换为具有不同误差分布的其他模型，可以将该过程推广到其他误差分布。同样，可以根据问题定义域和计算资源，对所列出的其他分布以及选择数量（10^6）进行调整。

损失函数。 我们采用索博列夫训练（Czarnecki 等人，2017年），在标准似然损失的基础上增加梯度匹配惩罚项。总损失为：

$$\mathcal{L}(\boldsymbol{\phi}) = \mathcal{L}_{\text{CE}}(\boldsymbol{\phi}) + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}(\boldsymbol{\phi}), \tag{20}$$

其中，$\lambda_{\text{grad}} > 0$ 控制梯度匹配项的权重。

交叉熵 项是多项负对数似然（忽略常数项）：

$$\mathcal{L}_{\text{CE}}(\boldsymbol{\phi}) = -\sum_{j=1}^{K} \hat{P}_j^{\text{sim}} \log \hat{P}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*; \boldsymbol{\phi}), \tag{21}$$

其中，$\hat{P}_j^{\text{sim}} = n_j / \sum_{k=1}^{K} n_k$ 是备选项 j 的模拟选择频率。

对于梯度匹配分量，我们采用Czarnecki 等人 (2017年提出的随机索博列夫训练方法，该方法通过匹配随机方向上的方向导数，避免计算完整的雅可比矩阵。令 $\boldsymbol{d} = (\boldsymbol{d}_v, \text{vech}(\mathbf{D}_\Sigma))$ 表示效用与协方差联合空间中的一个随机方向，其中 $\boldsymbol{d}_v \in \mathbb{R}^K$, $\mathbf{D}_\Sigma \in \mathbb{R}^{K \times K}$ 为对称矩阵，vech(·) 表示半向量化（包括对角线在内的上三角元素）。梯度匹配损失为：

$$\mathcal{L}_{\text{grad}}(\boldsymbol{\phi}) = \sum_{j=1}^{K} \hat{P}_j^{\text{sim}} \left(\nabla_{\boldsymbol{d}} \log \hat{P}_j - \nabla_{\boldsymbol{d}} \log P_j^{\text{target}} \right)^2, \tag{22}$$

其中，$\nabla_{\boldsymbol{d}}$ 表示沿 $\boldsymbol{d}$方向的方向导数。通过模拟选择频率$\hat{P}_j^{\text{sim}}$进行加权，可将学习重点放在经常被选择的备选项上，因为这些备选项的梯度对下游推断任务中的参数估计影响最大。

约束残差 **预期方向**。规范输入（$\boldsymbol{v}^*$, $\boldsymbol{\Sigma}^*$）位于一个受约束的

流形上：

$$\mathcal{X}_K = \left\{ (\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) : \sum_{j=1}^{K} v_j^* = 0,\ \boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0},\ \mathrm{tr}(\boldsymbol{\Sigma}^*) = K,\ \boldsymbol{\Sigma}^* \succeq 0 \right\}. \tag{23}$$

随机方向必须满足这些约束，才能保持在 $\mathcal{X}_K$ 的切空间内。我们按以下方式构造有效方向：

1. 采样 $\tilde{\boldsymbol{d}}_v \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ 并进行中心化： $\boldsymbol{d}_v = \tilde{\boldsymbol{d}}_v - \bar{d}_v \mathbf{1}_K$，其中 $\bar{d}_v = \frac{1}{K}\sum_j \tilde{d}_{v,j}$。

2. 采样一个对称矩阵 $\mathbf{D}_\Sigma$，其对角线及对角线上方的元素为独立同分布的标准正态分布元素，然后进行双重中心化：

$$\mathbf{D}'_\Sigma = \mathbf{M}\tilde{\mathbf{D}}_\Sigma \mathbf{M}, \tag{24}$$

其中 $\mathbf{M} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ 是第 4节中的中心化矩阵。

3. 去除迹分量，以确保 $\mathrm{tr}(\mathbf{D}_\Sigma) = 0$：

$$\mathbf{D}_\Sigma = \mathbf{D}'_\Sigma - \frac{\mathrm{tr}(\mathbf{D}'_\Sigma)}{K-1}\mathbf{M}. \tag{25}$$

4. 设置 $\boldsymbol{d} = (\boldsymbol{d}_v, \mathrm{vech}(\mathbf{D}_\Sigma))$ 并进行归一化： $\boldsymbol{d} \leftarrow \boldsymbol{d}/\|\boldsymbol{d}\|$。

计算方向导数。我们通过有限差分计算模拟器的方向导数：

$$\nabla_{\boldsymbol{d}} \log \hat{P}_j \approx \frac{\log \hat{P}_j(\boldsymbol{v}^* + \epsilon \boldsymbol{d}_v, \boldsymbol{\Sigma}^* + \epsilon \boldsymbol{d}_\Sigma; \boldsymbol{\phi}) - \log \hat{P}_j(\boldsymbol{v}^* - \epsilon \boldsymbol{d}_v, \boldsymbol{\Sigma}^* - \epsilon \boldsymbol{d}_\Sigma; \boldsymbol{\phi})}{2\epsilon}, \tag{26}$$

其中 $\epsilon > 0$ 是一个较小的步长（我们使用 $\epsilon = 10^{-5}$）。这种方法计算效率很高，对于每个方向只需额外进行两次前向传播，且不受输入维度影响。或者，也可以像Czarnecki 等人（2017年）那样，通过自动微分计算方向导数。

目标方向导数根据预先存储的雅可比矩阵计算。生成数据时，我们计算并存储完整的雅可比矩阵 $\nabla_{\boldsymbol{\theta}} \log P_j^{\text{target}}$，其相对于 $\boldsymbol{v}^*$ 和 $\boldsymbol{\Sigma}^*$。训练时，通过内积得到目标方向导数：

$$\nabla_{\boldsymbol{d}} \log P_j^{\text{target}} = \left\langle \frac{\partial \log P_j^{\text{target}}}{\partial \boldsymbol{v}^*}, \boldsymbol{d}_v \right\rangle + \left\langle \frac{\partial \log P_j^{\text{target}}}{\partial \boldsymbol{\Sigma}^*}, \mathbf{D}_\Sigma \right\rangle_F, \tag{27}$$

其中 $\langle \cdot, \cdot \rangle_F$ 表示弗罗贝尼乌斯内积。

目标梯度计算。索博列夫训练中的目标梯度计算需要对选择模拟过程进行求导。由于离散选择本质上不可微，我们采用温度缩放 softmax 进行平滑松弛。对于每次蒙特卡洛抽样 r，其效用向量为 $\boldsymbol{U}_{(r)} = \boldsymbol{v}^* + \boldsymbol{\epsilon}^{(r)}$，我们用以下方式近似硬选择指示函数：

$$\tilde{Y}_j^{(r)} = \frac{\exp(\tau U_j^{(r)})}{\sum_{k=1}^{K} \exp(\tau U_k^{(r)})}, \tag{28}$$

其中 $\tau > 0$ 是温度参数。当 $\tau \to \infty$时，该表达式收敛于硬 argmax；对于有限的 τ，它提供了一种可微近似。目标对数概率为：

$$\log P_j^{\text{target}} = \log \left(\frac{1}{R} \sum_{r=1}^{R} \tilde{Y}_j^{(r)} \right), \tag{29}$$

以及目标梯度 $\partial \log P_j^{\text{target}} / \partial \boldsymbol{v}^*$ 和 $\partial \log P_{\text{目标}_j} / \partial \boldsymbol{\Sigma}^*$ 均通过对该表达式进行自动微分获得。这种方法与神经网络梯度估计中常用的“Gumbel–softmax 技巧”密切相关（Jang 等人，2017年；Maddison 等人，2017年）。我们在实验中使用 $\tau = 3$，以兼顾梯度精度（低偏差）和计算稳定性。

训练细节。我们的训练流程使用两个重放缓冲区，这是深度学习中的一种常见技术（Mnih 等人，2015年）。第一个重放缓冲区存储预先计算的训练样本，包括输入（$\boldsymbol{v}^*$，$\boldsymbol{\Sigma}^*$）、模拟选择频率和目标雅可比矩阵。第二个缓冲区存储用于梯度匹配损失的随机方向。在每个训练步骤中，我们从重放缓冲区中随机抽取样本对，计算损失，并通过随机梯度下降更新神经网络参数。重放缓冲区通过重复利用计算成本高昂的高精度模拟和梯度计算，加快了训练速度。我们在实验中将梯度惩罚权重设为 $\lambda_{\text{grad}} = 10^{-8}$；根据经验，我们发现该值能够使导数保持足够良好的性质。

多重K 训练。值得注意的是，所有权重向量和矩阵的维数都不依赖于 K。维数 K 仅通过输入数据维度进入神经网络；因此，我们可以设计一个能够同时处理多个K 值的单一模拟器。随后，可以调整上述训练流程，使其遍历一系列 K值，例如 $K \in \{3, 4, 5\}$，这样得到的模拟器便能够学习近似选择数量变化时的选择概率（见第 6.3节）。在此配置下，我们建议将 K 作为 $\boldsymbol{s}_j$的一个元素。

4.5 使用训练后的模拟器进行推断

训练完成后，模拟器能够实现快速推断。给定一个新的数据集 $\{(y_i, \mathbf{X}_i)\}_{i=1}^n$ 以及候选参数向量 $\boldsymbol{\theta}$:

1. 对每个观测值，计算确定性效用向量 $\boldsymbol{v}_i(\mathbf{X}_i, \boldsymbol{\theta})$ 和尺度矩阵 $\boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\theta})$。

2. 应用预处理变换，得到 $(\boldsymbol{v}_i^*, \boldsymbol{\Sigma}_i^*)$。

3. 评估模拟器，得到所有备选项的估计选择概率 $\hat{P}_{ij}$。

4. 计算对数似然 $\hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log \hat{P}_{i,y_i}$。

由于预处理变换和神经网络都是平滑函数，因此可以通过自动微分精确计算模拟器对数似然相对于 $\boldsymbol{\theta}$ 的梯度。这使得我们能够使用基于梯度的优化器进行最大似然估计，也能够使用哈密顿蒙特卡洛等基于梯度的采样方法进行贝叶斯推断。模拟器的平滑性还有助于计算标准误，这些标准误可以通过梯度外积估计（参见第 5.2节）。

5 理论

本节将为模拟器架构以及基于模拟器的离散选择估计量的统计性质奠定理论基础。

5.1 通用逼近

我们首先证明，逐备选项编码器架构能够通用逼近选择概率。关键在于，编码器处理的多重集通常能够在相关对称群的作用下唯一确定输入。对于每个备选项 j，定义不变函数 $g_j: \mathcal{X}_K \to$ (多重集，多重集)，如下所示:

$$\mathcal{T}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) = \{\{(v_k^*, \Sigma_{kk}^*, \Sigma_{jk}^*) : k \neq j\}\}, \tag{30}$$

$$\mathcal{O}_j(\boldsymbol{\Sigma}^*) = \{\{\Sigma_{kl}^* : 1 \leq k < l \leq K, k \neq j, l \neq j\}\}, \tag{31}$$

$$g_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) = (\mathcal{T}_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \quad \mathcal{O}_j(\boldsymbol{\Sigma}^*))。 \tag{32}$$

多重集 $\mathcal{T}_j$ 由 $K-1$ 个三元组组成，每个三元组将备选项 k的效用和方差与其和备选项 j 的协方差关联起来。多重集 $\mathcal{O}_j$ 由非对角协方差组成

这些协方差对应于除 j之外的备选项之间的协方差。它们正是逐备选项编码器的对角 DeepSets 和非对角 DeepSets 所处理的基础输入。

定理 1 (一般性分离)。对于每个备选项 $j \in \{1, \dots, K\}$，存在一个闭的零测集 $\mathcal{B}_j \subset \mathcal{X}_K$，使得对于任意 $(\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*), (\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \in \mathcal{X}_K \setminus \mathcal{B}_j$:

$$g_j(\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*) = g_j(\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \implies (\boldsymbol{v}_1^*, \boldsymbol{\Sigma}_1^*) \text{ and } (\boldsymbol{v}_2^*, \boldsymbol{\Sigma}_2^*) \text{ are in the same } S_{K-1}^{(j)}\text{-orbit}, \tag{33}$$

其中， $S_{K-1}^{(j)}$ 是固定备选项 j的排列子群。

证明见附录 A.8，分为三个步骤:

步骤1：一般性条件。我们在 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 上定义两个一般性条件：(a) 对于 $k \neq j$，三元组 $(v_k^*, \Sigma_{kk}^*, \Sigma_{jk}^*)$ 两两不同，且属于 $\mathbb{R}^3$；(b) 对于每个$k \neq j$，都存在唯一一个非对角协方差子集，其和为 $-\Sigma_{kk}^* - \Sigma_{jk}^*$。条件 (b) 由行和约束 $\boldsymbol{\Sigma}^* \mathbf{1}_K = \mathbf{0}$ 得出：第 k 行中的非对角项（不包括 Σ_{ki}）之和为 $-\Sigma_{kk}^* - \Sigma_{jk}^*$。

步骤2：恢复过程。当两个条件均满足时，我们可以从g_j 中恢复（$\boldsymbol{v}^*$， $\boldsymbol{\Sigma}^*$），但除 j之外的备选项可以重新标记。首先，按字典序排列 $\mathcal{T}_j$ 中的三元组，为备选项分配标签。然后，对于每个备选项 k，将其非对角项所在的行识别为 $\mathcal{O}_j$ 中满足所需总和的唯一子集。最后，将行 k 和 l对应的子集的单元素交集作为共享项 Σ_{kl}^*进行恢复。

步骤3：零测例外集。满足以下条件的集合 $\mathcal{B}_j$（其中任一条件不成立）是仿射子空间 $\mathcal{X}_K$上非平凡解析函数的零点集的有限并。根据真解析子簇为零测集的原理（Mityagin, 2020年），$\mathcal{B}_j$ 是零测集。

这一分离结果结合 Stone–Weierstrass 定理，得到通用近似:

定理 2 （多层感知机通用逼近）。在假设 *1*下，对于每个备选项 j，选择概率函数 P_j 都可以在 $\mathcal{X}_K \setminus \mathcal{B}_j$ 的任意紧子集上由一个多层感知机进行一致近似，该多层感知机以 g_j 的组件作为输入。

证明概要。我们采用 Blum-Smith 等人 (2025) 的证明策略。选择概率 P_j 是连续且 $S_{K-1}^{(j)}$ 不变的。根据定理 1，对 g_j进行编码的函数——排序后的三元组和排序后的非对角项——在商空间中能够区分不同点 $(\mathcal{X}_K \setminus \mathcal{B}_j) / S_{K-1}^{(j)}$。由于 $S_{K-1}^{(j)}$ 是有限的（因此是紧的），标准结果表明该商空间是豪斯多夫空间

(Bredon，1972年）。根据斯通–魏尔施特拉斯定理，由这些函数生成的代数在商空间的紧子集上的连续函数中稠密。将其拉回到 $\mathcal{X}_K$，并应用多层感知机的通用逼近定理（Cybenko，1989年；Hornik 等人，1989年），即可得到结果。详情请参见附录 A.8。 □

推论1（架构通用逼近）。令 $\mathcal{B} = \cup_{j=1}^{K} \mathcal{B}_j$。在假设*1*下，对于每个备选项 j，选择概率函数 P_j 都可以由逐备选项编码器在 $\mathcal{X}_K \setminus \mathcal{B}_j$ 的任意紧子集上一致逼近。此外，选择概率向量 $(P_1, \ldots, P_K)$ 可以由第 $\mathcal{X}_K \setminus \mathcal{B}$节的神经网络架构在其任意紧子集上一致逼近 *4.3*。

证明。结合定理 2 的证明，以及 DeepSet（Zaheer等人，2017年）和等变架构（Segol和Lipman，2020年）的通用逼近结果即可。 □

例外集 $\mathcal{B}$ 的测度为零，因此在实践中不会影响近似：对于任何相对于勒贝格测度绝对连续、定义在 $\mathcal{X}_K$ 上的概率分布，输入落在 $\mathcal{B}$ 中的概率为零。

推论 1 保证仅使用逐备选项编码器就足以实现通用近似。加入最后的等变层可确保估计得到的概率之和为 1，并提升神经网络的表达能力。

对于较大的选择集，我们可以选择一种更简单的架构：仅对被选中的备选项应用逐备选项编码器，不使用任何等变层，从而得到单个输出 P_j。这一修改保留了通用逼近性质，但以放宽和为一约束为代价，提高了计算效率。

推论 1 也可以扩展为对有限个 K 取值同时成立，前提是对空间 $\mathcal{X}_K$ 及其例外集作适当扩展。第 6.3 节给出了数值结果，表明所提出的架构能够同时对多个 K 取值进行准确近似。

5.2 一致性与渐近正态性

现在，我们证明在适当条件下，基于模拟器的估计量继承了精确极大似然估计量的渐近性质。

令 $\{(y_i, \mathbf{X}_i)\}_{i=1}^{n}$ 为来自参数为 $\boldsymbol{\theta}_0 \in \Theta$ 的真实模型的独立同分布观测值。定义

真实与模拟器对数似然贡献为

$$m(y, \mathbf{X}; \boldsymbol{\theta}) = \log P_y \left\{ \boldsymbol{v}^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}^*(\mathbf{X}, \boldsymbol{\theta}) \right\}, \tag{34}$$

$$\hat{m}_n(y, \mathbf{X}; \boldsymbol{\theta}) = \log \hat{P}_y \left\{ \boldsymbol{v}^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}^*(\mathbf{X}, \boldsymbol{\theta}); \boldsymbol{\phi}_n \right\}, \tag{35}$$

其对应的得分函数

$$\psi(y, \mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} m(y, \mathbf{X}; \boldsymbol{\theta}), \qquad \hat{\psi}_n(y, \mathbf{X}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \hat{m}_n(y, \mathbf{X}; \boldsymbol{\theta}). \tag{36}$$

真实与模拟器（平均）对数似然分别为 $\ell_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n m(y_i, \mathbf{X}_i; \boldsymbol{\theta})$ 和 $\hat{\ell}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \hat{m}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})$。定义最大似然估计量和基于模拟器的估计量为

$$\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}), \qquad \hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} \hat{\ell}_n(\boldsymbol{\theta}). \tag{37}$$

我们要求模拟器对数似然收敛于真实对数似然。

假设2（模拟器近似质量）。对于某个 $\alpha \geq 0$，模拟器满足

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) \right| = o_p(n^{-\alpha}). \tag{38}$$

这里的 n 指假设2 中提到的观测数据样本量。因此，该假设涉及一种渐近制度：随着 $n \to \infty$ 增大，模拟器变得越来越准确。这可以通过逐步 (a) 增加模拟训练样本的数量，以及 (b) 提高神经网络复杂度（即增加层数和隐藏单元的数量）来实现。附录B.1 给出了假设2 更原始的充分条件，并引用了神经网络收敛速率及其在计量经济学理论中应用的密切相关研究。值得注意的是，对于具有解析选择概率的模型，我们预计该假设只需要 $O(n^{\alpha+\delta})$ 个训练样本，即可适用于任意 $\delta > 0$。

我们还要求真实模型具有充分的正则性。

假设3（正则性）。该模型满足：

(i) $\boldsymbol{\theta}_0$ 是 $\mathbb{E}\{m(y,\mathbf{X};\boldsymbol{\theta})\}$ 在紧集 Θ 上的唯一最大化点，其中 $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$且 $\mathbb{E}\{|m(y, \mathbf{X}; \boldsymbol{\theta}_0)|\} < \infty$；

(ii) $m(y, \mathbf{X}; \boldsymbol{\theta})$关于 $\boldsymbol{\theta}$ 二次连续可微，对所有$(y,\mathbf{X})$均成立；

(iii) $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\Theta} \|\psi(y,\mathbf{X}; \boldsymbol{\theta})\|^2\} < \infty$;

(iv) $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\Theta}\|\nabla^2_{\boldsymbol{\theta}} m(y,\mathbf{X};\boldsymbol{\theta})\|\} < \infty$;

(v) 费希尔信息$\mathbf{J}(\boldsymbol{\theta}_0) = \mathbb{E}\left\{\psi(y,\ \mathbf{X};\ \boldsymbol{\theta}_0)\psi(y,\ \mathbf{X};\ \boldsymbol{\theta}_0)^\top\right\}$是正定的。

定理 3 (一致性)。 在假设 2 和 3 下，$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ 当 $n \to \infty$时。

证明。根据假设2及 $\alpha \geq 0$，$\sup_{\boldsymbol{\theta}} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| \xrightarrow{p} 0$。此外，假设3的正则条件足以应用一致大数定律（ULLN），从而 $\sup_{\boldsymbol{\theta}} |\ell_n(\boldsymbol{\theta}) - \mathbb{E}\{m(y, \mathbf{X};\ \boldsymbol{\theta})\}| \xrightarrow{p} 0$。由三角不等式可得，$\ell_n$ 的估计量$\hat{}$一致收敛于 $\mathbb{E}\{m(y, \mathbf{X};\ \boldsymbol{\theta})\}$，根据假设3(i)，其在 $\boldsymbol{\theta}_0$ 处取得唯一最大值。根据极值估计量一致性定理（Newey 和 McFadden，1994 年，定理 2.1），$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$。

□

在更强的条件 $\alpha \geq 1$ 下，基于模拟器的估计量渐近正态。

定理 4 （渐近正态性）。 在假设 2 和 3 成立且满足 $\alpha \geq 1$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\right\}. \tag{39}$$

$\hat{}$ ˜附录 B.2中的证明表明， $\boldsymbol{\theta}_n$ 与 $\boldsymbol{\theta}_n$渐近等价。

备注 1（近似极大值点）。 如果最大值不是唯一取得的，可以采用满足以下条件的近似极大值点 $\hat{}\hat{}\hat{}$ $\ell_n(\boldsymbol{\theta}_n) \geq$上确界$_{\boldsymbol{\theta}}\, \ell_n(\boldsymbol{\theta}) - o_p(n^{-\alpha})$。一致性和渐近正态性结果仍然成立；参见 Newey 和 McFadden(1994 年，定理 2.1）及附录 B.2。

为估计费希尔信息，定义真实得分函数外积估计量和模拟器得分函数外积估计量：

$$\mathbf{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \mathbf{X}_i; \boldsymbol{\theta})\psi(y_i, \mathbf{X}_i; \boldsymbol{\theta})^\top, \tag{40}$$

$$\hat{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})\hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})^\top. \tag{41}$$

假设 4 （梯度近似质量）。 模拟器满足 s

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\hat{\mathbf{J}}_n(\boldsymbol{\theta}) - \mathbf{J}_n(\boldsymbol{\theta})\right\| \xrightarrow{p} 0 \quad \text{as } n \to \infty. \tag{42}$$

假设4 可通过索博列夫训练得到论证；这种训练会惩罚模拟器梯度与目标梯度之间的差异。之所以需要这种惩罚，是因为仅在对数似然上进行训练

值并不能保证梯度收敛。与假设 2一样，关于更基础的充分条件及更多讨论，请参见附录 B.1 。

命题 1 (费舍尔信息的一致估计)。 在假设 *2*、*3* 和 *4*下，$\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{J}(\boldsymbol{\theta}_0)$。

证明。根据三角不等式，$\|\hat{\mathbf{J}}_n(\boldsymbol{\theta}_n) - \mathbf{J}(\boldsymbol{\theta}_0)\| \leq \|\hat{\mathbf{J}}_n(\boldsymbol{\theta}_n) - \mathbf{J}_n(\boldsymbol{\theta}_n)\| + \|\mathbf{J}_n(\boldsymbol{\theta}_n) - \mathbf{J}(\boldsymbol{\theta}_n)\| + \|\mathbf{J}(\boldsymbol{\theta}_n) - \mathbf{J}(\boldsymbol{\theta}_0)\|$. 第一项根据假设 4 趋于零。对于第二项，假设3的正则条件足以对 $\mathbf{J}_n$ 应用一致大数定律；因此，该项也依概率收敛于零。最后，第三项根据 $\mathbf{J}$ 的连续性以及 $\boldsymbol{\theta}_n$ 的一致性趋于零。

□

模拟器得分函数 $\hat{\psi}_n(y_i, \mathbf{X}_i; \boldsymbol{\theta})$ 可通过对光滑神经网络进行自动微分计算得到。 $\boldsymbol{\theta}_n$ 的标准误可由$\hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n)^{-1}/n$的对角元素的平方根得到。

另一种基于海森矩阵的估计量，$\tilde{\mathbf{J}}_n = -\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\boldsymbol{\theta}} \hat{m}_n(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)$，需要达到一阶索波列夫训练并未直接针对的二阶导数精度。因此，更倾向于使用基于得分函数的估计量 $\hat{\mathbf{J}}_n(\boldsymbol{\theta}_n)$。

5.3 模型错设下的推断

当模拟器无法完美逼近真实似然时，我们仍可通过将模拟器视为工作模型，并遵循准最大似然框架（White，1982），获得有效推断。在本节中，我们将神经网络视为固定的，令 n，并省略观测对数似然中的 n 下标，$\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})$，以及得分函数 $\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta})$。将伪真参数定义为模拟器对数似然的总体最大化点：

$$\boldsymbol{\theta}_0^{\dagger} = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}. \tag{43}$$

当模拟器能够较好地逼近真实似然时，$\boldsymbol{\theta}_0^{\dagger} \approx \boldsymbol{\theta}_0$。渐近分布取决于两个矩阵：

$$\mathbf{A}(\boldsymbol{\theta}) = -\mathbb{E}\left\{\nabla^2_{\boldsymbol{\theta}} \hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}, \qquad \mathbf{B}(\boldsymbol{\theta}) = \mathbb{E}\left\{\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta})\hat{\psi}(y, \mathbf{X}; \boldsymbol{\theta})^{\top}\right\}. \tag{44}$$

在正确设定（即模拟器完美无误）下，信息矩阵等式给出$\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{B}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$。在模型错设下，这些矩阵并不相同，因此需要采用夹心协方差形式。在附录 B.3中给出的正则条件下，我们得到

以下结果。

定理 5 （模型设定错误下的一致性）。 在假设 *5* 和 *6*下，$\hat{}\boldsymbol{\theta}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$ 。**定理 6** （模型设定错误下的渐近正态性）。 在假设 *5* 和 *6*下，

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\mathbf{B}(\boldsymbol{\theta}_0^\dagger)\mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}\right\}. \tag{45}$$

证明遵循标准的 M 估计量论证；参见附录B.4。三明治协方差 $\mathbf{V} = \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\mathbf{B}(\boldsymbol{\theta}_0^\dagger)\mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}$ 可通过 $\hat{}$ 一致地估计为$\mathbf{V}_n = \hat{}\mathbf{A}_n^{-1}\ \hat{}\mathbf{B}_n\hat{}\mathbf{A}_n^{-\top}$，其中

$$\hat{\mathbf{A}}_n = -\frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\boldsymbol{\theta}}\hat{m}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n), \qquad \hat{\mathbf{B}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\psi}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)\hat{\psi}(y_i, \mathbf{X}_i; \hat{\boldsymbol{\theta}}_n)^\top. \tag{46}$$

$\hat{}$ 两个量都可以通过自动微分计算。注意，$\mathbf{B}_n$ 对应于上一节中的 $\hat{}\hat{}\mathbf{J}_n(\boldsymbol{\theta}_n)$，其渐近制度是假定模拟器固定。

定理 6 为 $\boldsymbol{\theta}_0^\dagger$ 提供有效推断，但不一定适用于 $\boldsymbol{\theta}_0$。如果我们采用模拟器随 n 变化的渐近制度，那么伪真参数也会随 n 变化，因此记作 $\boldsymbol{\theta}_n^\dagger$。当假设2中的 $\alpha \in [0, 1)$ 时，我们有 $\boldsymbol{\theta}_n^\dagger \xrightarrow{p} \boldsymbol{\theta}_0$ 随 $n \to \infty$ 收敛，但这种收敛速度相对于 $\sqrt{n}$ 的抽样变异可能过慢。三明治标准误对于 $\boldsymbol{\theta}_n^\dagger$ 仍然有效，但置信区间不应被解读为以 $\boldsymbol{\theta}_0$ 为目标。即使 $\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$ 和 $\mathbf{B}(\boldsymbol{\theta}_0^\dagger)$ 都近似于 $\mathbf{J}(\boldsymbol{\theta}_0)$，以 $\boldsymbol{\theta}_n^\dagger \neq \boldsymbol{\theta}_0$ 为中心仍会引入渐近偏差。对 $\boldsymbol{\theta}_0$ 进行有效推断需要 $\alpha \geq 1$（定理 4）。

同样，命题 1 给出了使 $\hat{}\mathbf{B}_n \xrightarrow{p} \mathbf{J}(\boldsymbol{\theta}_0)$ 成立的条件。然而，一阶索оболев训练过程并不保证模拟器海森矩阵会近似真实模型的海森矩阵；因此，$\mathbf{A}_n$ 不一定收敛到真实模型中的任何量。不过，定理 6 只要求 $\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$ 非奇异，并不要求它与真实模型相符。当模拟器还能提供准确的二阶导数时，夹心协方差矩阵会简化为 $\mathbf{J}(\boldsymbol{\theta}_0)^{-1}$。

6 模拟研究

我们评估了基于模拟器的多项式正态概率模型估计量在不同样本量、备选项数量和模型设定下的性能，并将其统计性能和计算需求与基于 GHK 模拟器的估计量进行比较。

6.1 模拟设计

模型设定。 我们考虑具有 $K \in \{3, 5, 10\}$ 个备选项和 $p = 2$ 个协变量的多元正态概率模型。备选项 K 作为参照项，其效用归一化为零，代表外部选项。对于每个非参照备选项 $j \in \{1, \ldots, K-1\}$，确定性效用为 $v_j = \boldsymbol{x}_j^\top \boldsymbol{\beta}$，其中 $\boldsymbol{x}_j \in \mathbb{R}^p$ 是备选项特定属性的向量，$\boldsymbol{\beta} = (0,\ 1)^\top$ 是真实系数向量。我们考虑 $\boldsymbol{\Sigma}$ 的两种设定，分别称为稠密和因子。

稠密设定假设 $\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}$（$\boldsymbol{\Sigma}$ 的上部（$K-1$）$\times$（$K-1$）子矩阵）具有稠密结构。为实现识别，我们设定 $\Sigma_{11} = 1$ 和 $\boldsymbol{\Sigma}_{\cdot K} = \boldsymbol{\Sigma}_{K\cdot}^\top = \mathbf{0}_K$；后一个方程将零效用选项的方差和协方差设为零。

因子设定假设 $\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}$ 具有如下形式的单因子结构：$\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)} = \boldsymbol{\Psi} + \boldsymbol{\gamma}\boldsymbol{\gamma}^\top$，其中 $\boldsymbol{\gamma} \in \mathbb{R}^{K-1}$ 和 $\boldsymbol{\Psi} \in \mathbb{R}^{(K-1)\times(K-1)}$ 是一个由独特性构成的对角线矩阵。为实现识别，我们施加锚定归一化，固定第一个独特性参数和第一个因子载荷：

$$\Psi_{11} = 0, \qquad \gamma_1 = 1. \tag{47}$$

与稠密设定一样，我们设定 $\boldsymbol{\Sigma}_{\cdot K} = \boldsymbol{\Sigma}_{K\cdot}^\top = \mathbf{0}_K$。这些约束意味着 $\boldsymbol{\Sigma}_{11} = 1$，从而固定模型的尺度；并且对于 $j \geq 2$，有 $\boldsymbol{\Sigma}_{1j} = \gamma_j$，因此因子载荷直接表示与第一个备选项的协方差。在这些约束下，我们用 $K-2$ 个自由唯一性参数（$\Psi_{22}, \ldots, \Psi_{K-1,K-1}$）和 $K-2$ 个自由载荷（$\gamma_2, \ldots, \gamma_{K-1}$）对模型进行参数化。这种锚定归一化旨在消除似然曲面中的平坦方向。

表 1 展示了稠密设定和因子设定中各类型的非冗余参数数量。在采用锚定归一化时，因子设定在 $K = 3$ 时恰好识别，在 $K \geq 4$时过度识别。对于 $K = 10$，因子设定的参数数量不到稠密设定的一半（18 对 46）。

数据生成。 对于每次重复抽样，我们为观测值 $\boldsymbol{x}_{ij}$ 独立同分布$\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$生成协变量 $i = 1, \ldots, n$ 和备选项 $j = 1, \ldots, K-1$，其中 $\boldsymbol{x}_{iK} = \mathbf{0}$ 对应参照项。

对于稠密设定，我们根据 Wishart$(\mathbf{I}_{K-1}, K+10)$ 分布抽取协方差矩阵，然后对其进行归一化，使得 $\Sigma_{11} = 1$，从而得到概率密度

协方差（$\boldsymbol{\Sigma}$）设定	K	系数	方差（独特性）	协方差（载荷）	总计
稠密	3	2	1	1	4
	5	2	3	6	11
	10	2	8	36	46
因子	3	2	1	1	4
	5	2	3	3	8
	10	2	8	8	18

表1：按 $\boldsymbol{\Sigma}$ 设定（稠密或因子）和 K划分的参数数量 .

函数（概率密度函数）如下

$$\frac{\Gamma(s)\cdot|\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)}|^5}{2^s\Gamma_{K-1}\left(\frac{K+10}{2}\right)\cdot\left\{\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{1:(K-1),1:(K-1)})\right\}^s}, \tag{48}$$

其中 $s=(K+10)(K-1)/2$ 和 $\Gamma_{K-1}(\cdot)$ 表示 $(K-1)$ 维多元伽马函数（参见附录 C.2）。

对于因子设定，我们在对数尺度上对自由唯一性参数进行抽样，对 $j=2,\ldots,K-1$ 抽取 $\log(\Psi_{jj}) \overset{iid}{\sim} \mathcal{N}(0,0.1^2)$，再取指数以得到 $\Psi_{jj}>0$。对于因子载荷，我们对 $j=2,\ldots,K-1$ 抽取 $\gamma_j \overset{iid}{\sim} \mathcal{N}(0,0.2^2)$。固定值 $\Psi_{11}=0$ 和$\gamma_1=1$ 完成了参数化。

给定协变量和真实参数，我们计算确定性效用，并根据两种数据生成过程（DGP）之一模拟选择 $y_i\in\{1,\ldots,K\}$：Probit 或模拟器。Probit 随机效用模型从精确的多项式正态概率模型中生成选择，抽取误差$\boldsymbol{\epsilon}_i\sim\mathcal{N}(0,\boldsymbol{\Sigma})$，并选择效用最大的备选项。模拟器随机效用模型则直接从模拟器预测的选择概率所定义的分类分布中抽样生成选择。采用模拟器随机效用模型时，基于模拟器的估计方法设定正确，因此可以比较该方法在正确设定（模拟器随机效用模型）和近似推断（Probit 随机效用模型）下的表现。

估计。为确保 $\boldsymbol{\Sigma}$ 正定，我们在乔莱斯基因子的尺度上估计稠密设定，并在对数尺度上优化对角元素以保证其为正。同样，我们在对数尺度上估计因子设定中的自由唯一性参数，并针对 $j=2$、.....、$K-1$ 优化 $\log(\Psi_{jj})$。

我们采用两种方法构造估计量：模拟器和 GHK，分别对应于使用预训练模拟器或 GHK 模拟器近似对数似然函数。我们使用 Probit 随机效用模型生成的数据拟合这两种方法；此外，我们还使用模拟器方法拟合模拟器随机效用模型生成的数据，以分析该方法在设定正确时的表现。我们采用 $R \in \{10, 50, 250\}$ 次抽样应用 GHK 方法，代表低、中、高三种计算投入。

Emulator 和 GHK 这两种方法都进行正则化最大似然估计。正则化由两个部分组成：屏障惩罚项和伪先验。对于因子设定，屏障惩罚项作用于 $\log(\Psi_{jj})$；对于稠密设定，则作用于对角线 Cholesky 参数。对于潜在参数 η，屏障惩罚项为 $\lambda_{\text{reg}} \cdot \log(\eta + 8)/n$。伪先验项的形式为 $\lambda_{\text{reg}} \cdot \log\{p(\boldsymbol{\theta})\}/n$，其中$p(\boldsymbol{\theta})$是数据生成过程中 $\boldsymbol{\theta}$ 的密度函数。我们将这些惩罚项相加，得到形式为 $\lambda_{\text{reg}} \cdot r(\boldsymbol{\theta})/n$ 的组合惩罚项，其中 $r(\boldsymbol{\theta})$是屏障惩罚项与对数先验惩罚项之和，λ_{reg} 是正则化超参数。

然后，我们将 $\boldsymbol{\theta}$ 估计为（近似）对数似然与$\lambda_{\text{reg}} \cdot r(\boldsymbol{\theta})/n$之和的最大化点。我们选择 $\lambda_{\text{reg}} = 0.1$，使得惩罚项较小且渐近可忽略。实际中，我们发现这些惩罚项能使模型拟合算法更可靠地收敛，并有助于解决协方差参数的弱识别问题；当K 较大且 n 较小时，就会出现这一问题。在计算标准误时，我们通过将 $\text{Cov}(\hat{\boldsymbol{\theta}}_n)$ 估计为 $\left\{n \cdot \hat{\mathbf{J}}_n(\hat{\boldsymbol{\theta}}_n) - \lambda_{\text{reg}} \cdot \nabla^2_{\boldsymbol{\theta}} r(\hat{\boldsymbol{\theta}}_n)\right\}^{-1}$，将这些惩罚项考虑在内。同样，惩罚项的影响在渐近意义下可忽略，因为项 $n \cdot \mathbf{J}_n(\boldsymbol{\theta}_n)$ 在渐近意义下占主导。

对于所有方法，我们都使用 L-BFGS 优化来最大化（近似）正则化对数似然（Liu 和 Nocedal，1989年）。实践中，我们发现有必要限制最小步长（我们使用 10^{-6}），并允许目标函数小幅下降，以便优化器能够跳出局部极小值。

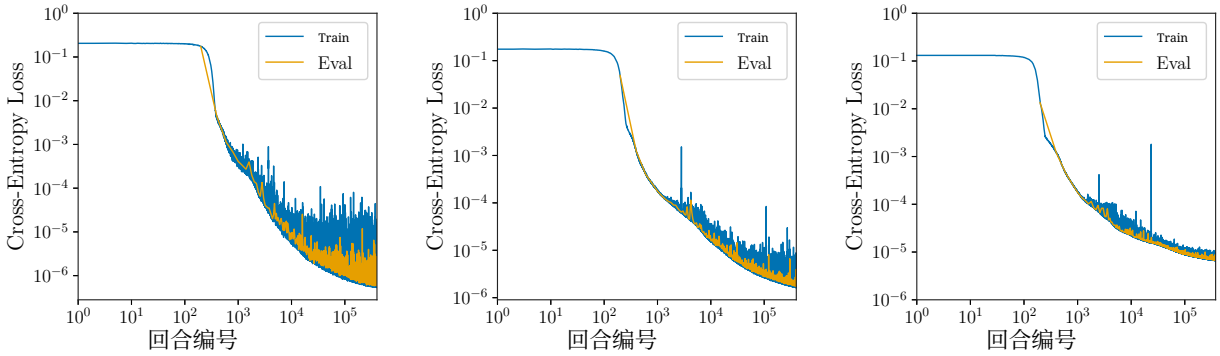
实验设计。我们考虑样本量 $n \in \{1{,}000, 10{,}000, 100{,}000\}$，并针对每种配置进行 100 次重复抽样。我们计算汇总所有参数类型的性能指标。我们使用按比例缩放的经验标准差，计算覆盖率和平均耗时的蒙特卡罗标准误。对于其余指标，我们通过自助重抽样（对重复抽样进行重抽样），使用 1,000 个自助样本来估计蒙特卡罗标准误。

计算环境。所有模拟均使用 PyTorch（Paszke 等人，2019年）实现，并在配备 48 GB 内存和 Apple M4 Pro

处理器的个人笔记本电脑上运行。我们使用 Apple Metal Performance Shaders（MPS）拟合仿真器。不过，为了获得可比较的估计耗时，我们在 CPU 上执行其余模拟。

6.2 仿真结果

我们使用 400，000 个预生成的训练样例来拟合模拟器，进行 400，000 个训练轮次，每个样例包含 10^6 个备选项。有关优化器、学习率调度和网络架构的详细信息，请参阅附录 C.1。对于 $K = 3$，5 和 10，该模型拟合流程分别耗时 2.1、4.6 和 16.6 小时，计算开销适中。



(a) $K = 3$　(b) $K = 5$　(c) $K = 10$

图 2：模拟器在 400，000 个训练轮次中的训练损失和评估损失，其中备选项数量为 $K \in \{3, 5, 10\}$。训练损失（蓝色）在整个训练过程中持续下降，评估损失（橙色）与训练损失紧密贴合，表明模拟器没有对模拟训练数据过拟合。

图 2 展示了仿真器的训练损失和评估损失随时间的变化。在每个 $K \in \{3, 5, 10\}$ 值下，即使在训练结束时，训练损失仍随时间持续下降，这表明继续训练很可能会带来进一步改进。评估损失紧跟训练损失的变化，说明模拟器具有良好的泛化能力，且对模拟训练概率的过拟合程度很低。我们每 200 个回合计算一次评估损失；与训练损失相比，这些定期计算使评估损失的变化不那么明显。

表2 展示了用于 $K = 3$ 比较模拟器方法与 GHK 估计方法性能的性能指标。这些性能指标包括均方根误差（RMSE）、均方根偏差（RMS Bias）、估计标准误与经验标准误之比（SE Ratio）、95% 置信区间覆盖率（覆盖率），以及以秒为单位的模型拟合耗时（时间（秒））。在所有样本量下，模拟器方法的表现都很好，

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.159 (0.016)	0.010 (0.009)	1.00 (0.05)	0.94 (0.01)	0.11 (0.00)
	概率单位	模拟器	0.206 (0.032)	0.049 (0.018)	0.95 (0.08)	0.95 (0.01)	0.10 (0.00)
	概率单位	GHK(10)	0.202 (0.031)	0.047 (0.019)	0.96 (0.08)	0.94 (0.02)	0.08 (0.00)
	概率单位	GHK(50)	0.203 (0.032)	0.048 (0.018)	0.96 (0.09)	0.95 (0.01)	0.14 (0.00)
	概率单位	GHK(250)	0.207 (0.033)	0.049 (0.019)	0.95 (0.08)	0.95 (0.01)	0.41 (0.01)
10,000	模拟器	模拟器	0.052 (0.006)	0.007 (0.004)	0.94 (0.07)	0.94 (0.01)	0.36 (0.00)
	概率单位	模拟器	0.053 (0.009)	0.008 (0.004)	0.99 (0.11)	0.94 (0.01)	0.36 (0.00)
	概率单位	GHK(10)	0.052 (0.008)	0.010 (0.003)	1.00 (0.09)	0.94 (0.01)	0.30 (0.00)
	概率单位	GHK(50)	0.052 (0.009)	0.007 (0.004)	1.00 (0.10)	0.95 (0.01)	0.44 (0.01)
	概率单位	GHK(250)	0.053 (0.009)	0.008 (0.005)	0.99 (0.11)	0.94 (0.01)	0.93 (0.01)
100,000	模拟器	模拟器	0.015 (0.002)	0.001 (0.001)	0.97 (0.08)	0.94 (0.01)	1.80 (0.01)
	概率单位	模拟器	0.016 (0.002)	0.001 (0.001)	1.06 (0.07)	0.98 (0.01)	1.82 (0.02)
	概率单位	GHK(10)	0.020 (0.002)	0.007 (0.001)	0.83 (0.06)	0.83 (0.02)	0.90 (0.01)
	概率单位	GHK(50)	0.016 (0.002)	0.002 (0.001)	1.03 (0.08)	0.97 (0.01)	1.90 (0.02)
	概率单位	GHK(250)	0.016 (0.002)	0.001 (0.001)	1.06 (0.07)	0.98 (0.01)	7.22 (0.09)

表 2：针对含有 $K = 3$ 的多项Probit模型，对比模拟器与GHK估计方法。模拟器的表现可与GHK(50)和GHK(250)相媲美，所需计算时间介于GHK(10)和GHK(50)之间。性能指标为所有参数（系数、方差和协方差）的平均值，括号中为蒙特卡罗标准误。

其统计性能与 GHK(50) 和 GHK(250) 相当，但计算成本更低。GHK(10) 在 $n = 100,000$ 时的覆盖率明显不足，其均方根偏差显著高于其他方法。这种较差的表现可能是 GHK 模拟器在对数尺度上因詹森不等式而产生有偏估计所致；当 n足够大时，相对于 $1/\sqrt{n}$抽样误差，这种偏差最终会变得不可忽略。

–

表3 显示了 $K = 5$的类似结果。模拟器在所有样本量下均可与 GHK(50) 和 GHK(250) 媲美。与 GHK(50) 相比，模拟器方法的计算时间在 $n = 1{,}000$ 时略短，在 $n = 10{,}000$ 时大致相同，在 $n = 100{,}000$ 时略长。

表4 展示了 $K = 10$的性能结果。模拟器的表现与 GHK(250) 相当，但计算需求低得多。在 $n = 1{,}000$ 时，模拟器平均只需不到 3 秒即可拟合模型，甚至快于 GHK(10)。在 $n = 10{,}000$ 时，模拟器的计算时间介于 GHK(10) 和 GHK(50) 之间；在 $n = 100{,}000$ 时，则介于 GHK(50) 和 GHK(250) 之间。与表2 和3相比，较小样本量下的标准误估计不太可靠，这可能是由于弱识别所致；省略对 $\boldsymbol{\Sigma}$的惩罚项时，这一问题会加剧。标准误有所改善

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.243 (0.017)	0.032 (0.012)	1.11 (0.06)	0.94 (0.01)	0.41 (0.01)
	概率单位	模拟器	0.281 (0.020)	0.045 (0.016)	1.04 (0.04)	0.93 (0.01)	0.42 (0.00)
	概率单位	GHK(10)	0.278 (0.015)	0.074 (0.017)	0.95 (0.03)	0.89 (0.01)	0.29 (0.00)
	概率单位	GHK(50)	0.280 (0.017)	0.051 (0.017)	1.02 (0.04)	0.92 (0.01)	0.62 (0.01)
	概率单位	GHK(250)	0.284 (0.020)	0.046 (0.017)	1.03 (0.04)	0.93 (0.01)	2.02 (0.03)
10,000	模拟器	模拟器	0.086 (0.006)	0.010 (0.004)	0.99 (0.05)	0.95 (0.01)	2.27 (0.02)
	概率单位	模拟器	0.084 (0.005)	0.009 (0.004)	1.06 (0.04)	0.96 (0.01)	2.26 (0.02)
	概率单位	GHK(10)	0.095 (0.006)	0.032 (0.006)	0.88 (0.03)	0.89 (0.01)	0.89 (0.01)
	概率单位	GHK(50)	0.086 (0.006)	0.013 (0.005)	1.02 (0.04)	0.95 (0.01)	2.27 (0.03)
	概率单位	GHK(250)	0.084（0.005）	0.009（0.004）	1.06（0.04）	0.96（0.01）	4.95（0.05）
100,000	模拟器	模拟器	0.028 (0.004)	0.003 (0.001)	0.97 (0.09)	0.95 (0.01)	18.09 (0.15)
	概率单位	模拟器	0.028 (0.002)	0.002 (0.001)	1.00 (0.06)	0.95 (0.01)	18.30 (0.17)
	概率单位	GHK(10)	0.056 (0.004)	0.024 (0.003)	0.49 (0.02)	0.60 (0.02)	3.75 (0.03)
	概率单位	GHK(50)	0.030 (0.002)	0.004 (0.002)	0.93 (0.05)	0.92 (0.01)	9.45 (0.11)
	概率单位	GHK(250)	0.028 (0.002)	0.002 (0.001)	1.00 (0.06)	0.95 (0.01)	42.96 (0.49)

表3：比较用于 $K = 5$的多项Probit模型模拟器与GHK估计方法。如表2所示，在计算成本相当或更低的情况下，模拟器的表现可与GHK(50)和GHK(250)相媲美。

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.324 (0.014)	0.195 (0.011)	1.56 (0.14)	0.93 (0.01)	2.24 (0.02)
	概率单位	模拟器	0.343 (0.016)	0.207 (0.012)	1.71 (0.24)	0.93 (0.01)	2.21 (0.02)
	概率单位	GHK(10)	0.388 (0.017)	0.217 (0.013)	0.35 (0.07)	0.22 (0.02)	5.86 (0.30)
	概率单位	GHK(50)	0.363 (0.016)	0.215 (0.013)	0.70 (0.07)	0.54 (0.03)	14.16 (0.95)
	概率单位	GHK(250)	0.351（0.017）	0.214（0.012）	1.43（0.11）	0.91（0.01）	14.87（0.41）
10,000	模拟器	模拟器	0.172 (0.010)	0.073 (0.007)	1.74 (0.19)	0.96 (0.01)	19.28 (0.26)
	概率单位	模拟器	0.200 (0.011)	0.080 (0.010)	1.81 (0.20)	0.95 (0.01)	19.04 (0.19)
	概率单位	GHK(10)	0.236 (0.014)	0.123 (0.010)	1.57 (0.52)	0.79 (0.02)	8.95 (0.30)
	概率单位	GHK(50)	0.203 (0.010)	0.090 (0.010)	1.29 (0.07)	0.91 (0.01)	33.54 (0.83)
	概率单位	GHK(250)	0.198 (0.010)	0.079 (0.010)	1.41 (0.07)	0.95 (0.01)	50.22 (0.70)
100,000	模拟器	模拟器	0.069 (0.004)	0.013 (0.003)	1.07 (0.05)	0.95 (0.01)	165.59 (1.15)
	概率单位	模拟器	0.082 (0.005)	0.020 (0.004)	0.92 (0.03)	0.93 (0.01)	164.88 (1.49)
	概率单位	GHK(10)	0.139 (0.009)	0.073 (0.006)	0.48 (0.02)	0.60 (0.01)	31.48 (0.35)
	概率单位	GHK(50)	0.085 (0.005)	0.027 (0.004)	0.89 (0.03)	0.90 (0.01)	93.23 (0.98)
	概率单位	GHK(250)	0.081 (0.005)	0.020 (0.004)	0.97 (0.03)	0.93 (0.01)	404.22 (5.59)

表4：针对 $K = 10$的多项Probit模型，比较模拟器与GHK估计方法。就统计性能而言，模拟器与GHK(250)相当，但所需计算时间少得多。参见表2和3。

随样本量显著变化（无论是否采用惩罚），在样本量为 $n = 100,000$ 时，模拟器和 GHK(250) 方法的标准误比率达到 0.92–1.07。

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.089 (0.004)	0.022 (0.006)	1.18 (0.05)	0.95 (0.01)	0.11 (0.00)
	概率单位	模拟器	0.095 (0.004)	0.007 (0.005)	1.10 (0.04)	0.96 (0.01)	0.10 (0.00)
	概率单位	GHK(10)	0.096 (0.004)	0.011 (0.005)	1.08 (0.04)	0.96 (0.01)	0.08 (0.00)
	概率单位	GHK(50)	0.094 (0.004)	0.007 (0.005)	1.11 (0.04)	0.96 (0.01)	0.15 (0.00)
	概率单位	GHK(250)	0.095 (0.004)	0.007 (0.005)	1.10 (0.04)	0.96 (0.01)	0.42 (0.01)
10,000	模拟器	模拟器	0.038 (0.002)	0.003 (0.002)	0.99 (0.04)	0.93 (0.01)	0.39 (0.01)
	概率单位	模拟器	0.038 (0.002)	0.007 (0.003)	1.00 (0.05)	0.96 (0.01)	0.38 (0.00)
	概率单位	GHK(10)	0.040 (0.002)	0.013 (0.002)	0.99 (0.05)	0.95 (0.01)	0.33 (0.00)
	概率单位	GHK(50)	0.038 (0.002)	0.008 (0.003)	1.00 (0.05)	0.96 (0.01)	0.48 (0.01)
	概率单位	GHK(250)	0.038 (0.002)	0.007 (0.003)	1.00 (0.05)	0.95 (0.01)	0.96 (0.01)
100,000	模拟器	模拟器	0.011 (0.001)	0.001 (0.001)	1.08 (0.05)	0.96 (0.01)	1.97 (0.02)
	概率单位	模拟器	0.011 (0.001)	0.001 (0.001)	1.04 (0.04)	0.96 (0.01)	1.91 (0.01)
	概率单位	GHK(10)	0.016 (0.001)	0.009 (0.001)	0.88 (0.04)	0.82 (0.02)	0.98 (0.01)
	概率单位	GHK(50)	0.012 (0.001)	0.002 (0.001)	1.04 (0.04)	0.95 (0.01)	2.07 (0.03)
	概率单位	GHK(250)	0.011 (0.001)	0.001 (0.001)	1.04 (0.05)	0.96 (0.01)	7.52 (0.09)

表5：在包含 $K = 3$ 的因子结构多项概率单位模型中，模拟器与 GHK 估计方法的比较。性能指标按因子参数（系数、独特性和因子载荷）取平均值。模拟器的性能与 GHK(250) 相当，但计算成本低得多。

表5 展示了因子设定下的类似结果，其中 $K = 3$。性能指标按 $\boldsymbol{\beta}$、 $\boldsymbol{\Psi}$和 $\boldsymbol{\gamma}$的非冗余元素取平均值。模拟器的性能与 GHK(250) 相当，但计算成本低得多。

表6 和7 分别展示了备选项数量为 5 和 10 时的类似结果。在这两种情况下，尽管模拟器并非专门针对因子模型训练，其表现仍与 GHK(50) 和 GHK(250) 相当。当 $n \in \{1{,}000, 10{,}000\}$时，模拟器的速度快于 GHK(50)。当 $n = 100{,}000$时，模拟器的计算时间介于 GHK(50) 和 GHK(250) 之间。

在所有情境下，模拟器方法的表现都与 GHK 不相上下，或显著优于 GHK。我们预计，随着 (a) 增加训练量，以及 (b) 在专用深度学习硬件（如 GPU）上实现，模拟器的相对表现将会提升。后者可能尤其有利于模拟器，因为其计算可轻松并行化，并且已针对现代 GPU 进行了高度优化。

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.128 (0.004)	0.016 (0.005)	1.20 (0.04)	0.97 (0.01)	0.32 (0.01)
	概率单位	模拟器	0.131 (0.003)	0.012 (0.004)	1.19 (0.03)	0.97 (0.01)	0.32 (0.00)
	概率单位	GHK(10)	0.134 (0.004)	0.017 (0.005)	1.12 (0.03)	0.95 (0.01)	0.23 (0.00)
	概率单位	GHK(50)	0.131 (0.003)	0.012 (0.004)	1.19 (0.03)	0.98 (0.01)	0.49 (0.01)
	概率单位	GHK(250)	0.132 (0.003)	0.012 (0.004)	1.19 (0.03)	0.97 (0.01)	1.61 (0.02)
10,000	模拟器	模拟器	0.059 (0.002)	0.007 (0.003)	1.01 (0.04)	0.95 (0.01)	1.92 (0.02)
	概率单位	模拟器	0.059 (0.003)	0.005 (0.002)	1.04 (0.04)	0.95 (0.01)	1.97 (0.02)
	概率单位	GHK(10)	0.062 (0.003)	0.011 (0.003)	0.93 (0.04)	0.93 (0.01)	0.77 (0.01)
	概率单位	GHK(50)	0.059 (0.003)	0.004 (0.003)	1.02 (0.04)	0.95 (0.01)	2.02 (0.03)
	概率单位	GHK(250)	0.059 (0.003)	0.007 (0.003)	1.04 (0.04)	0.95 (0.01)	4.67 (0.08)
100,000	模拟器	模拟器	0.019 (0.001)	0.002 (0.001)	1.02 (0.03)	0.96 (0.01)	15.46 (0.18)
	概率单位	模拟器	0.020 (0.001)	0.006 (0.001)	1.04 (0.03)	0.94 (0.01)	15.58 (0.18)
	概率单位	GHK(10)	0.029 (0.001)	0.017 (0.001)	0.75 (0.03)	0.75 (0.02)	3.21 (0.04)
	概率单位	GHK(50)	0.019 (0.001)	0.004 (0.001)	1.02 (0.03)	0.95 (0.01)	8.47 (0.12)
	概率单位	GHK(250)	0.019 (0.001)	0.002 (0.001)	1.04 (0.03)	0.96 (0.01)	37.30 (0.59)

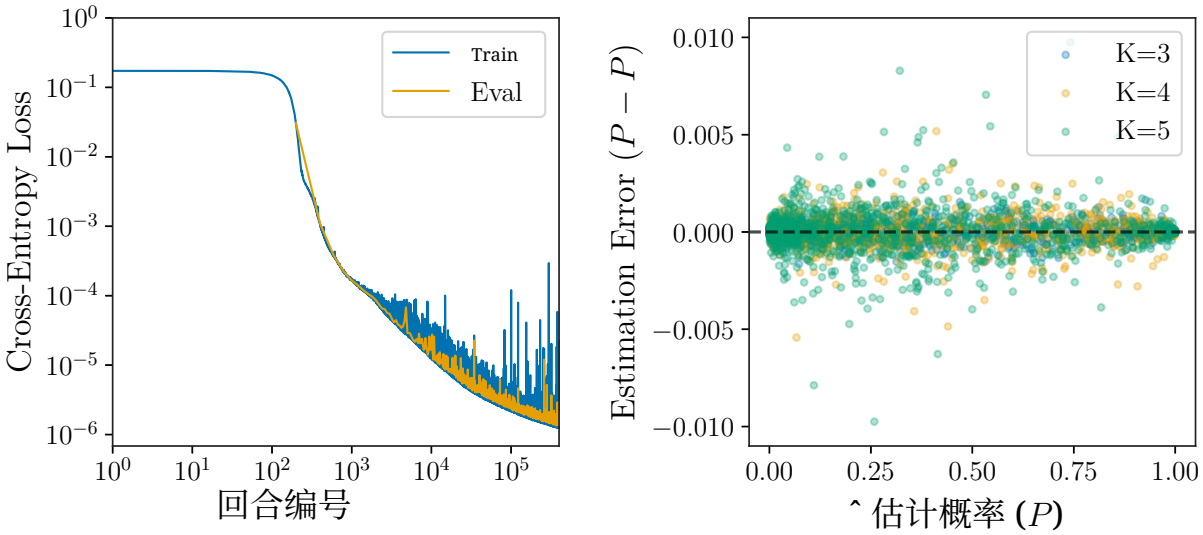
表6：模拟器与GHK估计方法在因子结构多元
名概率单位模型，具有 $K = 5$。与表 5一样，模拟器的表现不逊于
GHK(50) 和 GHK(250)，且计算成本相当或更低。

n	DGP	方法	RMSE	均方根偏差	标准误比率	覆盖率	时间（秒）
1,000	模拟器	模拟器	0.188 (0.004)	0.023 (0.004)	1.14 (0.02)	0.96 (0.00)	1.06 (0.01)
	概率单位	模拟器	0.198 (0.005)	0.030 (0.005)	1.10 (0.02)	0.96 (0.01)	1.01 (0.01)
	概率单位	GHK(10)	0.220 (0.005)	0.059 (0.006)	0.92 (0.02)	0.90 (0.01)	0.77 (0.01)
	概率单位	GHK(50)	0.205 (0.005)	0.039 (0.005)	1.06 (0.02)	0.95 (0.01)	2.44 (0.03)
	概率单位	GHK(250)	0.199 (0.005)	0.030 (0.005)	1.11 (0.02)	0.96 (0.01)	5.93 (0.08)
10,000	模拟器	模拟器	0.080 (0.002)	0.011 (0.003)	1.24 (0.04)	0.98 (0.00)	7.58 (0.06)
	概率单位	模拟器	0.082 (0.002)	0.012 (0.003)	1.22 (0.04)	0.97 (0.01)	7.48 (0.06)
	概率单位	GHK(10)	0.103 (0.002)	0.049 (0.003)	0.95 (0.03)	0.88 (0.01)	4.03 (0.04)
	概率单位	GHK(50)	0.087 (0.002)	0.019 (0.003)	1.14 (0.04)	0.96 (0.01)	14.83 (0.16)
	概率单位	GHK(250)	0.084 (0.002)	0.012 (0.003)	1.20 (0.04)	0.97 (0.01)	24.85 (0.25)
100,000	模拟器	模拟器	0.043 (0.002)	0.004 (0.002)	0.94 (0.04)	0.93 (0.01)	69.83 (0.57)
	概率单位	模拟器	0.036 (0.001)	0.004 (0.001)	1.11 (0.04)	0.97 (0.01)	69.05 (0.66)
	概率单位	GHK(10)	0.071 (0.002)	0.047 (0.002)	0.65 (0.02)	0.60 (0.02)	19.73 (0.20)
	概率单位	GHK(50)	0.039 (0.001)	0.012 (0.001)	1.04 (0.04)	0.93 (0.01)	55.94 (0.75)
	概率单位	GHK(250)	0.036 (0.001)	0.004 (0.001)	1.12 (0.05)	0.96 (0.01)	218.97 (2.94)

表7：因子结构多项式概率单位模型中模拟器与GHK估计方法的比较 $K = 10$。模拟器的统计性能与GHK(250)相当，但所需的计算时间要少得多。请与表5和6进行比较。

6.3 多-K 训练准确率

本节最后，我们给出训练神经网络模拟器以近似 $K \in \{3, 4, 5\}$ 的选择概率的数值结果。我们使用与上一节中用于 $K = 5$ 的相同神经网络。与上文一样，我们训练 400,000 个回合，但在每个回合中，我们循环处理包含 $K = 3$、4 个和 5 个备选项的训练样本。训练过程耗时 13.0 小时。



(a) 训练损失　　(b) 估计准确度

图 3: （a）具有 $K \in \{3, 4, 5\}$ 个备选项的仿真器在 400,000 个训练轮次中的训练损失和评估损失。训练损失（蓝色）在整个训练过程中持续下降，评估损失（橙色）紧随训练损失变化，表明过拟合程度很低。（b）仿真器概率估计与估计误差的对比。仿真器同时学习 $K = 3, 4, 5$ 的选择概率，且误差较小。

图 3a 显示，训练损失和评估损失在整个训练流程中稳步下降，接近图 2所示的水平。图 3b 绘制了 $K \in \{3, 4, 5\}$ 的估计选择概率与估计误差，每种情况各使用 400 个模拟样本。较小的误差表明，该模型能够同时准确近似 K 的三个取值下的选择概率。

7 结论

本文提出了一种适用于离散选择模型的摊销推断方法。该方法采用一种新颖的神经网络架构，在一般误差分布下近似计算选择概率，包括具有非平凡跨选项相关性的误差分布，

例如多元正态概率模型。这种通用性使得该方法能够灵活地对实证选择行为进行建模，包括复杂的替代模式，同时又不失随机效用模型的理论基础和可解释性。

由于模拟器针对通用选择问题运行，因此不受确定性效用所假设的参数形式或相应尺度（协方差）矩阵的限制。给定一个预训练模拟器，这种通用性使其能够直接采用，无需定制架构更改或额外训练。具体而言，从逻辑斯蒂模型推广到概率单位模型，只需将闭式软最大概率替换为模拟器评估即可。其他相关误差分布的处理也同样简单。

我们提出的架构和训练流程包含预处理变换、DeepSet 模块（Zaheer等人，2017年）、等变层、平滑激活函数以及索博列夫训练（Czarnecki 等人，2017年）。这些设计选择共同产生平滑的选择概率近似，既遵循其不变性，又通过参数共享加速学习。我们证明，该架构具有通用近似性质，将关于对称矩阵的最新理论结果（Blum-Smith 等人，2025）扩展到了效用向量与中心化尺度（协方差）矩阵的联合空间。

在蒙特卡洛模拟中，我们发现，在估计多元正态概率模型时，基于模拟器的估计量在统计性能和计算需求方面均不逊于基于 GHK 的估计量。这些发现支持我们的理论结果：在温和的近似条件下，基于模拟器的估计量具有精确最大似然估计量的性质。在模拟研究中，我们聚焦于多元正态概率模型，以便与 GHK 模拟器进行比较；该模拟器仅直接适用于高斯误差。摊销推断方法的计算优势还因模拟器评估易于并行化，且能在现代深度学习硬件（如 GPU）上高度优化而进一步凸显。

有必要讨论若干局限。实际应用者必须确保其应用中的输入分布与训练分布有充分的重叠。此外，我们的渐近理论要求模拟器近似的逼近误差以适当的速率趋于零，这意味着样本量非常大的应用可能需要训练更充分的仿真器。最后，误差相关的模型需要独立逻辑斯蒂模型不需要的识别约束。尽管我们的方法能够简单高效地计算这些模型的选择概率，但确保模型得到适当识别所需的数据要求和建模决策仍是一个独立的挑战。

该框架还支持多种扩展。我们可以针对特定情形调整模拟器架构和训练流程，例如因子结构尺度矩阵或混合分布。我们预计，可以通过重新配置架构中的模块，并在某些情况下推广预处理变换，来处理这些情形。具体而言，前述两个示例可能需要在因子或混合成分上额外执行一次 DeepSet 操作。我们的仿真器还可以与其他现代选择建模架构结合使用，例如第2节所述的架构。以往研究已将神经网络辅助估计方法应用于似然函数难以处理的其他计量经济学模型，例如特定参数模型中的动态离散选择（Norets，2012年）和消费者搜索（Wei 和 Jiang，2025）。本文提出的摊销推断方法可能也能推广到其中许多情形，从而无需修改或重新训练神经网络即可实现高效的摊销推断。

致谢

作者承认使用了人工智能工具来协助撰写这份手稿。所有内容均已人工审阅，以确保准确性。

参考文献

Aouad, A. 和 Desir, A.（2025）。使用神经网络表示随机效用选择模型。《管理科学》，预先发表文章。Bagheri, N.、Ghasri, M. 和 Barlow, M.（2025）。适用于任意误差分布的离散选择模型神经估计框架。《选择建模期刊》，57：100583。Bentz, Y. 和 Merunka, D.（2000）。用于品牌选择建模的神经网络与多项Logit：一种混合方法。《预测学报》，19(3)：177–200。Berbeglia, G. 和 Venkataraman, A.（2025）。广义随机偏好选择模型。*arXiv*预印本 *arXiv:1803.04244*。Bloch, H. 和 Marschak, J.（1960）。随机理性与显示性随机偏好。载于《概率与统计文集》。斯坦福大学出版社。Blum-Smith, B.、Huang, N. T.、Cuturi, M. 和 Villar, S.（2025）。通过伽罗瓦理论中的轻量级不变特征研究对称矩阵与点云上的函数。《工业与应用数学学会应用代数与几何期刊》，9(4)：902–938。Bredon, G. E.（1972年）。《紧变换群导论》，第46卷。学术出版社，纽约。Cai, Z.、Wang, H.、Talluri, K. 和 Li, X.（2022年）。用于选择建模的深度学习。*arXiv*预印本 *arXiv:2208.09325*。Chen, N.、Gallego, G. 和 Tang, Z.（2025）。用二元选择森林对离散选择进行建模与估计。*arXiv*预印本 *arXiv:1908.01109*。Chen, Y.-C. 和 Mi˘si´c, V. V.（2022年）。决策森林：一种对非理性选择进行建模的非参数方法。《管理科学》（*INFORMS*），68(10)：7090–7111。Cocola, J. 和 Hand, P.（2020年）。过参数化神经网络索博列夫训练的全局收敛。载于国际机器学习、优化与数据科学会议，第574–586页。施普林格。Cranmer, K.、Brehmer, J. 和 Louppe, G.（2020年）。基于模拟的推断前沿。《美国国家科学院院刊》，117(48)：30055–30062。Cybenko, G.（1989年）。通过S形函数的叠加进行近似。《控制、信号与系统数学》，2(4)：303–314。

Czarnecki, W. M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R.（2017）。神经网络的索博列夫训练。载于《神经信息处理系统进展》，第30卷，第4278–4287页。De Ryck, T., Lanthaler, S., and Mishra, S.（2021）。双曲正切神经网络对函数的近似。神经网络，143：732–750。Dugas, C., Bengio, Y., B´elisle, F., Nadeau, C., and Garcia, R.（2000）。纳入二阶函数知识以改进期权定价。载于 Leen, T., Dietterich, T., and Tresp, V.（编），《神经信息处理系统进展》，第13卷。麻省理工学院出版社。E, W. and Wang, Q.（2018）。深度神经网络对解析函数的近似呈指数收敛。《中国科学：数学》，61（10）：1733–1740。Farrell, M. H., Liang, T., and Misra, S.（2021）。用于估计和推断的深度神经网络。《计量经济学》，89（1）：181–213。Geweke, J.（1989）。使用蒙特卡罗积分对计量经济学模型进行贝叶斯推断。《计量经济学》，57（6）：1317–1339。Hajivassiliou, V. A. and McFadden, D. L.（1998）。用于有限因变量模型估计的模拟得分函数方法。《计量经济学》，66（4）：863–896。Han, Y., Pereira, F. C., Ben-Akiva, M., and Zegras, C.（2022）。神经网络嵌入式选择模型：通过强化学习偏好表征。《交通研究*B*辑》，163：166–186。Hornik, K., Stinchcombe, M., and White, H.（1989）。多层前馈网络是通用逼近器。神经网络，2（5）：359–366。Hruschka, H.（2007）。使用带有神经网络扩展的异质多项式Probit模型来建模品牌选择。《预测学报》，26（2）：113–127。Jang, E., Gu, S., and Poole, B.（2017）。使用Gumbel–softmax进行类别重参数化。载于国际学习表征会议。Jiang, Z., Li, J., and Zhang, D.（2025）。面向在线零售的高维选择模型。《管理科学》（*INFORMS*），71（4）：3320–3339。Keane, M. P.（1992）。关于多项式Probit模型识别的一则注记。《商业与经济统计学杂志》，10（2）：193–200。

Keane, M. P.（1994）。一种计算上实用的面板数据模拟估计量。《计量经济学》，62（1）：95–116。Kingma, D. P. and Ba, J. L.（2015）。Adam：一种随机优化方法。载于国际学习表征会议。Lh´eritier, A., Bocamazo, M., Delahaye, T., and Acuna-Agost, R.（2019）。使用机器学习进行航空行程选择建模。《选择建模期刊》，31：198–209。Liu, D. C. and Nocedal, J.（1989）。关于用于大规模优化的有限内存BFGS方法。《数学规划》，45（1）：503–528。Loshchilov, I. and Hutter, F.（2019）。解耦权重衰减正则化。载于国际学习表征会议。Lueckmann, J.-M., Bassetto, G., Karaletsos, T., and Macke, J. H.（2019）。借助模拟器网络进行无似然推断。载于 Ruiz, F., Zhang, C., Liang, D., and Bui, T.（编），第一届近似贝叶斯推断进展研讨会论文集，《机器学习研究论文集》第96卷，第32–53页。PMLR。Maddison, C. J., Mnih, A., and Teh, Y. W.（2017）。混凝土分布：离散随机变量的连续松弛。载于国际学习表征会议。McCulloch, R. and Rossi, P. E.（1994）。多项式Probit模型的精确似然分析。《计量经济学杂志》，64（1）：207–240。McFadden, D.（1974）。定性选择行为的条件Logit分析。载于 Zarembka, P.（编），《计量经济学前沿》，第105–142页。学术出版社，纽约。McFadden, D. and Richter, M.（1990）。随机理性与揭示的随机偏好。载于《偏好、不确定性与最优性：纪念利奥·赫维奇文集》。韦斯特维尤出版社。McFadden, D. and Train, K.（2000）。离散响应的混合MNL模型。《应用计量经济学杂志》，15（5）：447–470。Mityagin, B. S.（2020）。实解析函数的零点集。《数学札记》，107（3）：529–530。

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). 通过深度强化学习实现人类水平的控制。《自然》，518(7540):529–533。Newey, W. K. and McFadden, D. (1994). 大样本估计与假设检验。载于 Engle, R. F. and McFadden, D. L. 编，《计量经济学手册》第4卷，第2111–2245页。爱思唯尔。Norets, A. (2012). 使用人工神经网络近似估计动态离散选择模型。《计量经济学评论》，31(1):84–106。Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》，32。Pfannschmidt, K., Gupta, P., Haddenhorst, B., and H¨ullermeier, E. (2022). 学习上下文依赖选择函数。《国际近似推理杂志》，140:116–155。Ramachandran, P., Zoph, B., and Le, Q. V. (2018). 探索激活函数。载于《国际学习表征会议》。Rosenfeld, N., Oshiba, K., and Singer, Y. (2020). 通过集合依赖聚合预测选择。载于 III, H. D. and Singh, A. 编，《第37届国际机器学习会议论文集》，第119卷《机器学习研究论文集》，第8220–8229页。PMLR。Schmidt-Hieber, J. (2020). 使用带 ReLU 激活函数的深度神经网络进行非参数回归。《统计年鉴》，48(4):1875–1897。Segol, N. and Lipman, Y. (2020). 关于通用等变集合网络。载于《国际学习表征会议》。Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2023). 神经网络筛估计量的渐近性质。《非参数统计杂志》，35(4):839–868。PMID: 38169985。Sifringer, B., Lurkin, V., and Alahi, A. (2020). 通过表征学习增强离散选择模型。《交通研究B辑：方法论》，140:236–261。

Singh, A., Liu, Y., and Yoganarasimhan, H. (2023). 选择模型与置换不变性：差异化产品市场中的需求估计。*arXiv*预印本 *arXiv:2307.07090*。Train, K. E. (2009). 《离散选择模拟方法》。剑桥大学出版社，剑桥，第2版。van der Vaart, A. W. (1998). 渐近统计学。剑桥统计与概率数学丛书。剑桥大学出版社，英国剑桥。Wang, F. and Ross, C. L. (2018). 机器学习出行方式选择：比较极端梯度提升模型与多项Logit模型的表现。《交通研究记录》，2672(47):35–45。Wang, S., Mo, B., and Zhao, J. (2020). 用于选择分析的深度神经网络：采用备选项特定效用函数的架构设计。《交通研究*C*辑：新兴技术》，112:234–251。Wang, S., Mo, B., and Zhao, J. (2021). 基于理论的残差神经网络：离散选择模型与深度神经网络的协同。《交通研究*B*辑：方法论》，146:333–358。Wei, Y. M. and Jiang, Z. (2025). 使用神经网络估计结构模型的参数。《营销科学》，44(1):102–128。White, H. (1982). 错设模型的最大似然估计。《计量经济学》，50(1):1–25。Wong, M. and Farooq, B. (2021). Reslogit：一种用于数据驱动选择建模的残差神经网络逻辑斯蒂模型。《交通研究*C*辑：新兴技术》，126:103050。Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). 深度集合。载于《神经信息处理系统进展》，第30卷。

《基于等变神经网络的相关离散选择模型摊销推断》的补充材料

约翰斯·霍普金斯大学凯瑞商学院博士后研究员
Easton Huch

约翰斯·霍普金斯大学凯瑞商学院凯瑞杰出教授 Michael Keane

附录A 介绍一般性分离和通用近似的结果与证明。第A.8 节证明了第5.1节中的结果。附录B 给出了第5.2节中估计理论结果的证明。附录C 介绍模拟研究的详细内容。

一般性分离与通用逼近

本附录阐述了正文中所介绍的逐备选项编码器架构所依据的群论基础。我们证明，某些不变量通常能够分离置换群下的轨道，从而通过斯通–魏尔施特拉斯定理得到通用逼近结果。

令 $\mathbb{S}_K$ 表示对称实值 $K \times K$ 矩阵空间。我们考虑一个扩展群 G_K ，它通过置换、缩放以及加上双重中心化零空间中的矩阵来作用于 $\mathbb{S}_K$ 。A.1 节介绍所需的记号，并阐述中心化矩阵和归一化双重中心化函数 f 的基本性质。

A.2–A.4 节证明，复合函数 $h = g \circ f$先进行归一化双重中心化，再提取对角元素和非对角元素的多重集，它对 G_K-轨道在一般情形下具有分离性。分离失效的例外集是闭集、G_K-不变量，且为零测集。

第A.5 节讨论对中心化迹归一化子空间的限制 $\mathbb{S}_K^* = \mathrm{Im}(f)$ ，其中 g 通常能够区分 S_K-轨道，而斯通–魏尔施特拉斯定理给出了普适

近似。

第 A.6 节针对指定指标 j提出了一种变体，并证明不变量 g_j通常能够区分 $S_{K-1}^{(j)}$下固定 j 的子群的轨道。

A.7–A.8 节将理论扩展到纳入确定性效用，最终给出正文中定理 1 和定理 2 的证明。在全文中，$\mathbf{X}$ 和$\mathbf{Y}$ 表示尺度矩阵。如第 4.3 节所述，我们有时也将这些矩阵称为协方差矩阵，尽管在多项Probit模型之外，这种解释并不完全准确。

A.1 预备知识

全文中，令 $K \geq 2$ 表示备选项数量。令$\mathbf{I}_K$ 表示单位矩阵， $\mathbf{1}_K \in \mathbb{R}^K$ 表示全 1 向量。

定义 1（中心化矩阵）。中心化矩阵 $\mathbf{M} \in \mathbb{R}^{K\times K}$ 定义为$\mathbf{M} = \mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$。

命题 2（中心化矩阵的性质）。中心化矩阵$\mathbf{M}$ 是对称的（$\mathbf{M}^\top = \mathbf{M}$）、幂等的（$\mathbf{M}^2 = \mathbf{M}$），并且会消去常向量（$\mathbf{M}\mathbf{1}_K = \mathbf{0}$）。

证明。 对称性显而易见。对于幂等性，$\mathbf{M}^2 = \mathbf{I}_K - \frac{2}{K}\mathbf{1}_K\mathbf{1}_K^\top + \frac{1}{K^2}\mathbf{1}_K(\mathbf{1}_K^\top\mathbf{1}_K)\mathbf{1}_K^\top = \mathbf{I}_K - \frac{2}{K}\mathbf{1}_K\mathbf{1}_K^\top + \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top = \mathbf{M}$。最后，$\mathbf{M}\mathbf{1}_K = \mathbf{1}_K - \frac{1}{K}\mathbf{1}_K(\mathbf{1}_K^\top\mathbf{1}_K) = \mathbf{1}_K - \mathbf{1}_K = \mathbf{0}$。 □

定义 2（双重中心化函数）。 定义定义域 $\mathcal{U} = \{\mathbf{X} \in \mathbb{S}_K : \mathrm{tr}(\mathbf{MXM}) \neq 0\}$ 以及函数 $f : \mathcal{U} \to \mathbb{S}_K$，其定义为

$$f(\mathbf{X}) = \frac{K \cdot \mathbf{MXM}}{\mathrm{tr}(\mathbf{MXM})}.$$

命题 3（显式公式）。 对于 $\mathbf{X} \in \mathbb{S}_K$，令 $\bar{X}_{i\cdot} = \frac{1}{K}\sum_{k=1}^K X_{ik}$、 $\bar{X}_{\cdot j} = \frac{1}{K}\sum_{k=1}^K X_{kj}$ 和 $\bar{X}_{\cdot\cdot} = \frac{1}{K^2}\sum_{i,\,j} X_{ij}$。则：

(i) $[\mathbf{MXM}]_{ij} = X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot}$。

(ii) $\mathrm{tr}(\mathbf{MXM}) = \mathrm{tr}(\mathbf{X}) - K\bar{X}_{\cdot\cdot}$。

证明。 (i) 部分可通过直接展开得到。对于 (ii) 部分，利用幂等性以及迹的循环性质：$\mathrm{tr}(\mathbf{MXM}) = \mathrm{tr}(\mathbf{M}^2\mathbf{X}) = \mathrm{tr}(\mathbf{MX}) = \mathrm{tr}(\mathbf{X}) - \frac{1}{K}\mathbf{1}_K^\top\mathbf{X}\mathbf{1}_K = \mathrm{tr}(\mathbf{X}) - K\bar{X}_{\cdot\cdot}$。

□

命题 4 (定义域性质)。 集合 $\mathbb{S}_K \setminus \mathcal{U} = \{\mathbf{X} \in \mathbb{S}_K : \mathrm{tr}(\mathbf{MXM}) = 0\}$ 是 $\mathbb{S}_K$ 中的一个超平面，且其勒贝格测度为零。定义域 $\mathcal{U}$ 是开集且具有 G_K-不变性。

证明。 根据命题 3(ii)，$\mathrm{tr}(\mathbf{MXM})$ 是一个非平凡线性函数，因此其零点集是一个测度为零的超平面。由连续性可知其为开集。对于 G_K-不变性：缩放 $c \neq 0$ 可得 $\mathrm{tr}(\mathbf{M}(c\mathbf{X})\mathbf{M}) = c\mathrm{tr}(\mathbf{MXM}) \neq 0$；通过 $\mathbf{N} = \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top$ 进行零空间平移，满足 $\mathbf{MNM} = \mathbf{0}$，因为 $\mathbf{M}\mathbf{1}_K = \mathbf{0}$；置换根据循环性质保持迹不变。

□

定义 3 （扩展群作用）。 令 S_K 表示由 K 个元素组成的对称群，其对应的置换矩阵为 $\mathbf{P}_\pi$，并令 $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$。定义扩展群$G_K = \mathbb{R}^K \rtimes (\mathbb{R}^* \times S_K)$，它在 $\mathbb{S}_K$ 上的作用为：

$$(c, \boldsymbol{u}, \pi) \cdot \mathbf{X} = c\,\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top + \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top.$$

命题 5 (f 和 G_K)。 函数 f 满足：

(i) 置换等变性： $f(\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top) = \mathbf{P}_\pi f(\mathbf{X}) \mathbf{P}_\pi^\top$。

(ii) 缩放不变性： $f(c\mathbf{X}) = f(\mathbf{X})$ 对于 $c \neq 0$。

(iii) 零空间平移不变性： $f(\mathbf{X} + \boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top) = f(\mathbf{X})$.

证明。 对于 (i)，由于 $\mathbf{P}_\pi \mathbf{1}_K = \mathbf{1}_K$，我们有 $\mathbf{P}_\pi \mathbf{M} = \mathbf{M}\mathbf{P}_\pi$，因此 $\mathbf{M}(\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top)\mathbf{M} = \mathbf{P}_\pi \mathbf{MXM} \mathbf{P}_\pi^\top$，且迹保持不变。第 (ii) 部分由 $f(c\mathbf{X}) = \frac{cK\,\mathbf{MXM}}{c\,\mathrm{tr}(\mathbf{MXM})} = f(\mathbf{X})$ 得出。对于 (iii)，$\mathbf{M}(\boldsymbol{u}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{u}^\top)\mathbf{M} = \mathbf{0}$，因为 $\mathbf{M}\mathbf{1}_K = \mathbf{0}$。 □

引理 1 (原像刻画)。 若 $f(\mathbf{X}) = f(\mathbf{Y})$ 对于 $\mathbf{X}、\mathbf{Y} \in \mathcal{U}$，则存在$c \in \mathbb{R}^*$ 和 $\boldsymbol{w} \in \mathbb{R}^K$，使得 $\mathbf{X} = c\mathbf{Y} + \boldsymbol{w}\mathbf{1}_K^\top + \mathbf{1}_K\boldsymbol{w}^\top$。

证明。 令 $t_X = \mathrm{tr}(\mathbf{MXM})$ 和 $t_Y = \mathrm{tr}(\mathbf{MYM})$。假设 $f(\mathbf{X}) = f(\mathbf{Y})$ 意味着$\mathbf{MXM} = c\,\mathbf{MYM}$，其中 $c = t_X/t_Y \in \mathbb{R}^*$。

对于对称的 $\mathbf{Z}$，令 $\boldsymbol{\mu}_\mathbf{Z} = \mathbf{Z}\mathbf{1}_K/K$（行平均值），且 $m_\mathbf{Z} = \mathbf{1}_K^\top \mathbf{Z}\mathbf{1}_K/K^2$（总体均值）。展开 $\mathbf{MZM} = \mathbf{Z} - \boldsymbol{\mu}_\mathbf{Z}\mathbf{1}_K^\top - \mathbf{1}_K\boldsymbol{\mu}_\mathbf{Z}^\top + m_\mathbf{Z}\mathbf{1}_K\mathbf{1}_K^\top$，并将其应用于等式两边：

$$\mathbf{X} = c\mathbf{Y} + (\boldsymbol{\mu}_\mathbf{X} - c\boldsymbol{\mu}_\mathbf{Y})\mathbf{1}_K^\top + \mathbf{1}_K(\boldsymbol{\mu}_\mathbf{X} - c\boldsymbol{\mu}_\mathbf{Y})^\top - (m_\mathbf{X} - cm_\mathbf{Y})\mathbf{1}_K\mathbf{1}_K^\top.$$

令 $\boldsymbol{w} = \boldsymbol{\mu}_\mathbf{X} - c\boldsymbol{\mu}_\mathbf{Y} - \frac{m_\mathbf{X} - cm_\mathbf{Y}}{2}\mathbf{1}_K$ 即可得出结果。 □

定义 4 (中心化迹归一化子空间)。 定义$\mathbb{S}_K^* = \{\mathbf{Y} \in \mathbb{S}_K : \mathbf{Y}\mathbf{1}_K = \mathbf{0}, \ \mathrm{tr}(\mathbf{Y}) = K\}$。

定义 5 （不变函数 g）。 对于 $\mathbf{Y} \in \mathbb{S}_K^*$，定义 $g(\mathbf{Y}) = (\mathcal{D}_K(\mathbf{Y}) \ \mathcal{O}_K(\mathbf{Y}))$，其中$\mathcal{D}_K(\mathbf{Y}) = \{\{Y_{ii} : i = 1, \ldots, K\}\}$是对角元素的多重集，而 $\mathcal{O}_K(\mathbf{Y}) = \{\{Y_{ij} : 1 \leq i < j \leq K\}\}$ 是非对角项的多重集。

A.2 一般单射性

现在我们定义使得 g 能够区分轨道的一般性条件。**定义 6** （一般性条件）。 设 $\mathbf{Y} \in \mathbb{S}_K$。我们称 $\mathbf{Y}$ 满足：

（a）若所有元素互不相同，则满足条件 *(a)*： $Y_{ij} \neq Y_{kl}$ 对于 $(i, j) \neq (k, l)$，其中 $i \leq j, k \leq l$。

（b）若对于每个 i，恰好存在一个非对角位置的 $(K-1)$-子集 $\mathcal{S}$，使得 $\sum_{(k,l) \in \mathcal{S}} Y_{kl} = -Y_{ii}$，则满足条件 *(b)* 。

条件 (b) 利用了行和约束 $\mathbf{Y}\mathbf{1}_K = \mathbf{0}$：第 i 行的非对角项之和为 $-Y_{ii}$，因此规范子集 $\mathcal{S}_i^* = \{(i, \ j) : \ j \neq i\}$ 始终满足该等式。条件 (b) 要求该子集在所有可能的 $(K-1)$ 子集中是唯一的。

引理 2 (恢复引理)。 若$\mathbf{Y} \in \mathbb{S}_K^*$ 满足条件 *(a)* 和 *(b)*，则 $g(\mathbf{Y})$ 可确定$\mathbf{Y}$ 至多相差一个 S_K-轨道。

证明。 给定多重集 $\mathcal{D}_K$ 和 $\mathcal{O}_K$，我们可以在重新标记的意义下重构$\mathbf{Y}$ ：

步骤1： 将不同的对角元素按 $Y_{(1)(1)} > \cdots > Y_{(K)(K)}$ 排序，为备选项分配规范标签。

步骤2： 对于每个 (i)，将 $\mathcal{R}_{(i)} \subset \mathcal{O}_K$ 确定为和为 $-Y_{(i)(i)}$ 的唯一 $(K-1)$ 子集（由条件 (b) 可知其存在）。

步骤3： 对于 $i \neq j$，将 $Y_{(i)(j)}$ 恢复为 $\mathcal{R}_{(i)} \cap \mathcal{R}_{(j)}$ 中唯一的元素（根据条件 (a)，它是单元素集）。

任何满足条件 (a) 和 (b) 的$\mathbf{Y}'$ （ g（$\mathbf{Y}'$） $= g$（$\mathbf{Y}$）） 都会得到相同的重构结果，至多相差重新标记。 □

A.3 例外集

定义7 (例外集)。 定义 $\mathcal{E} = \mathcal{U}^C \cup \{\mathbf{X} \in \mathcal{U} : \ f(\mathbf{X})$ 违反 (a) 或 (b)$\}$。

引理 3 (约束的非平凡性)。 以下结论成立：

(i) 对于不同的位置 $(i,j) \neq (k,l)$，函数 $f(\mathbf{X})_{ij} - f(\mathbf{X})_{kl}$ 在 $\mathcal{U}$ 上并非恒为零。

(ii) 对于每个索引 i 和 $(K-1)$ 子集 $\mathcal{S} \neq \mathcal{S}_i^*$，函数 $\sum_{(k,l)\in\mathcal{S}} f(\mathbf{X})_{kl} + f(\mathbf{X})_{ii}$ 在 $\mathcal{U}$ 上并非恒为零。

证明。 考虑对角线 $\mathbf{X} = \text{diag}(\lambda_1, \ldots, \lambda_K)$，其中各项 λ_j 和 $\sum_j \lambda_j \neq 0$ 互不相同。于是 $[\mathbf{MXM}]_{ii} = \lambda_i(1-2/K) + \lambda/K$ 且 $[\mathbf{MXM}]_{ij} = (\lambda - \lambda_i - \lambda_j)/K$ 对于 $i \neq j$ 成立。对于一般的 λ_j，$f(\mathbf{X})$ 的所有条目都互不相同，从而证明了 (i)。对于 (ii)，$\mathcal{S} \neq \mathcal{S}_i^*$ 的线性组合在 λ_i 上的系数非零，因此它不是平凡组合。 □

命题 6 ($\mathcal{E}$ 的性质)。集合 $\mathcal{E}$ 是闭的、具有 G_K 不变性，且是勒贝格零测集。

证明。**闭性：** 条件 (a) 和 (b) 是开条件（严格不等式和唯一的子集和在小幅扰动下仍然成立），因此 $\mathcal{E}^C$ 是开集。

G_K**不变性：** 根据命题 5，缩放和零空间平移不会改变 $f(\mathbf{X})$。置换共轭会变换 $f(\mathbf{X})$，同时保持互异性和子集和结构。

零测性： 每个元素 $f(\mathbf{X})_{ij}$ 根据命题 3，在 $\mathcal{U}$ 上是有理（因而解析）函数。根据引理3，定义 (a) 和 (b) 失效条件的是非平凡解析函数。根据 Mityagin (2020年)，它们的零点集均为零测集，而 $\mathcal{E}$ 是这类集合的有限并。

□

A.4 全空间的基本结果

定理 7 (主定理)。对于任意 $\mathbf{X}$、$\mathbf{Y} \in \mathbb{S}_K \setminus \mathcal{E}$：若 $g\{f(\mathbf{X})\} = g\{f(\mathbf{Y})\}$，则 $\mathbf{X}$ 和 $\mathbf{Y}$ 属于同一个 G_K 轨道。

证明。由引理 2，$g\{f(\mathbf{X})\} = g\{f(\mathbf{Y})\}$ 意味着 $f(\mathbf{Y}) = \mathbf{P}_\pi f(\mathbf{X}) \mathbf{P}_\pi^\top$ 对某个 $\pi \in S_K$ 成立。由命题 5(i)，$f(\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top) = f(\mathbf{Y})$。由引理 1，$\mathbf{Y} = c\mathbf{P}_\pi \mathbf{X} \mathbf{P}_\pi^\top + \boldsymbol{w}\mathbf{1}_K^\top + \mathbf{1}_K \boldsymbol{w}^\top$ 对某个 $c \in \mathbb{R}^*$，$\boldsymbol{w} \in \mathbb{R}^K$ 成立。 □

A.5 限制到中心化、迹归一化矩阵

命题7 ($\mathbb{S}_K^*$**的性质)**。子空间 $\mathbb{S}_K^*$ 是一个非空、闭的仿射子空间，其维数为 $(K+1)(K-2)/2$，满足 $\mathbb{S}_K^* = \mathrm{Im}(f)$，并且是 S_K-不变的。

证明。维数由 $\dim(\mathbb{S}_K) - (K+1) = K(K+1)/2 - K - 1 = (K+1)(K-2)/2$ 得出。对于 $\mathbb{S}_K^* = \mathrm{Im}(f)$：若 $\mathbf{X} \in \mathcal{U}$，则 $f(\mathbf{X})\mathbf{1}_K = \mathbf{0}$ 且 $\mathrm{tr}(f(\mathbf{X})) = K$，因此 $\mathrm{Im}(f) \subseteq \mathbb{S}_K^*$；反之，若 $\mathbf{Y} \in \mathbb{S}_K^*$，则 $\mathrm{MYM}^\top = \mathbf{Y}$ 且 $f(\mathbf{Y}) = \mathbf{Y}$。对于 S_K-不变性：$(\mathbf{P}_\pi \mathbf{Y} \mathbf{P}_\pi^\top)\mathbf{1}_K = \mathbf{P}_\pi \mathbf{Y} \mathbf{1}_K = \mathbf{0}$，并且迹保持不变。 □

定义8 (受限例外集**)**。定义 $\mathcal{E}^* = \{\mathbf{Y} \in \mathbb{S}_K^* : \mathbf{Y}$ 违反 (a) 或 (b)$\}$。**定理 8 (**$\mathbb{S}_K^*$上的一般性分离**)**。集合 $\mathcal{E}^*$ 是闭的、S_K-不变的，且在 $\mathbb{S}_K^*$ 中测度为零。对于任意 $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathbb{S}_K^* \setminus \mathcal{E}^*$：$g(\mathbf{Y}_1) = g(\mathbf{Y}_2)$ 意味着 $\mathbf{Y}_1$ 和 $\mathbf{Y}_2$ 属于同一个 S_K-轨道。

证明。由命题 $\mathcal{E}^*$ 限制在 6 $\mathbb{S}_K^*$上的性质，运用引理 3即可得出（该引理通过 $\mathbb{S}_K^* = \mathrm{Im}(f)$ 给出了非平凡约束）。分离性由引理 2得出。 □

推论2 （在 $\mathbb{S}_K^*$ 上的通用逼近）。 任何连续的、关于 S_K 不变的函数，都可以在紧子集上由一个以排序后的对角线和非对角线元素为输入的多层感知机一致逼近。

证明。遵循 Blum-Smith 等人（2025），通过连续函数对 g 进行编码：$g_k^d(\mathbf{Y})$ 是第 k 大的对角线元素，$g_\ell^o(\mathbf{Y})$ 是第 ℓ 大的非对角线元素。这些函数关于S_K不变，并且根据定理 8，它们能区分 $(\mathbb{S}_K^* \setminus \mathcal{E}^*)/S_K$中的点。由于 S_K是有限的（因而是紧的），该商空间是豪斯多夫空间（Bredon，1972）。根据 Stone–Weierstrass 定理，由这些函数生成的代数是稠密的。根据通用逼近定理（Cybenko，1989；Hornik 等人，1989），多层感知机可以逼近这些输入的任意连续函数。

□

A.6 子矩阵限制

对于指定指标 j，我们构造一个用于区分 $S_{K-1}^{(j)}$-轨道的不变量，其中 $S_{K-1}^{(j)} = \{\pi \in S_K : \pi(j) = j\}$。

定义 9 (子矩阵不变量 g_j**)**。对于 $\mathbf{Y} \in \mathbb{S}_K^*$ 和索引 j，定义 $g_j(\mathbf{Y}) = (\mathcal{P}_j(\mathbf{Y})\ \mathcal{O}_j(\mathbf{Y}))$

其中：

$$\mathcal{P}_j(\mathbf{Y}) = \{\{(Y_{kk}, Y_{jk}) : k \neq j\}\},$$
$$\mathcal{O}_j(\mathbf{Y}) = \{\{Y_{kl} : 1 \leq k < l \leq K,\, k \neq j,\, l \neq j\}\}.$$

多重集 $\mathcal{P}_j$ 将每个对角线元素 Y_{kk} （对于 $k \neq j$）与协方差 Y_{jk}配对。多重集 $\mathcal{O}_j$ 包含 $\mathbf{Y}_{-j,-j}$的非对角项。注意， $Y_{jj} = K - \sum_{k\neq j} Y_{kk}$ 可以隐式恢复。

定义 10 （子矩阵的一般性条件）. 我们称$\mathbf{Y} \in \mathbb{S}_K^*$ 满足（a$_j$），如果除可能的 Y_{jj} 之外，所有元素都互不相同；满足（b$_j$），如果对于每个 $k \neq j$，都存在唯一的由 $K-2$ 个元素组成的子集 $\mathcal{S}$，它是$\mathbf{Y}_{-j,-j}$ 中非对角位置的子集，且满足 $\sum_{(k,m)\in\mathcal{S}} Y_{km} = -Y_{kk} - Y_{jk}$。目标和 $-Y_{kk} - Y_{jk}$ 源自行和约束：$Y_{kk} + Y_{kj} + \sum_{l\neq k, l\neq j} Y_{kl} = 0$。定理 *9*（通过 g_j实现一般性分离）。定义 $\mathcal{E}_j = \{Y \in \mathbb{S}_K^* : Y$ 违反（*a*$_j$）或（*b*$_j$）$\}$。则 $\mathcal{E}_j$ 是闭的、 $S_{K-1}^{(j)}$-不变的，且为零测集。对于 Y_1,$Y_2 \in \mathbb{S}_K^* \setminus \mathcal{E}_j$: $g_j(Y_1) = g_j(Y_2)$ 蕴含 Y_1,Y_2 属于同一个 $S_{K-1}^{(j)}$轨道。

证明。 证明思路沿用定理 8。对于可分性，恢复过程与引理 2类似：按字典序排列 $\mathcal{P}_i$ 中的各对，通过唯一的子集和识别每行的非对角项，并将共享项恢复为单元素交集。

□

推论 3 （通过 g_j实现通用逼近）。 在紧子集上，任何定义于 $\mathbb{S}_K^* \setminus \mathcal{E}_j$ 、关于 $S_{K-1}^{(j)}$ 不变的连续函数，都可以由一个多层感知机一致逼近，该感知机以来自 $\mathcal{P}_j$ 的排序后的各对以及来自 $\mathcal{O}_j$ 的排序后的各项作为输入。

证明。该 证明与推论 2类似，只需用各对替换对角元素。 □

A.7 效用—协方差对的扩展：全局不变量

我们将理论扩展到由效用向量和协方差矩阵组成的对 $(\boldsymbol{v},\ \mathbf{Y})$。**定义 11** （中心化效用—协方差空间）。 定义 $\mathcal{X}_K = \{(\boldsymbol{v},\ \mathbf{Y}) \in \mathbb{R}^K \times \mathbb{S}_K : \sum_{i=1}^K v_i = 0,\ \mathbf{Y}\mathbf{1}_K = \mathbf{0},\ \mathrm{tr}(\mathbf{Y}) = K\}$。

该空间的维数为 $(K-1) + (K+1)(K-2)/2 = (K^2 + K - 4)/2$。它是闭的、豪斯多夫的，并且在 $\pi \cdot (\boldsymbol{v}, \mathbf{Y}) = (\mathbf{P}_\pi \boldsymbol{v}, \mathbf{P}_\pi \mathbf{Y} \mathbf{P}_\pi^\top)$ 下具有 S_K-不变性。
定义 12 （全局联合不变量）。 对于 $(\boldsymbol{v},\ \mathbf{Y}) \in \mathcal{X}_K$，定义 $\tilde{g}(\boldsymbol{v},\ \mathbf{Y}) = (\mathcal{P}(\boldsymbol{v},\ \mathbf{Y})\ \mathcal{O}(\mathbf{Y}))$，其中 $\mathcal{P}(\boldsymbol{v},\ \mathbf{Y}) = \{\{(v_i,\ Y_{ii}) :\ i = 1,\ \ldots\ldots,\ K\}\}$，且 $\mathcal{O}(\mathbf{Y}) = \{\{Y_{ij} :\ 1 \leq i < j \leq K\}\}$。

关键的扩展在于， $\mathcal{P}$ 包含将效用与方差联系起来的配对 $(v_i,\ Y_{ii})$。

定义 13 （一般性条件）。 我们称 $(\boldsymbol{v},\ \mathbf{Y}) \in \mathcal{X}_K$ 满足条件 (˜a)：若各对 $(v_i,\ Y_{ii})$˜两两不同，且非对角项两两不同；并满足条件 (b)：若条件 (b) 对 $\mathbf{Y}$ 成立。

˜ ˜定理 *10*（ $\mathcal{X}_K$上的一般性分离）。定义 $\mathcal{B} = \{(\boldsymbol{v}, Y) \in \mathcal{X}_K :$ *(˜a)* 或 *(b)* 不成立}。˜ ˜那么 $\mathcal{B}$ 是闭集、 S_K-不变量，并且是零测集。对于 $(\boldsymbol{v}_1, Y_1)$、$(\boldsymbol{v}_2, Y_2) \in \mathcal{X}_K \setminus \mathcal{B}$：$\tilde{g}(\boldsymbol{v}_1, Y_1) = \tilde{g}(\boldsymbol{v}_2, Y_2)$ 意味着这两对属于同一 S_K-轨道。

证明。 证明沿用定理 8。恢复过程按字典序排列各对以分配标签，然后通过子集和交集恢复非对角项。由于 $v_i - v_j$ 在 $\mathcal{X}_K$上非平凡，零测度论证也同样适用。

□

A.8 效用—协方差对的扩展：逐备选项不变量

本节证明正文中的定理 1 （一般性分离）和定理 2 （多层感知机通用逼近）。

定义 14 （逐备选项不变量 g_j）。 对于 $(\boldsymbol{v},\ \mathbf{Y}) \in \mathcal{X}_K$和索引 j，定义 $g_j(\boldsymbol{v},\ \mathbf{Y})$ =($\mathcal{T}_j(\boldsymbol{v},\ \mathbf{Y})$ $\mathcal{O}_j(\mathbf{Y})$），其中：

$$\mathcal{T}_j(\boldsymbol{v}, \mathbf{Y}) = \{\{(v_k, Y_{kk}, Y_{jk}) : k \neq j\}\},$$
$$\mathcal{O}_j(\mathbf{Y}) = \{\{Y_{kl} : 1 \leq k < l \leq K,\, k \neq j,\, l \neq j\}\}.$$

这结合了第A.6 节和A.7节中的扩展： $\mathcal{T}_j$ 包含三元组 $(v_k,\ Y_{kk},\ Y_{jk})$，将每个备选项 k的效用和方差与其和 j 的协方差相耦合；而 $\mathcal{O}_j$ 包含$\mathbf{Y}_{-j,\ -j}$的非对角项。数值 $v_j = -\sum_{k\neq j} v_k$ 和 $Y_{jj} = K - \sum_{k\neq j} Y_{kk}$ 可以隐式恢复。

定义 15 （一般性条件）。 我们称 $(\boldsymbol{v},\ \mathbf{Y}) \in \mathcal{X}_K$ 满足 (˜a$_j$)，如果对于 $k \neq j$，三元组 $(v_k,\ Y_{kk},\ Y_{jk})$ 两两不同，且$\mathbf{Y}_{-j,\ -j}$ 的非对角项两两˜不同；若条件 (b$_j$) 成立，则满足 (b$_j$)。

˜**定义 16** (例外集)。定义 $\mathcal{B}_j = \{(\boldsymbol{v},\ \mathbf{Y}) \in \mathcal{X}_K :$ (˜a$_j$) 或 (b$_j$) 不成立}。

命题 8 ($\mathcal{B}_j$的性质)。集合 $\mathcal{B}_j$ 是闭集、 $S^{(j)}_{K-1}$-不变量，且在 $\mathcal{X}_K$ 中是零测集。

˜证明。 证明遵循命题 6。闭性成立，因为（˜a$_j$）和（b$_j$）是开条件。在 $S^{(j)}_{K-1}$ 下的不变性成立，因为固定 j 的置换会保留互异性

和子集和结构。对于零测集：两个三元组重合由三个超平面 $\{v_k = v_l\} \cap \{Y_{kk} = Y_{ll}\} \cap \{Y_{jk} = Y_{jl}\}$的交集定义，且每个超平面都非平凡；类似地，根据与引理 3（ii）相似的论证，子集和约束也都是非平凡的。 □

引理 4 (恢复引理)。 **</style}若(**$\boldsymbol{v}$**,**$\mathbf{Y}$**)** $\in \mathcal{X}_K$ 满足 *(*a_j*)* 和 *(*b_j*)*，则 $g_j(\boldsymbol{v},\mathbf{Y})$ 可确定$(\boldsymbol{v},\mathbf{Y})$，精确到 $S^{(j)}_{K-1}$-轨道。

证明。 该证明将引理2 推广到三元组。按 $\mathcal{T}_j$ 字典序排列三元组，以分配标签 **(1)** $\ldots$, $(K-1)$。对于每个 (k)，将 $\mathcal{R}_{(k)} \subset \mathcal{O}_j$ 识别为唯一一个求和等于 $-Y_{(k)(k)} - Y_{j(k)}$ 的 $(K-2)$-子集。恢复 $Y_{(k)(l)}$ 为单元素集 $\mathcal{R}_{(k)} \cap \mathcal{R}_{(l)}$。根据约束恢复 v_j 和 Y_{jj}。

□

现在我们证明第 5 节中的主要结果。

定理 1 的证明。根据命题 8， $\mathcal{B}_j$ 是闭的、 $S^{(j)}_{K-1}$-不变的，且为零测集。对于 $(\boldsymbol{v}_1, \mathrm{Y}_1)$、$(\boldsymbol{v}_2, \mathrm{Y}_2)$，若 $\in \mathcal{X}_K \setminus \mathcal{B}_j$ 满足 $g_j(\boldsymbol{v}_1, \mathrm{Y}_1) = g_j(\boldsymbol{v}_2, \mathrm{Y}_2)$，则由引理 4 可知，这两个配对均可由该共同值确定，精确到 $S^{(j)}_{K-1}$-轨道。 □

定理 *2* 的证明。 通过连续函数对 g_j 进行编码：按字典序排列三元组，得到 $3(K-1)$ 个函数 $g^t_{k,\,1} = v_{(k)}$， $g^t_{k,\,2} = Y_{(k)(k)}$， $g^t_{k,\,3} = Y_{j(k)}$；对非对角项排序，得到 $(K-1)(K-2)/2$ 个函数 g^o_ℓ。这些函数是连续的且具有 $S^{(j)}_{K-1}$不变性，并且根据定理 $\mathcal{X}_K \setminus \mathcal{B}_j)/S^{(j)}_{K-1}$，在 (1 上能够区分不同的点。

由于 $S^{(j)}_{K-1}$ 是有限的， $\mathcal{X}_K/S^{(j)}_{K-1}$ 是豪斯多夫空间（Bredon，1972年）。根据 Stone–Weierstrass 定理，由这些函数生成的代数在紧子集上稠密。根据通用逼近定理（Cybenko，1989年；Hornik 等人，1989年），多层感知机可以逼近这些输入的任意连续函数。由于选择概率是在假设 4 下根据式（1）给出的积分定义的，因此它们是连续函数；由此，结论成立。 □

B 估计理论

本附录补充介绍第5.2 节和5.3节中关于估计理论结果的内容。第B.1 节进一步讨论近似条件：假设2 和4。第B.2 节给出定理4的证明。第B.3节列出第5.3节的正则条件。第B.4节给出定理6的证明。

B.1 模拟器近似假设

本节进一步讨论假设2 和4，说明如何从关于模拟器的更原始条件推导出这些假设，并通过相关理论结果论证这些条件的合理性。

原始假设。 我们考虑一种筛选制度：随着样本量 n 增加，一系列模拟器 $\hat{}P_n$ 以越来越高的精度进行训练。具体而言，逐步（a）增加模拟训练样本 $s = s_n$ 的数量，并且（b）提高神经网络复杂度 $c = c_n$ （例如增加层数或隐藏单元数）。我们施加以下原始条件：

(P1) 紧的协变量支撑集：协变量空间 $\mathcal{C}$ 是紧的。

(P2) 效用和协方差映射的光滑性：对于所有 $\mathbf{X}_i \in \mathcal{C}$，映射 $\boldsymbol{\theta} \mapsto \boldsymbol{v}_i^*(\mathbf{X}_i, \boldsymbol{\theta})$ 和 $\boldsymbol{\theta} \mapsto \boldsymbol{\Sigma}^*(\mathbf{X}_i, \boldsymbol{\theta})$ 均连续可微。

(P3) 内部概率条件：存在一个紧集 $\mathcal{X}_K \subset \mathcal{X}_K$ 和 $p > 0$，使得 $\inf_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \bar{\mathcal{X}}_K} \min_j P_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \geq p$，且 $\mathcal{C} \times \Theta$ 在映射 $(\mathbf{X}, \boldsymbol{\theta}) \mapsto (\boldsymbol{v}^*(\mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\Sigma}^*(\mathbf{X}, \boldsymbol{\theta}))$ 下的像包含于 $\mathcal{X}_K$。

(P4) 概率尺度逼近：对于某个 $\alpha \geq 0$，模拟器满足

$$\sup_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \bar{\mathcal{X}}_K} \max_{j=1,\ldots,K} |\hat{P}_{n,j}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) - P_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)| = o_p(n^{-\alpha/2}). \tag{49}$$

(P5) 梯度逼近：模拟器梯度满足

$$\sup_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \bar{\mathcal{X}}_K} \max_{j=1,\ldots,K} \left\| \nabla_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)} \log \hat{P}_{n,j}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) - \nabla_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)} \log P_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \right\| = o_p(1). \tag{50}$$

接下来，我们将说明这些假设如何推出假设2和 4。随后，我们将利用逼近理论中的相关结果，为假设 **(P4)** 和 **(P5)** 提供依据。

从概率尺度误差到对数似然误差。现在我们证明，条件 **(P1)–(P4)** 蕴含假设 2。将对数似然误差写为

$$\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \xi_i, \quad \text{where} \quad \xi_i = \log \hat{P}_{n,y_i}(\boldsymbol{\theta}) - \log P_{y_i}(\boldsymbol{\theta}), \tag{51}$$

且 $y_i \in \{1, \ldots, K\}$ 表示观测选择。在给定协变量和训练后的模拟器的条件下， ξ_i 相互独立。我们将其分解为偏差和中心化分量：

$$\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\xi_i \mid \mathbf{X}_i)}_{\text{bias}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \{\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)\}}_{\text{centered term}}. \tag{52}$$

令 $\delta_n = \sup_{(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) \in \bar{\mathcal{X}}_K} \max_j |\hat{P}_{n,j} - P_j| = o_p(n^{-\alpha/2})$表示^ (P4)的概率尺度误差。对 $\log P_{n,j}$ 在 P_j 处进行泰勒展开，得到

$$\log \hat{P}_{n,j} - \log P_j = \frac{\hat{P}_{n,j} - P_j}{P_j} - \frac{(\hat{P}_{n,j} - P_j)^2}{2P_j^2} + O\left(\frac{\delta_n^3}{\underline{p}^3}\right). \tag{53}$$

ξ_i 的条件期望为

$$\mathbb{E}(\xi_i \mid \mathbf{X}_i) = \sum_{j=1}^{K} P_j(\log \hat{P}_{n,j} - \log P_j). \tag{54}$$

^ ^ 一阶项为 $\sum_j P_j \cdot (P_{n,j} - P_j)/P_j = \sum_j (P_{n,j} - P_j) = 0$，因为两个概率向量的元素之和均为一。因此，

$$\mathbb{E}(\xi_i \mid \mathbf{X}_i) = -\sum_{j=1}^{K} \frac{(\hat{P}_{n,j} - P_j)^2}{2P_j} + O\left(\frac{\delta_n^3}{\underline{p}^3}\right) = O\left(\frac{\delta_n^2}{\underline{p}}\right). \tag{55}$$

^ ^ 对于中心化项，由中值定理可得 $|\log P_{n,j} - \log P_j| \le |P_{n,j} - P_j|/p \le \delta_n/p$，因此 $|\xi_i| \le \delta_n/p$。条件方差满足 $\mathrm{Var}(\xi_i \mid \mathbf{X}_i) \le \mathbb{E}(\xi_i^2 \mid \mathbf{X}_i) \le \delta_n^2/p^2$。由于中心化项 $\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)$ 条件独立且均值为零，其样本均值的方差为

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\{\xi_i - \mathbb{E}(\xi_i \mid \mathbf{X}_i)\} \;\middle|\; \mathbf{X}_1, \ldots, \mathbf{X}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\xi_i \mid \mathbf{X}_i) \le \frac{\delta_n^2}{n\underline{p}^2}. \tag{56}$$

根据切比雪夫不等式，中心化项为 $O_p\left(\frac{\delta_n}{p\sqrt{n}}\right)$。

结合三角不等式：

$$|\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = O\left(\frac{\delta_n^2}{\underline{p}}\right) + O_p\left(\frac{\delta_n}{\sqrt{n}\underline{p}}\right). \tag{57}$$

对于 $\alpha \in [0,1]$，偏差项 $O(\delta_n^2/p) = o(n^{-\alpha})$ 占主导地位，而中心化项为 $O_p\left(\frac{\delta_n}{p\sqrt{n}}\right) = o_p\left(n^{-(\alpha+1)/2}\right)$。由于条件（P1）和（P2）保证从 $(\mathbf{X}, \boldsymbol{\theta})$ 到 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 的映射在紧集 $\mathcal{C} \times \Theta$ 上连续，因此该界对 $\boldsymbol{\theta} \in \Theta$ 一致成立：

$$\sup_{\boldsymbol{\theta}\in\Theta} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = o_p(n^{-\alpha}), \tag{58}$$

这就是假设 2。

从梯度逼近到费舍尔信息一致性。现在我们证明条件（P1）、（P2）、（P3）和（P5）蕴含假设 4。根据链式法则，

$$\nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,j}(\boldsymbol{\theta}) = \{\nabla_{\boldsymbol{\theta}}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)\}^\top \nabla_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log \hat{P}_{n,j}, \tag{59}$$

对于 $\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta})$，情况类似。根据条件（P2），雅可比矩阵 $\nabla_{\boldsymbol{\theta}}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 在紧集 $\mathcal{C} \times \Theta$ 上连续，因此一致有界：$\sup_{(\mathbf{X},\boldsymbol{\theta})\in\mathcal{C}\times\Theta} \|\nabla_{\boldsymbol{\theta}}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)\| \le C$，其中某个常数为 $C < \infty$。令

$$\varepsilon_n = \sup_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)\in\bar{\mathcal{X}}_K} \max_j \left\| \nabla_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log \hat{P}_{n,j} - \nabla_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log P_j \right\| = o_p(1) \tag{60}$$

表示（P5）中的梯度近似误差。那么

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\| \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,j}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta}) \right\| \le C\varepsilon_n = o_p(1). \tag{61}$$

回顾一下，模拟器和真实的得分外积估计量分别为

$$\hat{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,y_i}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,y_i}(\boldsymbol{\theta})^\top, \tag{62}$$

$$\mathbf{J}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta})^\top. \tag{63}$$

令 $\hat{\boldsymbol{s}}_i = \nabla_{\boldsymbol{\theta}} \log \hat{P}_{n,\,y_i}(\boldsymbol{\theta})$ 和 $\boldsymbol{s}_i = \nabla_{\boldsymbol{\theta}} \log P_{y_i}(\boldsymbol{\theta})$。恒等式

$$\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i^\top - \boldsymbol{s}_i\boldsymbol{s}_i^\top = (\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i)\hat{\boldsymbol{s}}_i^\top + \boldsymbol{s}_i(\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i)^\top \tag{64}$$

意味着

$$\|\hat{\boldsymbol{s}}_i\hat{\boldsymbol{s}}_i^\top - \boldsymbol{s}_i\boldsymbol{s}_i^\top\| \le \|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \cdot (\|\hat{\boldsymbol{s}}_i\| + \|\boldsymbol{s}_i\|). \tag{65}$$

真实得分函数 $\boldsymbol{s}_i$ 一致有界：(P3) 确保所有相关输入 $(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 都位于 $\mathcal{X}_K$ 中，其中 $P_j \geq p > 0$，从而保证梯度有界 $\nabla_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)} \log P_j$；(P2) 确保雅可比矩阵 $\nabla_{\boldsymbol{\theta}}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)$ 连续，因此在 $\mathcal{C} \times \Theta$上有界。令 $M = \sup_{\boldsymbol{\theta}\in\Theta} \max_j \|\nabla_{\boldsymbol{\theta}} \log P_j(\boldsymbol{\theta})\| < \infty$，则对所有 i均有 $\|\boldsymbol{s}_i\| \leq M$。由于 $\|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \leq C\varepsilon_n = o_p(1)$，模拟器得分满足 $\|\hat{\boldsymbol{s}}_i\| \leq \|\boldsymbol{s}_i\| + C\varepsilon_n \leq M + o_p(1)$。因此，

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} \|\hat{\mathbf{J}}_n(\boldsymbol{\theta}) - \mathbf{J}_n(\boldsymbol{\theta})\| &\leq \frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{s}}_i - \boldsymbol{s}_i\| \cdot (\|\hat{\boldsymbol{s}}_i\| + \|\boldsymbol{s}_i\|) \\ &\leq \frac{1}{n}\sum_{i=1}^{n} C\varepsilon_n \cdot \{2M + o_p(1)\} = o_p(1), \end{aligned} \tag{66}$$

即假设 4。

逼近理论方面的依据。 原始条件 (P4) 和 (P5) 得到了神经网络逼近与估计相关理论结果的支持，但针对我们这一特定设定的完整理论仍有待未来研究。

模拟器的总误差可分解为逼近误差和估计误差：

$$\sup_{(\boldsymbol{v}^*,\boldsymbol{\Sigma}^*)\in\bar{\mathcal{X}}_K} \max_{j=1,\ldots,K} |\hat{P}_{n,j}(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*) - P_j(\boldsymbol{v}^*, \boldsymbol{\Sigma}^*)| \leq \underbrace{a(c_n)}_{\text{approximation}} + \underbrace{e(s_n, c_n)}_{\text{estimation}}, \tag{67}$$

其中，$a(c)$ 表示真实选择概率函数与复杂度为 c 的网络类中最佳逼近函数之间的距离；$e(s, c)$ 表示训练后的网络因使用 s 个模拟样本进行训练而偏离该类最佳逼近函数的程度。

对于逼近误差，若选择概率在参数空间内部是关于（$\boldsymbol{v}^*$, $\boldsymbol{\Sigma}^*$）的实解析函数，则具有平滑激活函数的神经网络可以将其近似到指数级小的误差。De Ryck等人(2021年证明，具有两个隐藏层的 tanh 网络的逼近误差按以下速率衰减：$O\left\{N^{k/(d+1)}\exp(-C \cdot N^{1/(d+1)}\right.$ $\left.\log N)\right\}$（以 Sobolev范数计）$W^{k,\infty}$其中，N是网络宽度，d 是输入维度，$C > 0$ 取决于目标函数的解析性参数。早期研究为深层 ReLU 网络证明了类似的指数速率（E和 Wang，2018年），但其假设更强，要求目标函数存在绝对收敛的幂级数展开。鉴于这些指数收敛结果，要达到逼近误差 ε，网络复杂度只需随 $1/\varepsilon$呈多对数级增长。

对于估计误差，经验过程理论的标准结果给出了 $O\left\{\sqrt{c\log(sc)/s}\right\}$ 阶的 L^2 范数界，其中 c 是一个复杂度参数（Schmidt-Hieber，2020年；Farrell等人，2021年；Shen等人，2023年）。然而，对于我们的设定，我们需要一个 L^∞界。以下结果至少假设 $e(s,c) = O\left\{c^k \,\text{polylog}(s)/\sqrt{s}\right\}$，其中某个$k < \infty$成立。于是，当网络复杂度随目标精度呈多对数级增长时，对 c 的依赖可以忽略，估计误差实际上为 $O(1/\sqrt{s})$阶（相差对数因子）。

达到（P4）给出的速率，要求 $a(c_n) = o(n^{-\alpha/2})$ 和 $e(s_n,\ c_n) = o_p(n^{-\alpha/2})$；请注意，逼近误差是非随机的，因为它涉及纯函数近似。解析函数的指数近似速率意味着，第一个条件在 c_n 呈多对数级增长时即可满足。第二个条件要求 s_n 的增长速度快于 n^α，允许存在对数因子。实践中，由于从多项式正态概率模型中进行模拟的计算成本很低，因此生成包含 $s_n \gg n$ 的训练集是可行的。

这些情形 $\alpha = 0$和 $\alpha = 1$ 对应不同的推断目标。当 $\alpha = 0$ 时，假设2 仅要求对数似然误差依概率趋于零，这足以保证基于模拟器的估计量具有一致性（定理 3）。当 $\alpha = 1$ 时，更严格的要求是误差以快于 n^{-1} 的速度趋于零，这确保基于模拟器的估计量与真实的最大似然估计量渐近等价，从而具有其 $\sqrt{n}$-一致性和渐近正态性（定理 4）。后一种情形需要更大的训练集： s_n 的增长速度快于 n，而 s_n 的增长速度快于 $n^0 = 1$。

对于 (P5) 中的梯度逼近，Czarnecki 等人(2017年) 表明，神经网络是 Sobolev范数下的通用逼近器，而De Ryck等人(2021年) 的界适用于 $W^{k,\infty}$ 范数，从而证明光滑网络可以同时逼近函数值及其导数。此外，Cocola和Hand(2020年) 证明了在适当条件下，过参数化的双层网络进行 Sobolev 训练时，梯度流能够全局收敛，从而确保训练后的网络达到较小的 Sobolev损失。这些结果支持了 (P5) 的合理性，不过对于我们的设定，精确的收敛速率和条件仍值得进一步研究。

B.2 定理4的证明

^^^证明。 我们证明满足 $\ell_n(\boldsymbol{\theta}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) - \eta_n$ where $\eta_n = o_p(n^{-1})$ 的近似极大值点的结果。精确极大化者对应于 $\eta_n = 0$。证明分为三个

步骤：首先界定对数似然之差，然后证明这些估计量渐近等价，最后利用精确最大似然估计量的渐近正态性。

步骤1：界定对数似然之差。

根据假设 $\delta_n = \sup_{\boldsymbol{\theta}\in\Theta} |\hat{\ell}_n(\boldsymbol{\theta}) - \ell_n(\boldsymbol{\theta})| = o_p(n^{-1})$，其中 2 且 $\alpha \geq 1$。根据三角不等式，

$$|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| \leq |\ell_n(\hat{\boldsymbol{\theta}}_n) - \hat{\ell}_n(\hat{\boldsymbol{\theta}}_n)| + |\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)|. \tag{68}$$

根据假设 2，第一项的上界为 δ_n。对于第二项，我们给出 ˜的上界和下界 $\ell_n(\boldsymbol{\theta}_n$。由于 $\boldsymbol{\theta}_n$ 是 ℓ_n的最大化点：

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \leq \ell_n(\hat{\boldsymbol{\theta}}_n) + \delta_n \leq \ell_n(\tilde{\boldsymbol{\theta}}_n) + \delta_n. \tag{69}$$

对于下界 ˜，根据近似最大化性质以及以下事实 $\boldsymbol{\theta}_n \in \Theta$:

$$\hat{\ell}_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} \hat{\ell}_n(\boldsymbol{\theta}) - \eta_n \geq \hat{\ell}_n(\tilde{\boldsymbol{\theta}}_n) - \eta_n \geq \ell_n(\tilde{\boldsymbol{\theta}}_n) - \delta_n - \eta_n. \tag{70}$$

ˆ ˆ ˜结合这些界，$|\ell_n(\boldsymbol{\theta}_n) - \ell_n(\boldsymbol{\theta}_n)| \leq \delta_n + \eta_n$。因此，

$$|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| \leq 2\delta_n + \eta_n = o_p(n^{-1}), \tag{71}$$

ˆ ˜这意味着 $n|\ell_n(\boldsymbol{\theta}_n) - \ell_n(\boldsymbol{\theta}_n)| = o_p(1)$。

步骤 2：估计量的渐近等价。

我们证明 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^{-} - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$。固定 $\epsilon > 0$ 并定义邻域 d

$$B_{n,\epsilon} = \left\{ \boldsymbol{\theta} \in \Theta : \sqrt{n}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n\| \leq \epsilon \right\}. \tag{72}$$

ˆ我们证明，以概率趋于1，$\boldsymbol{\theta}_n \in B_{n,\ \epsilon}$。

˜ ˜对于任意 $\boldsymbol{\theta} \notin B_{n,\ \epsilon}$，我们有 $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| > \epsilon/\sqrt{n}$。由于 $\boldsymbol{\theta}_n$ 使 ℓ_n 达到最大值，得分函数为零：$\frac{1}{n}\sum_{i=1}^{n} \psi(y_i,\ \mathbf{X}_i;\ \tilde{\boldsymbol{\theta}}_n) = \mathbf{0}$。在 $\tilde{\boldsymbol{\theta}}_n$ 处进行泰勒展开，

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\tilde{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n)^\top \nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n), \tag{73}$$

其中 $\bar{\boldsymbol{\theta}}$ 位于 $\boldsymbol{\theta}$ 和 $\tilde{\boldsymbol{\theta}}_n$之间的线段上。由一致性可知，$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ 和 $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$,

因此 $\bar{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$。根据一致大数定律及假设 3 (ii) 和 (iv),

$$\nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}}) \xrightarrow{p} -\mathbf{J}(\boldsymbol{\theta}_0). \tag{74}$$

令 $\lambda_{\min} > 0$ 表示$\mathbf{J}(\boldsymbol{\theta}_0)$ 的最小特征值，根据假设3(v)，该值为正。当 n 足够大时，以概率趋于1的方式，海森矩阵 $\nabla^2_{\boldsymbol{\theta}} \ell_n(\bar{\boldsymbol{\theta}})$ 的所有特征值均小于等于 $-\lambda_{\min}/2$。因此，对于 $\boldsymbol{\theta} \notin B_{n,\epsilon}$,

$$n\left\{\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta})\right\} \geq \frac{\lambda_{\min}}{4} n\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}_n\|^2 > \frac{\lambda_{\min}}{4} \cdot \epsilon^2 = \kappa > 0 \tag{75}$$

当 $n \to \infty$ 时，该结论以概率趋于1成立。

ˆ ˆ 现在假设 $\Pr(\boldsymbol{\theta}_n \notin B_{n,\epsilon}) \not\to 0$。在 $\boldsymbol{\theta}_n \notin B_{n,\epsilon}$ 和海森矩阵界 ˜ ˆ 均成立的事件上，我们有 $n(\ell_n(\boldsymbol{\theta}_n) - \ell_n(\boldsymbol{\theta}_n)) > \kappa$。由于海森矩阵界以概率趋于1成立，因此该事件的概率有一个正的下界，这意味着 $\Pr\left[\left\{n(\ell_n(\tilde{\boldsymbol{\theta}}_n) - \ell_n(\hat{\boldsymbol{\theta}}_n)\right\} > \kappa\right] \not\to 0$。但这与ˆ步骤1中的 $n|\ell_n(\hat{\boldsymbol{\theta}}_n) - \ell_n(\tilde{\boldsymbol{\theta}}_n)| = o_p(1)$ 矛盾。因此，$\Pr(\boldsymbol{\theta}_n \in B_{n,\epsilon}) \to 1$。

由于 $\epsilon > 0$ 是任意的，我们得出 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$。

步骤3：渐近正态性。

根据标准极大似然理论（vander Vaart（1998年）中的定理5.39和引理7.6），假设3中的正则条件确保模型在˜ ˜处关于 $\boldsymbol{\theta}_0$ 具有二次均值可微性，且费希尔信息非奇异。由于 $\boldsymbol{\theta}_n$ 是一致的，$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\}$. 由于 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$，由斯卢茨基定理可得

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) + \sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)^{-1}\right\}. \tag{76}$$

□

B.3 模型错设下推断的正则条件

本附录陈述了第5.3节结果所需的正则条件。这些条件与 M 估计量的标准假设（van derVaart，1998，第5章）相似；对于具有平滑激活函数的神经网络模拟器而言，这些条件较为宽松。回顾一下，在第5.3节中，我们将神经网络视为固定的（不随 n变化），因此我们使用不带m 和不带ψ下标的 n ˆ符号。

假设5（识别与一致收敛）。满足以下条件：

（i）参数空间 Θ 是紧的。

(ii) 伪真参数 $\boldsymbol{\theta}_0^\dagger = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\{\hat{m}(y,\ \mathbf{X};\boldsymbol{\theta})\}$ 唯一存在，且位于 Θ 的内部。

(iii) 伪真参数是良好分离的：对于每个 $\epsilon > 0$,

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\geq\epsilon} \mathbb{E}\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta})\} < \mathbb{E}\left\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta}_0^\dagger)\right\}. \tag{77}$$

(iv) 样本准则函数一致收敛于其期望 :

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\boldsymbol{\theta}) - \mathbb{E}[\hat{m}(y,\mathbf{X};\boldsymbol{\theta})]\right| \xrightarrow{p} 0. \tag{78}$$

(v) 估计量 $\boldsymbol{\theta}_n$ 是样本准则函数的近似最大化点：

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\hat{m}(y_i,\mathbf{X}_i;\boldsymbol{\theta}) - o_p(n^{-1}). \tag{79}$$

假设6（渐近正态性的正则性）。以下条件成立：

(i)（可微性）对于每个 $(y,\ \mathbf{X})$，映射 $\boldsymbol{\theta}\mapsto\hat{m}(y,\ \mathbf{X};\boldsymbol{\theta})$ 在 Θ 上可微，其梯度为 $\hat{\psi}(y,\ \mathbf{X};\ \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\hat{m}(y,\ \mathbf{X};\ \boldsymbol{\theta})$。

(ii)（局部支配）存在 $\delta > 0$ 和一个可测函数 $\dot{m}(y,\ \mathbf{X})$，其具有 $\mathbb{E}[\dot{m}(y,\ \mathbf{X})^2] < \infty$，使得

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\leq\delta}\|\hat{\psi}(y,\mathbf{X};\boldsymbol{\theta})\| \leq \dot{m}(y,\mathbf{X}). \tag{80}$$

(iii)（二阶展开）映射 $\boldsymbol{\theta}\mapsto\mathbb{E}\{\hat{m}(y,\ \mathbf{X};\boldsymbol{\theta})\}$ 在 $\boldsymbol{\theta}_0^\dagger$ 处存在二阶泰勒展开：

$$\mathbb{E}\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta})\} = \mathbb{E}\left\{\hat{m}(y,\mathbf{X};\boldsymbol{\theta}_0^\dagger)\right\} - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger)^\top\mathbf{A}(\boldsymbol{\theta}_0^\dagger)(\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger) + o(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|^2), \tag{81}$$

其中 $\mathbf{A}(\boldsymbol{\theta}_0^\dagger) = -\mathbb{E}\left\{\nabla_{\boldsymbol{\theta}}^2\hat{m}(y,\ \mathbf{X};\ \boldsymbol{\theta}_0^\dagger)\right\}$ 是正定的。

B.4 定理5 和6 的证明

定理 5的证明。 定义样本准则函数和总体准则函数

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(y_i, \mathbf{X}_i; \boldsymbol{\theta}), \qquad M(\boldsymbol{\theta}) = \mathbb{E}\left\{\hat{m}(y, \mathbf{X}; \boldsymbol{\theta})\right\}. \tag{82}$$

根据假设5(iv)，$\sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{p} 0$。根据假设5(ii)–(iii)，$\boldsymbol{\theta}_0^\dagger$ 是 $M(\boldsymbol{\theta})$的唯一最大化点，并且满足如下意义上的良好分离性：对每个 $\epsilon > 0$，$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0^\dagger\|\geq\epsilon} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0^\dagger)$。根据假设5(v)，$M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(n^{-1})$，这意味着 $M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(1)$。由此，$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$ 的结论可由 van der Vaart（1998）的定理5.7得出。

□

定理6的证明。我们验证定理5.23的条件van der Vaart (1998)，并应用该定理得到结果。

根据假设 6(i)，映射 $\boldsymbol{\theta} \mapsto \dot{m}(y, \mathbf{X}; \boldsymbol{\theta})$ 在 $\boldsymbol{\theta}_0^\dagger$ 处对每个 $(y,\mathbf{X})$ 均可微，其导数为 $\dot{\psi}(y,\mathbf{X};\boldsymbol{\theta}_0^\dagger)$。根据假设 6(ii)，得分函数在局部受一个平方可积函数支配，这意味着满足利普希茨条件

$$|\hat{m}(y, \mathbf{X}; \boldsymbol{\theta}_1) - \hat{m}(y, \mathbf{X}; \boldsymbol{\theta}_2)| \leq \dot{m}(y, \mathbf{X})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{83}$$

根据中值定理，对于 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$，当其位于 $\boldsymbol{\theta}_0^\dagger$ 的一个邻域内时，满足利普希茨条件。假设6(iii)在极大值点$\boldsymbol{\theta}_0^\dagger$ 处，为 $\boldsymbol{\theta} \mapsto \mathbb{E}\{\dot{m}(y, \mathbf{X}; \boldsymbol{\theta})\}$ 提供所需的二阶泰勒展开，其二阶导数矩阵 $-\mathbf{A}(\boldsymbol{\theta}_0^\dagger)$ 非奇异。根据定理 5，$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0^\dagger$。根据假设5(v)，近似最大化条件 $M_n(\hat{\boldsymbol{\theta}}_n) \geq \sup_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta}) - o_p(n^{-1})$ 得以满足。

应用vanderVaart （1998年）中的定理5.23：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) = \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\psi}(y_i, \mathbf{X}_i; \boldsymbol{\theta}_0^\dagger) + o_p(1). \tag{84}$$

根据假设6(ii)，$\mathbb{E}\left\{\|\dot{\psi}(y, \mathbf{X}; \boldsymbol{\theta}_0^\dagger)\|^2\right\} \leq \mathbb{E}\{\dot{m}(y, \mathbf{X})^2\} < \infty$。中心极限定理和斯卢茨基引理给出

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0^\dagger) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-1}\mathbf{B}(\boldsymbol{\theta}_0^\dagger)\mathbf{A}(\boldsymbol{\theta}_0^\dagger)^{-\top}\right\}, \tag{85}$$

其中 $\mathbf{B}(\boldsymbol{\theta}_0^\dagger) = \mathbb{E}\left\{ \hat{\psi}(y,\ \mathbf{X};\ \boldsymbol{\theta}_0^\dagger)\, \hat{\psi}(y,\ \mathbf{X};\ \boldsymbol{\theta}_0^\dagger)^\top \right\}$。 □

C 模拟研究详情

本附录进一步介绍因子模型的参数化、模拟器训练流程，以及模拟研究中使用的缩放威沙特分布。

C.1 模拟器训练

我们使用 Adam 优化器（Kingma 和 Ba，2015年）及其 AdamW 权重衰减变体（Loshchilov 和 Hutter，2019年）训练模拟器。基础学习率为 0.003，动量参数为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$，数值稳定性常数为 $\epsilon = 10^{-8}$，权重衰减为 3×10^{-6}。

学习率调度。我们采用先线性预热、再进行平方根倒数衰减的学习率调度。在最初的 5，000 步中，学习率从零线性增加至基础学习率。预热结束后，学习率按 $\sqrt{t_{\text{warmup}}}/t$衰减，其中 t表示当前步数，$t_{\text{warmup}} = 5$，000。

Dropout 调度。我们在训练过程中应用 Dropout，以促进学习到的表示多样化。我们将丢弃率设为 $1/(10t)$，其中 t 表示训练轮次的索引，范围为 1 到 400，000。

训练流程。我们预先生成 400，000 个训练样本，每个样本包含一个规范化输入（$\boldsymbol{v}^*$，$\boldsymbol{\Sigma}^*$）、基于 10^6 次蒙特卡洛抽样得到的模拟选择频率，以及使用第 4 节所述的软松弛计算得到的目标梯度。训练期间，我们将这些样本与预先计算的切空间方向随机配对，用于方向导数损失。梯度会被裁剪至最大范数 0。01。我们训练 400，000 个优化器步数，批量大小为 10，000。

网络架构。表 8 总结了每个 值对应的神经网络架构。所有网络均使用 Swish 激活函数。的值越大，所需网络规模也越大，才能保持近似精度。

C.2 缩放 Wishart 概率密度函数的推导

本节推导 如第 6.1 节所述，给出缩放后的 Wishart随机矩阵的概率密度函数 .

设 $\mathbf{W} \sim W_d(\mathbf{V}, n)$ 表示一个尺度矩阵为 $\mathbf{V}$、自由度为 $n \geq d$ 的 $d \times d$ Wishart随机矩阵。我们推导 $\mathbf{\Sigma} = \mathbf{W}/W_{jj}$ 的分布，其中 W_{jj} 是 $\mathbf{W}$ 的 (j, j) 个元素，且 $j \in \{1, \ldots, d\}$ 为某个值。

Wishart密度为

$$f_{\mathbf{W}}(\mathbf{W}) = \frac{|\mathbf{W}|^{(n-d-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{W})\right\}}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)} \tag{86}$$

其中，$\Gamma_d(\cdot)$是多元伽马函数。

逆变换为$\mathbf{W} = W_{jj}\mathbf{\Sigma}$。注意，$\mathbf{\Sigma}$ 是正定的，且有 $\Sigma_{jj} = 1$，因此它有 $d(d+1)/2 - 1$ 个自由参数。加上 W_{jj}，这与$\mathbf{W}$的 $d(d+1)/2$ 个唯一元素相符。该变换的雅可比行列式为 $w_{jj}^{d(d+1)/2-1}$，其中 w_{jj}表示 W_{jj}的值。

将$\mathbf{W} = W_{jj}\mathbf{\Sigma}$ 代入Wishart密度并应用雅可比矩阵，可得联合密度

$$f_{W_{jj},\mathbf{\Sigma}}(w_{jj}, \mathbf{\Sigma}) = \frac{|\mathbf{\Sigma}|^{(n-d-1)/2}}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)} \cdot w_{jj}^{nd/2-1} \exp\left\{-\frac{w_{jj}}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{\Sigma})\right\}. \tag{87}$$

对 $w_{jj} > 0$ 积分并利用伽马积分，可得 $\mathbf{\Sigma}$的边缘密度：

$$f_{\mathbf{\Sigma}}(\mathbf{\Sigma}) = \frac{\Gamma(nd/2)}{2^{nd/2}|\mathbf{V}|^{n/2}\Gamma_d(n/2)} \cdot \frac{|\mathbf{\Sigma}|^{(n-d-1)/2}}{\left\{\frac{1}{2}\mathrm{tr}(\mathbf{V}^{-1}\mathbf{\Sigma})\right\}^{nd/2}} \tag{88}$$

对于对称正定的 $\mathbf{\Sigma}$，其中 $\Sigma_{jj} = 1$ 对某个固定的 $j \in \{1, \ldots, d\}$成立。

在模拟研究中，我们使用 $d = K - 1$、$n = K + 10$、$\mathbf{V} = \mathbf{I}_{K-1}$，以及 $j = 1$。代入这些值并化简后，可得第 6.1 节给出的密度。

组件	$K=3$	$K=5$	$K=10$
对角 *DeepSet*			
ϕ输出维度	8	24	32
ϕ隐藏层	0	1	1
ϕ隐藏维度	N/A	24	32
ρ输出维度	8	24	32
ρ隐藏层	0	0	0
参数	152	1,440	2,432
非对角 *DeepSet*			
ϕ输出维度	8	24	32
ϕ隐藏层	0	1	1
ϕ隐藏维度	N/A	24	32
ρ输出维度	8	24	32
ρ隐藏层	0	0	0
参数	128	1,368	2,336
组合多层感知机			
输出维度	8	24	32
隐藏维度	12	48	64
隐藏层	2	2	2
参数	572	6,312	10,976
等变层			
隐藏维度	4	12	16
隐藏层	1	1	1
参数	76	612	1,072
参数总数	928	9,732	16,816

表 8：按备选项数量划分的神经网络架构 K。所有配置均采用初始丢弃率 0.1 和 Swish 激活函数。当隐藏层数量为 0 时，隐藏维度不适用（N/A）。